Journal of Data-centric Machine Learning Research (2026) Submitted 3/26; Revised 7/26; Published 8/26

MObyGaze: a film dataset of multimodal objectification densely annotated by experts

Elisa Ancarani ELISA.ANCARANI@UNIV-COTEDAZUR.FR
Université Côte d'Azur, CNRS, I3S, France

Julie Tores JULIE.TORES@UNIV-COTEDAZUR.FR
Université Côte d'Azur, CNRS, Inria, I3S, France

Lucile Sassatelli LUCILE.SASSATELLI@UNIV-COTEDAZUR.FR
Université Côte d'Azur, CNRS, I3S, France

Hui-Yin Wu HUI-YIN.WU@INRIA.FR
Université Côte d'Azur, Inria, France

Rémy Sun REMY.SUN@INRIA.FR
Université Côte d'Azur, Inria, CNRS, I3S, France

Paul Mouret PAUL.MOURET@UNIV-COTEDAZUR.FR
Université Côte d'Azur, CNRS, I3S, France

Clément Bergman CLEMENT.BERGMAN@INRIA.FR
Université Côte d'Azur, Inria, France

Léa Andolfi LEA.ANDOLFI@SORBONNE-UNIVERSITE.FR
Sorbonne Université, GRIPIC, France

Victor Ecrement VICTOR.ECREMENT@SORBONNE-UNIVERSITE.FR
Sorbonne Université, GRIPIC, France

Thierry Devars THIERRY.DEVARS@SORBONNE-UNIVERSITE.FR
Sorbonne Université, GRIPIC, France

Magali Guaresi MAGALI.GUARESI@UNIV-COTEDAZUR.FR
Université Côte d'Azur, CNRS, BCL, France

Virginie Julliard VIRGINIE.JULLIARD@SORBONNE-UNIVERSITE.FR
Sorbonne Université, GRIPIC, France

Sarah Lécossais SARAH.LECOSSAIS@UNIV-PARIS13.FR
Université Sorbonne Paris Nord, LabSIC, France

Frédéric Precioso FREDERIC.PRECIOSO@UNIV-COTEDAZUR.FR
Université Côte d'Azur, CNRS, Inria, I3S, France

Reviewed on OpenReview: *https://openreview.net/forum?id=ai2PAWRIVA*

Editor: Hugo Jair Escalante

Abstract

Characterizing and quantifying gender representation disparities in audiovisual storytelling contents is necessary to grasp how stereotypes may perpetuate on screen. In this article, we consider the high-level construct of objectification and introduce a new AI task to the ML

©2026 Ancarani, Tores et al..

community: characterize and quantify complex multimodal (visual, speech, audio) temporal patterns producing objectification in films. Building on film studies and psychology, we define the construct of objectification in a structured thesaurus involving 5 sub-constructs manifesting through 11 concepts spanning 3 modalities. We introduce the Multimodal Objectifying Gaze (MObyGaze) dataset, made of 20 movies annotated densely by experts for objectification levels and concepts over freely delimited segments: it amounts to 6072 segments over 43 hours of video with fine-grained localization and categorization. We formulate new video interpretation tasks, show the feasibility of multimodal objectification detection, and analyze data and model bias to propose improvements. We exemplify two applications of MObyGaze, showing how rich concept annotation can improve model reliability and explainability. We make our code and our dataset available to the community and described in the Croissant format: `https://github.com/husky-helen/MObyGaze`

Keywords: video interpretation, objectification, multimodality, explainability

1 Introduction

While audiovisual storytelling contents have been shown to strongly shape our perception of sociological constructs (see, e.g., Ward and Grower (2020)), such as gender, race and others, disparities in on-screen representation persist, particularly between genders. Grasping subtle patterns of disparities in gender portrayal requires understanding how the content produces different perceptions of the characters, beyond quantifying gender presence. In film studies, this question has been the subject of numerous qualitative analyses, and the concept of *male gaze* was introduced by Mulvey (1975) and recently revisited by Brey (2020). Male gaze refers to the way the content can be composed to produce objectification, i.e., so that a character is perceived more as an object, often of desire, than a subject of action. Male gaze has a strong gendered component. Mulvey's 1975 account in introducing the idea of male gaze, is explicitly a feminist psychoanalytic argument about how Hollywood cinema structures spectatorship around a masculine heterosexual position. The asymmetry Mulvey identifies is not incidental — cinema, in her argument, routinely aligns three positions (the camera, the male protagonist, and the spectator) as the bearer of an active, desiring look, while coding the woman on screen for her "to-be-looked-at-ness": she is seen, he sees and acts. Brey's work sets out from the observation that the visual apparatus gives women the function of spectacle that interrupts or decorates that plot. These techniques are not gender-neutral formal devices — historically and statistically they are applied differentially to women's bodies, converting the female character from a subject the audience is invited to identify with into a surface the audience (and the male characters, and the camera-as-proxy-viewer) is invited to look at.

But **how is objectification produced by the content?** This involves deliberate filmmaking choices to compose the audiovisual content that unfolds over time, such as: What is the camera perspective? Who are the viewers looking at, whom does the camera embody, and how are characters portrayed? What are the dialogue dynamics? Who is talking, to whom, of whom, about what, and how? Fig. 1 shows an example of how objectification is produced through the combination of camera position, character gaze, posture, speech, and voice. Recent advances in multimodal media computing provide new approaches that may be wielded to characterize complex temporal and multimodal patterns of objectification, and quantify them.

Towards this goal, we introduce a new AI task: characterizing and quantifying how complex multimodal (visual, speech, audio) discursive patterns produce objectification in film. So far, interpretive tasks have only been thoroughly studied in the text modality, with approaches for hate speech detection and beyond, incorporating subtle aspects such as sexism (e.g., Samory et al. (2021)). Approaches for the visual modality are scarce and mostly limited to still images (Fersini et al. (2019); Kiela et al. (2020); Wang et al. (2025)). We therefore contribute the necessary elements to make this new interpretive multimodal task accessible to the machine learning community. **Our contributions are**:

• the Multimodal Objectifying Gaze (MObyGaze) dataset. For this, we devise a thesaurus of objectification by building on existing characterization in film studies and cognitive and social psychology. The thesaurus articulates visual, speech and audio components, which we denote as *concepts* involved in the production of objectification. The annotation process then consists in 2 experts densely annotating 20 movies: they manually delimit all the segments (unitization) they find relevant for objectification, and label each with a level of objectification (categorization). To allow fine-grained data and model analysis, they also annotate which objectification concepts are present, and indicate the classification difficulty (with a *hard negative* label). We verify the validity of the produced data with annotator agreement measures for both unitization and categorization. The resulting dataset comprises 6072 segments over 43 hours of 20 films each annotated by 2 experts.

• a new video interpretive task to detect objectification in films. We study and show the feasibility of the task for the individual concept modalities and with multimodal models. We analyze disaggregated model performance, unveil the contextuality of the annotation process, and show how to improve model generalization.

• two applications of MObyGaze where we show how rich concept annotation can improve multimodal models, how the new video interpretation task of objectification detection raises new challenges for explainable AI methods, and the approaches we propose.

We make our dataset available to the community, with Datasheet documentation (Gebru et al. (2021)) and Croissant metadata for file and recordset descriptions (Akhtar et al. (2024)). We believe this dataset is valuable for advancing computational approaches to make subtle patterns of bias in audiovisual content visible and more tangible, and quantify their prevalence.

The article is organized as follows. Sec. 2 positions our contributions in the context of related works. Sec. 3 presents the MObyGaze dataset, its creation and analysis. Sec. 4 presents the study of the feasibility of multimodal objectification detection, and data and performance analysis to improve models in terms of accuracy and explainability. Sec. 5 discusses the limitations and broader impact.

2 Related work

We position our contributions with respect to works on: analyses of biases in film datasets, annotation of audiovisual and multimodal contents, and dataset creation for interpretive tasks.

Bias analysis in film datasets Disparities in representing different groups of characters in films have been computationally quantified with analyses of low-level characteristics of the visual (Guha et al. (2015); Mazières et al. (2021); Jang et al. (2019)) or textual data.

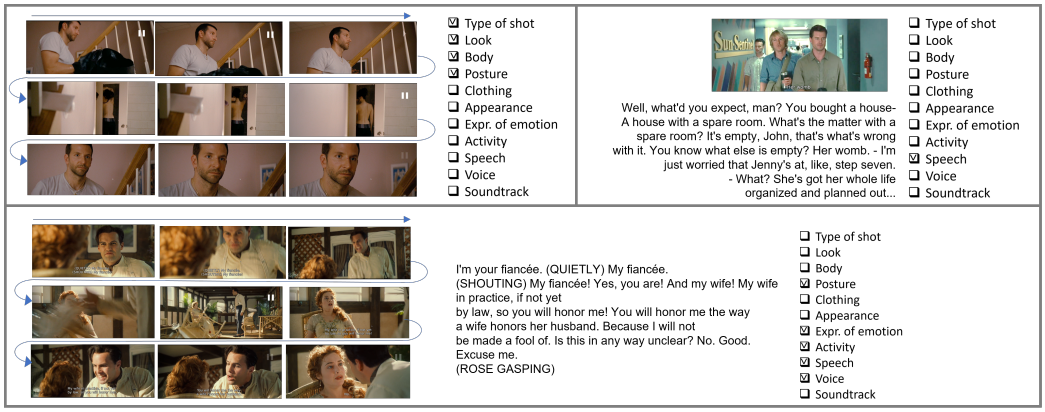


Figure 1: Examples of segments tagged with a *Sure* level of objectification. Top left: vision modality only. Top right: text modality only. Bottom: multimodal concepts producing objectification.

Jang et al. (2019) considered 20 films and show that women characters have a lower spatial and temporal occupancy, corroborating findings on vision datasets by Wang et al. (2022). Somandepalli et al. (2021) show in 1000 movie scripts that female characters appear more often as victims. Schofield and Mehr (2016) characterize differences in linguistic markers in dialogue utterances of female and male characters. Agarwal et al. (2015) propose a way to automate the Bechdel test from computationally analyzing 457 film scripts with their pre-existing annotations of Bechdel test results made by volunteers and hosted in a public website. Martinez et al. (2022) collected 912 movie scripts to investigate differences in how different genders are associated to different types of actions. They show that male characters are generally given more agency than female characters, and that female characters are more often the object of the gaze of male characters, with verbs reflecting their sexual objectification. These last two relied on non-expert human annotation only in the text modality, and without considering high-level interpretive constructs as we do in our work. Neither relied on the analysis of visual or audio data.

Annotation of audiovisual and multimodal content We present in Table 1 a synthetic view of the comparative discussion below. Video annotation is considered a heavier task than image annotation, and has therefore been mostly considered for short videos, notably for action recognition and video anomaly detection. For example, the ActivityNet benchmark (Heilbron et al. (2015)) comprises ca. 27000 videos lasting ca. 2 minutes in average, representing 203 activity classes, crowdworkers annotating the temporal boundaries of each action instance. Video anomaly is a more interpretive construct, with classes such as abuse, assault, or robbery. For example, Sultani et al. (2018) introduce the UCF-Crime video anomaly dataset of 1900 videos of ca. 4 minutes each, categorized into 13 anomaly classes. Temporal delimitation is costly and variable from annotator to annotator. For this reason, a lot of video anomaly detection data (including the training set of UCF-Crime) are only annotated for classes at the video level, hence requiring weakly-supervised learning approaches to localize and categorize anomaly.

Detecting and quantifying hateful multimodal content is crucial to train less harmful foundation models with large-scale image-text datasets, as investigated by Birhane et al. (2023). Yet, manual annotations of multimodal content remain scarce. Kiela et al. (2020) created a dataset of hateful memes, while Fersini et al. (2019) specifically considered sexist memes and advertisement imagery. More recently, Das et al. (2023) and Wang et al. (2025) have introduced dataset for multimodal hate speech detection in short videos. Specifically Wang et al. (2025) have created HateClipSeg, a dataset of 435 short videos (3 to 10 mins) segmented and annotated for Normal and Offensive content, with five sub-classes of Offensive including Sexual and Hateful. The dataset neither annotates individual modalities nor explanatory concepts, as we do with MObyGaze. Their work shows that existing models have significant room for improvement on such interpretive video tasks, as we do but on a more specific task (detecting objectification) defined structurally with a thesaurus. Bose et al. (2023) introduced MM-AU, a dataset of 8.4K advertisement videos annotated for three tasks including an interpretive one of so-called social message detection, but with a coarse annotation approach: only the presence of a social message is annotated, not its nature, and neither finer-grained concepts nor individual modalities are annotated. We, however, consider similar multimodal baselines in Sec. 4.4.2 and 4.5.1.

Movie datasets are usually not annotated manually. A prominent exception is MovieGraphs, introduced by Vicol et al. (2018), for which freelance workers were recruited to annotate 51 movies with time-grounded graph-based annotations of character relationships and interactions. Owing to the richness of MovieGraphs and the possible relevance of leveraging the original graph annotations to detect objectification, we select 20 out of the 51 movies from MovieGraphs reproducing the same distribution of genres, to be densely annotated for the construct of multimodal objectification.

Our endeavor of film annotation for objectification therefore has distinct objectives and challenges: 1) the nature of the data source is artistic and story-rich, on which we envision tasks beyond human activity recognition, 2) our tasks require the understanding of complex multimodal event cues (language, visual, audio, etc.) over long prediction horizons (a few minutes), 3) we investigate the social construct of objectification in films, that is not intended for other objectives (e.g., security) that require different types of data and annotations.

Dataset creation for interpretive tasks Like Kiela et al. (2020)) and Fersini et al. (2019) who considered hateful and sexist meme annotation, most of the works mentioned above do not provide a detailed definition of the high-level construct to annotate, rather giving annotators freedom to interpret the term. In contrast, systematic approaches for rigorous definition of high-level constructs are more common in NLP. Samory et al. (2021) identified how the lack of proper definition of a high-level construct such as sexism impedes proper data analysis. They therefore proposed to leverage questionnaires introduced and validated in social psychology to produce a codebook to assess different dimensions of sexism. They then employed crowdworkers, trained on the codebook, to annotate tweets. In a similar objective, Da San Martino et al. (2019) approached the difficulties of annotating propaganda in news articles by identifying 18 propaganda techniques from the existing literature. To avoid political views to excessively noise annotation, they had 4 experts localize and classify relevant text-spans. Dense annotation by a single expert has also been proposed for medical

images by Daneshjou et al. (2022), to assign skin lesion images with an end label of *benign* or *malignant*, along with which of the 48 clinical explanatory concepts are present.

In this article, we take inspiration from these codebook-based annotation of high-level constructs by Samory et al. (2021), Da San Martino et al. (2019) and Daneshjou et al. (2022) to approach the creation of data for the multimodal construct of objectification in a systematic and multi-disciplinary way. We leverage existing literature in psychology, cinematography and gender media studies to define a thesaurus, identifying concepts to be annotated by experts, who will annotate feature-length films (2h08min of average duration) with time delimitation (unitization) and categorization of the perceived level of objectification. To the best of our knowledge, this is the first time a dataset of audiovisual content is annotated for a high-level construct – objectification – defined in a thesaurus of multimodal concepts, with freely delimited timespans. In line with approaches advocated by, e.g., Paullada et al. (2021), our purpose is to produce a non-large scale but high-quality dataset enabling efficient model training and data analysis to contribute to unveiling how subtle representation disparities in audiovisual contents may persist.

Positioning with the works on this dataset A pre-print of this article has been on arxiv since early 2025. An early and preliminary version of the dataset was published at CVPR 2024 by Tores et al. (2024) and included only visual concepts and only 12 movies. The article consisted in showing the feasibility of the task from a simpler and somewhat noisier perspective. First, all annotations were aligned onto the MovieGraphs clip spans, aggregating the annotators' labels by max pulling the objectification levels and concatenating the concept annotations. Second, the splits were only across clips, meaning that different clips of the same movie could appear in train, validation and test. Besides, solely models based on MLP and CBMs were tested. After this preliminary work, we identified the need to augment the dataset to: (1) incorporate multimodal concepts such as speech and sound, and (2) obtain stronger representation of all concepts from which we can more easily sample for training various classification tasks and models. As mentioned above, we subsequently started to inverstigate ML challenges raised by the MObyGaze data: how to tackle interpretive video tasks, how to benefit from expert modality-wise annotations to improve multimodal models (Ancarani et al. (2025b), Ancarani et al. (2025a)), how to improve concept-based models (Tores et al. (2025)). While the core contribution of the present article is the complete dataset workflow (with its production, analysis, the principled proof of feasibility of multimodal objectification detection across movies, and corresponding model analyses), we believe the subsequent works above show the usefulness of MObyGaze as a dataset for ML research. That is why we have dedicated Sec. 4.5 to two of these works.

3 Dataset

We first present our definition of the construct of objectification in a structured thesaurus and describe the annotation process. We then analyze the obtained data by validating its consistency and showing key characteristics.

Table 1: Comparison of MObyGaze with other related datasets across domain, task interpretativity, scale, and annotation approach

Domain	Interpretive task	Avg. duration	Total hours	Annotators /content	Unitizing	Explanatory concepts
ActivityNet	−−	18 min	849	crowd, unspec.	✓	–
UCF-Crime	−	4 min	128	unspec.	✓	–
HateMM	+	4 min	43	4 novices trained by 2 experts	✓	–
HateClipSeg	+	4 min	26	2 novices	–	–
MObyGaze (us)	++	128 min	43	2 experts	✓	✓

3.1 Dataset composition

Thesaurus of multimodal objectification We set out from the concept of *male gaze* introduced in film studies to describe the filmmaking choices producing a perception of women characters as objects in male-driven actions, intentions and perspectives. While some formalizations of gaze stem from the gender of the director Malone (2018), we privilege the formalizations of Brey (2020) who defines male gaze only from the film content. Brey (2020) carries out a qualitative analysis of over 120 film and series scenes to describe complex temporal patterns involving filmic (framing, camera perspective and motion, etc.) and iconographic (whether the face is shown and how close, what body parts are shown, how the characters are dressed, what are their interactions) aspects, which either allow the audience to understand and engage with the experience of a character, or prevent the audience from doing so, hence de-humanizing, or *objectifying*, the character. Objectification has also been investigated in social and cognitive psychology. Results and validated questionnaires study how the perception of objectification depends on various elements such as gaze and appearance (Calogero (2004); Calogero et al. (2011); McKinley and Hyde (1996)), clothing and posture (Bernard et al. (2019)), body parts (Bernard et al. (2018)), sexualization (Denchik (2005); Bernard et al. (2020)), interactions (Gervais et al. (2020)), actions (Sap et al. (2017)). Put together, we identify 5 sub-constructs of objectification, shown in Fig. 2 (bottom right): Male Gaze (G), i.e., point of view of a man on a woman (Brey (2020); Mulvey (1975); Bernard et al. (2018)), Sexualization (S) (Bernard et al. (2020, 2019)), Surveillance of the feminine body (B) (Calogero (2004); McKinley and Hyde (1996); Denchik (2005)), Female inaction and male possession (P) (Gervais et al. (2020); Sap et al. (2017)), and Infantilization and animalization (I) (Mulvey (1975)). From the questionnaires, experiences and analyses of these above works in film studies and psychology, we enumerate representative instances of each sub-construct, which we group into 11 concepts spanning 3 modalities, vision, text and sound, as depicted in Fig. 2. The 5 sub-constructs manifest through several modalities. To align with the literature on explainable AI (Chen et al. (2020); Daneshjou et al. (2022); Zarlenga et al. (2022b)), we use the term *concepts* to denote the annotated factors motivating the rating of objectification. Annotating a segment with a concept means the annotator perceives an objectifying element, in this concept's dimension, that may contribute to objectification. The concept granularity is heterogeneous on purpose. Indeed, Vision Language Models (VLMs) and Multimodal LLMs (MLLMs) have been more

brittle at analyzing complex video content (Kesen et al. (2024); Asadi et al. (2026)). Given that a higher number of concepts entails a higher annotation effort, we decided to obtain a much needed supervision in the visual domain from a finer-grained concept annotation, while saving effort on the textual domain.

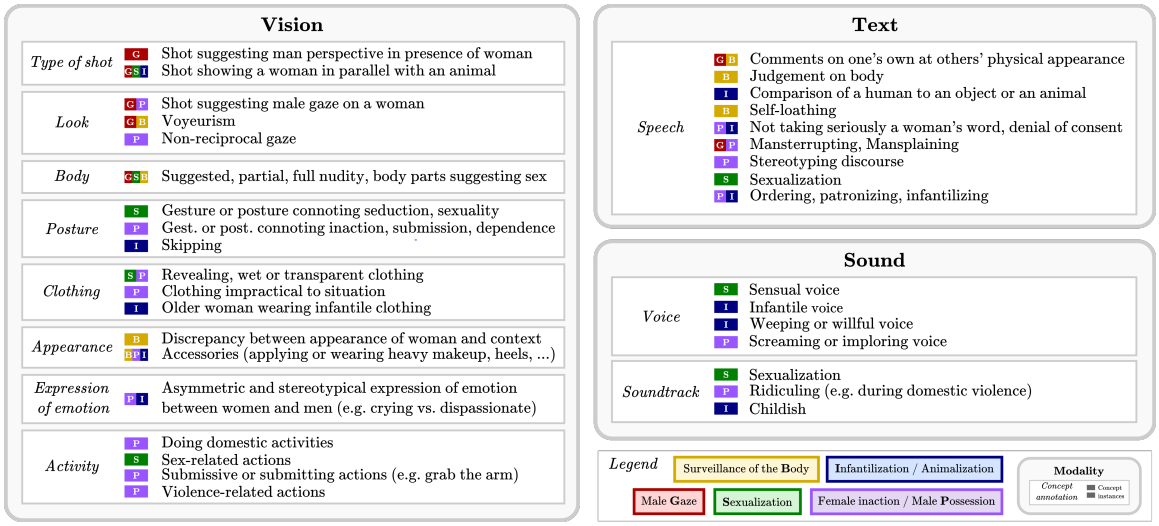


Figure 2: Thesaurus for the construct of objectification: 5 Sub-constructs (bottom right) manifested through 11 ***Concepts*** spanning 3 **Modalities**. Best seen in color. A color-blind version is available in App. A.2.4.

