Visual Objectification in Films: Towards a New AI Task for Video Interpretation

Julie Tores, Lucile Sassatelli, Hui-Yin Wu, Clement Bergman, Lea Andolfi, Victor Ecrement, Frédéric Precioso, Thierry Devars, Magali Guaresi, Virginie Julliard, et al.

Visual Objectification in Films: Towards a New AI Task for Video Interpretation

Julie Tores[1,2] Lucile Sassatelli[1,3] Hui-Yin Wu[4] Clement Bergman[4]
Léa Andolfi[6] Victor Ecrement[6] Frédéric Precioso[2] Thierry Devars[6]
Magali Guaresi[5] Virginie Julliard[6] Sarah Lecossais[7]
[1]Université Côte d'Azur, CNRS, I3S, France [2]Université Côte d'Azur, CNRS, Inria, I3S, France
[3]Institut Universitaire de France [4]Université Côte d'Azur, Inria, France
[5]Université Côte d'Azur, CNRS, BCL, France [6]Sorbonne Université, GRIPIC
[7]Université Sorbonne Paris Nord, LabSIC
julie.tores@univ-cotedazur.fr

Abstract

In film gender studies, the concept of "male gaze" refers to the way the characters are portrayed on-screen as objects of desire rather than subjects. In this article, we introduce a novel video-interpretation task, to detect character objectification in films. The purpose is to reveal and quantify the usage of complex temporal patterns operated in cinema to produce the cognitive perception of objectification. We introduce the ObyGaze12 dataset, made of 1914 movie clips densely annotated by experts for objectification concepts identified in film studies and psychology. We evaluate recent vision models, show the feasibility of the task and where the challenges remain with concept bottleneck models. Our new dataset and code are made available to the community.

1. Introduction

In film gender studies, the concept of "male gaze" [41] refers to the way the characters – especially women – are portrayed on-screen as objects of desire rather than subjects. Consider in Figure 1 how objectification is manifested in various ways such as camera placement and movement, the gaze interactions between characters, the choice of clothing, and arrangement of scene elements. Such disparities in how people are presented, depicted or addressed to in digital contents based on their gender has large-scale social implications such as the perpetuation of harmful stereotypes and hostile social situations.

These disparities have been the subject of an increasing number of studies at the intersection of social and computational sciences. In online social networks, computational approaches to sexism detection have been increasingly investigated for textual data as a part of hate speech detection.

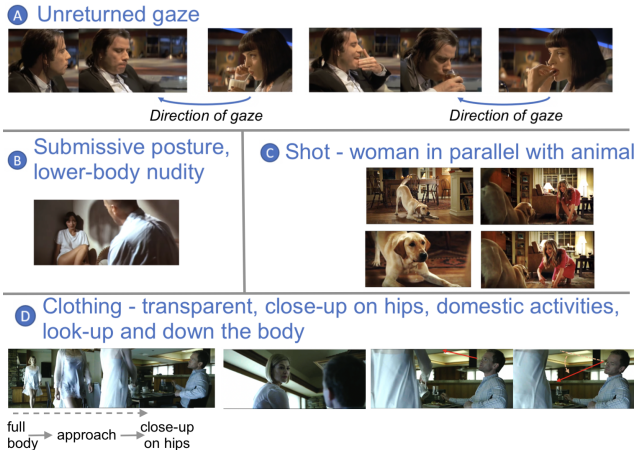

Figure 1. In modern film media, the unequal characterization of gender on screen frequently evokes concepts of objectification, such as (A) unequal gaze (*Pulp Fiction*, 1994), (B) Nudity and submissive postures (*Pulp Fiction*, 1994), (C) animalisation or infantilisation (*Marley and Me*, 2008), and (D) transparent clothing, camera framing, domestic gender roles, and voyeurism (*Gone Girl*, 2014).

As explained by Samory et al. [47], sexism is a complex sociological construct, whose high-level interpretive nature and subtle dimensions beyond offensive speech make for an unsolved challenge. In visual media such as films and TV series, the characters they depict shape our collective imagination and perception of sociological constructs, such as gender, race, and class. Currently, most large-scale approaches to understanding gender representation in these media have focused on quantifying the presence of women in the image and audio content (e.g., [35, 39, 51]). However, works in social sciences show that quantifying the presence of gender on screen is insufficient for grasping the issue of

gender inequalities in visual media. For films, the classic Bechdel test, although useful and simple, considers neither the visual modality, which is key to analyzing gender depiction [41], nor the textual constructs of speech and dialogue [49]. While a few works have investigated sexist memes [16, 19], sexist advertisement [18], and characterized the on-screen positioning and co-occurrence of certain groups with respect to scene types and objects (e.g., Wang et al. for still images [54], Jang et al. [26] for films), computational approaches to interpretive sexism in visual media remain very scarce.

In this article, we introduce a new challenging task for computer vision: detecting character objectification in films. Owing to the importance of this question, we consider it is critical to support the design of explainable methods and fine-grained model error analysis, which we address by densely annotating video data for theory-driven concepts. This presents a major step to tackle the question of subtle sexism in videos, operationalizing the popularly known concept of male gaze with the construct of objectification, and specifically considering the temporal dimension where such video patterns unfold. The end goal is to enable large-scale quantification and characterization of complex patterns producing on-screen objectification, and unveil possible correlations along the lines of the gender or race constructs.

