ConViS-Bench: Estimating Video Similarity Through Semantic Concepts

Benedetta Liberatori[1] **Alessandro Conti**[1] **Lorenzo Vaquero**[2]
Yiming Wang[2] **Elisa Ricci**[1,2] **Paolo Rota**[1]
[1]University of Trento [2]Fondazione Bruno Kessler
https://benedettaliberatori.github.io/convisbench/

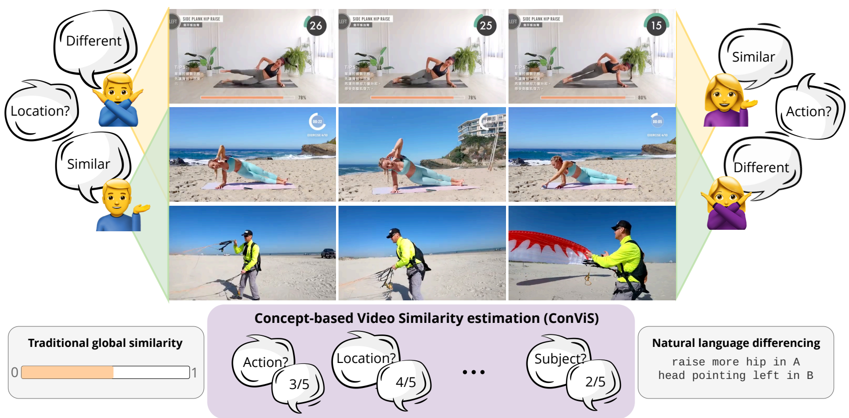


Figure 1: **Concept-based Video Similarity estimation (ConViS).** We introduce ConViS, a task that quantifies video similarity along specific semantic concepts (*e.g.*, location), and ConViS-Bench, a dataset of video pairs annotated with concept-level similarity scores (1-to-5) and free-form descriptions of similarities and differences. This bridges the gap between prior work focused solely on global similarity [18, 20, 21] or differences in natural language [4, 32].

Abstract

What does it mean for two videos to be similar? Videos may appear similar when judged by the actions they depict, yet entirely different if evaluated based on the locations where they were filmed. While humans naturally compare videos by taking different aspects into account, this ability has not been thoroughly studied and presents a challenge for models that often depend on broad global similarity scores. Large Multimodal Models (LMMs) with video understanding capabilities open new opportunities for leveraging natural language in comparative video tasks. We introduce Concept-based Video Similarity estimation (ConViS), a novel task that compares pairs of videos by computing interpretable similarity scores across a predefined set of key semantic concepts. ConViS allows for human-like reasoning about video similarity and enables new applications such as concept-conditioned video retrieval. To support this task, we also introduce ConViS-Bench, a new benchmark comprising carefully annotated video pairs spanning multiple domains. Each pair comes with concept-level similarity scores and textual descriptions of both differences and similarities. Additionally, we benchmark several state-of-the-art models on ConViS, providing insights into their alignment with human judgments. Our results reveal significant performance differences on ConViS, indicating that some concepts present greater challenges for estimating video similarity. We believe that ConViS-Bench will serve as a valuable resource for advancing research in language-driven video understanding.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

1 Introduction

Humans can effortlessly compare pairs of videos, rapidly identifying similarities and differences by attending to a range of semantic aspects, such as the depicted activity, the people involved, or the environment. This intuitive comparative ability relies on a rich understanding of events that unfold across space and time. Cognitive science research confirms this: humans naturally perceive, encode, and retrieve events along key semantic concepts, selectively attending to particular attributes of experience [38, 43]. As a result, perceived similarity between two videos is not a fixed quantity, but it depends on which concepts are being prioritized. For example, two videos may appear highly similar in terms of the activity being performed, but diverge substantially in the setting or the agents involved (see Fig. 1, top). This observation motivates a shift from holistic similarity to a more structured, concept-aware notion of video comparison.

Computational approaches to video similarity have traditionally focused on global similarity metrics, typically learned by comparing spatio-temporal embeddings [18, 20, 21]. The emergence of Large Multimodal Models (LMMs) [3, 24, 25, 49, 52] with video understanding capabilities has opened new possibilities for using natural language to describe and reason about differences between videos. Prior work has explored this by generating natural language descriptions of video differences, either through domain-specific cooking concepts [32] or fine-grained, action-specific skill differences [4]. However, these approaches remain limited to narrow domains and are purely descriptive, lacking *structured, quantitative assessments* of similarity across semantic concepts. As a result, comparative video understanding *via language* remains in its early stages, with existing benchmarks failing to capture the broad semantic diversity present in real-world scenarios.

To address this gap, we introduce a new task, *Concept-based Video Similarity estimation* (ConViS). Inspired by human cognition and grounded in semantic structure, ConViS aims to quantify how similar two videos are on specific concepts, *e.g.*, the activity, the location, or the order of actions (see Fig. 1, bottom). ConViS enables concept-specific video understanding, supporting applications like targeted video retrieval (*e.g.*, same activity with different subjects), anomaly detection based on particular factors (*e.g.*, unusual object presence or action sequence), and fine-grained model evaluation by isolating the conceptual sources of failure (*e.g.*, confusing similar-looking scenes with different actions). Building on the definition of ConViS, we introduce a novel benchmark, *ConViS-Bench*, to support model evaluation and foster further research. ConViS-Bench consists of video pairs spanning a broad range of domains, each annotated by multiple human evaluators with similarity scores conditioned on multiple semantic concepts and accompanied by textual descriptions.

Alongside introducing a novel dataset associated with the newly proposed task, we extensively benchmark several recent LMMs to assess their ability in predicting concept-based video similarities. Our analysis of their relevance to human judgment reveals significant performance differences across various LMMs on ConViS, highlighting that certain concepts are more challenging for models to judge in terms of video similarity. For instance, while some models can reliably identify visual similarities, they consistently struggle with more abstract notions such as the temporal structure of events, an issue also noted in prior work [2, 23]. Lastly, we demonstrate the utility of concept-aware similarity in downstream tasks such as concept-conditioned video-to-video retrieval, showing how ConViS can enable nuanced and interpretable video analysis.

Overall, our contributions are threefold:

- We introduce the ConViS task, a new formulation of video similarity that moves beyond traditional global scoring and computes interpretable similarity scores across semantic concepts.
- We release ConViS-Bench, a new benchmark dataset with human-annotated similarity judgments across multiple semantic concepts and diverse video domains.
- We conduct an extensive evaluation of state-of-the-art (video- and image-based) models on ConViS-Bench, analyzing their current strengths and limitations in concept-aware video comparison.

2 Related Work

Our work is related to previous research on comparing pairs of images and videos using natural language. We also discuss previous studies aimed at assessing the capabilities of LMMs in several video understanding tasks.

Visual Differences in Images. Image change captioning refers to the task of describing the differences between two images by generating a sentence caption [12, 16, 19, 48]. Recent advancements in LMMs have enabled the joint analysis and comparison of multiple images [3, 5, 17, 24, 49]. However, benchmarking studies indicate that model performance on multi-image understanding tasks remains relatively low, especially in tasks involving spatial understanding [31] and fine-grained visual distinctions [40]. Closely related to our work is [1], which defines a notion of *conceptual similarity* between pairs of images. This concept captures high-level relations, even between images that do not share visually similar elements. Our work tackles a related but more ambitious challenge, as analyzing videos is inherently harder than comparing images.

Visual Differences in Videos. Traditional approaches comparing videos typically compute a single, global similarity score [18, 20, 21]. While these methods are practical, the resulting scores are challenging to interpret. To address this, recent research has shifted towards natural language-based video differencing [4, 32]. Burgess et al. [4] introduces the task of Video Action Differencing and the associated VidDiffBench dataset, aimed at identifying subtle differences in how individuals perform the same action. Their approach focuses on textual descriptions of fine-grained skill variations, rather than quantifying similarity with explicit scores. Similarly, Nagarajan and Torresani [32] proposes the Difference Question Answering task through the StepDiff dataset, which targets cooking-related instructional videos. Here, the model compares two videos of the same procedural step and optionally ranks videos based on a common reference. While these approaches advance research in video pair comparison, they are not tailored to support applications that necessitate concept-specific similarity quantification, such as those discussed in Sec. 1. Moreover, the accompanying datasets are typically limited to a few domains. In contrast, we advocate the need for datasets and methods that generalize across a broader range of video types and domains.