Data selection As mentioned in Sec. 2, we select movies from the MovieGraphs dataset (Vicol et al. (2018)) owing to the richness of existing annotations on relationships and interactions. We hoped this choice makes it easier for the community to contribute to the task, as it has already been done by an external team (Serouis and Sèdes (2025)). We select 20 out of 51 movies, maintaining the distribution of genres (see App. A.2 in the supplementary material for details). This was a round number that allowed us to attain the above objectives while keeping the task reasonably feasible by the two annotators within a month.

Annotation Each movie is annotated by 2 experts (with background in computer science, film studies and cognitive psychology). This choice stems from two factors: existing works with interpretive multimodal tasks adopted a similar approach (see our new Table 1 comparing the datasets), and the annotation effort (80+ hours including pilot on the annotation workflow and expertise). The experts watch the movie entirely, and annotate the film by (1) setting temporal boundaries to create segments, and (2) for each such segment, rate the objectification on one of four levels:

- Easy Negative (EN): no objectifying concept is present;
- Hard Negative (HN): one or some concepts are present, are annotated, but are deemed insufficient to produce a perception of objectification;
- Sure (S): objectification is perceived and explained by the annotated concepts from the thesaurus;

• Not Sure (NS): objectification is perceived and concepts are annotated but the annotator considers they do not sufficiently explain the perception of objectification.
A custom annotation tool was made (see App. A.2). The annotation steps, including remediation and thesaurus refinement, are detailed in App. A.2 and summarized below. The annotators first annotated 2 movies. The obtained annotations were then aligned and colored for the annotators to identify their major divergences, exemplified in Fig. 19 in the appendix. They convened and identified that the agreement was generally high on the rating of objectification. Noticeable differences were in annotation of NS with levels of narratology spanning over more than a single segment. The thesaurus is indeed targeted at annotating salient concepts in segments, and we discuss this limitation in Sec. 5. They specifically identified under-determination of concepts Actitivies and Appearance. They expanded Activities to include all types of actions contributing to objectification, particularly momentary actions by a character onto another (including aspects of domination and violence). They trimmed the concept of Appearance, which initially consisted of instances deploying over the entire film (such as age of character not matching age or appearance of actress) to restrict it to scene-level features. Fig. 2 shows the resulting thesaurus. They then carried out individually the annotation over the rest of the 20 movies.

Data format The resulting dataset is made available, described in ML Croissant format with Responsible AI properties, and detailed in a Datasheet document (Gebru et al. (2021)) in App. A.1.

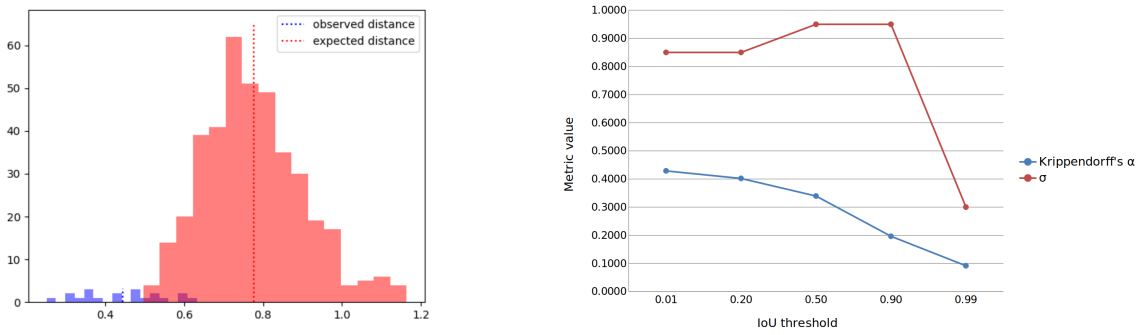
3.2 Dataset analysis

3.2.1 Validation of the data: internal consistency

We first compare the annotation data between each annotator. In the next section we compare them against external references.

IAA on objectification We adopt IAA measures recently introduced by Braylan et al. (2022) for complex multi-object labeling tasks. To assess how two movie annotation sequences are aligned, we consider two types of distance function. The first distance function between 2 movie annotations is 1-F1 score when a movie is considered as a set of tokens (equal to its number of frames) with 2 sets of labels. For each annotation considered as the gold/ground truth (GT) and the other as prediction, we obtain 2 F1 scores, then averaged. This distance therefore inherently incorporates agreement on unitizing. It has been introduced for Named Entity Recognition by Braylan et al. (2022). The second distance function we consider, and that may be more intuitive, is dist@IoUthresh, defined as the distance between two movie annotations when we consider the overlap between two segments must be greater than IoUthresh for these segments to be matched and their respective objectification or concept labels to be compared. A formal definition of this distance is provided in App. A.2.6.

Fig. 3a shows the distributions of the observed d_o (blue) and expected d_e (red) dist@IoUthresh distances for IoUthresh=0.9. d_o and d_e are the within-item (item = movie) and inter-item distances between 2 annotations, respectively. Chance correction is made by comparing d_o samples with d_e samples. To perform this comparison, we consider two different IAA metrics: Krippendorff's α and σ introduced by Braylan et al. (2022). Krippendorff's α is defined as $1 - \bar{d_o}/\bar{d_e}$ with the averages of d_o and d_e. Metric σ is the fraction of the



(a) (b)

Figure 3: (a) Distribution of distances dist@IoUthresh between both annotators' annotations within-movies (observed distances) and inter-movies (expected distances), for IoUthresh=0.9. Krippendorff's $\alpha = 0.20$, $\sigma = 0.95$. (b) Evolution of Krippendorff's α and σ as a function of IoUthresh.

observed distance samples that have a probability lower than 0.05 to have been drawned from the expected (chance) distance distribution.

First we observe in Fig. 3a that the distributions are significantly apart, which is confirmed by $\sigma = 0.95$. This is a strong indicator of the consitency of the annotators' annotations (also qualitatively exemplified in Fig. 19). However, corresponding Krippendorff's α is 0.20 owing to the range of average distances (x-axis), and while Fig. 3b also shows that α increases with lower IoU thresholds as expected, its highest value is 0.4. This is expected from the distance definition and the unitizing task itself, as investigated by Braylan et al. (2022), and we provide a formal intuition with two toy examples in App. A.2.6. They show how unitizing quickly noises the value of α and explain the partly opposite trends of α and σ we observe in Fig. 3b: increasing IoUthresh yields a monotonously decreasing α while σ stays constant or slightly increases, before decreasing sharply after some very high threshold (above the boundary agreement rate), as expected. Therefore, as an IAA metric should describe how much the overall data is better than chance, after observing the weakness of Kripendorff's α@IoU to describe this phenomenon, we opt for reporting σ instead. For the objectification labels (considering labels EN, HN and S and discarding NS amounting to less than 5% of the data) and both distance definitions:

- for distance NER based on F1 score of tokenized sequences: $\sigma = 0.74$
- for dist@IoUthresh=0.9, $\sigma = 0.95$.

These results show the consistency of objectification ratings between both expert annotators.

IAA on concepts We now look at the agreement on concept annotations. To ease interpretation, we report in Table 2 σ (and α) for dist@IoUthresh with IoUthresh=0.8. We first consider concepts individually with only two labels (either the concept is present in a segment annotation, or it is absent).

First, we observe that there is a high ($\sigma \geq 0.8$) or non-trivial ($\sigma \geq 0.6$) agreement for 5 in 11 concepts taken individually (Body, Posture, Clothing, Speech, Expression of an emotion), while the others have a significantly lower σ. Discussion between annotators after the annotation of two pilot movies made appear that the differences in annotated

Table 2: Inter-annotator agreement for individual concepts and concept groups, ordered by σ (descending), computed on the dist@IoUthresh distances, with IoUthresh=0.8.

Concept/Axis	Krippendorff's α	σ
Individual Concepts (ordered by σ, descending)		
Body	0.241	0.85
Posture	0.234	0.85
Clothing	0.271	0.85
Speech	0.240	0.65
Exp of emotion	0.155	0.60
Type of shot	0.180	0.15
Soundtrack	0.066	0.15
Activities	0.165	0.10
Look	0.214	0.05
Appearance	0.077	0.05
Voice	0.155	0.00
Concept groups		
Body, Clothing	0.287	0.95
Posture, Activities	0.235	0.90
Type of shot, Look	0.196	0.80
Speech	0.240	0.65
Voice, Soundtrack	0.134	0.00

concepts often do not correspond to disagreement, but rather to some concept samples being sometimes overlooked by either one of the annotators. As a side note, it is also interesting to observe (e.g., for Look) that α can be greater than σ.

Second, we build on the PCA presented below in Sec. 3.2.4 and analyze IAA not on individual concepts but on groups based on the PCA axes and on the modalities. The concept group is tagged present when at least one concept of the group is present. We consider 5 groups as shown in the lower part in Table 2. While grouping some well agreed-on concepts slightly increases IAA wrt to the individual case, it dramatically increases for concepts that have individually low IAA but which are important and well-represented concepts appearing both in the PCA and the decision tree below as driving objectification: Type of shot and Look have σ increasing from 0.15 and 0.05 when taken individually, to 0.8 when grouped. This shows that they are connected and the annotators agree to annotate the group consistently. Here for example, Type of shot is defined with camera placement while Look is defined with how one character is made to look at another. Cinematography often embodies the Type of Shot by placing the camera at a character's perspective for the viewer to adopt their Look. This explains the consistency of the annotation despite the non-systematic but frequent occurrence yielding possible ambiguity in what exact concept to annotate.

3.2.2 Validation of the data: external consistency

We further validate the annotations of the multimodal concepts against external measures. For speech and audio, we adopt the same approach as for the HateMM dataset by Das et al. (2023).

Visual concepts For the visual concepts, we define a concept score for each individual segment, for each of the 8 visual concepts. This score is a similarity between the X-CLIP

visual embedding of that segment and the textual descriptions of the concept. Textual descriptions of the concepts are provided in Fig. 24 in the appendix. Specifically, we do contrastive scoring precisely defined in App. A.2.7. For each of the 8 concepts, we analyze whether the concept score significantly differ between segments with that concept annotated present ("annotated") and segments without the concept annotated ("not annotated"). We use a two-sided Mann-Whitney U test to compare both groups and produce a p-value per concept. Since 8 separate tests are run (one per concept), a Benjamini-Hochberg False Discovery Rate (FDR) correction is applied across the 8 p-values, producing a q-value for each. This controls the expected proportion of false positives among the concepts ultimately called "significant".

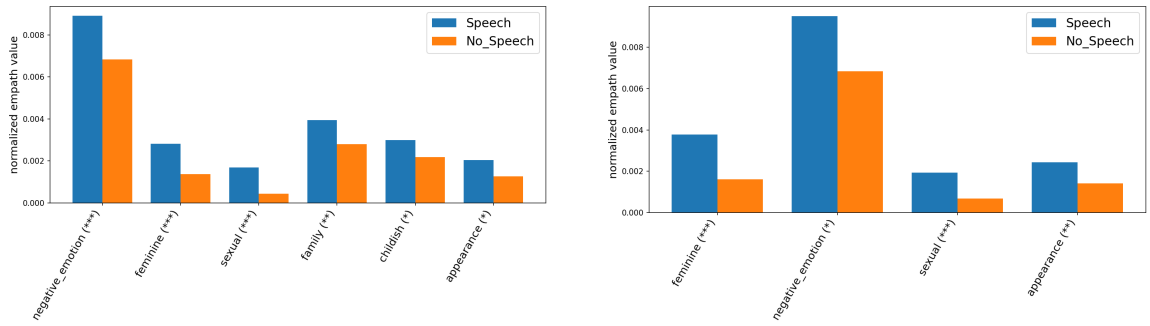
Table 3 reports the mean difference of scores, the rank biserial correlation r_{rb}, which can be interpreted as an effect size of how consistently one group's scores exceed the other's, ranging from -1 to $+1$. It is independent of the sample size. We also report the level of effect size as *negligible* ($r_{rb} < 0.1$), *small* ($0.1 \leq r_{rb} < 0.3$), *medium* ($0.3 \leq r_{rb} < 0.5$), and *large* ($r_{rb} \geq 0.5$). We observe that for both annotators, and for 7 visual concepts in 8, there is a statistically significant difference in concept score between segments annotated with or without the concept. The effect size is negligible for only one concept for both annotators (concept Look). This is hence another indicator of the consistency of the annotations, for the visual concepts.

Table 3: Comparison of the difference in visual concept score for the segments annotated with each concept and those not annotated, for each annotator (quantile 0.9 is used for intra-segment similarity).

	Annotator 1							Annotator 2						
Concept	diff	$\mathbf{r_{rb}}$	**effect size**	n_{ann}	$n_{\neg\text{ann}}$	q-value	sig.	diff	$\mathbf{r_{rb}}$	**effect size**	n_{ann}	$n_{\neg\text{ann}}$	q-value	sig.
Body	0.0245	0.402	medium	213	2056	3.38×10^{-21}	***	0.0179	0.271	small	833	2970	1.44×10^{-32}	***
Clothing	0.0125	0.201	small	433	1836	1.75×10^{-10}	***	0.0149	0.214	small	674	3129	4.41×10^{-18}	***
Type of shot	0.0131	0.197	small	165	2104	3.33×10^{-5}	***	0.0104	0.174	small	565	3238	5.51×10^{-11}	***
Look	0.0017	0.015	negligible	334	1935	0.654	–	0.0064	0.090	negligible	337	3466	0.00604	*
Posture	0.0181	0.293	small	248	2021	2.01×10^{-13}	***	0.0203	0.314	medium	677	3126	5.59×10^{-37}	***
Appearance	0.0136	0.230	small	127	2142	2.04×10^{-5}	***	0.0191	0.283	small	186	3617	7.66×10^{-11}	***
Exp. of emotion	0.0116	0.154	small	251	2018	8.02×10^{-5}	***	0.0317	0.346	medium	308	3495	1.28×10^{-23}	***
Activities	0.0223	0.235	small	178	2091	3.52×10^{-7}	***	0.0413	0.403	medium	438	3365	5.31×10^{-42}	***

Textual speech concept For speech, we use Empath by Fast et al. (2016) to analyze correlation of relevant Empath categories with concept Speech. We run Empath on the subtitle text of each segment, using a curated subset of Empath categories relevant to objectification (body/appearance, power/dominance, emotion, speech/communication), instead of the full 194 categories. The retained categories are shown in Fig. 25 in the appendix. For each Empath category, we compare the per-segment normalized Empath score between the two groups. As earlier, we use the Mann-Whitney U test and a Benjamini-Hochberg False Discovery Rate (FDR) correction since we are testing 71 categories at once.

Fig. 4 shows, for each annotator, the top N lexical categories with the most significant score differences. We observe that the categories related to feminine, negative emotion, sexual and appearance are among the ones with most significant scores across both annotator. We notice that annotator 1 has additional significant categories, family and childish, indicating sensible difference in sensitivity of annotation between both expert annotators.



(a) (b)

Figure 4: Comparative Empath scores for the curated categories for Speech vs. No-Speech concept (the FDR-corrected q-value stars are indicated. (a) Data of annotator 1. (b) Data of annotator 2.

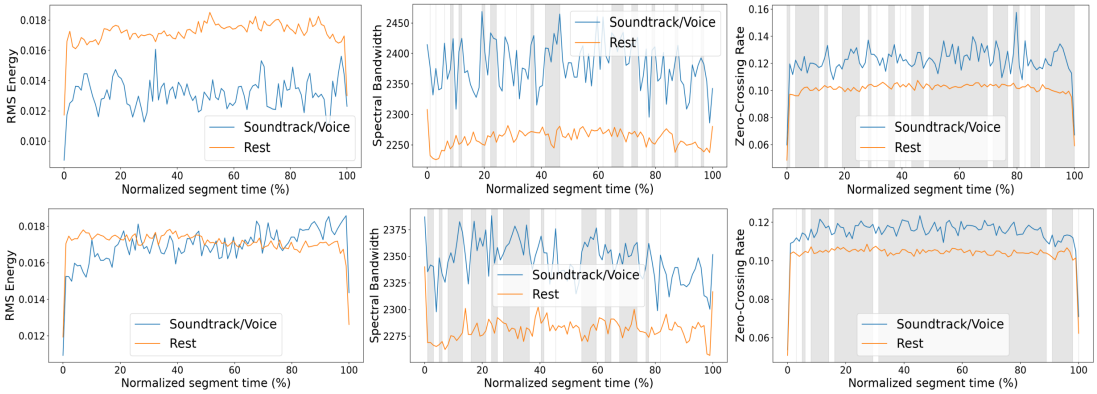


Figure 5: Audio features RMS (left), Spectral bandwidth (center), ZCR (right), shown over a normalized segment duration for each annotator: annotator 1 in top-row, annotator 2 in bottom-row (shaded = FDR q<0.01).

Audio concepts We then compare differences between segments annotated with an audio (voice or soundtrack) concept and those without, over three audio features: Zero-Crossing Rate (ZCR), Spectral Bandwidth, and Root Mean Square (RMS) Energy. The three features are computed with librosa as per-frame time series, then each time series is resampled onto a common normalized time axis (0-100% of the segment's duration) so that curves from segments of different durations can be averaged together. They are shown for both annotators (row-wise) in Fig. 5. A pointwise Mann-Whitney U test at each of the normalized-time bins, FDR-corrected across bins, with significant bins shaded on the time-series plot.

First, we observe with spectral bandwidth that annotators have selected audio concepts for significantly higher-pitched track. This may be due to aspect of sensual, feminine or infantile voices specified in the thesaurus. Second, we also observe that ZCR, that can be interpreted as a measure of the noisiness of a signal, is significantly higher in segments annotated with an audio concept. This may be due to "sensual", "weeping" voices in the thesaurus, that are more unvoiced and hence have a higher ZCR. Finally, RMS can help

estimate the average loudness of an audio track. For annotator 1, we observe that tracks with the audio concept are less loud than those without, while the difference is not significant for annotator 2.

In conclusion, this audio analysis also indicates significant consistency both between the annotators and with external indicators related to the thesaurus. As for Look and Type of shot, the consistency in features despite the low IAA for the audio concepts may suggest aggregating the annotations may be relevant to lower the false negative rate.

3.2.3 Analysis of objectification levels independently from concepts

The 20 films make a total of 43 hours of footage annotated by two annotators, yielding 6072 delimited and annotated segments. We first analyze the annotations of objectification levels, independently from concepts. Fig. 6 (left) shows the distribution of objectification levels in number of occurrences and time duration. EN segments represent 39.7% of segments and 60.1% of total duration, HN 31.4% and 20.2%, and S segments 24.2% and 15.7%, respectively. Fig. 6 (right) depicts the distribution of objectification levels over the film genres in the MObyGaze dataset (genres are shown in Table 17 of the supplementary material): comedy, family and romance are the genres with most S and overall non-EN segments; action, adventure and sci-fi are those with the least.

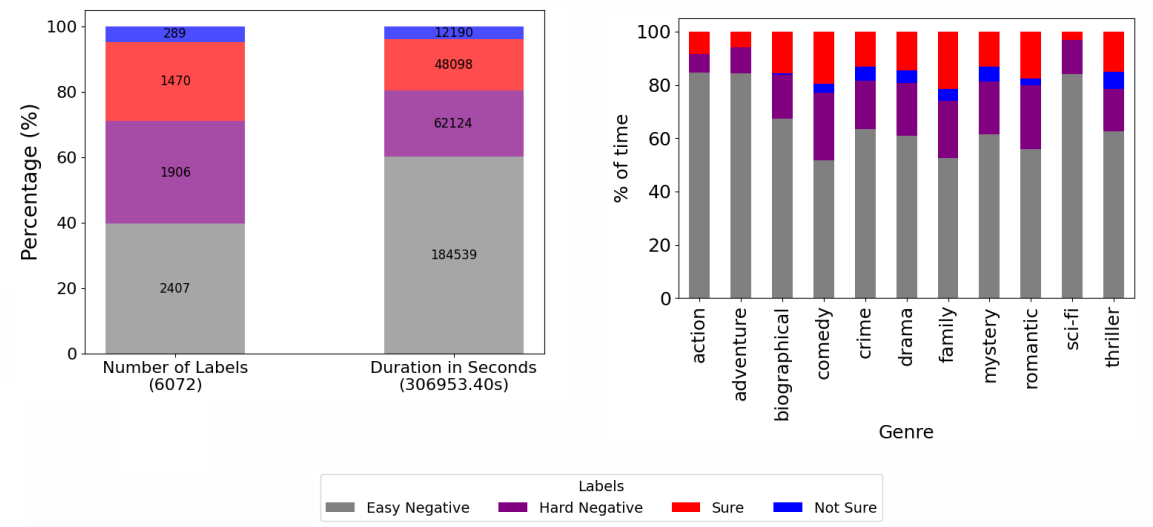
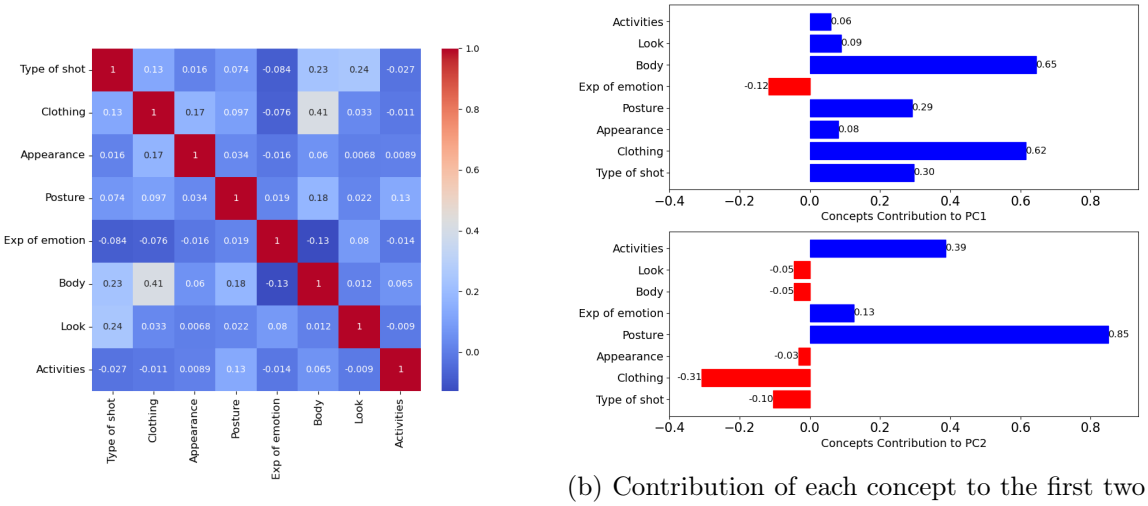


Figure 6: Descriptive analysis of the MObyGaze data. Left: distribution of label frequencies and corresponding durations. Right: Distribution of objectification levels over the film genres.