Our contributions are:

• We introduce a novel video-interpretation task to detect character objectification in films. This is an interpretive task, hence extending beyond the more classical yet still challenging video-understanding tasks, and involving a subjective judgement. In a team involving media studies experts, and building on results in cinematography and psychology, we design a thesaurus of visual objectification, defining coarse-grained concepts with exemplified instances. This thesaurus is then used to formulate precise annotation guidelines. We introduce the ***ObyGaze12*** dataset to the community, with 1914 clips from 12 films densely annotated by experts for concepts of objectification, including hard negative examples. It corresponds to 25% of the MovieGraphs dataset. We verify the consistency of the obtained data, and provide first analyses showing the compositional nature of objectification. The dataset is meant to explore the complex temporal patterns producing character objectification in films. To the best of our knowledge, it is the first work proposing a computational approach to this interpretive task in videos.

• We verify that the new task of objectification detection in videos is accessible by testing recent vision and vision-language models, and that hard negative examples improve classification. We also investigate the model weaknesses in representing each objectification concept. To do so, a thorough analysis is carried out with Concept Bottleneck Models (CBMs) which allows us to identify the individual concepts that pose the greatest challenge, namely: Type of Shot, Look, Posture, and Appearance.

To the best of our knowledge, this is one of the few video datasets with dense concept-based annotations for a high-level construct, and the first for objectification. The dataset and code used in this article is entirely provided in https://github.com/husky-helen/ObyGaze12.

The article is organized as follows. In Sec. 2, we first provide a review of the relevant works on visual biases in films, dataset creation, and models for video understanding. We then introduce dataset creation for *ObyGaze12* and present first analyses in Sec. 3. Sec. 4 presents the evaluation of models on the new task, and the analysis of the difficulty of concept representation with CBMs. Finally, we provide discussions on ethical aspects and challenges in Sec. 5, and conclusions in Sec. 6.

2. Related works

In this section, we position our contributions with respect to the relevant existing work. First, we introduce biases in visual datasets and computational approaches to the analysis of visual gender representation in films. We then discuss interpretive-level tasks, increasingly common in natural language processing, and the approaches to dataset creation to instantiate them for ML approaches. Here we highlight the scarcity of visual datasets made for high-level interpretive tasks, in particular for video data. Finally, we introduce the video understanding approaches that we consider to benchmark on our new interpretive task, specifically focusing on explainable concept-based approaches to locate the challenges ahead in video interpretation for this new task.

2.1. Visual biases in film datasets

The task we introduce is connected to the general problem of bias detection. As exposed by Fabbrizzi et al. [15], biases in visual datasets can be classified into selection bias (how subjects are included in a dataset), framing bias (how the visual content has been artificially composed) and label bias (errors or disparities in the labelling data). Our contribution is closely related to the framing and labelling biases. In film datasets, the first studies of biases in gender representation were from a presence quantification perspective. Guha et al. [21] and Somandepalli et al. [51] automatically estimate the screen time (from video) and speaking time (from audio) of male and female characters in Hollywood movies. They show that women are seen (36% screen time) and heard (41% screen time) significantly less than male characters. Mazieres et al. [39] consider a movie dataset spanning three decades, and show a temporal trend towards general fairer representation between both binary genders.

They however show simultaneously that the applied framings remain unfair, with only 40% of one-face frames featuring a female (60% for males). Jang et al. specifically analyze the qualitative framing differences in gender portrayal in a dataset of 20 Hollywood movies and 20 Korean movies [26]. They find that female characters are portrayed with lower emotional diversity, spatial occupancy, temporal occupancy, intellectual image or mean age.

In contrast in this article, we take a first step towards detecting bias in gender representation from a high-level construct, objectification, qualitatively described in various disciplines such as cinematography [9, 41], social psychology [1, 32], and neuroscience [5–7]. Objectification is produced by complex temporal patterns never analyzed computationally in videos until now.

2.2. Interpretive-level tasks and dataset creation

At the same time that biases in visual datasets are uncovered and analyzed, other approaches aim to detect bias in human data. Detecting highly interpretive constructs, such as hate speech, propaganda, sexism, and racism, has been a long-standing endeavor in NLP. Da San Martino et al. [12] recruited to 4 experts to annotate news articles with text spans associated to 18 possible propaganda techniques. Samory et al. uncover the challenge of construct complexity for sexism detection [47]. They observe that multiple articles for automating this task consider widely different definitions of sexism, often referring to sub-dimensions of the broader construct. They consider existing works in social psychology where sexism dimensions have long been operationalized with sub-scale questions tested for consistency. To approach dataset creation for on-screen objectification, we inspire from these two last works, and on dense annotation approaches of image datasets recently proposed for the medical domain [13]. To the best of our knowledge, approaches for sexism detection from visual content are scarce, and almost none existing for video data. Two main types of visual content have been considered so far: hateful or sexist memes [16, 29], and sexist advertisements [18] which can also relate to symbolic advertisement understanding [25, 28].

In films, analysis of biases in gender representation has been also automated with NLP approaches applied to film scripts [2, 37]. Specifically, Martinez et al. [37] propose a RNN-based model to automatically extract agent-verb-patient triplets. From 912 movie scripts, they show that male characters are associated with a higher agency while female characters are more frequently the object of gaze. Su et al. introduce the more abstract task of trope understanding in movies [52]. Our work is close to this last one as we also introduce a dataset of films annotated for a higher-level construct beyond event and story understanding. However, we provide dense annotations of sequences with constitutive concepts for the high-level construct of character objectification. Also, while Su et al. exploit an existing online base contributed to by the community, we design and carry out a strictly defined annotation process by experts.