Our work is also related to composed video retrieval [15, 42], where a query consists of a reference video plus a textual modification, and the goal is to retrieve a target video that reflects that modification. However, our task differs in two principal ways: (i) rather than focusing on retrieving a matching video, we aim to compute explicit, quantitative similarity scores; and (ii) we support exploration along multiple conceptual dimensions, enabling comparisons not limited to a single textual modification.

Benchmarking LMMs for videos. The rapid advancement of LMMs have driven a surge in the number of benchmarks used to evaluate their video understanding capabilities across spatial, temporal, and semantic dimensions. For instance, MMBench-Video [8], MVBench [26], and VideoVista [27] assess LMMs' ability to reason temporally, semantically, and causally across varied video content. Similarly, Video-MME [10] and Video-MMMU [13] provide full-spectrum evaluations over a wide range of video domains, durations, and task complexities, with the latter explicitly modeling cognitive stages such as perception, comprehension, and adaptation. Several benchmarks also probe specific aspects of video understanding. TempCompass [30] targets fine-grained temporal perception by constructing videos with identical static content but controlled temporal variations, while Perception-Test [34] examines high-level reasoning across physical and counterfactual dimensions. Benchmarks like LVBench [44] and ActivityNet-QA [50] assess long-form video understanding by testing models' memory and narrative coherence. These benchmarks primarily assess video understanding capabilities through multi-choice question answering. While this is effective for standardized evaluation, recent work has shown that such questions are often overly informative, leading to biased evaluations [6, 47]. Furthermore, our focus here is on assessing the ability of different LMMs to capture a set of specific concepts that cognitive studies [38, 43] have shown to be crucial for human understanding of videos.

3 Concept-based Video Similarity estimation

Humans naturally compare videos by focusing on specific semantic aspects rather than their entire content. For example, given the top two clips shown in Fig. 1, one viewer may judge them to be highly similar, as both depict the same action (*i.e.*, side plank), while another viewer may judge them to be completely different, as the locations vary significantly (*i.e.*, house interior *vs.* beach). Such concept-specific video comparison is crucial for applications that require the isolation of individual concepts, as discussed in Sec. 1.

Formally, let $\mathcal{C} = \{C_1, \dots, C_K\}$ be a set of predefined concepts, each expressed in natural language (*e.g.*, *main action* or *location*). Given two videos V_1, V_2 and a concept $C_i \in \mathcal{C}$, we define the

Table 1: **Comparing ConViS-Bench with prior work.** Video pairs in ConViS-Bench are annotated by humans with similarity scores across fine concepts and free-form text descriptions of similarities and differences. †StepDiff provides scores only for a subset of its domain-specific categories.

Dataset	Pairs	Domains	Focus	Scores	Text	Split	Avg. duration
StepDiff [32]	6,292	1	5 cooking categories	✓†	differences	train/test	12.0 s
VidDiffBench [4]	557	5	1 concept (action)	✗	differences	test	8.8 s
ConViS-Bench (Ours)	610	16	5 broad concepts	✓	similarities/differences	test	28.2 s

concept-based similarity:

$$s(V_1, V_2 \mid C_i) \in \mathbb{R} \tag{1}$$

which measures how similar the two videos are to the concept C_i. This formulation forces the comparison to attend only to details relevant to C_i, ignoring other information.

Since concepts are expressed through natural language, our approach offers two key advantages. The first aspect is flexibility, allowing users to introduce any semantic dimension without being confined to a fixed taxonomy. The second is composability, where users can select a set of concepts and assign non-negative weights λ_i to each concept. The individual scores can be aggregated into an overall similarity score:

$$s(V_1, V_2) = \sum_{i=1}^{K} \lambda_i \, s(V_1, V_2 \mid C_i) \quad \text{with} \sum_{i=1}^{K} \lambda_i = 1, \tag{2}$$

which enables fine-grained control over which aspects drive the final judgment.

Our concept-based formulation bridges two extremes of prior research on video pair comparison. Traditional *global similarity* approaches [20, 21] compute an unconditioned score $s(V_1, V_2)$ (*e.g.*, via average-pooled video embeddings), but do not indicate which factors influence the similarity. Conversely, *video differencing* techniques describe qualitative differences in natural language [4, 32] without yielding per-concept quantitative scores. By contrast, our formulation provides both *structured conditioning* on C_i and *interpretable quantitative similarities*, facilitating comparisons and direct evaluation of a model's alignment on user-defined concepts.

4 The ConViS-Bench Dataset

We seek to advance video similarity estimation by introducing a new dataset that captures the complex, concept-dependent nature of how humans compare videos. Existing datasets describing video pairs with language are limited in scope: some focus narrowly on a single concept, such as *action* [4], while others are restricted to domain-specific categories, such as *ingredients* [32] (see Tab. 1). To address this gap, we introduce ConViS-Bench, a benchmark for multi-concept video similarity estimation.

We follow a structured protocol to build and curate ConViS-Bench (see Fig. 2). ConViS-Bench contains 610 video pairs drawn from 543 videos, each annotated with human-judged similarity scores across five general-purpose concepts. In addition to quantitative similarity scores, each pair is accompanied by free-form descriptions highlighting shared and differing elements, offering qualitative insight into human reasoning. It covers the broadest range of domains to date, *i.e.*, 16 in total, significantly more than the 5 covered in [4], and features longer videos on average.

4.1 Collecting and annotating video pairs for ConViS

Here, we outline the key steps involved. Further details about the dataset statistics and the annotation process are in the Appendix.

Choice of Concepts. Our approach to computing video similarity is based on defining key semantic concepts. This design is motivated by research in cognitive science showing that memory representations are organized around semantic and temporal features, and that people update their understanding of events by comparing such structured elements via mechanisms like prediction error and recursive reminding [38, 43]. Guided by these findings, we identify five core concepts that reflect how people naturally interpret events: *main action*, *main subjects*, *main objects*, *location*, and *order of actions*. These constitute the set $\mathcal{C}$ as defined in Sec. 3, which is designed to be domain-agnostic, ensuring consistent estimation of video pair similarity across diverse domains.

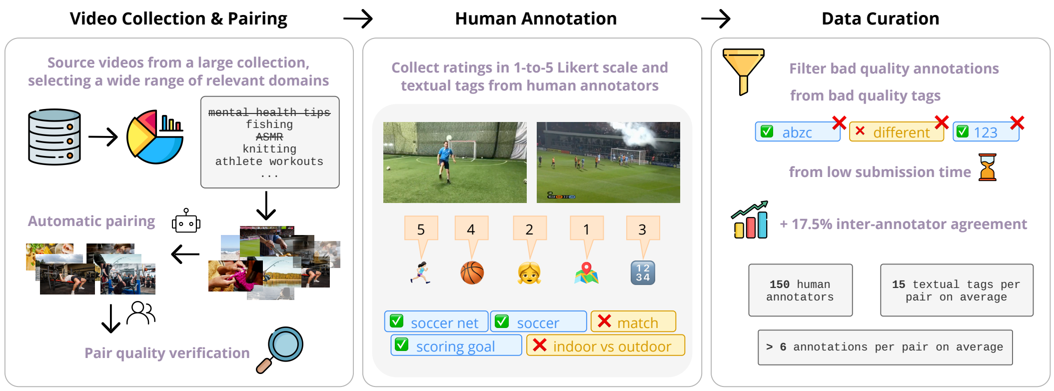


Figure 2: **The construction ConViS-Bench.** We source videos from FineVideo [9], a large-scale collection of videos, and perform automatic trimming and pairing. We then collect human ratings on a 1-to-5 Likert scale across multiple concepts (*main action*, *main subjects*, *location*, *main objects*, *order of actions*) and free-form tags describing key similarities and differences. Finally, we systematically curate the data to ensure quality.