3.2.4 Analysis of concepts independently from objectification levels

Figure 7a shows the cross-correlation matrix derived from the annotations of visual concepts and reveals distinct patterns of co-occurrence. For instance, the *Body* and *Clothing* concepts are frequently annotated together, as are *Type of Shot* and *Look*. Fig. 7b shows the contribution of each concept to the first two principal components of the principal component



(a) Cross-correlation matrix of concepts in the MObyGaze data.

(b) Contribution of each concept to the first two principal components (PCA), explaining most variance.

Figure 7: (a) and (b): Overview of concept cross-correlation and principal component contributions in MObyGaze data.

analysis (PCA). This analysis reveals which concept associations account for the greatest variance in the data. Notably, the *Body-Clothing* pair contributes most significantly to the first principal component, while *Activities* and *Posture* contributes significantly to the second component.

3.2.5 Analysis of the connection between objectification levels and concepts

The distribution of concepts in connection with the objectification levels is presented in Fig.8, which shows the number of occurrences of each concept, disaggregated over each of the non-EN levels. We can see that the most prominent concepts are *Look*, *Posture*, *Clothing*, *Body* and *Speech*. All concepts have significant representation except for Soundtrack. It is notable that the average number of concepts annotated per segment increases significantly with the level of objectification: the number of concepts for S segments (3.1) is almost twice that of HN segments (1.7). The fact that objectification is a multi-factorial phenomenon interestingly corroborates with results in neuro-psychology where Bernard et al. (2019) showed that clothing alone is not sufficient to produce objectification. Fig. 9 depicts, for each level S and HN, the distribution of concept modalities. We see that the visual modality is most frequent, and relatively more frequent than text and sound in S compared to HN segments. Visual concepts are also split into filmic (gathering *Type of shot* and *Look*) and iconographic properties (the rest of the visual concepts). Fig. 10 represents, for each concept modality, the ratio of occurrences of S (resp. HN) to HN (resp. S) and NS segments. We observe that visual concepts are significantly more specific to S than to HN segments (in the visual filmic modality, Sure segments are ca. 50% more numerous than the other levels).

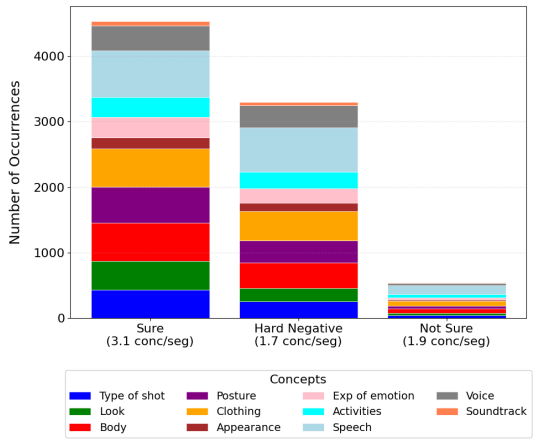


Figure 8: Number of concept occurrences for each objectification level

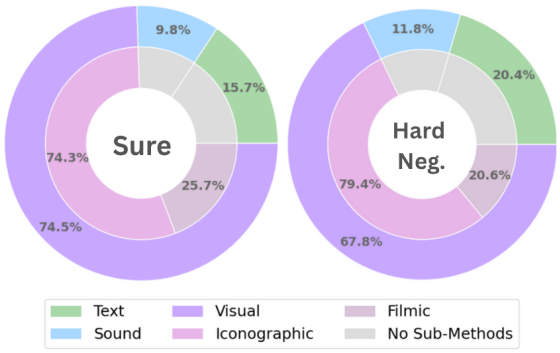


Figure 9: Distribution of concept modalities for S (left) and HN (right) segments, neglecting NS $< 5\%$.

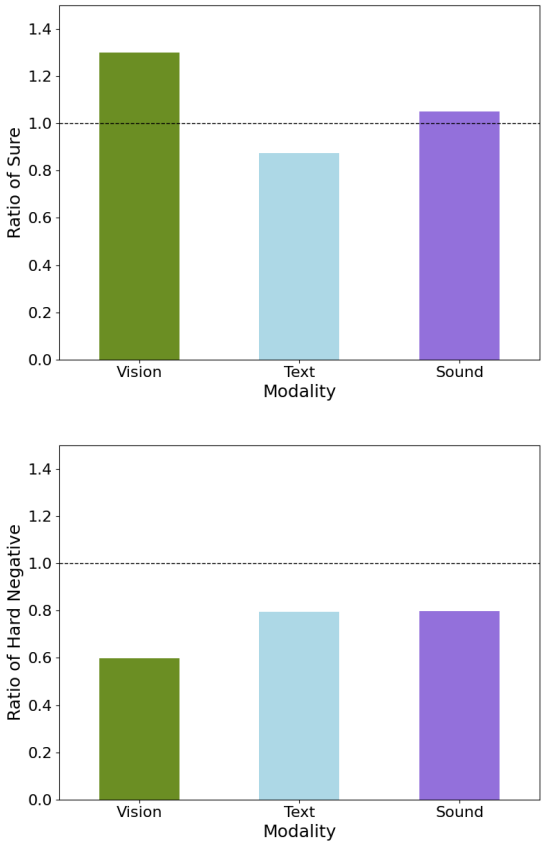


Figure 10: Ratios of concept occurrences. Top: ratio of S to HN and NS segments. Bottom: ratio of HN to S and NS segments for each modality.

We finally examine whether the annotated concepts alone allow us to classify a video clip as showing objectification (S) or as a hard negative (HN), since easy negatives (EN) contain no concepts by definition. Fig. 11 depicts a decision tree to classify a clip between HN and S. It shows that concept patterns more specific to S segments are: *Look*, *Posture*, *Look+Posture*, *Body+Posture*, *Speech+Posture*, *Speech+Body*.

4 New multimodal interpretive tasks for ML: feasibility, model assessment and analysis, benefits from concept annotations

The objective of this section is threefold. After formulating the learning tasks in Sec. 4.1 and describing the common experimental settings in Sec. 4.2, in Sec. 4.3, we study the feasibility of detecting objectification at the level of each concept modality. In Sec. 4.4, we analyze more precisely model design: (i) how hard negative samples affect them, (ii) how multimodality can be approached, (iii) what are the model biases in connection with our

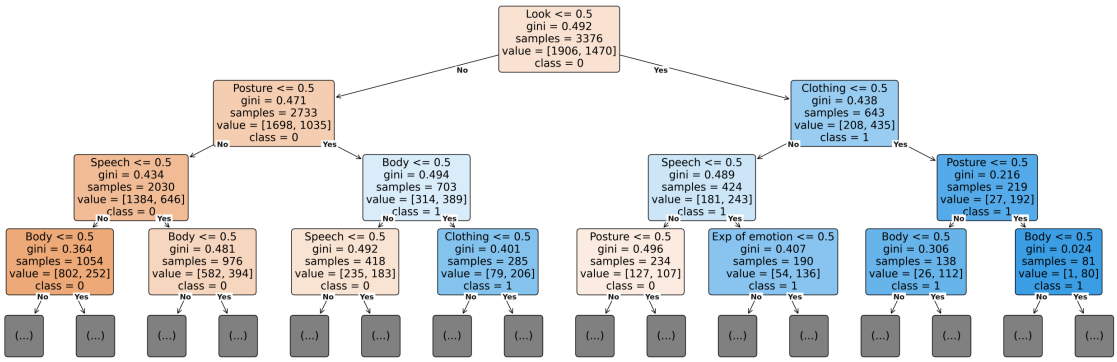


Figure 11: Decision tree to classify between HN and S segment based only on the list of concepts present for each segment. Produced with scikit-learn. Nodes report the splitting rule (feature ≤ 0.5), Gini impurity, sample count, and class distribution. Since all features are binary indicators (1 = concept present, 0 = absent) derived from annotations, splitting at the midpoint 0.5 is equivalent to splitting on presence/absence (not a hyper-parameter). Left branches correspond to the concept being absent, right branches to the concept being present. Node fill color indicates the majority class (orange = class 0, HN; blue = class 1, S), with color intensity proportional to node purity — the proportion of samples belonging to the majority class. Accuracy: 0.768, F1 Score: 0.684 , Weighted F1 Score: 0.759

specific annotation process and how to design more sound models for MObyGaze. In Sec. 4.5, we exemplify how to leverage the rich concept annotation of MObyGaze to improve multimodal and explainable models.

4.1 Formulation of the tasks

The MObyGaze data consists of temporal segments with annotated boundaries, levels of objectification, and associated concepts, for which we define a formal notation as follows. A set of movies $M = \{M_m\}_{m=1}^{20}$ is annotated by labellers $L = \{L_l\}_{l=1}^{2}$, each producing a sequence of annotations $A_{m,l} = \{a_j^{m,l}\}_{j=1}^{N_{m,l}}$ for labeller L_l having delimited $N_{m,l}$ segments for movie M_m. An annotation is a tuple $a_j^{m,l} = (s_j^{m,l}, e_j^{m,l}, l_j^{m,l}, \mathbf{c}_j^{m,l})$ corresponding to start frame $s_j^{m,l}$, end frame $e_j^{m,l}$, annotated level of objectification $l_j^{m,l}$ and list of concepts $\mathbf{c}_j^{m,l}$.

We can hence consider several tasks: (i) detect the presence of objectifying concepts in given video segments, (ii) detect for each video segment the presence (or even level S or HN) of objectification corresponding to the task label, or (iii) localize segments with objectification. In the rest of the article, we first consider task (i) of detecting the presence of specific objectifying concepts (Sec. 4.3.1), then task (ii) of detecting the presence of objectification (Sec. 4.4). We therefore use the term "clip" instead of "video segment" to emphasize that we do not make use of the segmenting information in what follows, instead considering each video segment as an individual video clip (except for the split definition discussed below). We show the feasibility of the localization task (iii) in App. A.5.

4.2 Common experimental settings

We provide below the main aspects of our experimental protocol. All the model details and data preparation for multimodal feature extraction are provided in App. A.4.

Train/validation/test splits To properly test the generalization of the models, we consider no overlap in the movies of the train, validation and test sets. To maximize the size of the training set, we proceed by leave-4-movies-out cross-validation, creating 5 folds each with 4 movies for test, 2 movies for validation and the remaining 14 movies for train (exact movie split in Table 16 in App. A.1). To have a cleaner dataset, we choose to discard NS samples, which correspond to cases where the labeller is uncertain and only represent 4.8% of the samples (see Fig. 6). To account for the class imbalance, data balancing is performed in the training set by oversampling the minority class, and we report PR AUC from the perspective of the positive class, of the negative class and the macro average.

Baselines To assess the feasibility of a task by a ML model, we compare with a trivial random baseline where the positive class is predicted with probability 0.5. We adopt the recommended practices in AI experimental design and analysis (Mozer et al. (2024)) and following Loftus and Masson (1994), we report the "adjusted confidence intervals" for the metric scores when several models are compared. We consider 5 seeds, additionally to the 5 folds, and the adjustment is meant to remove the (fold, seed) variability and enables a more reliable visualization of model effect. Multiple models can then be compared by considering that non-overlapping confidence intervals over a certain metric indicate a statistically significant difference between the models' performance. When a single model is considered, we report the standard (Student) confidence interval. We give more details on our approach of hypothesis testing for model comparison in App. A.3.

Annotator variety For interpretive tasks, we often cannot assume the existence of a single gold label (absolute single truth being accessible or to be approximated from the label statistics) and Bucarelli et al. (2023); Uma et al. (2021) have studied how to best consider label diversity (i.e., different annotators assigning different labels for the same item) in both model training and evaluation. When the number of labels per item is insufficient, or the level of noise/disagreement is high, Wei et al. (2023) show that label separation is preferable to label aggregation for training. For this reason, and to avoid any confusion due to data aggregation in the analyses, we generally consider only one annotator after having verified model performance consistency across annotators. Specifically, we verify in Tables 5 and 12 that:
• we obtain similar results over both annotators when we aim to generalize across movies only and have the same annotator in train and test;
• the models can generalize across different annotators in train and test.

4.3 Concept detection

Over each of the three modalities taken independently, we assess the feasibility of detecting each concept defined in the thesaurus. The model choices presented below build on the extensive literature in multimodal tasks, e.g. Das et al. (2023); Koushik et al. (2025).

Table 4: Performance for the task of detecting components of visual objectification, i.e., the presence of at least one visual concept (with adjusted confidence intervals)

Model	PR AUC Positive	PR AUC Negative	PR AUC Macro
X-CLIP+MLP	0.703[±0.005]	0.731[±0.008]	0.717[±0.006]
X-CLIP+Transf	0.719[±0.005]	0.758[±0.008]	0.739[±0.006]
random	0.47	0.53	0.50

4.3.1 Vision modality

We first investigate whether objectifying concepts can be detected from the MObyGaze dataset (task (i) in Sec. 4.1). For the vision modality, we consider the following models.
X-CLIP+MLP: We adapt X-CLIP (Ni et al. (2022)), an extension of CLIP for videos (Radford et al. (2021)). We keep the pre-trained model frozen and extract a 512-dimensional feature vector on every window of 16 frames, with a stride of 16, for each input video segment. The obtained vectors are mean-pooled, and the resulting vector fed to an MLP with 2 hidden layers with a final softmax layer of classification. This model is trained using fully-supervised learning with a cross-entropy loss.
X-CLIP+Transf: The X-CLIP feature vectors above are first projected to a 128-dimensional space, and appended with a CLS (classification) token to be learned. This concatenation is then passed through two self-attention layers, each followed by two linear layers with a hidden dimension of 1024 and GeLU activations. Finally, a classification head composed of two linear layers with ReLU activations is applied to the CLS token. The model is trained using cross-entropy loss.
Task 1 (Coarse-grained/Overall concept detection): We first consider the binary classification task of detecting the presence of any visual concept, in which case the positive samples correspond to the clips annotated with at least one visual concept.

Table 4 shows the performance of both models. The first observation is that both models outperform the random baseline statistically significantly and with a very large effect size – large Cohen's d's around 1 for the three metrics considered (as $\hat{\sigma^{2\prime}}$, defined in A.3, is systematically smaller than 0.001, resulting in narrow confidence intervals and very large Cohen's d). The second observation is that the Transformer model is statistically significantly better than the MLP model, with an absolute difference in PR AUC of about 0.02.

We show in Table 5 that these results hold for both annotators, and when we train on one annotator and test on the other. It is another indicator of the consistency of the MObyGaze dataset.

These results show that it is feasible for vision models to tackle the detection of visual objectification, even if they also show there is significant room for improvement and interesting research challenges.
Task 2 (Fine-grained/Individual concept detection): Second, we consider the task of detecting each individual visual concept (as many tasks as concepts), in which case the positive samples correspond to the clips annotated with this visual concept.

Table 6 shows the performance of detecting each individual visual concept. We observe that several concepts can be detected with non-trivial performance, but only 3 visual concepts out of 8 with a large gain in Macro PR AUC (>0.65) of X-CLIP+Transf over the random baseline (for which Macro PR AUC is 0.5). These concepts are Body, Clothing and Posture

Table 5: Performance for the task of detecting components of visual objectification, intra- and inter-annotator comparison, results of the X-CLIP+Transf model on PR AUC Macro (with adjusted confidence intervals)

Train annotator \Test annotator	Annotator 1	Annotator 2
Annotator 1	0.697[±]0.006]	0.719[±0.005]
Annotator 2	0.7[±0.006]	0.739[±0.005]

Table 6: Performance of model X-CLIP+Transf for the task of detecting individual visual concepts (with standard confidence intervals)

	Activity	Appearance	Body	Clothing
PR AUC Positive	0.293[±0.107]	0.078[±0.08]	0.516[±0.113]	0.375[±0.101]
PR AUC Negative	0.936[±0.053]	0.956[±0.061]	0.912[±0.084]	0.926[±0.056]
PR AUC Macro	0.614[±0.08]	0.517[±0.054]	0.714[±0.07]	0.65[±0.071]
	Expr. emotion	Look	Posture	Type of shot
PR AUC Positive	0.222[±0.088]	0.143[±0.096]	0.391[±0.11]	0.27[±0.114]
PR AUC Negative	0.958[±0.048]	0.937[±0.062]	0.917[±0.062]	0.9[±0.08]
PR AUC Macro	0.59[±0.059]	0.54[±0.062]	0.654[±0.071]	0.585[±0.079]

and, interestingly, are the most common in the dataset (see Fig. 8). Other well-represented concepts are Activity, Expr. of emotion and Type of shot : they are detected with non-trivial performance (statistical significance), but with a small difference compared to the random baseline. We hypothesize that such concepts are more subtle (see the Thesaurus in Fig. 2), and relevant information may not be (sufficiently) present in the X-CLIP features for Transformer layers to be trained to properly detect them with the low amount of positives. It also shows the limits of positive oversampling to counter the strong class imbalance. Lastly, Appearance and Look are poorly detected, and scarcely represented in the dataset.

4.3.2 Textual modality

For the textual modality, we consider the following models.

BERT+Transf: We choose to benchmark a masked language model of type bidirectional encoder. We specifically consider a frozen BERT model to extract contextualized embeddings, which are then passed into a lightweight Transformer-based classifier similar to the one presented for the vision modality. This model is referred to as BERT+Transf. We only present these results for the sake of concision, but we also considered a distilled version of RoBERTa Liu et al. (2019), namely DistillRoBERTa[1], with two fine-tuning strategies (one keeping frozen only the embedding layers and the other tuning only the 769 parameters for the classification head applied to the CLS token). Results were quantitatively similar to BERT+Transf.

BERT+MLP: Same BERT features as above are fed into the MLP similar to the one for the visual modality.

1. `https://huggingface.co/distilbert/distilroberta-base`

Table 7: Performance of model BERT+Transf for the task of detecting the (textual) speech concept (with adjusted confidence intervals)

Model	PR AUC Positive	PR AUC Negative	PR AUC Macro
BERT+MLP	0.411[±0.07]	0.895[±0.003]	0.653[±0.004]
BERT+Transf	0.407[±0.07]	0.893[±0.003]	0.650[±0.004]
random	0.214	0.786	0.500

Table 8: Performance of model AST+Transf for the task of detecting the audio concept (with adjusted confidence intervals)

Model	PR AUC Positive	PR AUC Negative	PR AUC Macro
AST+MLP	0.190[±0.05]	0.835[±0.006]	0.513[±0.005]
AST+Transf	0.183[±0.05]	0.833[±0.006]	0.508[±0.005]
random	0.174	0.826	0.500

Task (Textual Speech concept detection): We consider the task of detecting the textual concept (when speech is annotated as objectifying owing to what is said), in which case the positive samples correspond to the clips annotated with the speech concept.

Table 7 shows both models outperform the random baseline significantly. However, there is no significant difference between having a Transformer or an MLP processing BERT features. Again, this demonstrates the feasibility of detecting objectification in text but also shows room for improvement.

4.3.3 Audio modality

For the audio modality, we consider the following models.
AST+Transf: Audio Spectrogram Transformer (AST) Gong et al. (2021) is used to extract audio features (see App. A.4 for details). These features are then fed as input into a Transformer model similar to the ones for the visual and textual tasks.
AST+MLP: Same AST features as above are fed into the MLP similar to the one for the visual modality.
Task: We consider the task of detecting the presence of any audio concept (voice or soundtrack), in which case the positive samples correspond to the clips annotated any of each.

The results reported in Table 8 show no performance improvement compared to the random baseline. This negative result must however be put in perspective with the choice of the feature extractor: while we wanted to have a homogeneous choice of pre-extractor architecture over the modalities, whereby the choice of AST feature, the results in Sec. 3.2.2 indicate that lower-level features, albeit simpler, would provide more useful information compared to AST (which has been trained on non-human sounds).

4.4 Objectification classification

We now investigate the classification of the ojectification levels from the MObyGaze dataset (task (ii) in Sec. 4.1). We first work with the objectification rating labels (EN, HN, S) to analyze the impact of HN (Sec. 4.4.1), then we study multimodal objectification detection

Table 9: Performance for classification of objectification levels from the visual modality only (with adjusted confidence intervals)

Task	Model	PR AUC Positive	PR AUC Negative	PR AUC Macro
EN∪HN vs S	X-CLIP+Transf	0.337[±0.008]	0.890[±0.006]	0.613[±0.006]
	random	0.198[±0.008]	0.802[±0.006]	0.500[±0.006]
EN vs HN vs S	X-CLIP+Transf	-	-	0.489[±0.003]
	random	-	-	0.333[±0.003]

(Sec. 4.4.2), and finally biases and model failures (Sec. 4.4.3) to propose improvements (Sec. 4.4.4).

4.4.1 Impact of hard negative samples on objectification classification

In Sec. 4.3.1, we have considered binary classification for the visual concepts where any clip annotated with a visual concept is considered positive. However, a number of such clips are labeled Hard Negative (HN), meaning that objectifying visual elements are present but are not judged sufficient by the annotator to consider a Sure level of objectification. Here we focus on the X-CLIP+Transf visual model and analyze the impact on model performance of assigning HN clips to the negative class, leaving the positive clips being S only with at least a visual concept.