In this article, we set out from the concept of male gaze defined in various ways in film gender studies [9, 36, 41]. Mulvey [41] characterizes the concept of gaze by the three relations between the camera and the characters, between the characters, but also between the spectator and the characters. While some formalizations of gaze set out from the gender of the director [36], Brey defines female and male gaze only from the film content, with a corpus analysis focusing on aesthetics [9] and revolving around the construct of objectification. We consider this gaze analysis of Brey, which does not significantly rely the socio-historical context of production and reception. We build the operationalization objectification in both cinematography and psychology, to create the conditions to make a new challenge accessible to the CV community: we produce a strict annotation process to obtain densely annotated video data and analyze where the new challenges lie ahead. We generally position our approach producing a non-large scale but high-quality dataset that we hope will be useful within the lines of the call of Paullada et al. for such data-centric AI approaches [44].

2.3. Approaches to video and movie understanding

Cross-modal foundation models, such as CLIP [45] and ALIGN [27], learn aligned image and text representations through contrastive pre-training on large-scale closed datasets. Generalizations of the CLIP model to video data have included VideoCLIP [57] and two X-CLIP models [34, 42]. In particular, X-CLIP [42], which we employ in this article, expands CLIP with video temporal modeling and video-adaptive textual prompts.

Legacy and large-scale pre-trained vision-language models have been leveraged for movie-related tasks. Bose et al. consider the difficulty of visual scene recognition in movies due to domain mismatch and create the MovieCLIP dataset obtained from weakly labeling movie shots from scene categories using the CLIP model [8]. An important vision-language movie-related task is audio-description for the visually-impaired, for which a major dataset introduced recently is MAD [50], gathering sparse natural language sentences grounded in over 1200 hours of movie videos. AutoAD [23] and AutoADII [22] are two recent approaches to generate audio-description from the video, both leveraging CLIP to learn to prompt GPT. To design approaches to learn human-level constructs, such as emotions, interactions or relationships, datasets generated with human supervision are also instrumental. A prominent representative is the MovieGraphs dataset, providing detailed annotations of clips of 51 movies with emotional states, character interac-

tions and relationships, and other scene reasoning elements [53]. It has prompted works tackling such recognition tasks [17, 33]. In this article, we build on the MovieGraphs dataset to annotate a selection with the construct of objectification, to later analyze it in connection with the other annotated social elements.

In this article, we aim to assess the capacity of CLIP-based methods to provide relevant embeddings for the concept of objectification. We do so with a direct evaluation of classification results when an adapter (MLP) fed by X-CLIP embeddings is learnt. We analyze the results with a concept-based approach by building on Concept Bottleneck Models (CBM) [31]. Concept-based models are an active area of research in XAI, with works tackling the accuracy-explainability tradeoff [60] and the need for user-defined concepts [58]. We specifically employ Post-hoc CBM (PCBM) [59], which consists in learning a concept subspace (made of Concept Activation Vectors [30]) in the embedding space of the pre-trained model. Data samples are then projected in this concept subspace, from where the classification task can be performed with an interpretable classifier.

3. Data and methods

This section presents our approach to create the first dataset for visual objectification in videos, specifically in films. We name this dataset *ObyGaze12*, short for *ObjectifyingGaze12*, which has the following **highlights**:

• It considers the multiple dimensions of the construct of visual objectification, made of filmic (framing and editing over successive shots, camera motion, etc.) and iconographic properties (visible objects, body parts, attire, character interactions, etc.).

• It is based on a thesaurus articulating five sub-constructs identified from multidisciplinary literature (film studies and psychology) from which we define typical instances, then grouped into coarse-grained visual concepts.

• The data is annotated densely with concepts, and shows the multi-factorial property of objectification, corroborating with some recent developments in cognitive psychology [6].

• A hard negative category is included with the goal to perform fine-grained error analysis and improve model generalization.

3.1. A thesaurus of objectification

We first formalize the construct of visual objectification and derive key concepts to annotate in film scenes. Together with media studies experts, we identify five sub-constructs of objectification from literature on film cognition and film gender studies, and social and cognitive psychology: male gaze (point of view of a man on a woman) [5, 9, 41], sexualization [6, 7], surveillance of the feminine body [11, 14, 40], female inaction / male possession [20, 48], and infantilism / animalization [41].

These sub-constructs come with typical instances and examples from filmmaking techniques ([9, 41]) or validated questionnaires ([11, 40]), as shown in the middle and right-most columns in Table 1. These typical instances are then grouped into eight coarse-grained visual concepts, corresponding to the possible means of production of visual objectification. They are shown in the left-most column in Table 1: type of shot (framing and gaze of camera), look (gaze of characters on the other), body (partial or full nudity, and sexually suggestive body parts), posture (connoting, e.g., childhood, submission or inaction), clothing (in relation to context and activities), appearance (age and makeup), expression of emotion (restrained or exaggerated according to gender role), and activities (linked to gender roles). Visual examples are provided in Fig. 1, where we show video samples of objectification concepts *Look*, *Posture*, *Type of shot*, *Clothes* and *Activity*.