Video Collection. As shown in the left part of Fig. 2, ConViS-Bench's data collection pipeline begins by identifying an existing dataset to serve as the primary video source. We select the FineVideo dataset [9], a large-scale collection of YouTube videos that spans a diverse range of human activities. The dataset is densely annotated, recently compiled, and available under the Attribution (CC-BY) license. It includes 126 fine-grained categories organized within a two-level taxonomy, comprising 9 coarse-grained categories. However, not all of them are equally informative for assessing similarity in video content. In particular, some categories, *e.g.*, *Mental Health Tips* or *Poetry Reading*, exhibit limited visual diversity and temporal dynamics, as they predominantly consist of static shots of people speaking. Since our goal is to analyze visual similarity across semantically grounded dimensions (*i.e.*, the concepts), we restrict our focus to categories with rich, varied visual content and exhibit sufficient temporal dynamics. Accordingly, we select a subset of relevant coarse- and fine-grained categories from FineVideo, as illustrated in Fig. 2. A complete list can be found in the Appendix. As a result, we include videos from 16 diverse domains, making ConViS-Bench the most semantically and visually diverse dataset of video pairs to date, as illustrated in Tab. 1.

Video Pairing. In FineVideo [9], the average video length is 4.7 minutes, with each video potentially containing various shots and events. First, to ensure compatibility, we trim the original videos using the provided timestamp annotations to isolate clips corresponding to individual events. From this curated pool, we construct video pairs. To promote diversity, we match videos based on either semantic similarity (guided by textual descriptions) or visual similarity. To this end, we compute visual embeddings for the trimmed clips and textual embeddings for their corresponding ground-truth descriptions, using DINOv2 [33] and Sentence-BERT [37] as the respective encoding methods. We then form candidate video pairs by selecting those with a high cosine similarity computed considering a single modality, either video-to-video or text-to-text, but explicitly not both. The rationale behind this procedure is to ensure that each pair of videos shares certain common patterns while also exhibiting distinctive elements. Finally, we perform manual filtering and verification, obtaining 610 video pairs that exhibit varying degrees of similarity across the five considered concepts.

Human Annotation. We collect human annotations for all the video pairs, as shown in the center of Fig. 2. The goal of these annotations is to obtain reliable similarity scores for the concepts in $\mathcal{C}$, along with textual descriptions highlighting both the similarities and differences between the paired videos. We collect annotations from a mix of volunteers and paid workers (totaling 150 annotators) recruited via Prolific, a crowdsourcing platform selected for its demonstrated quality and reliability compared to existing alternatives [7, 35]. We instruct annotators (all college-educated English speakers) to evaluate each video pair along the five predefined semantic concepts in $\mathcal{C}$. For each concept, similarity is rated on a 5-point Likert scale [28], with detailed scoring guidelines provided to ensure clarity and consistency across annotators. To further promote thoughtful engagement with the task, annotators must assign at least one *similarity tag* and one *difference tag* to each pair, describing specific aspects in which the videos align or differ.

Acknowledging that our predefined concepts may not fully encompass all relevant aspects of similarity, we encourage annotators to introduce new concepts when applicable. Ultimately, only 5.28% of the annotations involve a new concept, supporting our initial assumption that the predefined ones adequately capture the key aspects of similarity across videos. The annotation interface, statistics on inter-annotator agreement, and additional details on this core step are available in the Appendix.

Data Curation. As shown in the right part of Fig. 2, the collected annotations undergo a final curation process. Specifically, we use the free-form text tags provided by annotators to assess annotation quality, removing samples containing non-English or irrelevant content, while allowing for correctable typographical errors. We also flag and manually review annotations submitted in unusually short timeframes, which may indicate rushed or low-effort responses. As a result, we discard 7.75% of the original annotations, resulting in a cleaner and more reliable dataset. The final collection contains video pairs collected from 147 annotators, with an average of 6.2 annotations per pair.

5 Benchmarking Models on ConViS-Bench

A key contribution of this work is a comprehensive evaluation of state-of-the-art models on the proposed ConViS-Bench benchmark. In this section, we report the results of our experiments.

We first consider ten state-of-the-art LMMs and evaluate their ability to compare videos explicitly accounting for specific concepts (Sec. 5.1). Then, we analyze different approaches that assess video similarity based on a global score, (Sec. 5.2). Finally, we assess models on the task of concept-based video retrieval (Sec. 5.3). All experiments were conducted using up to two NVIDIA H100 GPUs.

5.1 Assessing LMMs performance on Concept-based Video Similarity estimation

In this section, we consider state-of-the-art LMMs and evaluate their capabilities to compare videos based on specific concepts. We select nine open-source, state-of-the-art models with strong performance in video understanding, including mPLUG-Owl3 [49], LLaVA series [24, 51], Qwen-VL series [3], and InternVL series [5, 52], and the closed source model Gemini (2.0-Flash) [39]. These models can be prompted with questions explicitly targeting particular aspects of a video. Following the approach in [4], we provide each model with concatenated frames from both videos in a pair and ask it to *"[...] output a single similarity score between 1 and 5, where 1 means completely different and 5 means perfectly similar in terms of <concept>"*. For Gemini, which natively supports multi-video within a single prompt, we supply the complete videos. The exact prompt used and additional implementation details are described in the Appendix. The pre-training data of Gemini is private, and we do not know if the source data used for our video pairs (*i.e.*, FineVideo) was included. On the contrary, we know that the training set of InternVL [5] includes it. Nevertheless, we still decide to include these models in the evaluation to assess their performance on the new task of Concept-based Video Similarity estimation. We use Spearman's rank correlation ρ and Kendall's τ as our evaluation metrics, following prior work evaluating alignment with human judgment [1, 14, 29]. Both coefficients take values in $[-1, +1]$ (here reported $\times 100$) with -1 denoting complete disagreement with human evaluators, and $+1$ perfect agreement.

Can LMMs judge video similarity conditioned on specific concepts? Table 2 presents the results of our concept-based evaluation. Unsurprisingly, larger models consistently outperform their smaller counterparts, with LLaVA-OV-7B achieving the highest correlations overall. Interestingly, certain concepts, such as *order of actions*, are inherently more challenging for all models. While some models perform better on other concepts, such as *main objects* or *location*, with LLaVA-OV-7B scoring above $58/47\%$ in ρ/τ, these correlations with human judgment are nevertheless more difficult to capture than those related to global assessments. Surprisingly, InternVL models, despite having seen the evaluation videos during pre-training, struggle to align with human judgments. This indicates that pre-training with interleaved video-text data alone does not equip models to accurately estimate similarity concerning specific semantic concepts. Rather than task-specific skill, our results may reflect how well general-purpose LLMs are compatible with human notions of conceptual similarity.

Does ConViS-Bench challenge models on temporal understanding? To assess whether our benchmark poses a temporal challenge and how the temporal dimension impacts ConViS, we repeat the analysis in Tab. 2 while varying the number of input frames per video. As shown in Fig. 3, models not pre-trained on FineVideo (LLaVA-OV/LLaVA-Video/Qwen2.5-VL), consistently

Table 2: **Alignment performance of LMMs on ConViS-Bench.** We report Spearman's ρ and Kendall's τ correlations with human judgment ($\times 100$). Top-performing scores are shown in **bold**, and second-best scores are underlined, proprietary models in gray.