Table 9 shows that the difference between the X-CLIP+Transf model and the random baseline is still statistically significant both for a binary task where HN clips are in the negative class, and a 3-class problem where HN makes up a separate class. Comparing the results with the case when HN are in the positive class (since we consider the presence of any visual concept as a positive in Table 4), we however observe that HN are strong confusers, which shows the importance of fine-grained annotation for interpretive tasks, as also shown by Samory et al. (2021).

4.4.2 Multimodal objectification detection

We now consider whether combining information from multiple modalities could benefit objectification detection.

Models: We consider the single-modality models for vision (named V) and text (named T) introduced in Sec. 4.3. We consider a multimodal model named VT where, for each clip, the visual (X-CLIP) and textual (BERT) token vectors are projected to a common feature dimension, concatenated along the token axis and fed to a Transformer model similar to that also described in Sec. 4.3 for the individual modalities. We consider feature concatenation following Koushik et al. (2025) who found that attention-based fusion never obtains better performance than concatenation for multimodal hate detection. More details are provided in App. A.4.5.

Task (Multimodal Objectification detection EN vs HN∪S): The task is to detect HN or S samples from both the visual and textual modality. All three models, which differ in their inputs, are trained and tested on the dataset (still leaving movies out between folds) where the positive class is made of the samples with at least a visual or textual concept present.

Table 10: Performance of multimodal models on the task EN vs HN∪S (with adjusted confidence intervals)

Model	PR AUC Positive	PR AUC Negative	PR AUC Macro
V	0.781[±0.006]	0.701[±0.009]	0.741[±0.006]
T	0.696[±0.006]	0.636[±0.009]	0.666[±0.006]
VT	0.769[±0.006]	0.711[±0.009]	0.740[±0.006]

Table 10 shows the performance of the models. We observe that model VT outperforms model T and obtains results similar to model V. We also observe with VT that such an early fusion of modalities does not bring a quantitative advantage over having the visual modality only. This is a known challenge in multimedia fusion that properly weighing the importance of each modality for each sample is not trivial to learn (Zheng et al. (2023); Du et al. (2025)). We indeed show in Sec. 4.5.1 how per-modality training based on the additional concept information can be leveraged to improve multimodal models.

4.4.3 Analysis of annotator and model biases

We now further analyze the performance of the vision model with disaggregated quantitative and qualitative analyses. We then propose model improvements in Sec. 4.4.4.

Analysis of annotator bias Fig. 12a shows the performance of model X-CLIP+Transf on task EN vs HN∪S for visual concepts only, disaggregated over the 10 movies used in the validation sets of all the folds. Colored bars further depict the accuracy disaggregated over the positive and negative classes. We observe that certain movies have highly heterogeneous results, such as *Crash*, *Forrest Gump* and *Up in the air* for which positive clips are more often misclassified as negative, while for others like *The Ugly Truth*, negative clips are frequently misclassified as positive. Let us note that *The Ugly Truth* is highly stereotypical with the highest fraction of positive clips (75%, see Table 19 in App. A.2).

Fig. 12b shows qualitative examples of such misclassifications in these movies. Two clips, which, taken independently may seem similar, but have been annotated (by the same annotator) differently (positive for *Up in the Air* and negative for *The Ugly Truth*). From there, we can suppose that the annotators, who have annotated an entire movie at once while watching it, make contextual and contrastive annotations: they tend to self-regulate by trying to balance the number of positive and negative labels they assign within a movie, limiting the number of positives in stereotypical movies in comparison of strongly objectifying scenes in the same movie, while annotating as positive in the less stereotypical movies scenes that are less obviously positive when taken independently but are labelled positive in contrast with other scenes within that movie. This finding of movie-dependent annotation aligns well with important knowledge established in psychophysics according to which humans naturally assess quality through comparison rather than in absolute terms, and this has been a central consideration in several domains based on human assessments, specifically driving the methodologies in image/video quality assessment (IQA/VQA). This supposition is ground for our hypothesis on how to improve model in Sec. 4.4.4.

Analysis of model bias We make a second observation of how the movie context may bias the model performance. In Fig. 13, a clip from the movie *The Help* is annotated as



(a) (b)

Figure 12: (a) Accuracy of the X-CLIP+Transf model, disaggregated by movie and by true label, on the movies in the validation sets (note that the accuracy on positive corresponds to recall). (b): Illustration of the dependence of annotations on the movie context: clips from different movies annotated differently while reverse labels may have been assigned if the clips were considered outside their respective movies.



Figure 13: Illustration of context bias learned by the model. Top: The original clip is annotated positive but the predicted probability for the positive class is 0.231. Bottom: Replacing frames where visual objectification does not occur (that depict the character of color) with semantically equivalent frames from a train movie (movie with the least fraction of positive clips) significantly increases the probability of the positive class.

positive, but the predicted probability of the positive class is only 0.231, while being typically positive and not a corner case according to the thesaurus. To check whether the model may be confused by elements of context not seen in train, we engineer (bottom part of the figure) a new clip preserving the objectifying frames and replacing the non-objectifying frames with semantically equivalent frames from a train movie (movie with the least fraction of positive clips). This operation increases the probability of the positive class from 0.231 to 0.498. We consider this another indication that differences in context, aside from the very type of objectification, between train and validation/test can (expectedly) confuse the model.

4.4.4 Improving objectification detection with differential representation

Based on the observation of biases above, we propose the following model improvement.

Hypothesis: We hypothesize that a model that can contrast a clip feature against the entire movie to then make the class prediction will be able to better generalize over movies.
Model: To test this hypothesis, we design a new model that we call **Differential**, defined as follows. Let f be the representation component of the X-CLIP+Transf model, returning the CLS-token output that is fed to the MLP classifier. For a clip x, $f(x) \in \mathbb{R}^{128}$ thus denotes the task features extracted by the trained Transformer from input x. For a given movie M, we define the movie context as:

$$c_M = \text{mean}_{\{x \in M\}} f(x)$$

The differential paradigm simply consists in introducing an additional operation before the classification head, outputting a *differential representation*

$$f_{\text{diff}}(x, c_M) := f(x) - c_M$$

that is fed to the MLP instead of $f(x)$. During the training stage, c_M is replaced by its batch counterpart:

$$c_M(\text{batch}) = \text{mean}_{\{x' \in M \cap \text{batch}\}} f(x')$$

In practice, in order to ensure that $c_M(\text{batch})$ is close enough to c_M, we construct the batches so that they contain 32 clips per movie: with 2 different movies per batch, this represents a total batch size of 64.
Results: Table 11 shows the test results in terms of classification performance (PR AUC) but also in terms of heterogeneity in performance between movies (right-most column), defined as the average over the folds of the standard deviation of the accuracy over all four movies of a fold test set. The larger the heterogeneity, the more differently the model performs from a movie to another; thus, a model with lower heterogeneity is likely to better generalize to new movies. We observe that the differential paradigm significantly improves the performance over the standard X-CLIP+Transformer model in terms of PR AUC, but also significantly decreases the movie performance heterogeneity, hence verifying the hypothesis we have emitted from our previous observations on the validation sets. Table 12 shows that this holds for both annotators (we omit confidence intervals for the sake of clarity). Under this new paradigm, we verify that models maintain cross-annotator generalization. Table 12 shows that this is the case: a model trained to suppress the movie bias for one annotator relieves the movie bias for another annotator in test: the differential model maintains or improves the average classification performance in every case, and decreases movie performance heterogeneity in 3 cases in 4. Fig. 14 further shows the performance evolution between both models, over the different movies and classes, clearly showing the improved heterogeneity of the differential model without sacrificing classification accuracy. The better performance of the differential model supports the assumption underlying it, according to which human annotations are contextual.

4.4.5 MLLM

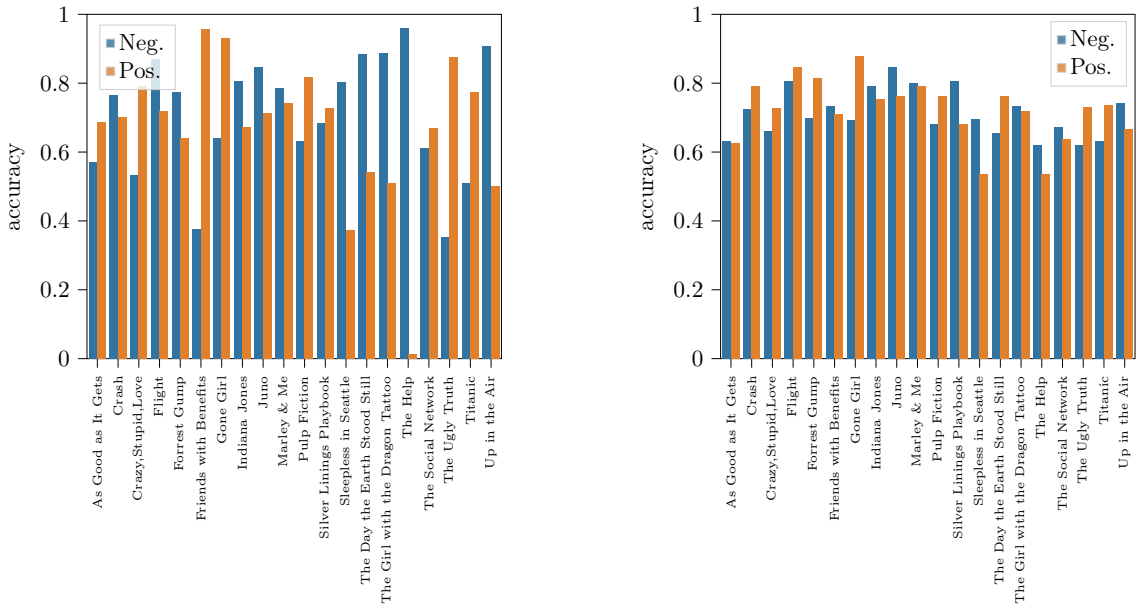
We now provide preliminary results with Multimodal Large Language Models (MLLMs). We consider the following open-source models: `InternVideo2.5`, `InternVL3`, `VideoLLaMA 3`, `LLaVA-Video` and `Qwen2.5-VL`. We provide the details on prompts, formatting and hyper-parameter choices in App. A.4.6.

Table 11: Performance for visual objectification detection (adjusted confidence intervals)

Model	PR AUC Pos. (↑)	PR AUC Neg. (↑)	PR AUC Macro (↑)	Heterogeneity (↓)
X-CLIP+Transf	0.719[±0.005]	0.758[±0.007]	0.739[±0.003]	0.073[±0.005]
Differential X-CLIP+Transf	0.730[±0.005]	0.784[±0.007]	0.757[±0.005]	0.061[±0.005]

Table 12: Performance for visual objectification detection, model comparison with intra- and inter-annotator comparison, results as PR AUC Macro (↑)| Heterogeneity (↓)

Train annotator \Test annotator	Annotator 1		Annotator 2	
Annotator 1	Non-diff 0.697 | 0.079	Diff 0.704 | 0.068	Non-diff 0.719 | 0.063	Diff 0.721 | 0.055
Annotator 2	Non-diff 0.7 | 0.057	Diff 0.711 | 0.064	Non-diff 0.739 | 0.073	Diff 0.757 | 0.061



(a) Standard X-CLIP+Transf

(b) Differential model

Figure 14: Accuracy on the test set disaggregated by movie and by true label. (a) Original model X-CLIP+Transf. (b) Differential X-CLIP+Transf.

Zero-shot inference Table 13 reports Macro F1 (as the answers provide discrete classification results) and shows the performance obtained for the task EN vs HN∪S. For comparison, the first line shows the performance of a X-CLIP+Transf model. While this task is balanced (48% of samples are annotated as positive), we observe a general tendency of MLLMs to predict the negative class. In particular, the models `Qwen2.5-VL` and `VideoLLaMA 3` almost systematically predict the negative class, leading to a F1-macro performance worse than random. `LLaVA-Video` is the only MLLM model predicting positive in a similar proportion to the true labels, but performs hardly better than random in terms of Macro F1. The best MLLMs models are `InternVideo2.5` and `InternVL3`, which achieve non-trivial performance;

Table 13: Zero-shot performance on the task EN vs HN∪S

Model	Accuracy	Macro F1	proportion of predicted positive
X-CLIP+Transf	0.685	0.677	0.51
`InternVL3`	0.640	0.612	0.28
`InternVideo2.5`	0.615	0.554	0.17
`LLaVA-Video`	0.548	0.547	0.49
`Qwen2.5-VL`	0.538	0.383	0.025
`VideoLLaMA 3`	0.535	0.381	0.03

Table 14: Comparison of zero-shot performance of `Qwen2.5-VL` for different input modalities

Modalities	Accuracy	Macro F1	proportion of predicted positive
text	0.79 [± 0.025]	0.537 [± 0.02]	0.052 [± 0.013]
text+video	0.454 [± 0.025]	0.361 [± 0.02]	0.052 [± 0.013]
video	0.538 [± 0.025]	0.383 [± 0.02]	0.025 [± 0.013]

however, because of the aforementioned tendency, they still perform worse. Table 14 compare results of `Qwen2.5-VL` when fed with different modalities, without changing the above conclusion.

Few-shot inference We inspire on recent works by Fujii et al. (2026) and Kim et al. (2025) to test few-shot learning with MLLMs. The choice of few-shot demonstrations is characterized by the initial demonstration pool, and the selection strategy from this demonstration pool. For the demonstration pool, we consider the training and validation set of each fold. For the selection strategy, we use the X-CLIP embeddings of the clips to compute the cosine similarity between the test sample and all observations from the demonstration pool. We select the 4 demonstration examples with the largest similarity under the condition the example classes are balanced. The order in which examples are presented is random. In few-shot, compared to the zero-shot setting, for a given task description there is more flexibility in the structure of the prompt: whether few-shot examples are presented at the beginning, or at the end, and how they are described.

Table 15 shows that few-shot learning does not readily improve the results. Qualitative analysis shows that, in few-shot settings, models can hardly segment the different video clips that they are given. As a consequence, it is not surprising that their Macro F1 is hardly better than random. An interesting observation, however, is that while `VideoLLaMA 3` systematically predicts negative, in the few-shot setting the proportion of positives predicted by `Qwen2.5-VL` increases.

Table 15: Performance on EN vs HN∪S with few-shot inference, with examples first

Model	Accuracy	Macro F1	proportion of predicted positive
`Qwen2.5-VL`	0.595	0.552	0.22
`InternVideo2.5`	0.543	0.517	0.30
`VideoLLaMA 3`	0.527	0.371	0.03

While MLLMs are interesting models for the new video interpretive task we introduce, this section shows that they do not readily provide satisfying results and require more work. In the line of recent works for video anomaly detection by Huang et al. (2025), approaches with supervised fine-tuning and reinforcement learning leveraging rich concept supervision, are promising perspectives.

4.5 Interest of MObyGaze to improve multimodal and explainable models

We now demonstrate two applications of interest for MObyGaze. In Sec. 4.5.1, we show how rich concept annotations can be leveraged to improve multimodal models for our video interpretation task. In Sec. 4.5.2, we show how MObyGaze raises new challenges for explainable AI and exemplify a contribution to achieve better accuracy-interpretability trade-off for our task.

4.5.1 Improving multimodal models with concept supervision

The objective here is to show that concept supervision can help multimodal training. For this, we first introduce the idea of Concept Modality Specific Dataset (CMSD) below. The models used for the experiments below are the same as in Sec. 4.3 with the same unimodal pre-extractors. Only, for the sake of computational intensity and to relate to the approach taken by Bose et al. (2023), we consider a Perceiver IO architecture instead of the usual Transformer in X-CLIP+Transf or BERT+Transf (we highlight that Perceiver IO is hence not fed with raw data but with pre-extracted features). This work is more detailed in Ancarani et al. (2025a).

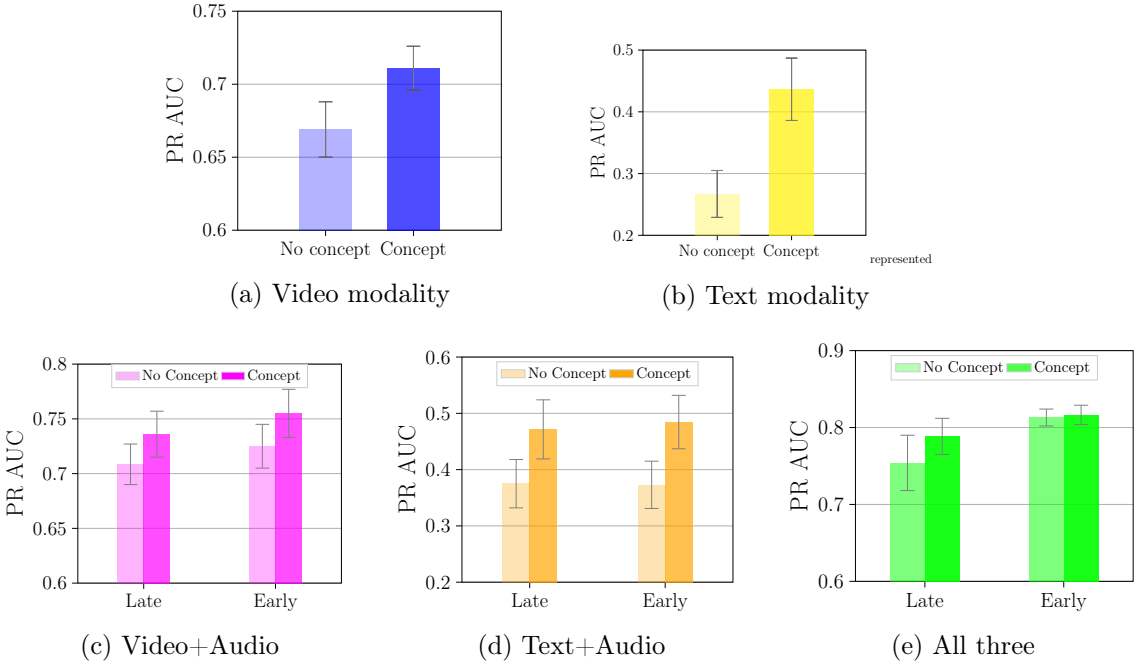
As we have seen in Sec. 4.4.2, benefiting from several modalities with varying levels of relevance is challenging. Here we show that the concept annotations can be used to improve multimodal models. Depending on how concept information is considered, we define two dataset versions: the Concept Modality Agnostic Dataset (CMAD) and the Concept Modality Specific Dataset (CMSD). In CMAD, a sample is labeled as positive regardless of the modality in which the concept appears. In contrast, CMSD labels a sample as positive only if the annotation contains at least one concept that is relevant to the specific modality. We study the interest of training unimodal and multimodal models on CMAD or CMSD, while always testing on CMSD. Indeed, CMSD can be seen as the truthful label we desire the model to output as, e.g., we want a vision model to predict the negative class if a clip is labelled as objectifying but not owing to visual factors.

When considering a single modality (vision or text), we verify in Fig. 15a and Fig. 15b that unimodal models trained on CMSD ("Concept" in the figure) perform significantly better than those trained on CMAD ("No concept" in the figure). This already highlights the interest of concept annotation in multimodal tasks.

Fig. 15c-15d present the performance of bimodal models, and Fig. 15e that of a trimodal model. We consider two major fusion strategies: mid-Early fusion ("mid-Early") where modality tokens are concatenated at input as for our model in Sec. 4.4.2, and Late fusion ("Late") where unimodal models are first pre-trained (on CMAD and CMSD), with their individual decision later fused with a simple max operation. For the bimodal models (Fig. 15c-15d), we again observe the interest to use concept information (CMSD) at training time, regarless of the fusion type. For the trimodal model with mid-Early fusion (Fig. 15e) can

only be trained on CMAD. However, Late fusion, while performing worse that mid-Early fusion when trained on CMAD, is able to catch up to mid-Early fusion performance when its composing unimodal models are trained on CMSD. This is all the more interesting that Late fusion architectures are more explainable that mid-Early fusion because the information on what modality drives the decision is readily accessible.

While these results show the interest of exploiting concept annotation in the design of multimodal models for MObyGaze, we further the question of designing explainable models for MObyGaze in the next section.



(a) Video modality (b) Text modality

(c) Video+Audio (d) Text+Audio (e) All three

Figure 15: Performance of unimodal (a, b) bimodal (c, d) and trimodal (e) models trained without or with concept supervision. Results are on the test set of CMSD.

4.5.2 Improving explainable models for challenging video interpretation

In this section we exemplify some challenges our MOByGaze dataset and task raise for existing explainable AI approaches. The results below are more extensively presented in Tores et al. (2025). We also show another example of how concept supervision can help for the downstream classification task, and which challenge remain.

The use of auxiliary conceptual information in image classification tasks is at the heart of important work in explainable deep learning, especially with so-called Concept Bottleneck Models (CBMs) introduced by Koh et al. (2020). They have given rise to a line of subsequent works, such as Concept Embedding Models (CEMs) by Zarlenga et al. (2022a). These methods are classically developed for image classification tasks and datasets with simple conceptual information (e.g., the color of a bird's beak). Ramaswamy et al. (2023) have analyzed the factors that participate in the success of concept-based explanations: explanatory concepts should be limited (no more than 32) to be truly useful for human interpretation,

and they must be easy to learn. MObyGaze, which centers on a complex multimodal video interpretation task aided with high-level concepts, therefore challenges this validity domain for concept-based models because the complexity of the end label would require numerous concepts if these were as simple as color or texture areas (such as in Wah et al. (2011) where there are 112 concepts). Instead, we choose with MObyGaze to have a limited number of concepts (11 in total, 8 visual) by having each concept itself interpretive and hence possibly not easy to learn. We investigate the ensuing challenges for concept-based models.