3.2. Data selection

Over the various existing movie datasets (see Sec. 2 and [4, 24, 39, 43, 46, 50, 53], [51, Table 1]), many have overlapping titles and only a few have rich human supervision. Amongst these, the MovieGraphs dataset [53] includes rich, high-level human annotations of 7637 clips of 51 movies, with emotional states, interactions and relationships, and other social reasoning elements. These elements are important and valuable in exploring social concepts of objectification in visual media. These movies are also frequently adopted for media research in existing work in [4, 24, 39]. The movie clips have also a short duration – mostly within 1-5 minutes – facilitating dense and granular annotations (over 100 clips for an average 2 hour film) while preserving the possibility to observe longer-term interactions and story development across a number of shots. From the 7637 clips of the MovieGraphs dataset, we select 1914 clips to annotate for objectification, amounting to 25% of the dataset and 12 complete movies, which were selected to approximately reproduce the fraction of genres in the original dataset. The list of selected films, year of release, and genre can be found in Table 3 in Appendix 7.

3.3. Data annotation

Every selected movie is annotated by at least two experts for objectification level and concepts over the movie scenes. Specifically, the annotators were asked to repeat a three-step process for every scene they deemed interesting from an objectification perspective: (1) watch the movie entirely and when they identify a scene worth annotating, (2) delimit the clip by using the cutting function in our annotation tool, and (3) assigning an objectification level and annotate the concept(s) involved in the objectification rating. We define

Table 1. Thesaurus of the typical instances and examples of visual objectification in films, grouped into eight main visual concepts used for annotation. Examples are possible means to produce one of the five sub-constructs of objectification (male gaze, sexualisation, surveillance of the feminine body, female inaction/male possession, infantilism/animalisation), to be assessed by annotators.

Concept	Concept instances	Examples
Type of shot	Shot suggesting man perspective in presence of woman	close-up on a man's face; body parts of woman
	Shot suggesting man gaze on woman	camera takes the perspective of a male character with first close-up on the face followed by camera motion looking a woman from bottom up
	Shot showing a woman in parallel with an animal	woman at same level and position with a dog
Look	Voyerism	character watching another one without their knowledge
	Non-reciprocal gaze	woman looking at man who does not look back
Body	Suggested nudity	clothing on floor; silhouette behind shower curtain; nude shadow on wall
	Partial nudity	nude upper or lower body; partially open clothing or draping; in underwear
	Full nudity	nude person fully or partially shown
	Body parts suggestive of sex	close-up shots on breast, buttocks, hips, or lips
Posture	Gesture or posture connoting seduction	lip-biting; hip roll; twisting or tucking hair
	Gesture or posture connoting sexuality	eating phallic symbols; arching back
	Gesture or posture connoting inaction	being undressed by someone
	Gesture or posture connoting submission	leaning on a man
	Gesture or posture connoting dependence	following a man
	Skipping	skipping gait
Clothing	Wet or transparent clothing	thin shirt soaked in rain
	Clothing impractical to situation	wear pumps for running, a skirt when gardening
	Color code associated to character	woman with pink clothing and accessories
	Older woman wearing infantile clothing	woman wearing an Alice band or high socks
Appearance	Discrepancy between appearance of woman and context or biographical elements	perfect makeup when waking up; mother of heroine too young; young girl played by older actress
Exp. of emotion	Asymmetric expression of emotion	boys don't cry; woman being hysterical
Activity	Doing domestic activities	doing laundry, cooking, cleaning, being constantly in the kitchen

four levels of objectification:

- **Easy Negative**: there are no elements suggestive of objectification. No concept can be annotated. Default value for watched but non-selected scenes.
- **Hard Negative**: the scene contains elements of objectification from the thesaurus, but their presence does not result in an objectification effect.
- **Not Sure**: the scene indicates objectification, but does not completely fit the definition in the thesaurus.
- **Sure**: the scene contains elements of objectification from the thesaurus, and their presence results in an objectification effect.

Four expert annotators were recruited to annotate the dataset. A first presentation session was held to introduce our annotation tool and the annotation procedure. All four annotators were then given the same two films – *Juno* and *Silver Linings Playbook* – to annotate using the proposed methodology. A second meeting was then set up after the annotation of the two films to analyze the reasons for divergence and remedy them by identify which elements of the annotation guidelines to clarify and how. We then randomly assigned two annotators to each of the remaining 10 films to annotate separately.

Data processing and fusion The data processing is described in detail in Appendix 7. Following the annotation step, the annotations are then projected from the delimitations provided by each annotator onto the delimitations of the MovieGraphs clips. Since multiple annotations may overlap the same MovieGraphs clip, the annotation that is projected on the MovieGraphs clip corresponds to the annotation (including objectification level and concepts) with (1) the highest level of objectification that has at least 20% overlap with the MovieGraphs clip, and (2) when multiple annotations exist at the same level of objectification, the annotated concepts for these annotations are aggregated. The same process is used to aggregate the annotations of the annotators of a same clip: the maximum objectification level is kept, with possible aggregation of concepts in case both annotators chose the same level but annotated different concepts. The merged data is shared and used in the remaining of this article.

3.4. Analysis of the *ObyGaze12* dataset

Here we comment on some interesting statistics of the resulting annotations and concepts of the 1914 clips originally delimited in the MovieGraphs dataset.