Model	Main Action		Main Subjects		Main Objects		Location		Order of Actions	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
mPLUG-Owl3-7B [49]	30.64	25.12	20.59	16.84	28.53	23.47	21.00	17.00	23.11	18.74
LLaVA-OV-0.5B [24]	1.95	1.54	-5.05	-3.91	-4.00	-3.20	5.66	4.59	1.30	0.90
LLaVA-OV-7B [24]	**51.76**	**42.48**	**48.43**	**39.14**	**58.64**	**48.17**	**58.94**	**47.72**	41.02	33.31
LLaVA-Video-7B [51]	44.17	36.53	39.81	32.81	45.85	37.96	55.96	46.33	**41.25**	**34.05**
Qwen2.5-VL-3B [3]	21.84	18.10	8.62	6.70	15.20	12.61	13.44	11.14	12.79	10.60
Qwen2.5-VL-7B [3]	37.88	31.28	17.53	14.43	26.97	22.26	23.63	19.17	23.85	19.61
InternVL2.5-4B [44]	13.23	10.13	16.71	12.89	14.99	11.71	13.93	10.71	9.88	7.54
InternVL2.5-8B [5]	28.70	22.42	28.60	22.08	25.06	19.40	19.64	15.18	18.15	14.07
InternVL3-8B [52]	40.69	31.31	36.54	27.80	42.50	32.88	45.47	34.98	32.74	24.88
Gemini-2.0-Flash	44.91	34.09	41.16	31.56	38.36	29.21	42.12	34.54	32.36	25.00

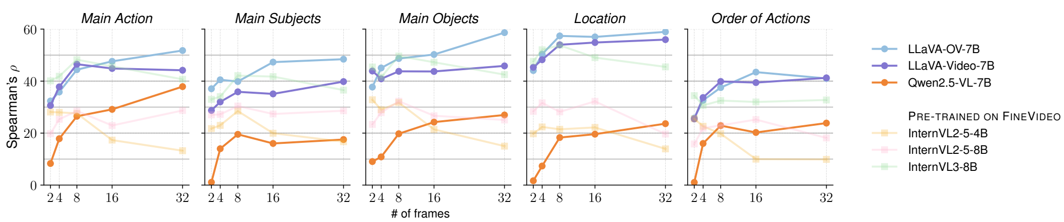


Figure 3: **Concept-based video similarity at varying number of input frames.** We report the Spearman's ρ correlations with human judgement ($\times 100$) for different numbers of frames per pair.

suffer from reduced temporal context, with performance dropping sharply as fewer frames are provided (average drops: $-16.36/10.61/21.74\%$). In contrast, models from the InternVL family, exposed to FineVideo during pre-training, exhibit remarkably flat performance trends, suggesting limited reliance on temporal cues. These results indicate a likely memorization effect in these latter models, as their performance remains stable even with severely reduced temporal input. Furthermore, we observe that the *main action* concept benefits the most from additional frames, with gains of $+29.57/19.40/13.56\%$ compared to $+16.45/11.37/11.09\%$ for *main subjects* in Qwen2.5-VL, LLaVA-OV, and LLaVA-Video. These findings show that ConViS-Bench poses a temporal challenge, where model performance significantly depends on access to rich temporal context.

5.2 Probing concept focus in global video representations

In this section, we analyze which semantic concepts are implicitly prioritized by models that compute global video similarity scores. We focus on models that extract global video representations and compare them in an embedding space, without explicit conditioning on specific concepts. We categorize these models into three groups based on how they process and compare inputs. ① *Video-to-video* approaches like VideoMAE [41] and DINOv2 [33] use unimodal vision encoders and compute similarity directly in the embedding space. ② *Text-to-text* approaches first map videos into text format using different video captioning models [3, 5, 49], and then compare the resulting text representations in a shared embedding space using Sentence-BERT [37]. ③ *Cross-modal* approaches rely on vision-language alignment methods. To compute the similarity between pairs of videos, we first determine the alignment score of each video with the caption generated from its counterpart using [49]. The final similarity score for a video pair is obtained by averaging these two alignment scores. Two alignment metrics are employed: CLIPScore [11], which measures cosine similarity via contrastive vision-language models, and VQAScore [29], which assesses alignment through visual question-answering between captions and visual content. For CLIPScore, both CLIP [36] and InternVideo [45] serve as vision-language encoders. In image-based approaches, a video's representation is obtained by averaging the representations of its individual frames.

Do global video representations implicitly align with specific concepts? The semantic concepts introduced in our ConViS-Bench benchmark enable more detailed and granular analysis and com-

Table 3: **Alignment performance computing global video representations on ConViS-Bench.** We report Spearman's ρ and Kendall's τ correlations with human judgement ($\times 100$). *Video* indicates the use of a video encoder. Best-performing scores are in **bold**, and second-best scores are underlined.

Model	Method	Video	Main Action ρ/τ	Main Subjects ρ/τ	Main Objects ρ/τ	Location ρ/τ	Order of Actions ρ/τ
Video-to-video							
VideoMAE [41]	Cosine	✓	13.0 / 8.7	23.1 / 15.6	13.2 / 8.8	37.8 / 26.4	15.1 / 10.2
DINOv2 [33]	Cosine		33.3 / 22.6	40.9 / 28.6	37.4 / 25.8	**57.4 / 41.7**	34.6 / 24.0
Text-to-text							
mPLUG-Owl3 [49]	SBERT [37]	✓	**52.1 / 36.2**	45.5 / 31.1	<u>55.1</u> / <u>38.9</u>	28.4 / 19.6	**49.9 / 34.9**
InternVL2.5 [5]	SBERT [37]	✓	39.4 / 27.4	31.2 / 21.4	45.0 / 31.1	29.9 / 20.5	42.0 / 29.5
Qwen2.5-VL [3]	SBERT [37]	✓	42.3 / 29.5	45.4 / 31.7	45.4 / 31.8	38.5 / 26.7	40.9 / 28.4
Cross-modal							
CLIP [36]	CLIPScore [11]		35.7 / 24.6	39.0 / 27.0	39.0 / 27.1	36.0 / 24.9	30.9 / 21.1
InternVideo [45]	CLIPScore [11]	✓	43.1 / 30.1	<u>54.8</u> / <u>38.7</u>	54.5 / 38.5	40.9 / 28.4	40.2 / 28.1
LLaVA-OV [24]	VQAScore [29]	✓	<u>51.1</u> / <u>36.1</u>	**55.8 / 39.6**	**58.3 / 41.4**	<u>46.5</u> / <u>32.3</u>	<u>48.1</u> / <u>33.9</u>

parison across models, a distinctive feature of our proposed benchmark. Results in Tab. 3 report the outcome of our analysis. Among the evaluated models, video-to-video approaches exhibit a stronger focus on the *location* concept, with DINOv2 achieving the highest correlation with humans on this concept ($57.4/41.7$ of ρ/τ). This indicates that spatial scene information is more effectively captured through purely visual representations. However, these models consistently show weaker correlations in the remaining categories. In contrast, text-to-text models excel at capturing action-related concepts ($44.6/31.0$ and $44.3/30.9$ of ρ/τ on average for *action* and *main action*) but show limited ability to compare videos with respect to *location* ($32.3/22.3$ of ρ/τ on average). Notably, across all model types, the concepts of *order of actions* and *location* consistently yield the lowest correlations with human judgment. This indicates that existing models, regardless of modality, tend to overlook or inadequately model spatial context and temporal structure, two key components in understanding video semantics. Although no single model achieves the highest alignment with human judgments across all concepts, VQAScore stands out by attaining the highest average correlation overall.

5.3 Concept-conditioned video-to-video retrieval

In this section, we evaluate whether, given an anchor video V and a concept of interest $C_j \in \mathcal{C}$, models can retrieve the most conceptually similar videos from a set of related target videos $T_V = \{V_i\}_{i=1}^{N}$. Importantly, we focus on LMMs, as these methods can be conditioned on specific concepts. Following standard evaluation protocols [46], we report retrieval performance using recall, precision, and F_1 score at rank 1 (R@1, P@1, and F_1@1). R@1 measures the percentage of queries for which a relevant video is retrieved at the top rank, while P@1 denotes the proportion of top-ranked videos that are truly relevant. The F_1@1 score summarizes the trade-off between R@1 and P@1.