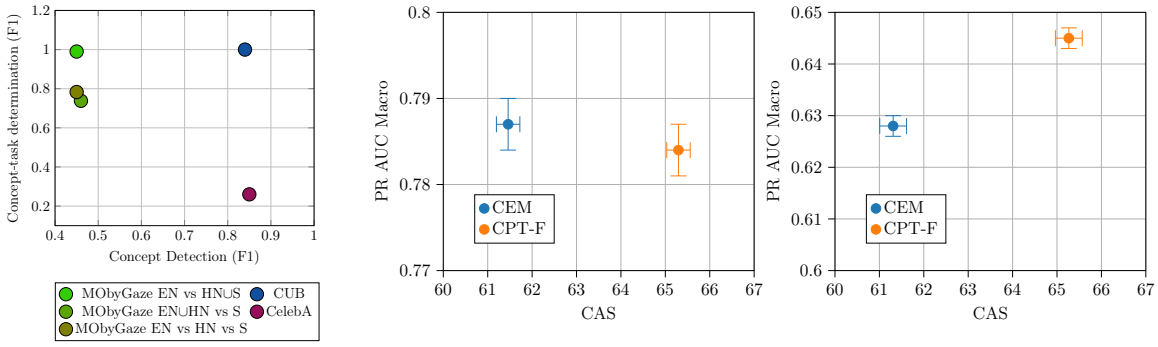
Analysis of MObyGaze for existing concept-based models We first position MObyGaze with respect to two classical concept-based image datasets on the dimensions of *Concept-task determination* and *Concept detection*, shown in the y-axis and x-axis of Fig. 16.a, respectively. The *Concept-task determination* quantifies how much the knowledge of which concepts are present is sufficient to determine the end label, and is measured by the performance of a reference CBM fed with the ground-truth presence of each concept. The *Concept detection* score is the mean F1-score of concept detectors jointly trained within the reference CEM. We consider three datasets: two reference datasets annotated with concepts, CUB (Wah et al. (2011)) and CelebA (Liu et al. (2015)), and MObyGaze with visual concepts only. CUB is a dataset of bird images annotated for their species and with 112 curated concepts (such as beak color). CelebA is an image dataset of human faces annotated with their ID and a set of 40 concepts (including "eyeglasses" but also ethically questionable ones such as "attractive" or "narrow eyes").

Fig. 16a shows that, while CEM achieves good concept detection on CUB and CelebA, its detection performance on MObyGaze (despite being adapted with proper loss weights to balance heterogeneous concept representations) is significantly lower, and is worse than concept detectors trained outside CEM and shown in Sec. 4.3.1. This illustrates a difficulty of legacy concept-based models such as CBM and CEM to deal with complex tasks with relatively more complex concepts.

A new concept-based model for MObyGaze To tackle this challenge, in Tores et al. (2025), we introduce a novel CEM-type architecture that we name Concept Pre-trained Transformer Frozen (CPT-F). It replaces MLP-based layers by Transformer-based layers and leverages frozen pre-trained concept detectors, specifically incorporated to address the severe class imbalance present in the MObyGaze dataset.

Fig. 16b and Fig. 16c show the **End-task performance - Interpretability trade-off**. Performance (y-axis) is measured in PR AUC. Interpretability (x-axis) is measured with the Concept Alignment Score (CAS) introduced by Zarlenga et al. (2022a). CAS represents the homogeneity of concept labels amongst clusters of samples in the latent space used for end-task decision. High values of CAS indicate a latent space shaped by concepts, hence a decision relying heavily on concept presence but represented in a latent space, without constraining the decision to a binary concept bottleneck like with CBMs.

We observe that for the end-task of objectification classification EN vs HN∪S (Center (b)), which is the easiest in that it boils down to detecting the presence of at least one concept, CPT-F does not allow to gain in task performance compared with CEM, but it increases interpretability thanks to improved concept detection. For the harder end-task of objectification classification EN∪HN vs S (Left (c)), CPT-F allows to gain both in



(a) Datasets (b) EN vs. HN∪S (c) EN∪HN vs. S

Figure 16: (a): Comparison of the datasets on the concept detection difficulty (x-axis) and determination of the downstream task (y-axis). (b) and (c): End-task performance-Interpretability trade-off in terms of PR AUC and CAS for the tasks of objectification detection.

interpretability and performance, as the difference between HN and S requires finer concept counts.

It is worth noting that, despite the concepts being high-level and hence harder to detect on an individual basis, they nonetheless help for the downstream classification task with the proposed architecture: results in PR AUC Macro in Table 4 and Fig. 16 can be compared and

- on the EN vs HN∪S task, PR AUC Macro: X-CLIP+Transf 0.739, CEM 0.785,
- on the EN∪HN vs S task, PR AUC Macro: X-CLIP+Transf 0.613, CEM 0.628.

We have therefore shown that MObyGaze represents a valuable asset to design better and more effective explainable models for more interpretive tasks.

5 Broader Impact Statement

Limitations The main limitation of the MObyGaze dataset is that the annotation has been performed at scene-level, not allowing for supervision to learn longer-term objectification patterns. These are known to relate to narratology and occur recurrently throughout a movie, such as tropes as studied by Su et al. (2021). We intend to extend the dataset with such longer-term patterns thanks to the annotation tool, which allows for a multi-level and evolving thesaurus. It will also be possible by the availability of the experts who have developed fine-grained memory of the 20 movies. Another limitation is the low number of annotators, albeit their expertise, and the limited set of movies tied to the original definition of male gaze within the Hollywood film industry.

Applications We have shown that detecting objectification, as defined for the MObyGaze dataset, is accessible to current vision and text models, with significant room for improvement. Immediate challenges are hence in designing models able to learn better representations of complex concept instances. This includes in particular designing explainable models to better characterize complex temporal and multimodal objectification patterns, which can

in turn enrich qualitative studies by media scholars. Sub-constructs detection is also an an interesting future in this interdisciplinary context. The MObyGaze dataset can also be used to study the fairness of existing computer vision models: person detectors and human pose estimators may miss the presence of characters onscreen, compromising the study of how certain patterns correlate with certain human groups, if the humans are often mis-detected for these patterns (e.g., shots with headless body parts as shown by Wu et al. (2022)). Finally, the social purpose of this work is to feed public debate and reflection on the tangibility of subtle widespread audiovisual patterns conveying biased gender representations. Designing and making publicly available models to detect objectification patterns can help raise awareness, but also serve for filmmaking training.

Potential misuse and mitigation measures As we expand in the datasheet in App. A.1 under section Uses, the MObyGaze dataset should not be used for tasks such as content filtering, defining regulatory standards, or censorship of media content. Another potential negative impact would lie in the interpretation of the outputs of models, being it by social science researchers assisting their qualitative studies with models trained on MObyGaze to detect objectification, or by amateurs. We detail these cases we envision, and the corresponding proactive measures, which involve the licensing prohibiting commercial usage and the dataset documentation highlighting that the models trained on the MObyGaze dataset must be used in a clear ethical framework: that of making subtle patterns of representation disparities visible and contributing responsible usages.

Acknowledgments and Disclosure of Funding

This work was supported by the French National Research Agency through the ANR TRACTIVE project ANR-21-CE38-0012-01, and through the France 2030 investment plan under reference number ANR-15-IDEX-01. This work was partly supported by EU Horizon 2020 project AI4Media, under contract no. 951911 (https://ai4media.eu/). The authors are grateful to the Université Côte d'Azur's Center for High-Performance Computing (OPAL infrastructure) for providing resources and support.

References

Apoorv Agarwal, Jiehan Zheng, Shruti Kamath, Sriramkumar Balasubramanian, and Shirin Ann Dey. Key Female Characters in Film Have More to Talk About Besides Men: Automating the Bechdel Test. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 830–840, Denver, Colorado, 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1084. URL `http://aclweb.org/anthology/N15-1084`.

Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Pieter Gijsbers, Joan Giner-Miguelez, Nitisha Jain, Michael Kuchnik, Quentin Lhoest, Pierre Marcenac, Manil Maskey, Peter Mattson, Luis Oala, Pierre Ruyssen, Rajat Shinde, Elena Simperl, Goeffry Thomas, Slava Tykhonov, Joaquin Vanschoren, Jos van der Velde, Steffen Vogler, and Carole-Jean Wu. Croissant: A metadata format for ml-ready datasets. In *Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning*, SIGMOD/PODS '24. ACM, June 2024. doi: 10.1145/3650203.3663326. URL `http://dx.doi.org/10.1145/3650203.3663326`.

Elisa Ancarani, Julie Tores, Lucile Sassatelli, Rémy Sun, Hui-Yin Wu, and Frédéric Precioso. Leveraging multimodal explanatory annotations for video interpretation with modality specific dataset. *arXiv preprint arXiv:2504.11232*, 2025a. URL https://arxiv.org/abs/2504.11232.

Elisa Ancarani, Julie Tores, Rémy Sun, Lucile Sassatelli, Hui-Yin Wu, and Frederic Precioso. Leveraging concept annotations for trustworthy multimodal video interpretation through modality specialization. In *Proceedings of the 1st International Workshop on Cognition-Oriented Multimodal Affective and Empathetic Computing*, CogMAEC '25, page 11–19, New York, NY, USA, 2025b. Association for Computing Machinery. ISBN 9798400720598. doi: 10.1145/3746277.3760415. URL https://doi.org/10.1145/3746277.3760415.

Mohammad Asadi, Jack W. O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Fardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. MIRAGE: The Illusion of Visual Understanding, 2026. URL https://arxiv.org/abs/2603.21687. Version Number: 2.

Philippe Bernard, Sarah J. Gervais, and Olivier Klein. Objectifying objectification: When and why people are cognitively reduced to their parts akin to objects. *European Review of Social Psychology*, 29(1):82–121, January 2018. ISSN 1046-3283. doi: 10.1080/10463283.2018.1471949. URL https://doi.org/10.1080/10463283.2018.1471949. Publisher: Routledge _eprint: https://doi.org/10.1080/10463283.2018.1471949.

Philippe Bernard, Florence Hanoteau, Sarah Gervais, Lara Servais, Irene Bertolone, Paul Deltenre, and Cécile Colin. Revealing Clothing Does Not Make the Object: ERP Evidences That Cognitive Objectification is Driven by Posture Suggestiveness, Not by Revealing Clothing. *Personality and Social Psychology Bulletin*, 45(1):16–36, January 2019. ISSN 0146-1672, 1552-7433. doi: 10.1177/0146167218775690. URL http://journals.sagepub.com/doi/10.1177/0146167218775690.

Philippe Bernard, Carlotta Cogoni, and Andrea Carnaghi. The Sexualization–Objectification Link: Sexualization Affects the Way People See and Feel Toward Others. *Current Directions in Psychological Science*, 29(2):134–139, April 2020. ISSN 0963-7214, 1467-8721. doi: 10.1177/0963721419898187. URL http://journals.sagepub.com/doi/10.1177/0963721419898187.

Abeba Birhane, Vinay Prabhu, Sanghyun Han, Vishnu Boddeti, and Sasha Luccioni. Into the LAION's den: Investigating hate in multimodal datasets. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=6URyQ9QhYv.

Digbalay Bose, Rajat Hebbar, Tiantian Feng, Krishna Somandepalli, Anfeng Xu, and Shrikanth Narayanan. Mm-au:towards multimodal understanding of advertisement videos. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 86–95, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085. doi: 10.1145/3581783.3612371. URL https://doi.org/10.1145/3581783.3612371.

Alexander Braylan, Omar Alonso, and Matthew Lease. Measuring Annotator Agreement Generally across Complex Structured, Multi-object, and Free-text Annotation Tasks. In *Proceedings of the ACM Web Conference 2022*, pages 1720–1730, April 2022. doi: 10.1145/3485447.3512242. URL http://arxiv.org/abs/2212.09503. arXiv:2212.09503 [cs].

Iris Brey. *Le regard féminin-Une révolution à l'écran*. Média Diffusion, 2020.

M. Bucarelli, L. Cassano, F. Siciliano, A. Mantrach, and F. Silvestri. Leveraging inter-rater agreement for classification in the presence of noisy labels. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3439–3448, Los Alamitos, CA, USA, jun 2023. IEEE Computer Society. doi: 10.1109/CVPR52729.2023.00335. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.00335.

Rachel M. Calogero. A Test of Objectification Theory: The Effect of the Male Gaze on Appearance Concerns in College Women. *Psychology of Women Quarterly*, 28(1):16–21, March 2004. ISSN 0361-6843, 1471-6402. doi: 10.1111/j.1471-6402.2004.00118.x. URL http://journals.sagepub.com/doi/10.1111/j.1471-6402.2004.00118.x.

Rachel M. Calogero, Stacey Tantleff-Dunn, and J. Kevin Thompson. Operationalizing self-objectification: Assessment and related methodological issues. In *Self-objectification in women: Causes, consequences, and counteractions.*, pages 23–49. American Psychological Association, Washington, 2011. ISBN 978-1-4338-0798-5 978-1-4338-0799-2. doi: 10.1037/12304-002. URL http://content.apa.org/books/12304-002.

Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept Whitening for Interpretable Image Recognition. *Nature Machine Intelligence*, 2(12):772–782, December 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00265-z. URL http://arxiv.org/abs/2002.01650. arXiv:2002.01650 [cs, stat].

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. Fine-Grained Analysis of Propaganda in News Article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5635–5645, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1565. URL https://www.aclweb.org/anthology/D19-1565.

Roxana Daneshjou, Mert Yuksekgonul, Zhuo Ran Cai, Roberto A. Novoa, and James Zou. Skincon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=gud0qopqJc4.

Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. HateMM: A Multi-Modal Dataset for Hate Video Classification. *Proceedings of the International AAAI Conference on Web and Social Media*, 17:1014–1023, June 2023. ISSN 2334-0770, 2162-3449. doi: 10.1609/icwsm.v17i1.22209. URL https://ojs.aaai.org/index.php/ICWSM/article/view/22209.

Angela Denchik. *Development and Psychometric Evaluation of the Interpersonal Sexual Objectification Scale*. PhD thesis, The Ohio State University, 2005.

Quanqi Du, Loic De Langhe, Els Lefever, and Veronique Hoste. LDW: Label Divergence Weighting for Multimodal Sentiment Analysis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12342–12351, Dublin Ireland, October 2025. ACM. ISBN 979-8-4007-2035-2. doi: 10.1145/3746027.3758160. URL https://dl.acm.org/doi/10.1145/3746027.3758160.

Ethan Fast, Binbin Chen, and Michael S. Bernstein. Empath: Understanding Topic Signals in Large-Scale Text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4647–4657, San Jose California USA, May 2016. ACM. ISBN 978-1-4503-3362-7. doi: 10.1145/2858036.2858535. URL https://dl.acm.org/doi/10.1145/2858036.2858535.

Elisabetta Fersini, Francesca Gasparini, and Silvia Corchs. Detecting Sexist MEME On The Web: A Study on Textual and Visual Cues. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 226–231, 2019. doi: 10.1109/ACIIW.2019.8925199.

Ryo Fujii, Hideo Saito, and Ryo Hachiuma. VIOLA: Towards Video In-Context Learning with Minimal Annotations, 2026. URL https://arxiv.org/abs/2601.15549. Version Number: 1.

David C Funder and Daniel J Ozer. Evaluating effect size in psychological research: Sense and nonsense. *Adv. Methods Pract. Psychol. Sci.*, 2(2):156–168, June 2019.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Commun. ACM*, 64(12):86–92, nov 2021. ISSN 0001-0782. doi: 10.1145/3458723. URL https://doi.org/10.1145/3458723.

Sarah J. Gervais, Gemma Sáez, Abigail R. Riemer, and Olivier Klein. The Social Interaction Model of Objectification: A process model of goal-based objectifying exchanges between men and women. *British Journal of Social Psychology*, 59(1):248–283, January 2020. ISSN 0144-6665, 2044-8309. doi: 10.1111/bjso.12339. URL https://onlinelibrary.wiley.com/doi/10.1111/bjso.12339.

Yuan Gong, Yu-An Chung, and James R. Glass. AST: audio spectrogram transformer. *CoRR*, abs/2104.01778, 2021. URL https://arxiv.org/abs/2104.01778.

Tanaya Guha, Che-Wei Huang, Naveen Kumar, Yan Zhu, and Shrikanth S. Narayanan. Gender Representation in Cinematic Content: A Multimodal Approach. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 31–34, Seattle Washington USA, November 2015. ACM. ISBN 978-1-4503-3912-4. doi: 10.1145/2818346.2820778. URL https://dl.acm.org/doi/10.1145/2818346.2820778.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, Boston, MA, USA, June 2015. IEEE. ISBN 978-1-4673-6964-0. doi: 10.1109/CVPR.2015.7298698. URL http://ieeexplore.ieee.org/document/7298698/.

Chao Huang, Benfeng Wang, Wei Wang, Jie Wen, Chengliang Liu, Li Shen, and Xiaochun Cao. Vad-r1: Towards video anomaly reasoning via perception-to-cognition chain-of-thought. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=6eLlczxbrp.

Ji Yoon Jang, Sangyoon Lee, and Byungjoo Lee. Quantification of Gender Representation Bias in Commercial Films based on Image Analysis. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–29, November 2019. ISSN 2573-0142. doi: 10.1145/3359300. URL https://dl.acm.org/doi/10.1145/3359300.

Ilker Kesen, Andrea Pedrotti, Mustafa Dogan, Michele Cafagna, Emre Can Acikgoz, Letitia Parcalabescu, Iacer Calixto, Anette Frank, Albert Gatt, Aykut Erdem, and Erkut Erdem. ViLMA: A Zero-Shot Benchmark for Linguistic and Temporal Grounding in Video-Language Models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=liuqDwmbQJ.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Kangsan Kim, Geon Park, Youngwan Lee, Woongyeong Yeo, and Sung Ju Hwang. VideoICL: Confidence-based Iterative In-context Learning for Out-of-Distribution Video Understanding . In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3295–3305, Los Alamitos, CA, USA, June 2025. IEEE Computer Society. doi: 10.1109/CVPR52734.2025.00313. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52734.2025.00313.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5338–5348. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Girish A. Koushik, Diptesh Kanojia, and Helen Treharne. Towards a Robust Framework for Multimodal Hate Detection: A Study on Video vs. Image-based Content. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2014–2023, May 2025. doi: 10.1145/3701716.3718382. URL http://arxiv.org/abs/2502.07138. arXiv:2502.07138 [cs].

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Geoffrey R. Loftus and Michael E. J. Masson. Using confidence intervals in within-subject designs. *Psychonomic Bulletin amp; Review*, 1(4):476–490, December 1994. ISSN 1531-5320. doi: 10.3758/bf03210951. URL http://dx.doi.org/10.3758/BF03210951.

Alicia Malone. *The Female Gaze: Essential Movies Made by Women*. Mango Publishing Group, 2018.

Victor R. Martinez, Krishna Somandepalli, and Shrikanth Narayanan. Boys don't cry (or kiss or dance): A computational linguistic lens into gendered actions in film. *PLOS ONE*, 17(12): e0278604, December 2022. ISSN 1932-6203. doi: 10.1371/journal.pone.0278604. URL https://dx.plos.org/10.1371/journal.pone.0278604.

Michael E. J. Masson and Geoffrey R. Loftus. Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale*, 57(3):203–220, September 2003. ISSN 1196-1961. doi: 10.1037/h0087426. URL http://dx.doi.org/10.1037/h0087426.

Antoine Mazières, Telmo Menezes, and Camille Roth. Computational appraisal of gender representativeness in popular movies. *Humanities and Social Sciences Communications*, 8(1):137, December 2021. ISSN 2662-9992. doi: 10.1057/s41599-021-00815-9. URL http://www.nature.com/articles/s41599-021-00815-9.

Nita Mary McKinley and Janet Shibley Hyde. The objectified body consciousness scale: Development and validation. *Psychology of women quarterly*, 20(2):181–215, 1996.

Michael Mozer, Katherine Hermann, and Jennifer Hu. Experimental Design and Analysis for AI Researchers. *NeurIPS Tutorial*, December 2024. URL https://neurips.cc/virtual/2024/tutorial/99528.

Laura Mulvey. Visual Pleasure and Narrative Cinema. *Screen*, 16(3):6–18, 10 1975. ISSN 0036-9543. doi: 10.1093/screen/16.3.6. URL https://doi.org/10.1093/screen/16.3.6.

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision (ECCV)*, 2022.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, November 2021. ISSN 26663899. doi: 10.1016/j.patter.2021.100336. URL https://linkinghub.elsevier.com/retrieve/pii/S2666389921001847.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Vikram V. Ramaswamy, Sunnie S. Y. Kim, Ruth Fong, and Olga Russakovsky. Overlooked Factors in Concept-Based Explanations: Dataset Choice, Concept Learnability, and Human Capability. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10932–10941, Vancouver, BC, Canada, June 2023. IEEE. ISBN 979-8-3503-0129-8. doi: 10.1109/CVPR52729.2023.01052. URL https://ieeexplore.ieee.org/document/10203934/.

Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. "Call me sexist, but..." : Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):573–584, May 2021. doi: 10.1609/icwsm.v15i1.18085. URL https://ojs.aaai.org/index.php/ICWSM/article/view/18085.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2329–2334, 2017.

Shlomo S Sawilowsky. New effect size rules of thumb. *J. Mod. Appl. Stat. Methods*, 8(2):597–599, November 2009.

Alexandra Schofield and Leo Mehr. Gender-Distinguishing Features in Film Dialogue. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*, pages 32–39, San Diego, California, USA, 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-0204. URL http://aclweb.org/anthology/W16-0204.

Ibrahim Serouis and Florence Sèdes. Towards context and domain-aware algorithms for scene analysis. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=JQGmbVK4Fr.

Krishna Somandepalli, Tanaya Guha, Victor R. Martinez, Naveen Kumar, Hartwig Adam, and Shrikanth Narayanan. Computational Media Intelligence: Human-Centered Machine Analysis of Media. *Proceedings of the IEEE*, 109(5):891–910, May 2021. ISSN 0018-9219, 1558-2256. doi: 10.1109/JPROC.2020.3047978. URL https://ieeexplore.ieee.org/document/9319168/.

Hung-Ting Su, Po-Wei Shen, Bing-Chen Tsai, Wen-Feng Cheng, Ke-Jyun Wang, and Winston H. Hsu. TrUMAn: Trope Understanding in Movies and Animations. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, pages 4594–4603, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8446-9. doi: 10.1145/3459637.3482018. URL https://doi.org/10.1145/3459637.3482018. event-place: Virtual Event, Queensland, Australia.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2018. doi: 10.1109/CVPR.2018.00678.