First, we verify data consistency by computing the inter-annotator agreement (IAA). Given the task of annotating timespans, we choose the γ agreement measure introduced in [38] (and used for, e.g., annotating text spans [12]) owing to its consideration of temporal alignment, multiple annotators, and label classification at the same time. It attributes a score between 1 (complete agreement) and $-\infty$. A value of

$\gamma \leq 0$ indicates no agreement. The computation details of the γ metric is provided in Appendix 7. Considering all four categories EN, HN, NS, S, we obtain and average $\gamma = 0.42$. Not considering the clips annotated Not Sure (NS), which is the uncertain and "noisy" class in human annotations, the IAA increases to $\gamma = 0.69$. This shows the consistency of the obtained annotations despite the interpretive nature of the task. Let us also mention that recent works improve learning approaches by considering explicitly the IAA in case of low number of annotators with moderate agreement [10, 55, 56].

Second, we analyze the obtained annotations in Fig 2. The Sure category is the least represented with 16%, the Easy Negative being, as expected, the most represented class with 52% of clips. It is interesting to note that every concept is approximately annotated with the same rate throughout the Hard Negative, Not Sure and Sure levels of objectification. Finally, it is very interesting to observe that the average number of concepts annotated per clip increases with the level of objectification: 1.26 concepts on average per Hard Negative clip, 1.71 for Not Sure, up to 2.6 for Sure. We verify that this trend is observable for every single annotator. It gives an important insight into our video interpretation data: that objectification is a compositional process. This corroborates with recent findings in neuroscience experiments that found that a single element, such as clothing on its own, is not sufficient for people to perceive a character as an object [6].

4. Experiments

The experiments have two objectives: to verify that the new classification task is feasible, and to identify the challenges of designing efficient models. To tackle these objectives, we consider pre-trained vision models and specifically address the following research questions:

- **Task accuracy –** What are the baseline performances by pre-trained vision models on the objectification detection task? How does the performance vary with hard negative examples?
- **Concept representation –** Can we implement interpretable models of objectification using concepts? What is the quality of representation of every concept, and what are the objectification concepts poorly captured by current models?

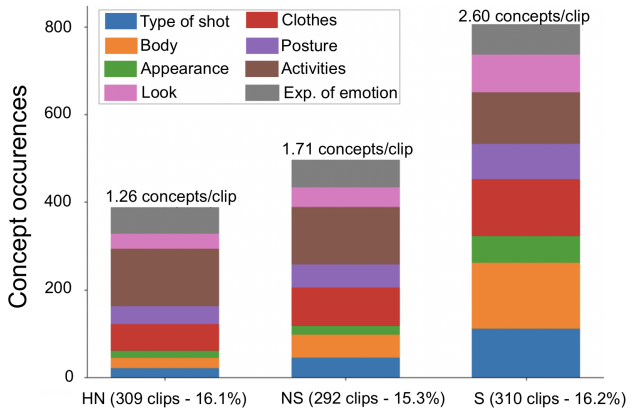

Figure 2. Distribution of visual factors annotated for each level of objectification (HN = Hard negative, NS = Not sure, S = Sure). The percentage of the dataset for each level of objectification as well as the average number of concepts per clip are also shown. (Best viewed in colors)

4.1. Task accuracy

Setup We discard the Not Sure (NS) class from the *ObyGaze12* dataset, as it gathers by definition samples highly uncertain for humans, and consider the Easy Negative (EN, 62% of the clip samples), Hard Negative (HN, 19%) and Sure (S, 19%) classes. We approach binary classification in a progressive way, the positive class being made of the S samples. We consider two levels of classification difficulty by composing the negative class either with EN samples, or with HN samples. The implementation details of cross-validation and data balancing are provided in Appendix 8. The average performance over the test set of the best models on validation folds are shown in Table 2 with standard deviations.

Baselines We consider video embeddings obtained from pre-trained models owing to their zero-shot classification capabilities on video tasks. We select ViViT-B/16 [3], and the available X-CLIP model, trained on Kinetics [42]. We also re-train a X-CLIP model [34] on the LSMDC [46] film dataset, and refer to Appendix 10 for corresponding results, where all implementation details are described. We keep the pre-trained models frozen and perform an adaptative max pooling of the resulting frame tokens, and feed the output to an MLP made of 2 dense layers, the hidden layer with 128 neurons and ReLU activations, the last with 2 softmax neurons.

Results To assess the quality of the models on possibly imbalanced data with a minority of positive samples, we report the F1-scores in Table 2. First, by comparing with trivial classifiers (random and an all-positive, see App. 8), we observe that **the task is indeed feasible**, warranted by the data consistency described in Sec. 3 despite the interpretive nature of the task. Second, we observe that **the inclusion of Hard Negative examples improves** the classification results, showing the importance of a fine-grained annotation for highly-interpretive tasks. Results of X-CLIP on other configurations, specifically when the movies of clips in test are different from those in train, are shown in App. 8. The best results based on existing models are of moderate qual-

Table 2. F1-score on the binary task of objectification detection for models trained with easy or with hard negatives and tested on easy or all negative samples, with standard deviations.