To conduct this analysis, we identify as anchors any video V appearing in multiple annotated pairs within ConViS-Bench, and construct for each anchor a fixed candidate set of four videos, $T_V = \{V_i\}_{i=1}^{4}$. To ensure the reliability of these rankings and to assign similarity scores to any newly included videos, we engage three human annotators to validate or supplement the existing annotations. We exclude the concept of *order of action* from this analysis due to challenges in establishing consistent similarity rankings across video sets, despite robust pairwise annotation agreement. While temporal ordering is clear in video pairs, aligning actions across sets introduces ambiguities, *e.g.*, added or missing atomic actions. Empirically, annotators found it significantly more challenging to agree on consistent partial orderings across sets than within pairs. The result is a collection of 532 concept-based partial orderings, *i.e.*, some videos may be considered equally relevant, allowing for ties in the ranking. For each anchor video V and concept C_j, we compute similarity scores between V and each target video $V_i \in T_V$, denoted as $s(V, V_i \mid C_j)$. These scores are then used to induce a ranking over T_V for each concept.

Table 4: **Concept-based retrieval performance.** We report recall, precision, and F_1 score at rank 1 (R@1, P@1, F_1@1). The random chance baseline is in highlighted in gray.

Model	Main Action			Main Subjects			Main Objects			Location		
	R@1	P@1	F_1@1	R@1	P@1	F_1@1	R@1	P@1	F_1@1	R@1	P@1	F_1@1
mPLUG-Owl3-7B [49]	89.8	44.2	54.8	85.2	52.2	59.0	91.0	48.9	56.9	89.6	59.8	66.2
LLaVA-OV-0.5B [24]	73.4	32.3	41.2	69.9	44.1	48.5	74.7	35.9	45.4	85.3	51.2	59.7
LLaVA-OV-7B [24]	84.2	54.8	**61.9**	82.0	66.4	**68.7**	81.8	58.8	63.0	80.1	68.7	**69.3**
LLaVA-Video-7B [51]	88.8	51.1	60.7	86.6	63.1	68.3	88.7	58.2	**65.1**	78.8	68.4	67.5
Qwen2.5-VL-3B [3]	88.0	42.2	51.8	94.5	51.9	61.6	96.9	41.7	54.4	96.6	55.8	66.6
Qwen2.5-VL-7B [3]	63.9	42.5	46.8	63.5	55.8	54.0	63.5	44.1	47.9	38.5	70.7	47.1
InternVL2.5-8B [5]	58.8	47.6	48.7	57.8	60.2	53.6	58.5	49.9	49.4	55.2	65.0	53.4
InternVL2.5-4B [5]	60.5	44.6	46.5	57.8	55.7	50.6	59.9	47.2	48.4	55.6	59.8	52.0
InternVL3-8B [52]	63.0	58.3	56.4	55.8	66.8	56.0	57.8	59.7	54.5	54.8	71.0	56.5
Random Choice	25.0	34.4	27.8	25.0	44.9	30.4	25.0	50.0	31.5	25.0	36.6	28.4
Average	74.5	46.4	52.6	72.5	57.4	58.9	74.7	49.4	54.6	70.5	63.3	61.1

Do LMMs retrieve similarly to humans under conceptual constraints? In Tab. 4 we evaluate LMMs' performance on concept-conditioned retrieval. Across all models, R@1 consistently exceeds P@1, indicating that, while they often retrieve the true relevant video at rank 1, they also produce more false positives. Performance varies across concepts: the median P@1 is higher for *location* (with InternVL3-8B, LLaVA-OV-7B, LLaVA-Video-7B and Qwen2.5-VL-7B achieving P@1 $>$ 68%), while *main action* tends to be more challenging, with a mean P@1 of 46.4%. *Random choice* references a model that randomly picks one sample. Reported results take into consideration partial ordering. All evaluated models consistently outperform this random baseline in the F_1@1 score, despite some having a worse precision. This indicates that the model retrieves more relevant samples overall, albeit including more irrelevant ones. Overall, these results indicate that while current LMMs exhibit some capabilities, concept-based video-to-video retrieval remains a limitation. In Fig. 4, we compare the retrieval performance of the best-performing LMM with that of the models described in Sec. 5.2, which are not explicitly conditioned on the concept. While these approaches consistently underperform, the gap is smaller for *main objects*, where CLIPScore achieves comparable performance, highlighting its bias toward object-level representations. This further highlights that tailored, concept-aware methods are essential to complement general-purpose ones for accurate specialized retrieval.

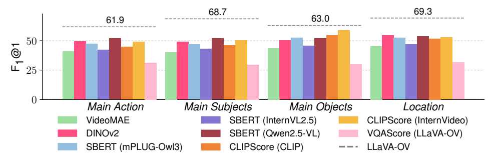


Figure 4: **Retrieval performance.** F_1@1 score reported for models discussed in Sec. 5.2, against LLaVA-OV-7B [24].

6 Conclusions

We introduced ConViS, a task for structured video similarity estimation, and ConViS-Bench, a novel human-annotated dataset of video pairs designed to evaluate video similarity along core semantic concepts, covering the widest range of domains to date. We extensively benchmark models, including ten state-of-the-art LMMs, on ConViS-Bench, revealing substantial variability in how they align with human judgments, for the first time, exposing their biases toward particular concepts. By enabling concept-conditioned evaluation and retrieval, ConViS offers a path towards more interpretable, controllable, and user-aligned video understanding systems.

Limitations and Future Work. In this work, we focused on a small set of general-purpose concepts to enable consistent and interpretable comparisons across diverse video domains. While this choice, grounded in cognitive science, supports broad applicability, it also introduces a limitation: the current concept set may not capture all domain-specific or fine-grained aspects of video similarity. We believe the same framework can be extended to incorporate additional concepts, enabling richer analyses in specialized settings. Moreover, ConViS-Bench currently includes a relatively modest number of video pairs. Although limited in scale, this is a consequence of prioritizing annotation quality over quantity. Expanding the dataset, provided quality standards are maintained, would further increase its value and support broader evaluations.

Acknowledgements

We acknowledge the EuroHPC Joint Undertaking for awarding this project access to the EuroHPC supercomputer MareNostrum5, hosted by BSC (Spain) through an EuroHPC Development Access call. This work was partially supported by the EU Horizon projects ELIAS (No. 101120237), ELLIOT (No. 101214398), and IAMI (No. 101168272). This work was supported by Ministero delle Imprese e del Made in Italy (IPCEI Cloud DM 27 giugno 2022 – IPCEI-CL-0000007) and European Union (Next Generation EU). We also thank the Deep Learning Lab of the ProM Facility for the GPU time.

References

[1] Alessandro Achille, Greg Ver Steeg, Tian Yu Liu, Matthew Trager, Carson Klingenberg, and Stefano Soatto. Interpretable measures of conceptual similarity by complexity-constrained descriptive auto-encoding. In *CVPR*, 2024.

[2] Piyush Bagad, Makarand Tapaswi, and Cees GM Snoek. Test of time: Instilling video-language models with a sense of time. In *CVPR*, 2023.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv*, 2025.

[4] James Burgess, Xiaohan Wang, Yuhui Zhang, Anita Rau, Alejandro Lozano, Lisa Dunlap, Trevor Darrell, and Serena Yeung-Levy. Video action differencing. In *ICLR*, 2025.

[5] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv*, 2024.

[6] Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees GM Snoek, and Yuki M Asano. Tvbench: Redesigning video-language evaluation. *arXiv*, 2024.

[7] Benjamin D Douglas, Patrick J Ewell, and Markus Brauer. Data quality in online human-subjects research: Comparisons between mturk, prolific, cloudresearch, qualtrics, and sona. *Plos one*, 2023.

[8] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *NeurIPS*, 2024.

[9] Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. Finevideo. https://huggingface.co/datasets/HuggingFaceFV/finevideo, 2024.

[10] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv*, 2024.

[11] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

[12] Mehrdad Hosseinzadeh and Yang Wang. Image change captioning by learning from an auxiliary task. In *CVPR*, 2021.

[13] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv*, 2025.

[14] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *ICCV*, 2023.