Julie Tores, Lucile Sassatelli, Hui-Yin Wu, Clement Bergman, Lea Andolfi, Victor Ecrement, Frederic Precioso, Thierry Devars, Magali Guaresi, Virginie Julliard, and Sarah Lecossais. Visual objectification in films: Towards a new ai task for video interpretation. In *2024 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, June 2024. IEEE.

Julie Tores, Elisa Ancarani, Rémy Sun, Lucile Sassatelli, Hui-Yin Wu, and Frederic Precioso. Reexamining concept-based explainable models for multimodal interpretative tasks. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 12437–12445, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3758170. URL https://doi.org/10.1145/3746027.3758170.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72: 1385–1470, December 2021. ISSN 1076-9757. doi: 10.1613/jair.1.12752. URL https://jair.org/index.php/jair/article/view/12752.

Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. MovieGraphs: Towards Understanding Human-Centric Situations from Videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset, 2011.

Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. Revise: A tool for measuring and mitigating bias in visual datasets. *Int. J. Comput. Vision*, 130(7):1790–1810, jul 2022. ISSN 0920-5691. doi: 10.1007/s11263-022-01625-5. URL https://doi.org/10.1007/s11263-022-01625-5.

Han Wang, Zhuoran Wang, and Roy Ka-Wei Lee. HateClipSeg: A Segment-Level Annotated Dataset for Fine-Grained Hate Video Detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 13304–13310, Dublin Ireland, October 2025. ACM. ISBN 979-8-4007-2035-2. doi: 10.1145/3746027.3758289. URL https://dl.acm.org/doi/10.1145/3746027.3758289.

L. Monique Ward and Petal Grower. Media and the Development of Gender Role Stereotypes. *Annual Review of Developmental Psychology*, 2(1):177–199, December 2020. ISSN 2640-7922, 2640-7922. doi: 10.1146/annurev-devpsych-051120-010630. URL https://www.annualreviews.org/doi/10.1146/annurev-devpsych-051120-010630.

Jiaheng Wei, Zhaowei Zhu, Tianyi Luo, Ehsan Amid, Abhishek Kumar, and Yang Liu. To Aggregate or Not? Learning with Separate Noisy Labels. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pages 2523–2535, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599522. URL https://doi.org/10.1145/3580305.3599522. event-place: Long Beach, CA, USA.

Hui-Yin Wu, Luan Nguyen, Yoldoz Tabei, and Lucile Sassatelli. Evaluation of deep pose detectors for automatic analysis of film style. In *EUROGRAPHICS Workshop on Intelligent Cinematography and Editing*, page 9, Reims, France, 2022.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lio, and Mateja Jamnik. Concept embedding models: beyond the accuracy-explainability trade-off. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022a. Curran Associates Inc. ISBN 9781713871088.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lio, and Mateja Jamnik. Concept embedding models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=HXCPA2GXf_.

Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, volume 13664 of *LNCS*, pages 492–510, 2022.

Xiao Zheng, Chang Tang, Zhiguo Wan, Chengyu Hu, and Wei Zhang. Multi-Level Confidence Learning for Trustworthy Multimodal Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):11381–11389, June 2023. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v37i9.26346. URL https://ojs.aaai.org/index.php/AAAI/article/view/26346.

Appendix A. Supplementary material for "MObyGaze: a film dataset of multimodal objectification densely annotated by experts"

A.1 Datasheet documentation for the MObyGaze dataset

The datasheet documentation below provides the necessary information required in the checklist, specifically:

- Dataset documentation and intended uses.
- URL to website where the dataset can be downloaded by the reviewers and
- URL to Croissant metadata record documenting the dataset:
  The MObyGaze dataset artifacts are provided along with a Croissant description at `https://github.com/husky-helen/MObyGaze`
- Author statement: We bear all responsibility for the MObyGaze dataset, which is shared under a CC BY-NC-SA licence.
- Hosting, licensing, and maintenance plan are described in the datasheet below.

Motivation

For what purpose was the dataset created?

The purpose of the MObyGaze dataset is to advance computational approaches to help make subtle patterns of bias in audiovisual content visible and more tangible, and quantify their prevalence. We create the MObyGaze dataset to enable the AI community to design computational approaches to characterize and quantify complex temporal and multimodal patterns of character objectification in films. For this, we devise a thesaurus of objectification by building on existing characterization in film studies and cognitive and social psychology. The thesaurus articulates visual, speech and audio components, which we denote as *concepts* involved in the production of objectification. The annotation process then consists in 2 experts densely annotating 20 movies: they manually delimit all the segments they find relevant for objectification, and label each with a level of objectification. To allow for fine-grained data and model analysis, they also annotate which objectification concepts are present.

Who created this dataset?

The multidisciplinary team of authors composed of gender and media studies researchers, data scientists, and AI researchers from multiple research institutes and universities.

Who funded the creation of the dataset?

The project was supported through public research funds mentioned in the acknowledgments section: the French National Reasearch Agency (ANR) and the European Commission (EC).

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

The dataset consists of annotations of 20 feature-length films, for which we consider the video track, the sound track and associated subtitles. Each movie is annotated by 2 experts for a freely determined number of segments per movie. A dataset instance

is therefore a video segment identified with its indices of start and end frame and start and end time-stamps, annotated with the objectification rating and thesaurus concepts tagged as present in the segment by the annotator. The annotator considered the image, sound dialogue modalities to annotate, and the dialogue transcript is also provided for each segment. Fig. 17 shows an example of two dataset instances. The files we provide are:

• the list of films (mobygaze_movielist.csv), also reported in Table 16;

• the objectification thesaurus (objectification-thesaurus.json) listing the concepts and their instances the annotators used to annotate the films;

• the entire table of annotated segments (mobygaze_dataframe.csv). One segment corresponds to an interval of a movie delimited by a given annotator. Fig. 17 shows an example;

• the SQL description of the dataset as a database, with tables annotations, movies and subtitles (Neurips.sql).

How many instances are there in total (of each type, if appropriate)?

There are 20 films annotated by two annotators, yielding in total 6072 segments delimited and annotated.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

The dataset contains all the instances produced by both annotators who have entirely annotated each movie. The 20 movies of MObyGaze are a subset of the 51 movies of the pre-existing MovieGraphs dataset. The 20 movies have been chosen to maintain the ratio of represented genres and diversify to the maximum the actors/directors represented. Table 16 provides details on the movies.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features?

Each instance consists of the start and end frames and time stamps of each segment, along with the dialog transcript from the subtitle file, and annotator ratings of level of objectification and thesaurus concepts. The text of the subtitle was down-cased and HTML tags were removed.

Is there a label or target associated with each instance?

Each instance is a segment associated with an objectification rating (Easy Negative - EN, Hard Negative - HN, Sure - S, Not Sure - NS) and a set of concepts selected by the annotator as present, and chosen from the thesaurus. An additional free text box may be used by the annotator (column 'comment' in Fig. 17).

Is any information missing from individual instances?

The visual and sound data are not provided in the dataset, but exact frame and time stamping are provided to correctly use the annotations of the MObyGaze dataset and compare to the benchmarked models accompanying the dataset.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?

Any relationship between annotated segments is assured by the consistency in movie identifiers, frame and time stamping as well as by annotator identifiers.

Are there recommended data splits (e.g., training, development/validation, testing)?

In order to assess generic objectification patterns while maximizing the amount of data available for training and testing, we recommend to use leave-4-movies-out cross-validation, where no movie in train is used

clip_index	imdb_key	movie_title	annotator	start_frame	end_frame	concepts	label	label_vision	label_audio	label_text	framerate	comment
id_107	tt0120338	Titanic	annotator_1	104666	105069	['Clothing', 'Speech']	Sure	1	0	1	23.976	[]
id_102	tt0120338	Titanic	annotator_2	83681	84756	['Type of shot', 'Look']	Sure	1	0	0	23.976	[]

Figure 17: Example of two dataset instances.

Table 16: List of films annotated in the MObyGaze dataset (sorted by genre)

IMDB key	Movie Title	Duration	Year	Genre	Test Fold	Validation Fold
tt0097576	Indiana Jones and the Last Crusade	2h 6min	1989	adventure, action	1	5
tt1454029	The Help	2h 26min	2011	drama	2	
tt1285016	The Social Network	2h 0min	2010	drama, biographical	3	4
tt0467406	Juno	1h 36min	2007	drama, comedy	4	
tt0110912	Pulp Fiction	2h 34min	1994	drama, crime	5	3
tt0822832	Marley & Me	1h 55min	2008	drama, family	1	
tt1568346	The Girl with the Dragon Tattoo	2h 38min	2011	drama, mystery, crime	2	
tt2267998	Gone Girl	2h 29min	2014	drama, mystery, thriller	3	2
tt0109830	Forrest Gump	2h 22min	1994	drama, romantic	4	1
tt0120338	Titanic	3h 14min	1997	drama, romantic	5	
tt0108160	Sleepless in Seattle	1h 45min	1993	drama, romantic, comedy	1	5
tt0119822	As Good as It Gets	2h 18min	1997	drama, romantic, comedy	2	
tt1193138	Up in the Air	1h 49min	2009	drama, romantic, comedy	3	4
tt1570728	Crazy,Stupid,Love.	1h 58min	2011	drama, romantic, comedy	4	
tt1045658	Silver Linings Playbook	2h 2min	2012	drama, romantic, comedy	5	3
tt0970416	The Day the Earth Stood Still	1h 43min	2008	drama, sci-fi, adventure	1	
tt1907668	Flight	2h 18min	2012	drama, thriller	2	
tt0375679	Crash	1h 55min	2004	drama, thriller, crime	3	2
tt1142988	The Ugly Truth	1h 35min	2009	romantic, comedy	4	1
tt1632708	Friends with benefits	1h 49min	2011	romantic, comedy	5	

in test (even for different segments). The dataset therefore comes with 5 different folds. Each fold is made of a train, validation and test split, each composed of 14, 2 and 4 movies, respectively. There is no overlap between the test sets, so that every movie appears exactly once in test. The fold indices where each movie appears in test or validation are indicated in Table 16. We recommend to use the same folds to generate results comparable with the models benchmarked in the article introducing MObyGaze (present article submitted to NeurIPS dataset and benchmarks track, to be edited after review). Results should be reported as average of model results over 5 folds, along with standard deviations. The definition of the ML task on the MObyGaze dataset must be fully specified (classification, localization, constitution of the positive and negative classes from the objectification ratings and tagged concepts).

Are there any errors, sources of noise, or redundancies in the dataset?

N/A

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The dataset only relies on the films, not shared for itellectual property reasons. Table 16 and file mobygaze_dataframe.csv provide the necessary information for anyone to acquire the films and align the

MObyGaze annotations onto it. We integrate the subtitles in the dataset.

Does the dataset contain data that might be considered confidential?
N/A

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?
The annotated movies can contain offensive and otherwise disturbing content. Detailed information on the movie age suitability rating per country is available on the IMDB page under parental guide [2]This page also lists the types of scenes under each category on non-mild content. The free text field of the annotations provided in MObyGaze may contain speech segments or descriptions of visual content linked to the labels that can be offensive.

Does the dataset relate to people?
All annotated films are fictitious, and do not depict real persons. The characters are played by human actors. The dataset is meant to train models to study disparities in gender representation in cinema.

Does the dataset identify any subpopulations (e.g., by age, gender)?
The MObyGaze dataset does not provide annotation of gender or other demographics of the characters. However, character gender can be obtained from the MovieGraphs dataset for the corresponding movies, or easily inferred from the actor cast and face recognition technologies applied to the content to detect the actor.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?
The actors can be identified from their appearance in the movies, but the MObyGaze dataset is not the enabler.

Does the dataset contain data that might be considered sensitive in any way?
N/A

Collection Process

How was the data associated with each instance acquired?
Each movie is annotated by 2 experts (with background in computer science, film studies and cognitive psychology), who watch it entirely, setting temporal boundaries of each segment where at least one objectifying concept is deemed present. For each such segment, they rate objectification on one of four levels:
• Easy Negative (EN): no objectifying concept is present;
• Hard Negative (HN): one or some concepts are present, are annotated, but are deemed insufficient to produce a perception of objectification;
• Sure (S): objectification is perceived and explained by the annotated concepts from the thesaurus;
• Not Sure (NS): objectification is perceived and concepts are annotated but the annotator considers they do not sufficiently explain the perception of objectification.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?
Fig. 18 shows the tool specifically designed for densely annotating objectification levels and concepts. It can be seen that the tool provides a free-text field that the annotators can choose to use. The annotators first annotated 2 movies. The obtained annotations

2. www.imdb.com/title/[IMDBkey]/parentalguide

were then aligned and colored for the annotators to identify their major divergences. They convened and identified that the agreement was generally high on the rating of objectification. Analyzing the annotation differences for the 11 concepts, the annotators specifically identified under-determination of concepts Actitivies and Appearance. They expanded Activities to include all types of actions contributing to objectification, particularly momentary actions by a character onto another (including aspects of domination and violence). They trimmed the concept of Appearance, which initially consisted of instances deploying over the entire film (such as age of character not matching age or appearance of actress) to restrict it to scene-level features. Fig. 2 shows the resulting thesaurus. They then carried out individually the annotation over the rest of the 20 movies.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
The dataset contains all the instances produced by both annotators having entirely annotated each movie. The 20 movies of MObyGaze are a subset of the 51 movies of the pre-existing MovieGraphs dataset. The 20 movies have been chosen to maintain the ratio of represented genres and diversify to the maximum the actors/directors represented. Table 16 provides details on the movies.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?
Both experts that produced all the annotations are project members (tenured scholars) and worked on annotation as part of their research tasks.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?
The annotations took place between May 2023 and May 2024. The annotated films were produced between 1989 and 2014.

Were any ethical review processes conducted (e.g., by an institutional review board)?
The data collection neither involved intervention nor interpersonal contact with subjects, or collection of data on subjects. The dataset creation therefore did not require an institutional review board or ethical committee review.

Does the dataset relate to people?
The annotations provided do not relate to people. The annotated films depict fictitious characters.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
The data has been collected directly by project members using the annotation tool on their local computers.

Were the individuals in question notified about the data collection?
N/A

Did the individuals in question consent to the collection and use of their data?
The annotators were project members. The film data have been lawfully used as per the French and European regulations.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?
N/A

Has an analysis of the potential impact of the dataset and its use on data subjects

(e.g., a data protection impact analysis) been conducted?
N/A

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?
The annotations created and provided in MObyGaze are the raw transcription from the annotation tool, which generates, for each annotator annotating a movie, json files sharing indices of start and end frame of each segment, objectification level, objectification concepts, and free text. These are re-formatted without any approximation in the mobygaze_dataframe.csv provided in the dataset.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?
The raw data is saved but does not bring additional information compared to mobygaze_dataframe.csv and is not shared to preserve anonymity. Indeed, the json files produced by the annotation tool contain folder paths of the local machines of the annotators.

Is the software used to preprocess/-clean/label the instances available?
Yes, the python scripts used to create the database from the json files generated by the annotation tool, and the python scripts used to generate mobygaze_dataframe.csv from the database, will be made available along with the publication of the annotation tool to the community.

Uses

Has the dataset been used for any tasks already?
The dataset has been used for classification of objectification knowing the true segment boundaries, and localization of objectififcation in fixed-length segments. All vision, speech and audio modalities have been used separately.

Is there a repository that links to any or all papers or systems that use the dataset?
`https://github.com/husky-helen/MObyGaze`

What (other) tasks could the dataset be used for?
The MObyGaze dataset is also meant to design explainable models to better characterize complex temporal and multimodal objectification patterns, which can in turn enrich qualitative studies by media scholars. The MObyGaze dataset can also be used to study the fairness of existing computer vision models: person detectors and human pose estimators may miss the presence of characters onscreen, compromising the study of how certain patterns correlate with certain human groups, if the humans are often mis-detected for these patterns (e.g., shots with headless body parts).

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?
Any future user must be aware that the movies selected for annotating objectification were in no case chosen for their specific crew or other production affiliation, but on the sole basis of preserving the genre distribution when sampling from the pre-existing MovieGraphs dataset.

The MObyGaze dataset should not be used for tasks such as content filtering, defining regulatory standards, or censorship of media content.

Another potential negative impact would lie in the interpretation of the outputs of models, being it by social science researchers assisting their qualitative studies with models trained on MObyGaze to detect objectification, or by amateurs. For example, if the models to detect objectification trained on filmic fictional content are used on real-user videos (such as those uploaded on online social networks), one may
(1) establish and highlight correlations between the level of objectification and demographics of users that put on a performance and upload their videos, then
(2) claim that the users of these demographic groups (most likely historically under-privileged or submitted, e.g., women) are the sole responsible for choosing to objectify themselves, somehow discarding and making invisible the social root causes leading to self-objectification.
While the risk of fallacious interpretation already exists from qualitative analyses not assisted by AI models, a misleading narrative can be reinforced with an AI model allowing larger-scale analysis and a veneer of objectivity.

The first proactive measure is the choice of the dataset license prohibiting commercial usage, to address the risk of censorship mentioned first. The second proactive measure is to highlight in the dataset documentation that the models trained on the MObyGaze dataset must be used in a clear ethical framework: that of making subtle patterns of representation disparities visible and contribute responsible usages. These may be discussions on the extent of these patterns in certain contexts and their causes (discussions in social sciences, or by associations raising awareness on representation disparities in the media). This can also be for training purpose, e.g., to assist students in audiovisual production in the analysis of their production trials and develop reflexivity in their practice. We will also advertise this message in the public website under construction where we will demo the dataset and possible models. We will indicate that every further developed model should come with proper standardized documentation for this task, relaying the need for responsible interpretation and usage.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?
Yes, the dataset will be made publicly available.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)
The dataset is made available on Github: `https://github.com/husky-helen/MObyGaze`], and will also be made available on Zenodo, from where a DOI will be obtained, and long-term storage ensured.

When will the dataset be distributed?
The dataset is already available online, but will be hosted on Zenodo and attributed a DOI after the double-blind review process.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?
The dataset will be made available under an open CC BY-NC-SA license.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances?
N/A

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?
N/A

Maintenance

Who will be supporting/hosting/maintaining the dataset?
The dataset will be hosted on permanent public storage Zenodo. The dataset will be supported and maintained by the project team. The team is led by tenured researchers who will dedicate the necessary resources, after project funding ends, to maintain the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
By institutional email lucile.sassatelli@univ-cotedazur.fr

Is there an erratum?
N/A

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?
New versions will be added to the dataset in the case of inclusion of a completely new session of annotation with modifications to the thesaurus, film set, and/or annotators. The dataset will be updated in the case of adding individual annotations on new or existing films using the same thesaurus. Regular updates (every 3-6 months) to address minor issues in the dataset will be provided based on requests. Next version previewed is in Dec 2026 from another annotation session on television series and historical drama. By 2027, we intend to make the thesaurus evolve to annotate tropes both at the film and at the segment levels, creating a new version of the dataset.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?
N/A

Will older versions of the dataset continue to be supported/hosted/maintained?
The older version of the dataset will continue to be hosted and maintained, thanks to the resources described above where tenured researchers responsible for the research funding will dedicate the necessary resources to maintenance.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?
Under the terms of the chosen license CC BY-NC-SA, any user can fork the dataset and extend, modify, and share it as desired.

A.2 Details on the MObyGaze dataset

A.2.1 Films

Table 17 shows the list of films in the MObyGaze dataset, while Table 18 shows how this list reproduces the genre distribution of the original MovieGraphs dataset Vicol et al. (2018).

Table 17: List of films annotated in the MObyGaze dataset (sorted by genre)

IMDB key	Movie Title	Duration	Year	Genre	Test Fold	Validation Fold
tt0097576	Indiana Jones and the Last Crusade	2h 6min	1989	adventure, action	1	5
tt1454029	The Help	2h 26min	2011	drama	2	
tt1285016	The Social Network	2h 0min	2010	drama, biographical	3	4
tt0467406	Juno	1h 36min	2007	drama, comedy	4	
tt0110912	Pulp Fiction	2h 34min	1994	drama, crime	5	3
tt0822832	Marley & Me	1h 55min	2008	drama, family	1	
tt1568346	The Girl with the Dragon Tattoo	2h 38min	2011	drama, mystery, crime	2	
tt2267998	Gone Girl	2h 29min	2014	drama, mystery, thriller	3	2
tt0109830	Forrest Gump	2h 22min	1994	drama, romantic	4	1
tt0120338	Titanic	3h 14min	1997	drama, romantic	5	
tt0108160	Sleepless in Seattle	1h 45min	1993	drama, romantic, comedy	1	5
tt0119822	As Good as It Gets	2h 18min	1997	drama, romantic, comedy	2	
tt1193138	Up in the Air	1h 49min	2009	drama, romantic, comedy	3	4
tt1570728	Crazy,Stupid,Love.	1h 58min	2011	drama, romantic, comedy	4	
tt1045658	Silver Linings Playbook	2h 2min	2012	drama, romantic, comedy	5	3
tt0970416	The Day the Earth Stood Still	1h 43min	2008	drama, sci-fi, adventure	1	
tt1907668	Flight	2h 18min	2012	drama, thriller	2	
tt0375679	Crash	1h 55min	2004	drama, thriller, crime	3	2
tt1142988	The Ugly Truth	1h 35min	2009	romantic, comedy	4	1
tt1632708	Friends with benefits	1h 49min	2011	romantic, comedy	5	

Table 18: Distribution of movie genres between MovieGraphs (51 movies) Vicol et al. (2018) and MObyGaze (subset of 20 movies). Genre source: IMDB (note: a movie has several genres).

Genre	MoviGraphs	MObyGaze
Action	0.02	0.05
Adventure	0.08	0.1
Biography	0.06	0.05
Comedy	0.43	0.4
Crime	0.2	0.15
Drama	0.76	0.85
Family	0.04	0.05
Fantasy	0.02	0
Film noir	0.02	0
History	0.02	0
Mystery	0.12	0.1
Romance	0.49	0.45
Sci-Fi	0.08	0.05
Thriller	0.16	0.1

A.2.2 Annotation tool

Fig. 18 shows the tool specifically designed for densely annotating objectification levels and concepts. It can be seen that the tool allows for a free text field that the annotators are free to use.