Test	EN vs. S		(EN U HN) vs. S	
Train	EN vs. S	HN vs. S	EN vs. S	HN vs. S
ViViT-B/16	0.53 (0.18)	0.62 (0.13)	0.54 (0.24)	0.73 (0.1)
X-CLIP	**0.79** (0.05)	0.71 (0.05)	0.66 (0.05)	**0.82** (0.03)
Random	0.54		0.55	
All positive	0.08		0.06	
PCBM-DT	0.68	0.44	0.58	0.38
PCBM-LR	0.64	0.43	0.50	0.37

ity, which calls for more investigation into where the difficulties lie.

4.2. Concept accuracy

To infer on-screen objectification, it is key for the model to detect the means of its production, which correspond to the eight concepts listed in Table 1. We reiterate that in the *ObyGaze12* dataset, every clip annotated with a level of objectification S, NS or HN is also annotated with the presence of instances of the eight concepts. For example, if the *Body* concept is annotated, it means that some level of nudity and/or suggestive body parts are shown on screen, that could contribute to the production of objectification. The means of producing objectification through the eight concepts can be subtle to detect, making it difficult to provide the final interpretation. To investigate this difficulty, we implement Post-hoc Concept Bottleneck Models (PCBMs) [59], which allow us to approach a classification task with pre-trained models in an interpretable way when concept-annotated data is available. In our case, from the X-CLIP embedding space where our video clips are represented, we identify a Concept Activation Vector (CAV) [30] for every concept. We then project the X-CLIP embedding of every clip onto the subspace defined by the eight CAVs. The representation of the clip that is the output of this bottleneck is a low-dimensional vector with number-of-concepts components. This vector can then be fed to an interpretable classifier for the objectification detection task.

CAV computation For every concept i, we collect two sets of samples: positive samples where concept i is present, hence made of S and HN samples with the concept annotated as present, and negative with EN, S and HN without concept i. We then train a linear SVM for each concept i, the CAV of concept i being the normal vector of the SVM hyperplan. To train the SVM, we split the data in 10 folds, and reserve the last fold for test. We then perform an 8-fold cross-validation to select the SVM (choice of margin tolerance c), every fold training set being balanced with different draws of negative sub-sampling.

Interpretable classifier We then train a decision tree (DT) and a logistic regression (LR) classifier on the same 8-fold cross-validation to classify the level of objectification, the classifiers being fed with the projection of every clip onto the CAVs.

Task accuracy with PCBM We first verify the quality of objectification detection with F1-scores of PCBM-DT and PCBM-LR shown in Table 2. The results lower than X-CLIP are expected owing to the known accuracy-interpretablity tradeoff of CBMs [59, 60]. They are above random and all-positive predictions when training on EN vs. S, which is indicative of the relevance of information held in the concepts. However, the low results obtained when training the DT and LR to distinguish between S and HN reveals the **low quality of some CAVs**, where the X-CLIP embeddings cannot be linearly well-separated for these concepts. We investigate this point next. The resulting DT is discussed in App. 9.

Concept accuracy We now analyze the quality of each obtained CAV by plotting its capability to classify whether the concept is present in a test sample. We consider a positive similarity between the X-CLIP embedding and CAV of concept i indicative of the presence of concept i. F1-score on the test set are shown in Fig. 3. Plots correspond to CAVs obtained from classifying the presence of concept i against EN only (solid bars) and against EN with S and HN without concept i (hatched bars). The former is used for PCBM-DT in Table 2. We first observe that concept detection is harder when negative samples also include S and HN samples (without the concept). This is expected considering that scenes tagged EN have by definition no element possibly conducive to objectification, and are hence likely to differ visually more from scenes where the concept is present, than do S and HN scenes without the concept. However within S and HN clips with and without the concept, such shortcuts cannot be exploited anymore. We observe in this case that **the X-CLIP embedding related to concepts *Type of shot*, *Posture*, *Look* and *Appearance* are harder to separate linearly**. This can be correlated with the analysis of factors of error detailed next. These subtler means of on-screen objectification therefore warrant future work to be properly captured and detected.

Error Analysis We analyze the factors impacting objectification classification errors corresponding to the results shown in Table 2. We select an average X-CLIP-based adaptation model trained on HN vs. S and tested on (ENUHN) vs. S, and consider its predictions on the clips in the test set. We label each test clip with 0 if the model fails to predict the correct label of the clip, and with 1 otherwise. We describe the clip with a one-hot encoding vector

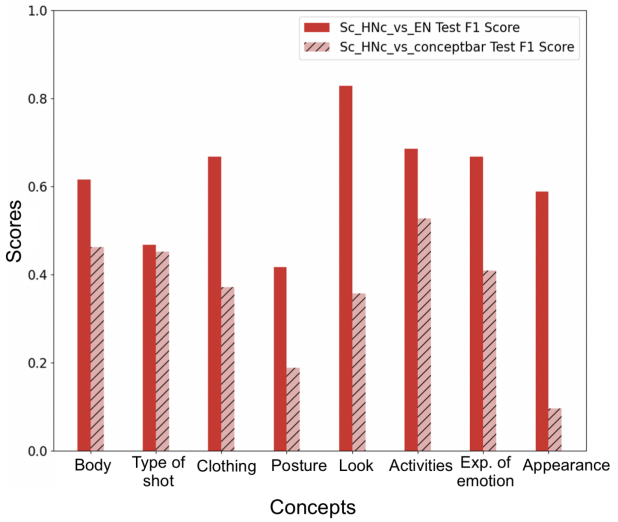

Figure 3. For every concept, F1-score of the best linear SVM selected to define the CAV of this concept. Positive samples (S and HN with the concept) must be separated from: [non-hatched bars] negative samples made of EN only, or [hatched bars] negative samples made of EN and S and HN without the concept.