[15] Thomas Hummel, Shyamgopal Karthik, Mariana-Iuliana Georgescu, and Zeynep Akata. Egocvr: An egocentric benchmark for fine-grained composed video retrieval. In *ECCV*, 2024.

[16] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. In *EMNLP*, 2018.

[17] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *TMLR*, 2024.

[18] Won Jo, Geuntaek Lim, Gwangjin Lee, Hyunwoo Kim, Byungsoo Ko, and Yukyung Choi. VVS: video-to-video retrieval with irrelevant frame suppression. In *AAAAI*, 2024.

[19] Hoeseong Kim, Jongseok Kim, Hyungseok Lee, Hyunsung Park, and Gunhee Kim. Agnostic change captioning with cycle consistency. In *CVPR*, 2021.

[20] Giorgos Kordopatis-Zilos, Christos Tzelepis, Symeon Papadopoulos, Ioannis Kompatsiaris, and Ioannis Patras. Dns: Distill-and-select for efficient and accurate video indexing and retrieval. *IJCV*, 2022.

[21] Giorgos Kordopatis-Zilos, Giorgos Tolias, Christos Tzelepis, Ioannis Kompatsiaris, Ioannis Patras, and Symeon Papadopoulos. Self-supervised video similarity learning. In *CVPR*, 2023.

[22] Klaus Krippendorff. *Content analysis: An introduction to its methodology*. Sage publications, 2018.

[23] Jie Lei, Tamara Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In *ACL*, 2023.

[24] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *TMLR*, 2025.

[25] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-interleave: Tackling multi-image, video, and 3d in large multimodal models. In *ICLR*, 2024.

[26] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, 2024.

[27] Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. Videovista: A versatile benchmark for video understanding and reasoning. *arXiv*, 2024.

[28] Rensis Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 1932.

[29] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *ECCV*, 2024.

[30] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv*, 2024.

[31] Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Tianshuo Yang, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. MMIU: Multimodal multi-image understanding for evaluating large vision-language models. In *ICLR*, 2025.

[32] Tushar Nagarajan and Lorenzo Torresani. Step differences in instructional video. In *CVPR*, 2024.

[33] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024.

[34] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *NeurIPS*, 2023.

[35] Eyal Peer, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. Data quality of platforms and panels for online behavioral research. *Behav. Res. Methods*, 2022.

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[37] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP/IJCNLP*, 2019.

[38] Viviana Sastre Gomez, Rebecca Defina, Paul Michael Garrett, Jeffrey M Zacks, and Simon Dennis. What binds non-contiguous events together? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2023.

[39] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv*, 2023.

[40] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, 2024.

[41] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*. Curran Associates, Inc., 2022.

[42] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. Covr: Learning composed video retrieval from web video captions. In *AAAI*, pages 5270–5279, 2024.

[43] Christopher N Wahlheim and Jeffrey M Zacks. Memory updating and the structure of event representations. *Trends Cogn. Sci.*, 2024.

[44] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. Lvbench: An extreme long video understanding benchmark. *arXiv*, 2024.

[45] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv*, 2022.

[46] Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. Cap4video: What can auxiliary captions do for text-video retrieval? In *CVPR*, 2023.

[47] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *CVPR*, 2024.

[48] Linli Yao, Weiying Wang, and Qin Jin. Image difference captioning with pre-training and contrastive learning. In *AAAI*, 2022.

[49] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mPLUG-owl3: Towards long image-sequence understanding in multi-modal large language models. In *ICLR*, 2025.

[50] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, 2019.

[51] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv*, 2024.

[52] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv*, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: This work presents a novel task and dataset for assessing the similarity of videos based on semantic concepts. The task is described in Sec. 3, the dataset is presented in Sec. 4, and various LMMs are benchmarked in Sec. 5. These are the main claims in the abstract and the introduction, which are appropriately reflected in the main text.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper contains a "Limitations and Future Work" section (see Sec. 6). It discusses the limitations of our ConViS-Bench dataset and potential extensions of the ConViS task.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Implementation details regarding the benchmarking setup are discussed in Sec. 5. Furthermore, our code and dataset are released along with this submission and are considered core contributions of the work, as indicated in the NeurIPS Datasets and Benchmarks Guidelines. Overall, the paper, code, and dataset contain all the details for experiment reproducibility.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The dataset and code are considered core contributions of the work and are publicly available, as described in the paper. The code repository contains all the necessary information to replicate the reported results.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Dataset composition and implementation details about the proposed benchmark are described in depth in Sec. 4 and Sec. 5, respectively.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Although our approach allows for multiple evaluations and statistical significance analyses, we choose not to report these results to remain consistent with related work [4, 32].

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In Section 5, we explicitly tell which and how many GPU devices were used.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have carefully inspected the code of ethics, and believe our work conforms to it.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The broader impact of the proposed ConViS task and the ConViS-Bench dataset is described the Appendix.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The proposed dataset was sourced from an existing, publicly available dataset that already includes appropriate safeguards. As our work does not involve the release of high-risk models or data, additional safeguards were not required.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have properly referenced all the software, data, and related works that supported the development and validation of ConViS-Bench. Specifically, in Section 4, we describe how our source of raw videos is the FineVideo dataset [9] (CC-BY 4.0 license). In

Section 5, we cite all the models used in our experiments. Moreover, we publicly release our dataset and code.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The dataset, code, and instructions for running our ConViS-Bench benchmark are released alongside the manuscript and are available during the reviewing process.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The Appendix includes the full text of the annotation instructions, screenshots of the interface, details about the crowdsourcing platform and compensation, as well as a playable example of the annotation website.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: The study received IRB approval, which confirmed that the research does not involve risks to the psycho-physical well-being of participants, nor does it affect their rights to confidentiality, informed consent, or autonomy in decision-making. These details are provided in the Appendix.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The method development and manuscript writing in this research do not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix

In this Appendix, we first provide further details about the interface used for the annotation of ConViS-Bench in Sec. A. In Sec. B, we provide additional statistics and insights about our proposed dataset. In Sec. C we show the prompts used for the evaluation of the LMMs and in Sec. D we provide additional implementation details. In Sec. E we extend the experiments provided in the main paper, and in Sec. F we show additional qualitative examples from ConViS-Bench. Finally, in Sec. G we discuss the broader impacts of our work, in Sec. H we declare that our institutional review board approved the present study, and list the licenses and URLs for all external assets used in Section I.

A Annotation interface

In this section, we provide all the details about the human annotation process. We show screenshots of our interface, a matrix outlining the scoring guidelines, and information about annotator screening procedures and compensation.

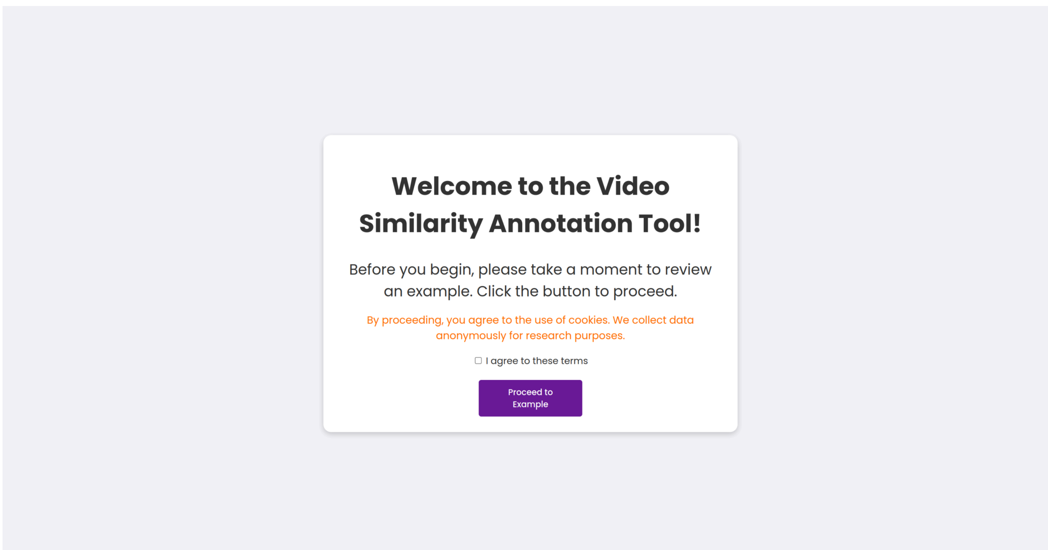


Figure 5: **Welcome page of the annotation interface.**

As a first step, annotators are asked to accept the use of cookies (see Fig. 5). We do not collect any personally identifiable information, such as email addresses, names, or nicknames. However, we keep track of the reviewed samples to ensure that the same annotation pairs are not shown to the same annotator multiple times. Next, we present an example to illustrate how to complete the task correctly (see Fig. 6). Finally, Fig. 7 shows the main annotation interface used to carry out the task. The annotation interface includes task instructions, which can be viewed at any time, as well as embedded guidelines that appear when hovering over relevant buttons. This appendix also contains a video demonstrating the interface in use to provide a clearer understanding of the annotation process. A scoring matrix summarizing these guidelines is also displayed for reference in Tab. 5.