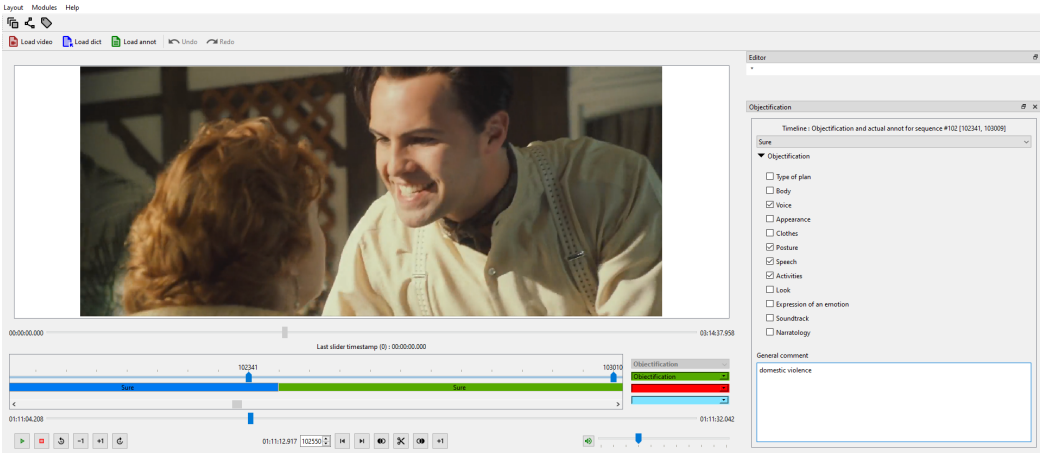
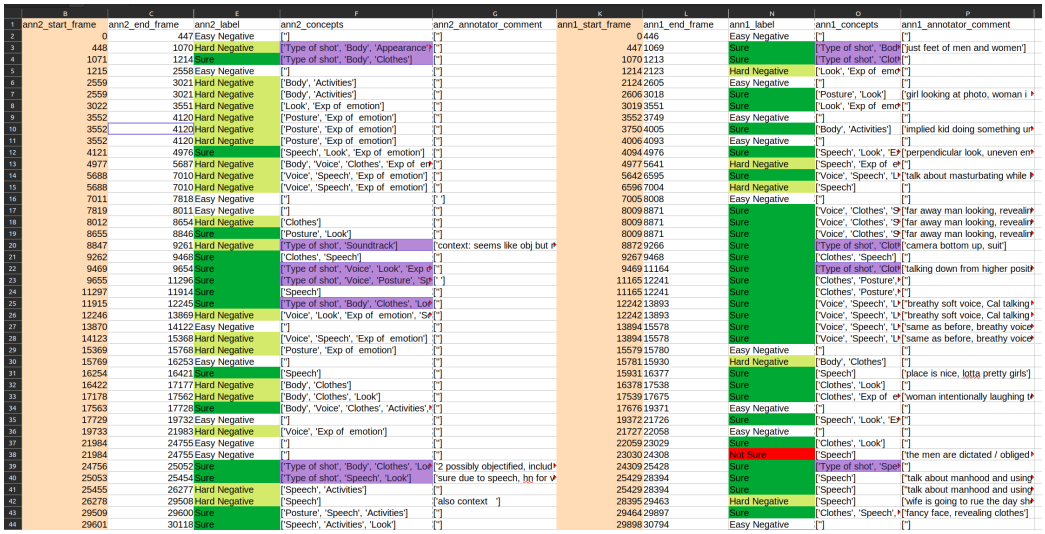


Figure 18: The annotation interface.

A.2.3 Details on the annotation procedure

The annotators are experts who are both tenured scholars at public institutions and are the instigators of the nationally-funded project funding this work through the involvement of 3 laboratories in social sciences and 3 laboratories in computer science. They have coordinated the multidisciplinary team for every phase of the work for this article. One annotator holds a B.A. in English literature, a M.Sc in Communications and a PhD in computer science with expertise in cinematic discourse, the other holds a PhD in computer science, has led projects at the intersection of CS and social psychology for gender inequalities, and is the project PI. They both identify as women, in their late 30s-early40s, they both work in Europe, one is from Asian descent and the other from European descent.

The annotators first annotated 2 movies. The obtained annotations were then aligned and colored for the annotators to identify their major divergences. They convened and identified that the agreement was generally high on the rating of objectification. Noticeable differences were in annotation of NS with levels of narratology spanning over more than a single segment. The thesaurus is indeed targeted at annotating sailient concepts in segments, and we discuss this limitation in Sec. 5. They specifically identified under-determination of concepts Actitivies and Appearance. They expanded Activities to include all types of actions contributing to objectification, particularly momentary actions by a character onto another (including aspects of domination and violence). They trimmed the concept of Appearance, which initially consisted of instances deploying over the entire film (such as age of character not matching age or appearance of actress) to restrict it to scene-level features. Fig. 2 shows the resulting thesaurus. They then carried out individually the annotation over the rest of the 20 movies.



	B	C	E	F	G	K	L	N	O	P
1	ann2_start_frame	ann2_end_frame	ann2_label	ann2_concepts	ann2_annotator_comment	ann1_start_frame	ann1_end_frame	ann1_label	ann1_concepts	ann1_annotator_comment
2	0	447	Easy Negative	['']	['']	0	446	Easy Negative	['']	['']
3	448	1070	Hard Negative	['Type of shot', 'Body', 'Appearance']	['']	447	1069	Sure	['Type of shot', 'Bod	['just feet of men and women']
4	1071	1214	Sure	['Type of shot', 'Body', 'Clothes']	['']	1070	1213	Sure	['Type of shot', 'Cloth	['']
5	1215	2558	Easy Negative	['']	['']	1214	2123	Hard Negative	['Look', 'Exp of emo	['']
6	2559	3021	Hard Negative	['Body', 'Activities']	['']	2124	2605	Easy Negative	['']	['']
7	2559	3021	Hard Negative	['Body', 'Activities']	['']	2606	3018	Sure	['Posture', 'Look']	['girl looking at photo, woman i
8	3022	3551	Hard Negative	['Look', 'Exp of emotion']	['']	3019	3551	Sure	['Look', 'Exp of emo	['']
9	3552	4120	Hard Negative	['Posture', 'Exp of emotion']	['']	3552	3749	Easy Negative	['']	['']
10	3552	4120	Hard Negative	['Posture', 'Exp of emotion']	['']	3750	4005	Sure	['Body', 'Activities']	['implied kid doing something ur
11	3552	4120	Hard Negative	['Posture', 'Exp of emotion']	['']	4006	4093	Easy Negative	['']	['']
12	4121	4976	Sure	['Speech', 'Look', 'Exp of emotion']	['']	4094	4976	Sure	['Speech', 'Look', 'Ex	['perpendicular look, uneven en
13	4977	5687	Hard Negative	['Body', 'Voice', 'Clothes', 'Exp of em	['']	4977	5641	Hard Negative	['Speech', 'Exp of em	['']
14	5688	7010	Hard Negative	['Voice', 'Speech', 'Exp of emotion']	['']	5642	6595	Sure	['Voice', 'Speech', 'L	['talk about masturbating while
15	5688	7010	Hard Negative	['Voice', 'Speech', 'Exp of emotion']	['']	6596	7004	Hard Negative	['Speech']	['']
16	7011	7818	Easy Negative	['']	[' ']	7005	8008	Easy Negative	['']	['']
17	7819	8011	Easy Negative	['']	['']	8009	8871	Sure	['Voice', 'Clothes', 'S	['far away man looking, revealin
18	8012	8654	Hard Negative	['Clothes']	['']	8009	8871	Sure	['Voice', 'Clothes', 'S	['far away man looking, revealin
19	8655	8846	Sure	['Posture', 'Look']	['']	8009	8871	Sure	['Voice', 'Clothes', 'S	['far away man looking, revealin
20	8847	9261	Hard Negative	['Type of shot', 'Soundtrack']	['context: seems like obj but n	8872	9266	Sure	['Type of shot', 'Cloth	['camera bottom up, suit']
21	9262	9468	Sure	['Clothes', 'Speech']	['']	9267	9468	Sure	['Clothes', 'Speech']	['']
22	9469	9654	Sure	['Type of shot', 'Voice', 'Look', 'Exp o	['']	9469	11164	Sure	['Type of shot', 'Cloth	['talking down from higher posit
23	9655	11296	Sure	['Type of shot', 'Voice', 'Posture', 'Sp	[' ']	11165	12241	Sure	['Clothes', 'Posture', '	['']
24	11297	11914	Sure	['Speech']	['']	11165	12241	Sure	['Clothes', 'Posture', '	['']
25	11915	12245	Sure	['Type of shot', 'Body', 'Clothes', 'Loo	['']	12242	13893	Sure	['Voice', 'Speech', 'L	['breathy soft voice, Cal talking
26	12246	13869	Hard Negative	['Voice', 'Look', 'Exp of emotion', 'Sp	['']	12242	13893	Sure	['Voice', 'Speech', 'L	['breathy soft voice, Cal talking
27	13870	14122	Easy Negative	['']	['']	13894	15578	Sure	['Voice', 'Speech', 'L	['same as before, breathy voice
28	14123	15368	Hard Negative	['Voice', 'Speech', 'Exp of emotion']	['']	13894	15578	Sure	['Voice', 'Speech', 'L	['same as before, breathy voice
29	15369	15768	Hard Negative	['Posture', 'Exp of emotion']	['']	15579	15780	Easy Negative	['']	['']
30	15769	16253	Easy Negative	['']	['']	15781	15930	Hard Negative	['Body', 'Clothes']	['']
31	16254	16421	Sure	['Speech']	['']	15931	16377	Sure	['Speech']	['place is nice, lotta pretty girls']
32	16422	17177	Hard Negative	['Body', 'Clothes']	['']	16378	17538	Sure	['Clothes', 'Look']	['']
33	17178	17562	Hard Negative	['Body', 'Clothes', 'Look']	['']	17539	17675	Sure	['Clothes', 'Exp of em	['woman intentionally laughing t
34	17563	17728	Sure	['Body', 'Voice', 'Clothes', 'Activities',	['']	17676	19371	Easy Negative	['']	['']
35	17729	19732	Easy Negative	['']	['']	19372	21726	Sure	['Speech', 'Look', 'Ex	['']
36	19733	21983	Hard Negative	['Voice', 'Exp of emotion']	['']	21727	22058	Easy Negative	['']	['']
37	21984	24755	Easy Negative	['']	['']	22059	23029	Sure	['Clothes', 'Look']	['']
38	21984	24755	Easy Negative	['']	['']	23030	24308	Not Sure	['Speech']	['the men are dictated / obliged
39	24756	25052	Sure	['Type of shot', 'Body', 'Clothes', 'Loo	['2 possibly objectified, includ	24309	25428	Sure	['Type of shot', 'Spee	['']
40	25053	25454	Sure	['Type of shot', 'Speech', 'Look']	['sure due to speech, hn for v	25429	28394	Sure	['Speech']	['talk about manhood and using
41	25455	26277	Hard Negative	['Speech', 'Activities']	['']	25429	28394	Sure	['Speech']	['talk about manhood and using
42	26278	29508	Hard Negative	['Speech']	['also context '']	28395	29463	Hard Negative	['Speech']	['wife is going to rue the day sh
43	29509	29600	Sure	['Posture', 'Speech', 'Activities']	['']	29464	29897	Sure	['Clothes', 'Speech', '	['fancy face, revealing clothes']
44	29601	30118	Sure	['Speech', 'Activities', 'Look']	['']	29898	30794	Easy Negative	['']	['']

Figure 19: Excerpt of an alignment file created after the pilot annotation by both experts so support the discussion leading to invalidate or modify certain concepts or aligning their interpretations.

A.2.4 Alternate representation of the thesaurus

A color-blind version of the thesaurus is shown in Fig. 20.

A.2.5 Examples of annotations

Examples of annotated segments are detailed in the case where objectification is produced by visual concepts mainly in Fig. 21, by textual concept only in Fig. 22, and by a multimodal combination of visual, textual and audio concepts in Fig. 23.

A.2.6 Details on inter-annotator agreement analysis

We provide the code for computing IAA metrics in the Github repository.

Definition of dist@IoUthresh As expected distances are used for chance correction, we take them as distances between any annotations (regardless of annotator) between 2 different movies. Between 2 movie annotations **a** and **b**, we defined the distance itself as follows. Considering **a** as GT, for each segment $s_a \in \mathbf{a}$, we select the subset S_b of segments in **b** so that IoU≥IoUthresh. For each such segment s_b, we set dist(s_a, s_b)=D_0 if the objectification labels are the same, $= D_1$ if the labels are one hop apart (we set the order EN - HN- S), $= D_2$ if 2 hops apart. We discard all NS segments that represent less than 5%. We then take dist(s_a, S_b) as the minimum of dist(s_a, s_b) for all $s_b \in S_b$, then dist(S_a, S_b) as the average of dist(s_a, S_b) for all $s_a \in S_a$, and finally $D(S_a, S_b) = (dist(S_a, S_b) + dist(S_b, S_a))/2$. D_0, D_1 and D_2 are hyper-parameters we set to 0, 1 and 2, respectively. We verify their values do not impact the IAA metrics provided the order remains the same (as it can be understood from the following explanation).

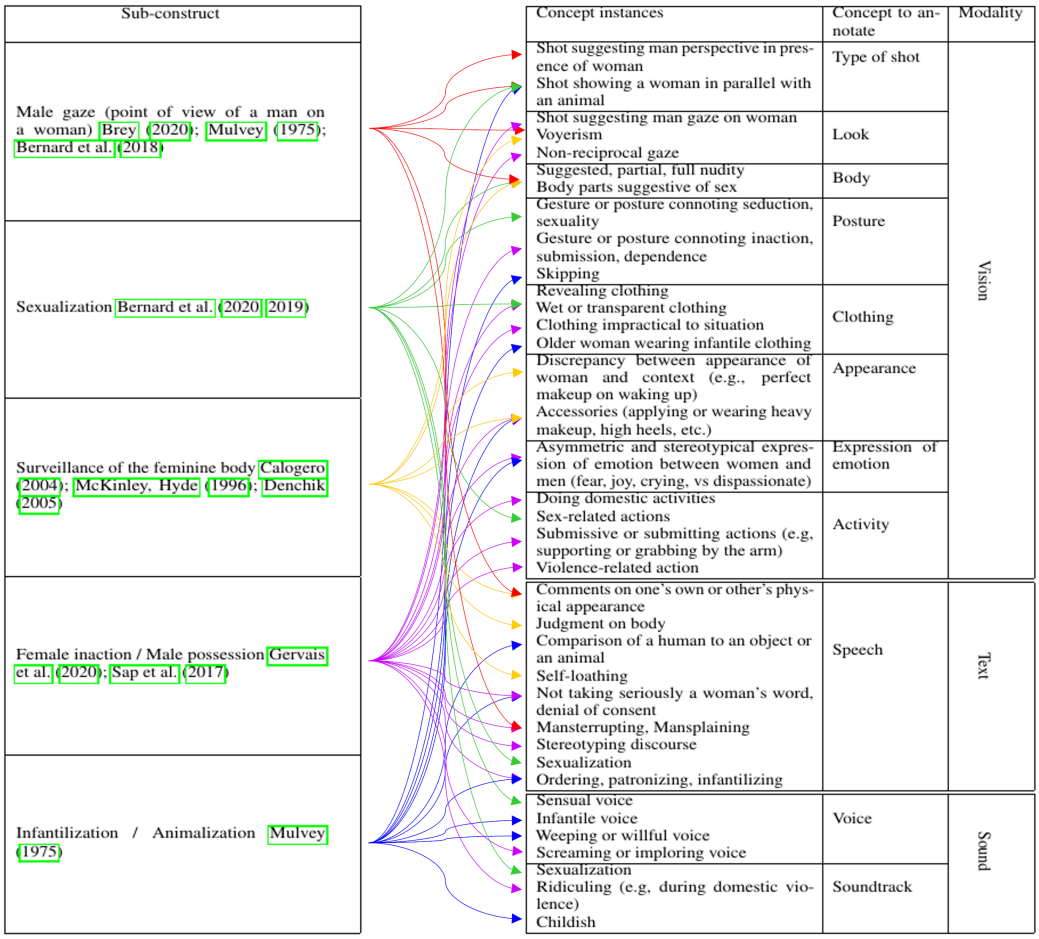


Figure 20: Thesaurus for the construct of objectification: 5 sub-constructs (left table) manifested through 11 concepts spanning 3 modalities (right table).

Formal intuition for the difference between Krippendorff's α and σ with unitizing Let us consider that each segment label can take $V = 2$ values, instead of $V = 3$ (EN, HN, S, and we discard distance value D_2). Let us assume for now that every segment s_a of an annotation sequence can be matched with exactly $K = 1$ other annotation segment s_b. Let us assume that the annotators agree on a fraction F_A of the samples, and chance correction is made by matching with another random segment with a label picked uniformly at random. Then $\Pr(s_a \neq s_b) = 1 - F_A$ if s_a and s_b are from the same item, otherwise $\Pr(s_a \neq s_b) = V(1/V)(1 - V)$. Taking $V = 2$ and $F_A = 0.8$ yields $\alpha = 1 - \Pr(s_a \neq s_b|\text{same item})\ D_1/(\ \Pr(s_a \neq s_b|\text{different items})\ D_1) = 1\text{-}0.2/0.5 = 0.6$ as expected. Let us now consider a non-perfect agreement on the segment boundaries, and that under a certain IoUthresh constraint, we have exactly $K = 2$ segments s_b^1 and s_b^2 that can be matched with s_a for different items, while this is possible only with probability f for same items (to consider a non-trivial random agreement on boundaries for a given movie). Then we can express: $\Pr(s_a \neq s_b^1 \text{ and } s_a \neq s_b^2|\text{same item}) = (1-f)[V(1/V(1-F_A))] + f[V(1/V(1-F_A)(1-1/V))] = (1 - F_A)(1 - 1/Vf)$,



Figure 21: Examples of segments delimited and tagged with a Sure level of objectification, produced by only or mainly visual concepts.

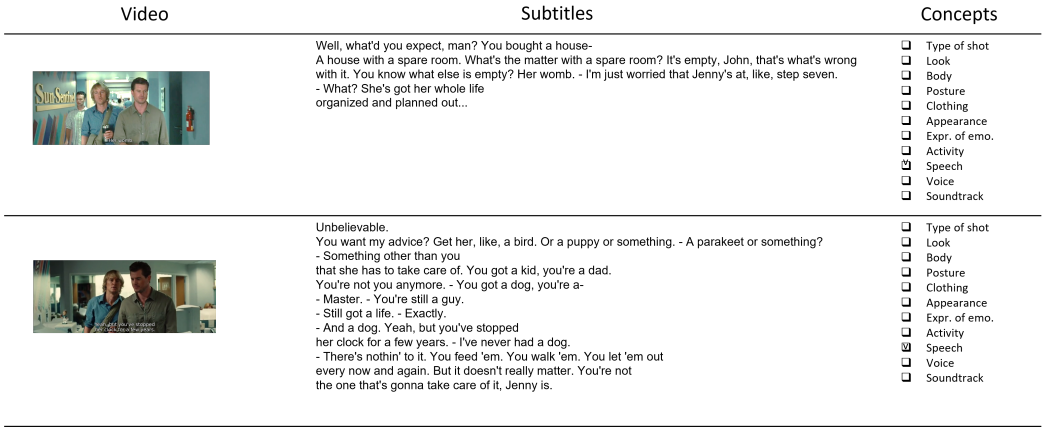


Figure 22: Examples of segments delimited and tagged with a Sure level of objectification, produced by only textual concept.



Figure 23: Examples of segments delimited and tagged with a Sure level of objectification, produced by a combination of concepts of different modalities.

and $\Pr(s_a \neq s_b^1 \text{ and } s_a \neq s_b^2 | \text{different items}) = V(1/V(1-1/V)^K) = (1-1/V)^K$
Taking $V = 2$ and $K = 2$:
$\alpha = 1 - (1 - F_A)/0.5(2 - f)$
so for $f = 1$: $\alpha = 0.6$ again
but for $f < 1$, e.g., $f = 0.7$: $\alpha = 0.22$.

This explains the partly opposite trends of α and σ we observe in Fig. 3 (remind that there in the real case V=3): increasing IoUthresh yields a monotonously decreasing α while σ stays constant or slightly increases, before decreasing sharply after some high IoUthresh (above the boundary agreement rate), as expected. Therefore, as an IAA metric should describe how much the overall data is better than chance, after observing the weakness of Kripendorff's α@IoU to describe this phenomenon, we opt for reporting σ instead.

A.2.7 Details on computation for external consistency verification

For visual concepts, we do contrastive scoring precisely as follows:
1. Cosine similarity is computed between every window embedding in the segment and every prompt embedding for that concept, giving a *(windows × prompts)* similarity matrix.
2. This is aggregated across windows — either by the mean, or by a chosen quantile (e.g., the 90th percentile, capturing "peak" rather than "average" presence across the segment) — giving one similarity value per prompt.
3. This is then aggregated with max-pooling across the concept's multiple prompts, giving a single raw concept score for the segment.
4. CLIP-family text embeddings are known to be anisotropic. We correct this with embedding centering and similarity offset. First for embedding centering, a single "background" vector is computed as the mean of all prompt embeddings (every concept's prompts and the negative prompts, pooled together), and this background is subtracted from every individual prompt embedding before re-normalizing to unit length. The same idea is applied on the video side: each movie's window embeddings are centered by that movie's own average window embedding (removing shared, movie-specific bias such as color grading or cinematographic style) before re-normalizing. Second for similarity offset, the identical two-step aggregation above is applied to the negative/background prompts, giving a single background score for the segment.
5. The final concept score = raw concept score - background score.