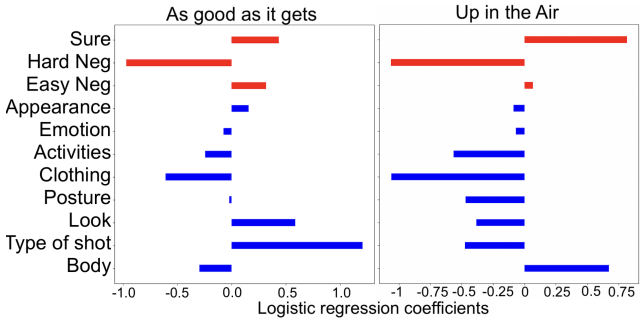

Figure 4. Analysis of the factors of error for the objectification detection task: weights of a logistic regressor predicting whether the test set examples are well classified or not. Positive (resp. negative) weights indicate a positive (resp. negative) contribution to classification success. Left: *As good as it gets*, Right: *Up in the Air*.

corresponding to all 11 factors shown the y-axis of Fig. 4: every of the 8 concepts, and the Sure (S), Hard negative (HN) and Easy negative (EN) labels. We then train a logistic regression model on these clip descriptors to predict the failure/success labels. Fig. 4 shows the error regression weights associated with each factor, for two film examples. A negative weight indicates a contribution of the factor to a classification failure. We first observe that the HN characteristic contributes to a classification failure of the model, while S and EN contribute to classification success. Second, we observe that there is variability over the movies on the presence of which concept strongly influences the classification success. However, the presence of the *Clothing* concept seems to be a strong confounder. This can be due to the frequency of appearance of this concept in HN samples, and to the subtlety of the description of this concept (provided in Table 1), which should make it difficult for a pre-trained model to discriminate between an objectifying and non-objectifying overall label on the basis of *Clothing*. Third, it is worth noting that the concepts shown to be poorly linearly separable when described with the X-CLIP embeddings (see CAV analysis in Sec. 4.2), are also those with a non-stable contribution to the model errors over the film examples: *Type of shot*, *Look*, *Posture* and *Appearance*.

5. Discussion

Ethical aspect This work has an explicit societal motivation in its purpose to tackle, with the help of AI, the analysis of complex temporal patterns operated in cinema that produce the perception of certain characters as objects. This is a challenging but valuable task that aims to uncover and quantify differences in how various identities may be portrayed on screen.

Limitations and challenges A distinctive element of our work is the subjective judgement involved in annotating granular video elements for objectification. Video annotation is tedious, and approaching data annotation for such an interpretive task in a rigorous way is even more so, and difficult to scale. We therefore believe that pursuing high-quality, dense annotations with well-defined concepts goes a long way to tackle this new video interpretation task, which represents a valuable new challenge for the computer vision community.

6. Conclusion

In this article, we have introduced a new video interpretation task to detect character objectification in films. We have introduced the *ObyGaze12* dataset, densely annotated by experts for objectification concepts defined from five sub-constructs identified in film studies and psychology. *ObyGaze12* is made available to the community. We evaluate recent vision models, show the feasibility of the task and where the challenges remain with concept bottleneck models. We show that the representation learning of the concepts of Type of shot, Look, Posture and Appearance need to be improved.

Acknowledgements. This work has been partly supported by the French National Research Agency through the ANR TRACTIVE project ANR-21-CE38-0012-01. This work was partly supported by EU Horizon 2020 project AI4Media, under contract no. 951911 (https://ai4media.eu/).

References

[1] 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. 3

[2] 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. 3

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. ViViT: A Video Vision Transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6816–6826, Montreal, QC, Canada, 2021. IEEE. 6

[4] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed Movies: Story Based Retrieval with Contextual Embeddings. Technical Report arXiv:2005.04208, arXiv, 2020. arXiv:2005.04208 [cs] type: article. 4

[5] 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, 2018. Publisher: Routledge _eprint: https://doi.org/10.1080/10463283.2018.1471949. 3, 4

[6] 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, 2019. 4, 6

[7] 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, 2020. 3, 4

[8] Digbalay Bose, Rajat Hebbar, Krishna Somandepalli, Haoyang Zhang, Yin Cui, Kree Cole-McLaughlin, Huisheng Wang, and Shrikanth Narayanan. Movieclip: Visual scene recognition in movies. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2082–2091, 2023. 3

[9] Iris Brey. *Le regard féminin-Une révolution à l'écran*. Média Diffusion, 2020. 3, 4

[10] 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, 2023. IEEE Computer Society. 6, 2

[11] 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, 2004. 4

[12] 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. 3, 5

[13] 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. 3

[14] Angela Denchik. *Development and Psychometric Evaluation of the Interpersonal Sexual Objectification Scale*. PhD thesis, The Ohio State University, 2005. 4

[15] Simone Fabbrizzi, Symeon Papadopoulos, Eirini Ntoutsi, and Ioannis Kompatsiaris. A survey on bias in visual datasets. *Computer Vision and Image Understanding*, 223: 103552, 2022. 2