Annotators are screened to ensure they are college-educated and reside in English-speaking countries. They are compensated at a rate of £8 per hour.

B Additional dataset analysis

Annotation quality assessment. We assess annotation quality by measuring inter-annotator agreement (IAA) using Krippendorff's α [22]. After removing low-quality samples through our end-stage curation process, the resulting α scores for each concept are: *main subjects* (0.244), *main action* (0.322), *location* (0.361), *main objects* (0.288), and *order of actions* (0.319). On average, this curation step improved inter-annotator agreement by 17.5%, highlighting the effectiveness of our quality control process.

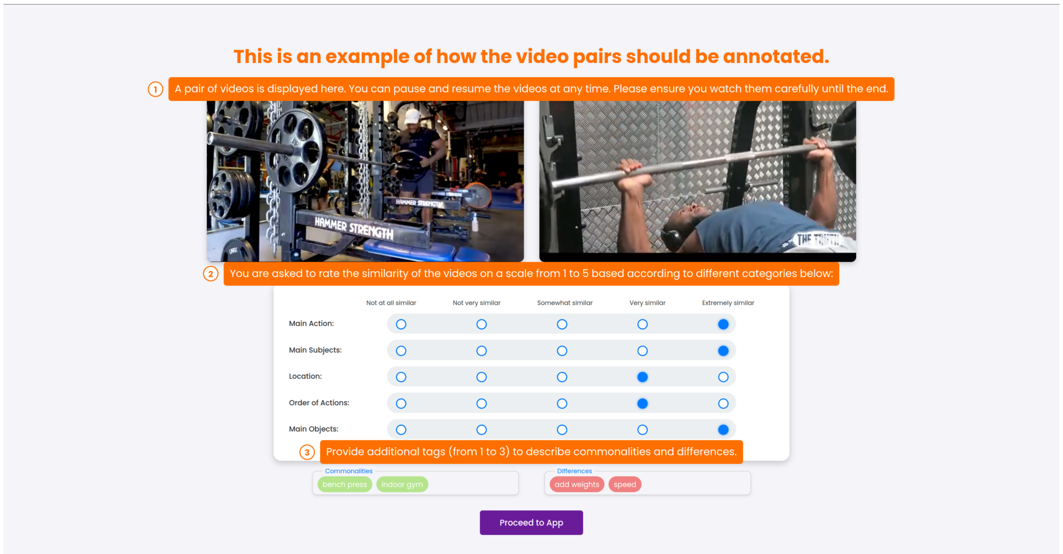


Figure 6: **Provided example in the annotation interface.**

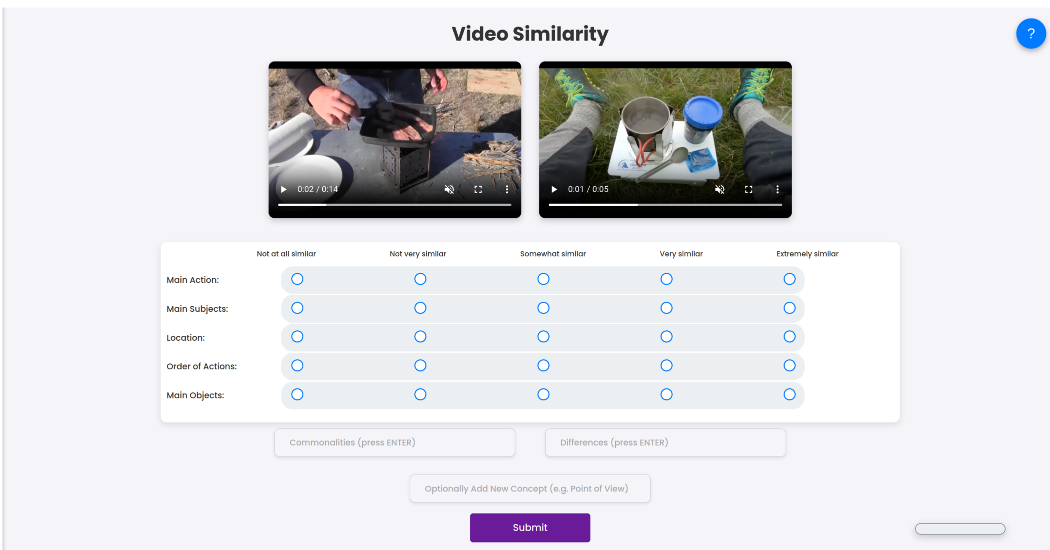


Figure 7: **Annotation interface.**

Statistics on tags. We obtain, on average, 15 textual tags for each video pair. Figure 8 shows wordclouds of prominent tags for similarities and differences. From these wordclouds, we observe that most tags relate to actions (*e.g.*, *gym*, *cooking*, *boxing*, *fishing*), objects (*e.g.*, *food*, *bike*, *gun*), subjects (*e.g.*, *people*, *gender*), or locations (*e.g.*, *location*, *beach*, *mountain*). A smaller portion of tags describe aspects of time, such as *order* and *speed*.

User-defined concepts. Figure 9 reports the frequencies of concepts proposed by annotators beyond the predefined set. The most commonly added concepts include point of view, skill level, and lighting condition. However, even the most frequently mentioned concept appears in fewer than 16% of video pairs, indicating that such additions are relatively sparse. These user-defined concepts highlight directions for further development of the concept taxonomy.

Domains The videos in ConViS-Bench span 16 distinct categories, as shown in Fig. 10. The category names are retained from the source dataset [9], where they are grouped into top-level categories. These include *Sports*, which encompasses *Athlete Workouts, Training Techniques, Game Highlights, and Match Replays*; *Lifestyle*, which includes *Recipe Videos, Gardening Tips, Workout*

Table 5: **Scoring guidelines provided in the annotation interface.**

Concepts	**1**	**2**	**3**	**4**	**5**
`Main Action`	Completely different types of action (*e.g.*, cooking and sports domains)	Actions are different but related (*e.g.*, both sports)	Actions are similar with significant differences in how they are carried out	Actions are similar with slight differences in how they are carried out	No difference in the main action performed
`Main Subjects`	Totally different cardinality, gender and age (*e.g.*, a man and a group of women)	Different cardinality but some similar in gender or age (*e.g.*, a man and a group of men)	Same cardinality but significant difference in age or gender (*e.g.*, a man and a boy)	Same cardinality but minor difference in age (*e.g.*, a young male and a boy)	Extremely similar (*e.g.*, two young women in both)
`Location`	Completely different places	Different but similar type of place (*e.g.*, indoor and outdoor soccer field)	Similar locations still with significant differences	Similar locations with slight differences	Extremely similar (*e.g.*, rooms with a similar size and layout)
`Order of Actions`	Completely different, or no sub-actions are in common	The order of actions has significant changes, and some missing steps	Some actions are reordered, with a few steps missing, but key steps remain intact	Minor changes in the order of actions	Actions occur in the same order, without missing steps
`Main Objects`	Main objects are entirely different	A few main objects are shared, but most are different	Most main objects are similar, with some not in common	Minor differences in main objects (*e.g.*, color or shape)	Main objects are the same or extremely similar