A.3 Experimental analysis

Below we define formally the confidence intervals (CIs) we use for model comparison, building on Mozer et al. (2024); Loftus and Masson (1994); Masson and Loftus (2003).

We want to compare K types of models on a given metric X, each type of model being trained and evaluated on n (fold, seed) pairs, referred to as *subjects*. (Here $n = 5 \times 5 = 25$.) In total, we have $N = nK$ observations. For $k \in [\![1, K]\!]$, for $i \in [\![1, n]\!]$, we note x_{ik} the metric for the i-th instance of model k. For $k \in [\![1, K]\!]$, we note $\bar{x}_k$ the average metric for model k, while $\bar{x}$ represents the average metric over the whole population of nK models.

While still corresponding to the outcome of Student tests, the following confidence intervals rely on the usual instrumental ANOVA quantities named sums of squares (SS):

```
VISUAL_CONCEPT_PROMPTS = {
    "Body": [
        "suggested, partial, or full nudity",
        "body parts suggesting sex",
        "a photo of a woman with visible cleavage",
        "a low-cut top exposing cleavage",
        "a close-up of a chest with cleavage showing",
        "a person wearing a v-neck shirt",
        "a photo focused on the buttocks",
        "a close-up of a person's butt",
        "a rear view showing the buttocks",
        "a photo focused on the buttocks",
        "a close-up of a person's butt",
        "a rear view showing the buttocks",
        "a photo of a bare midriff or torso",
        "an exposed stomach and torso",
        "a cropped top showing the torso",
        "a close-up of full, plump lips",
        "a photo emphasizing large lips",
        "a close-up of lips wearing red lipstick",
        "glossy lipstick on lips",
        "a person with visible lipstick makeup",
        "a sexy woman",
        "a sexy man",
        "a bare torso",
        "a nude person",
        "bare arms",
        "bare legs"
    ],
    "Clothing": [
        "revealing, wet, or transparent clothing",
        "clothing impractical for the situation",
        "an older woman wearing infantile clothing",
        "a photo of a woman with visible cleavage",
        "a low-cut top exposing cleavage",
        "a close-up of a chest with cleavage showing",
        "a person wearing a v-neck shirt",
        "a top with a v-shaped neckline",
        "a v-neck sweater",
        "a person wearing a sleeveless tank top",
        "a tank top exposing the shoulders",
        "a fitted tank top",
        "a person wearing tight-fitting pants",
        "skinny jeans hugging the legs",
        "leggings showing body contour",
        "a woman wearing a short mini-skirt",
        "a skirt ending above mid-thigh",
        "a short skirt exposing the thighs",
        "a woman wearing high heel shoes",
        "high-heeled stiletto shoes"
    ],
    "Type of shot": [
        "a point-of-view shot looking at a woman's body",
        "a low angle shot looking up a woman's body",
        "a camera panning slowly up and down a woman's body",
        "a shot from a man's point of view looking at a woman",
        "a woman filmed in parallel with an animal",
        "a woman and an animal shown side by side",
        "a woman posed like an animal",
        "a shot visually comparing a woman to an animal",
    ],

    "Look": [
        "a man staring intently at a woman's body",
        "an objectifying gaze directed at a woman",
        "a person's eyes fixed on a woman's chest or body",
        "a voyeuristic shot of a woman undressing",
        "someone secretly watching a woman through a window",
        "a hidden or concealed camera view of a woman",
        "spying on a woman through a gap or opening",
        "a woman being stared at without noticing",
        "a woman unaware she is being watched",
        "one-sided staring at an unaware woman",
    ],

    "Posture": [
        "a woman posing seductively",
        "a person in a sexually suggestive pose",
        "a woman arching her back in a suggestive pose",
        "a provocative, sexualized body pose",
        "a woman looking passive and submissive",
        "a person cowering in a submissive posture",
        "a woman in a docile, passive posture",
        "a person looking helpless and dependent on someone else",
        "a woman skipping playfully",
        "a person skipping like a child",
    ],

    "Appearance": [
        "a woman overdressed for a casual, ordinary setting",
        "a woman in glamorous clothing in an everyday setting",
        "a mismatch between a woman's clothing and the situation",
        "a woman dressed inappropriately for the occasion",
        "a close-up of a woman applying makeup",
        "a woman putting on lipstick in a mirror",
        "a woman wearing heavy, elaborate makeup",
        "a woman applying mascara or eyeliner",
        "a close-up of a woman putting on high heel shoes",
        "a woman doing an elaborate beauty routine",
    ],

    "Exp of emotion": [
        "a woman crying",
        "a woman in tears, sobbing",
        "a woman being emotionally expressive and distraught",
        "a man with a completely emotionless expression",
        "a man showing no emotion, stoic and unmoved",
        "a calm, dispassionate man next to a distressed woman",
        "an emotionally expressive woman contrasted with an unemotional man",
    ],

    "Activities": [
        "a woman cooking in the kitchen",
        "a woman cleaning the house",
        "a woman doing laundry or domestic chores",
        "domestic housework being performed by a woman",
        "a couple kissing passionately",
        "an intimate physical act between two people",
        "a suggestive sexual gesture",
        "a person kneeling submissively before someone",
        "a woman being physically restrained or held down",
        "a person grabbing another's arm forcefully",
        "a woman being pulled by the arm against her will",
        "a person being hit or struck",
        "a violent physical altercation",
        "a person being pushed forcefully",
        "domestic violence between two people",
    ],
}
```

Figure 24: The text prompt used for the X-CLIP similarities.

```
CURATED_CATEGORIES = [
    # Body / appearance
    "attractive", "masculine", "feminine", "sexual", "body", "appearance",
    "beauty", "ugliness", "clothing", "fashion", "youth", "exotic",
    "shape_and_size", "lust", "strength", "weakness",
    # Power / dominance / agency
    "dominant_heirarchical", "dominant_personality", "power", "leader",
    "independence", "order", "competing", "achievement", "gain",
    "stealing", "deception", "trust", "negotiate",
    # Emotion
    "cheerfulness", "pride", "shame", "fear", "nervousness", "horror",
    "disgust", "anger", "rage", "sadness", "joy", "contentment",
    "positive_emotion", "negative_emotion", "envy", "surprise",
    "disappointment", "timidity", "irritability", "exasperation",
    "torment", "suffering", "optimism", "zest", "confusion", "love",
    "affection", "warmth", "sympathy",
    # Speech / communication / conflict
    "politeness", "ridicule", "swearing_terms",
    "aggression", "dispute", "fight", "violence", "kill", "injury",
    "pain", "hate", "neglect", "childish",
]
```

Figure 25: The semantic categories used for the Empath analysis of Speech concept.

Table 19: Proportion of positives (in EN vs HN∪S perspective) in each movie

Movie title	fraction of positive (in duration)	fraction of positive (in nb of clips)
As Good as It Gets	0.30	0.68
Crash	0.32	0.59
Crazy,Stupid,Love	0.68	0.77
Flight	0.31	0.64
Forrest Gump	0.23	0.57
Friends with Benefits	0.38	0.67
Gone Girl	0.33	0.58
Indiana Jones and the Last Crusade	0.13	0.54
Juno	0.12	0.68
Marley & Me	0.36	0.52
Pulp Fiction	0.35	0.59
Silver Linings Playbook	0.30	0.64
Sleepless in Seattle	0.55	0.65
The Day the Earth Stood Still	0.15	0.34
The Girl with the Dragon Tattoo	0.32	0.58
The Help	0.56	0.66
The Social Network	0.26	0.59
The Ugly Truth	0.63	0.75
Titanic	0.27	0.57
Up in the Air	0.42	0.64

$$SS_{total} = \sum_{k=1}^{K}\sum_{i=1}^{n}(x_{ik} - \bar{x})^2$$

$$SS_{between} = \sum_{k=1}^{K} n(\bar{x}_k - \bar{x})^2$$

$$SS_{within} = \sum_{k=1}^{K}\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2$$

In particular, the following relationship holds : $SS_{total} = SS_{between} + SS_{within}$.

When the subjects are paired across the groups, which is the case here, we can further decompose the variance: $SS_{within} = SS_{subject} + SS_{error}$, where

$$SS_{subject} = K\sum_{i=1}^{n}(\bar{x}_i - \bar{x})^2 \text{ , and}$$

$$SS_{error} = SS_{within} - SS_{subject}$$

Corresponding mean sums of squares are obtained by dividing sums of squares by their corresponding degrees of freedom. In particular:

$$MS_{within} = \frac{SS_{within}}{df_{within}} \text{ , where } df_{within} = N - K$$

Assuming the homogeneity of variance across groups (this has to be verified in practice), MS_{within} is an estimate of $\hat{\sigma}^2$ the (common) variance of X in each group. Similarly:

$$MS_{error} = \frac{SS_{error}}{df_{error}} \text{ , where } df_{error} = (K-1)(n-1)$$

MS_{error} is nothing else but an estimate of $\hat{\sigma'}^2$ the (common) variance of each group on the normalized data, i.e., the data we removed the inter-subject variability of (for more details, see Loftus and Masson (1994)).

As proven in Loftus and Masson (1994), under the assumption that the (fold, seed) variability plays no role in the comparison of models, the $(1-\alpha)$-confidence interval corresponding to the Student test comparing models k and j is the following:

$$CI_{kj} = \bar{x}_k \pm \sqrt{\frac{2\ MS_{error}}{n}}\ t_{1-\alpha/2}(df_{error})$$

In other words : if $\bar{x}_j$ does not belong to CI_{kj}, with a level of confidence $(1-\alpha)$, the models k and j are significantly different. Because the width of this confidence interval is the same for all (k, j) pairs, we can derive, for each model k, the confidence interval:

$$CI_{diff\ k} = \frac{1}{2}CI_{kj} = \bar{x}_k \pm \sqrt{\frac{MS_{error}}{2n}}\ t_{1-\alpha/2}(df_{error})$$

so that, for any two models k and j, non-overlapping intervals $CI_{diff\ k}$ and $CI_{diff\ j}$ can be interpreted as a significant difference.

We also comment on effect sizes (between such populations k and j) considered as Cohen's d, which is nothing else than their standardized difference. Consistently with the work above, we compare the normalized populations, i.e. the population we removed the inter-subject variability of, so we have:

$$d = \frac{\bar{x}_k - \bar{x}_j}{\hat{\sigma'}}$$

An effect size larger than 0.9 is considered to be very large (Sawilowsky (2009), Funder and Ozer (2019)).

A.4 Details on the models

A.4.1 Common elements of the training procedure over all the models

We proceed with cross-fold validation with 5 folds also shown in Table 17. The folds are made so as to have an even representation of the genres. The classes for the learning tasks are determined as described in the main article (Sec. 4.1). Training is always done with random oversampling on the minority class. The validation set is used for early stopping with patience of 10.

A.4.2 X-CLIP+Transf.

Data Preparation We adapt X-CLIP (Ni et al. (2022)), an extension of CLIP for videos (Radford et al. (2021)). We keep the pre-trained model frozen and extract a 512-dimensional feature vector on every window of 16 frames, with a stride of 16, for each input video segment. The obtained vectors are mean-pooled

Model Our baseline architecture consists of two main components: a self-attention encoder and a classification head. Input visual features are first adjusted to a fixed sequence length of 160 tokens using either zero-padding (for shorter sequences) or uniform down-sampling (for longer ones). A corresponding binary mask is generated to indicate valid token positions within each sequence. The self-attention encoder begins with a patch embedding module that projects the 512-dimensional input features into a 128-dimensional embedding space through a sequence of normalization and linear layers. A special classification token (CLS token) is appended to each input sequence and is treated as a learnable embedding. Then the resulting embeddings, including the CLS token, are given as input to a Transformer encoder comprising 2 layers, each with 8 attention heads. Each layer includes a multi-head self-attention mechanism, followed by a feed-forward network that expands the hidden dimension to 1024. GELU activations and dropout are applied within the feed-forward sublayers. The classification head processes the latest CLS token through two linear layers with a ReLU activation in between. The final output is a set of logits, to which a sigmoid activation is applied for binary classification.

Training The model is trained with a batch size of 32 for 100 epochs. Early stopping and learning rate scheduling (with reduction on plateau) are applied based on validation

loss performance. Training on a single fold takes approximately 30 minutes on an NVIDIA GeForce GTX 1080 Ti GPU.

A.4.3 Language model Bert Embeddings + Transf

Data Preparation We consider binary classification where we want to detect whether there was an objectifying element in the textual transcription of the speech of the characters. All the annotated segments are associated with the corresponding span of subtitles, and the positive class is made of the segments with the speech concept annotated. The text of the subtitles was downcased and HTML tags were removed.

Model To extract textual representations, subtitle text is tokenized using a bert-base-uncased tokenizer with padding to 512 tokens and truncation. In the absence of subtitle text, a placeholder sequence of 512 [MASK] tokens is used to ensure consistent input length. The tokenized inputs, along with the attention masks, are processed through a frozen BERT model to obtain contextual token embeddings (last hidden state) and the corresponding attention mask. These embeddings are then processed by a transformer-based architecture, which serves as the backbone of our model and follows the design detailed in Subsec. A.4.2.This architecture matches the vision baseline in both structure and parameters (approximately 800k).

Training We introduce an initial warmup phase stabilise the learning process. Specifically, we increased incrementally the learning rate from zero to the base value (0.00001) over 10% of the total training steps. During warmup, the learning rate increases linearly with each step, controlled by a custom scheduler. After the warmup, a ReduceLROnPlateau scheduler adjusts the learning rate based on validation loss performance, reducing it when improvement plateaus with a patience of 5 epochs. Early stopping with a patience of 10 epochs is used to halt training. Maximum number of epochs is set to 100 and the batch size to 32.

A.4.4 Audio model

Data preparation To encode the audio track to capture aspects of voice and soundtrack, we select the speech audio encoder Audio Spectrogram Transformer (AST) [3] We process each clip by segmenting it into 2-second audio chunks. We extract feature representation for each chunk using AST kept frozen.

Model AST features are first adjusted to a fixed sequence length of 15 tokens, corresponding to 30 seconds, using either zero-padding (for shorter sequences) or uniform down-sampling (for longer ones). A binary mask is generated to indicate valid token positions within each sequence. The baseline architecture follows the same Transformer-based design used for the vision and text modality, with a comparable number of trainable parameters (approx. 770k).

Training The training configuration for this Transf. architecture is the same as BERT+Transf, described in A.4.3. Specifically, the dropout rate is set to 0.1. The batch size is 32 and the warmup phase similar to that for BERT+Transf in Sec. A.4.3 above.

3. `https://huggingface.co/docs/transformers/model_doc/audio-spectrogram-transformer`.

A.4.5 Multimodal model

The multimodal model described below is generalizable to any combination of modalities.

Model For a given combination of modalities, we use the corresponding features described above. Formally, for a given observation: a video feature is a 160-token long sequence of 512-dimensional vectors; a text feature is a 768-token long sequence of 512-dimensional vectors, and an audio feature is a 160-token long sequence of 768-dimensional vectors. First, all modalities are projected to a common feature dimension (equal to the smaller feature dimension from all modalities); then they are concatenated along the temporal dimension, i.e., their tokens are concatenated.

This new dimensionality aside, the architecture of the multimodal model is identical to the Transformers model described in A.4.2. For a text-vision multimodal model, this represents approximately 1,500k parameters.

Training For a given objectification task, we only keep as positive the examples that present at least one concept from the chosen modalities; others are treated as negative. For instance, for training a text-vision model, a Hard Negative example whose only annotated concept is Soundtrack is treated as a negative.

The new dimensionality and relabeling aside, the training of the multimodal model is identical to the training of the Transformers model described in A.4.2.

A.4.6 Details on MLLMs

The following hyperparameters are common to all our models.

When using MLLMs, video inputs are actually provided in the form of sampled frames: from a given video clip, we uniformly sample 8 frames that are compressed so they do not exceed 50,176 pixels. This definition is sufficient to preserve semantic information of the frames.

MLLMs perform inference by sampling, based on their predicted output distribution for each output token. This sampling is fully characterized by the parameters `temperature` and `top_p`. We use standard values `temperature` $= 0.1$ and `top_p` $= 0.9$, while enforcing reproducibility thanks to a random seed.

For our classification task, we require the answer to be enclosed in `<answer></answer>` tags. If this formatting is not respected and recovery strategies (such as searching at the end of the output text) are unsuccessful, we consider the default prediction to be the negative class. In practice, the arbitrary choice of the default class has little influence on the final performance, given that only 1% of samples require the default prediction.

The following prompt is used for the task EN vs HNS, that is detecting the presence of objectifying concepts.

```
Your task is to analyze whether the given video presents any objectifying concept.
Objectification can be due to the presence of various visual elements, including:
voyeuristic gaze, suggested or partial nudity, non-reciprocal gaze,
gesture or posture connoting submission,
discrepancy between appearance of woman and context (e.g., perfect makeup on
    waking up),
stereotypical activities, etc.
```

```
Think before answering, and generate:
1. A structured reasoning process enclosed in <think></think> tags
2. A final answer enclosed in <answer></answer> tags

For objectifying videos, the reasoning should be based on a structured 2-step
    process:
<think> must include the following steps enclosed in corresponding tags:
<step1>: Scene Description  Provide an objective overview of the environment and
    behaviors in the scene.
<step2>: Objectifying Element Description  Describe the objectifying elements and
    which characters are involved.
<answer> must be a single answer labeling the video as \"Objectifying\"

For normal videos, the reasoning should be based on a structured 2-step process:
<think> must include the following two steps:
<step1>: Scene and Object Description  Provide a concise and objective overview of
     the environment and typical behaviors.
<step2>: Normal Element Explanation  Explain why the video is not objectifying.
<answer> must be a single answer labeling the video as \"Normal\"
%\end{Verbatim}
```

Prompt for few-shot inference:

```
Your task is to analyze whether the given video is objectifying or normal.
Objectification occurs when a person or character is represented as an object of
    desire or service rather than a subject of action.
Objectification can be due to the presence of various visual elements, including:
voyeuristic gaze, suggested or partial nudity, non-reciprocal gaze,
gesture or posture connoting submission,
discrepancy between appearance of woman and context (e.g., perfect makeup on
    waking up),
stereotypical activities, etc.

Think before answering, and generate:
1. A structured reasoning process enclosed in <think></think> tags
2. A final answer enclosed in <answer></answer> tags

For objectifying videos, the reasoning should be based on a structured 2-step
    process:
<think> must include the following steps enclosed in corresponding tags:
<step1>: Scene Description - Provide an objective overview of the environment and
    behaviors in the scene.
<step2>: Objectifying Element Description - Describe the objectifying elements and
     which characters are involved.
<answer> must be a single answer labeling the video as \"Objectifying\"

For normal videos, the reasoning should be based on a structured 2-step process:
<think> must include the following two steps:
<step1>: Scene and Object Description - Provide a concise and objective overview
    of the environment and typical behaviors.
<step2>: Normal Element Explanation - Explain why the video is not objectifying.
<answer> must be a single answer labeling the video as \"Normal\"
```

```
<|video_token|>

Below, you will find example videos and corresponding desired answers.
<|video_token|>The video above is <answer>Normal</answer>. It contains no
    objectifying concept.
<|video_token|>The video above is <answer>Objectifying</answer>. More precisely,
    it contains the following concepts : Type of shot, Look, Exp of emotion.
<|video_token|>The video above is <answer>Normal</answer>. It contains no
    objectifying concept.
<|video_token|>The video above is <answer>Objectifying</answer>. More precisely,
    it contains the following concepts : Type of shot, Body, Clothing.
```

A.5 The task of localizing objectification

We consider Actionformer, a reference model by Zhang et al. (2022) for action detection, that we adapt to the new tasks of classifying (named TClassif) and localizing (named TLoc) objectification. We provide all the implementation details below. Table 20 shows results of Actionformer-Obj on both TLoc (involving localization) and TClassif. Results on TLoc are comparable to those of original Actionformer on EpicKitchen (Zhang et al., 2022, Table 2), showing again the accessibility of the task with the visual modality.

We re-use the code provided by Zhang et al. (2022) in their Github repository [4]. For task TLoc of temporal objectification localization, we use both original regression and classification branches of Actionformer. For task TClassif of classification only, we replace the regression branch by the true segment boundaries and only predict the segment class.

Localization settings The dataset needs to be adapted to task TLoc of temporal objectification localization. To do so, each film is cut into 5-minute clips. All clips not overlapping a positive segment are discarded, to reproduce the data filtering in Zhang et al. (2022). We consider this 5 minute duration to hit a trade-off: we do not consider a smaller value to limit the number of clips we discard because they do not overlap any positive segment, and we do not consider a higher value to limit the difficulty to scale attention. The resulting clips overlap in average two positive segments (with presence of objectification).

Hyperparameters A certain number of hyperparameters need to be adapted to our own dataset. The hyperparameters we adapt are *sequence length* (maximum length of a video in terms of number of features), *window size* (for the attention mechanism) and *regression ranges*. The latter are connected to the possible duration of action detected at each layer. Owing to various constraints including divisibility, we considered (450,15) and (512,17) for sequence length-window size pairs. We selected the latter from performance on a validation set. We also compared the validation performance obtained with the original regression ranges and regression ranges we set from the distribution of objectifying segment durations in our data. The latter gave best results in validation. The ranges are: [[0, 11], [11, 22], [22, 36], [36, 47], [47, 10000]].

4. https://github.com/happyharrycn/actionformer_release

Table 20: Performance on Actionformer-Obj on TLoc and TClassif on the visual modality. Baselines are on TClassif. Class configuration is EN vs S.

Task	Model	tIoU	MAP 0.3	MAP 0.4	MAP 0.5	Average MAP	Recall@1 0.3	Recall@1 0.4	Recall@1 0.5	Average Recall@1
TLoc	Actionformer-Obj		0.252	0.169	0.093	0.171	0.385	0.296	0.198	0.293
TClassif	Actionformer-Obj (w/o reg.)			N/A		0.587		N/A		0.673
	random			N/A		0.281		N/A		0.447
	allpos			N/A		0.195		N/A		0.262
	allneg			N/A		0.364		N/A		0.384