[16] 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. 2, 3

[17] Bei Gan, Xiujun Shu, Ruizhi Qiao, Haoqian Wu, Keyun Chen, Hanjun Li, and Bohan Ren. Collaborative noisy label cleaner: Learning scene-aware trailers for multi-modal highlight detection in movies. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18898–18907, 2023. 4

[18] Francesca Gasparini, Ilaria Erba, Elisabetta Fersini, and Silvia Corchs. Multimodal Classification of Sexist Advertisements:. In *Proceedings of the 15th International Joint Conference on e-Business and Telecommunications*, pages 399–406, Porto, Portugal, 2018. SCITEPRESS - Science and Technology Publications. 2, 3

[19] Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *arXiv:2106.08409 [cs]*, 2021. arXiv: 2106.08409. 2

[20] 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, 2020. 4

[21] 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, 2015. ACM. 2

[22] Tengda Han, Max Bain, Arsha Nagrani, Gul Varol, Weidi Xie, and Andrew Zisserman. Autoad ii: The sequel - who,

when, and what in movie audio description. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13645–13655, 2023. 3

[23] Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. AutoAD: Movie Description in Context. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18930–18940, Vancouver, BC, Canada, 2023. IEEE. 3

[24] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. MovieNet: A Holistic Dataset for Movie Understanding. In *Computer Vision – ECCV 2020*, pages 709–727. Springer International Publishing, Cham, 2020. Series Title: Lecture Notes in Computer Science. 4

[25] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic Understanding of Image and Video Advertisements. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1100–1110, Honolulu, HI, 2017. IEEE. 3

[26] 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, 2019. 2, 3

[27] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 2021. 3

[28] Nasrin Kalanat and Adriana Kovashka. Symbolic image detection using scene and knowledge graphs. Technical Report arXiv:2206.04863, arXiv, 2022. arXiv:2206.04863 [cs] type: article. 3

[29] 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*, Red Hook, NY, USA, 2020. Curran Associates Inc. 3

[30] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, 2017. 4, 7

[31] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020. 4

[32] Holly B. Kozee, Tracy L. Tylka, and Casey L. Augustus-Horvath. Interpersonal Sexual Objectification Scale, 2011. 3

[33] Anna Kukleva, Makarand Tapaswi, and Ivan Laptev. Learning Interactions and Relationships Between Movie Characters. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9846–9855, Seattle, WA, USA, 2020. IEEE. 4

[34] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 638–647, New York, NY, USA, 2022. Association for Computing Machinery. 3, 6

[35] Sarah Macharia. *Global Media Monitoring Project (GMMP)*, pages 1–6. John Wiley Sons, Ltd, 2020. 1

[36] Alicia Malone. *The Female Gaze: Essential Movies Made by Women*. Mango Publisher, 2018. 3

[37] 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, 2022. 3

[38] Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479, 2015. 5, 1

[39] Antoine Mazières, Telmo Menezes, and Camille Roth. Computational appraisal of gender representativeness in popular movies. *Humanities and Social Sciences Communications*, 8 (1):137, 2021. 1, 2, 4

[40] Nita Mary McKinley and Janet Shibley Hyde. The objectified body consciousness scale: Development and validation. *Psychology of women quarterly*, 20(2):181–215, 1996. 4

[41] Laura Mulvey. Visual Pleasure and Narrative Cinema. *Screen*, 16(3):6–18, 1975. 1, 2, 3, 4

[42] 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. 3, 6

[43] Alejandro Pardo, Fabian Caba Heilbron, Juan León Alcázar, Ali Thabet, and Bernard Ghanem. MovieCuts: A New Dataset and Benchmark for Cut Type Recognition. In *Computer Vision – ECCV 2022*, pages 668–685. Springer Nature Switzerland, Cham, 2022. Series Title: Lecture Notes in Computer Science. 4

[44] 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, 2021. 3

[45] 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 *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3

[46] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie Description. *International Journal of Computer Vision*, 123(1):94–120, 2017. 4, 6

[47] Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. "Call me sexist, but..." : Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples. page 12. 1, 3

[48] 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. 4

[49] 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. 2

[50] M. Soldan, A. Pardo, J. Alcazar, F. Heilbron, C. Zhao, S. Giancola, and B. Ghanem. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5016–5025, Los Alamitos, CA, USA, 2022. IEEE Computer Society. 3, 4

[51] 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, 2021. 1, 2, 4

[52] 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*, pages 4594–4603, New York, NY, USA, 2021. Association for Computing Machinery. event-place: Virtual Event, Queensland, Australia. 3

[53] 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. 4, 1

[54] Angelina Wang, Arvind Narayanan, and Olga Russakovsky. REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets. In *Computer Vision – ECCV 2020*, pages 733–751. Springer International Publishing, Cham, 2020. Series Title: Lecture Notes in Computer Science. 2

[55] Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. *ArXiv*, abs/2110.12088, 2021. 6, 2

[56] 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*, pages 2523–2535, New York, NY, USA, 2023. Association for Computing Machinery. event-place: Long Beach, CA, USA. 6, 2

[57] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 3

[58] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a Bottle: Language Model Guided Concept Bottlenecks for Interpretable Image Classification. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19187–19197, Vancouver, BC, Canada, 2023. IEEE. 4

[59] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *ICLR 2022 Workshop on PAIR^2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*, 2022. 4, 7

[60] 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 *Advances in Neural Information Processing Systems*, 2022. 4, 7