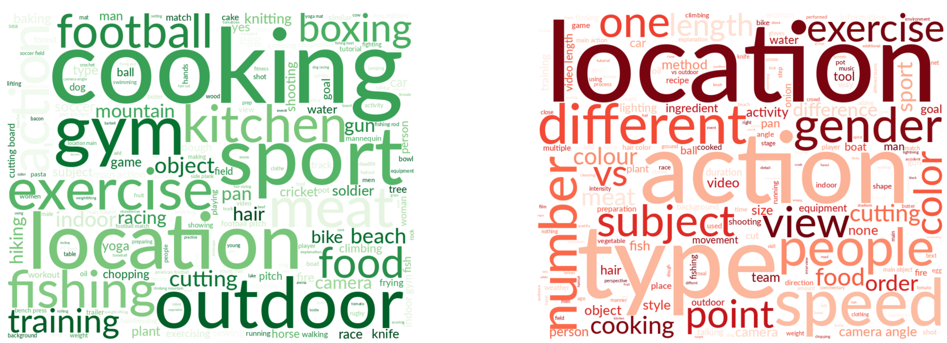


Figure 8: **Wordclouds of collected tags.** Visualization of tags for similarities and differences, displayed as wordclouds.

Routines, and Destination Guides; *Education*, consisting of *Engineering Projects, Cooking Tutorials, Documentaries, and DIY & Crafts*; and finally, *Hobbies & Interests*, which covers *Knitting, Hiking, Camping, and Fishing*.

Statistics on scores. Figure 12 reports the frequency distribution of the collected similarity scores for each concept. The distributions vary across concepts, with *main objects* exhibiting the highest number of scores equal to 1. Furthermore, Figure 11 shows the inter-correlation matrix between the scores assigned to different concepts. These results highlight that, according to human annotators, most visual concepts are not strongly correlated. The only notable exception is the pair *main action* and *order of action*, which exhibit a correlation higher than 80.

C Prompts for LMMs evaluation

Figure 13 illustrates the prompts used to evaluate LMMs on ConViS. These prompts, for all the concepts, are used in the evaluation setup described in Sec. 5.1.

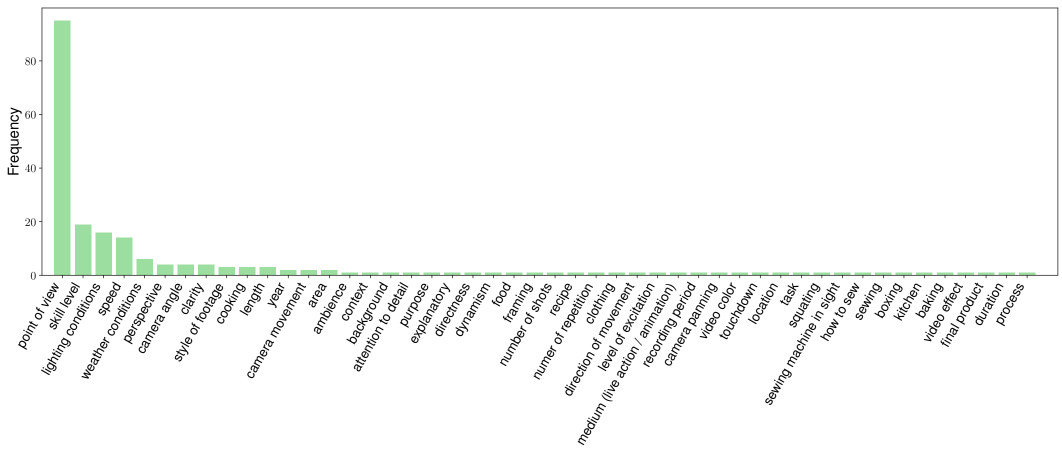


Figure 9: **User-defined concepts.**

Figure 10: **Video distribution per domain in ConViS-Bench.**

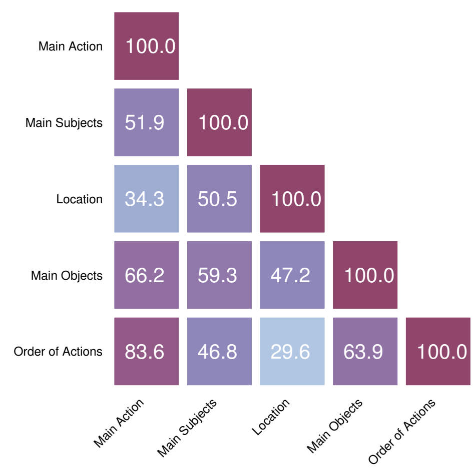


Figure 11: **Concept inter-correlation.** We report Spearman's ρ correlation ($\times 100$) between different concepts across the dataset.

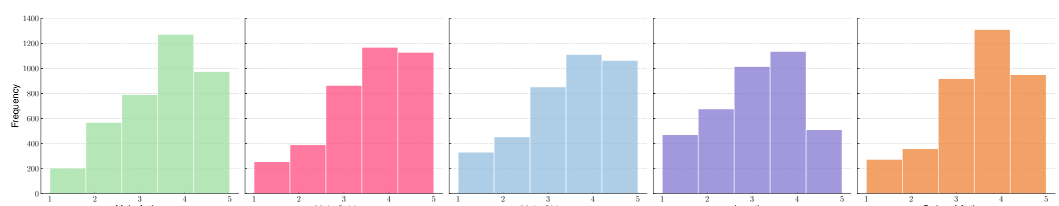


Figure 12: **Frequency of scores assigned by human annotators, divided by concepts.**

D Implementation details

Regarding the specific models used in Sec 5.2, we use DINOv2 [33], CLIP [36], and InternVideo [45] with a ViT-L/14 backbone, VideoMAE [41] with ViT-L, and Sentence-BERT [37] with the MiniLM-L6 (*i.e.*, `all-MiniLM-L6-v2` variant), which is a widely adopted model for text-to-text semantic similarity tasks.

System Prompt

```
You are an AI designed to compare two videos based on the visual concept of
<concept>.

The input consists of a sequence of concatenated frames:  the first half
represents Video 1, and the second half represents Video 2.
Your task is to evaluate how similar these two videos are with respect to the
concept <concept>.
Output a single similarity score between 1 and 5, where 1 means completely
different and 5 means perfectly similar in terms of <concept>.
Do not explain your reasoning.  Only output the numerical score.
```

Figure 13

E Additional experiments

Table 6 reports the performance differences when using a single frame per video instead of the full sequence of frames. It shows detailed results for all the concepts, with negative values indicating degraded alignment with human judgments. Tab. 7 reports the extended results across varying numbers of frames per video, illustrating how human-alignment scales with temporal input.

Table 6: **Alignment performance delta when using a single frame per video.** We report the difference (Δ) in Spearman's ρ and Kendall's τ correlations with human judgement ($\times 100$), computed as the change in performance when models receive a single frame per video instead of N frames. A negative value indicates performance degradation with fewer frames. A dash (–) indicates that the correlation could not be computed because the model outputs constant predictions.

	Main Action		Main Subjects		Main Objects		Location		Order of Actions	
Model	$\Delta\rho$	$\Delta\tau$	$\Delta\rho$	$\Delta\tau$	$\Delta\rho$	$\Delta\tau$	$\Delta\rho$	$\Delta\tau$	$\Delta\rho$	$\Delta\tau$
mPLUG-Owl3-7B [49]	-6.45	-5.13	-1.86	-1.35	-10.89	-8.85	-11.14	-8.89	+0.57	+0.78
LLaVA-OV-0.5B [24]	-3.84	-3.11	+0.62	+0.24	+3.50	+2.79	-6.11	-4.96	+4.39	+3.81
LLaVA-OV-7B [24]	-19.40	-16.24	-11.37	-9.05	-20.93	-17.69	-14.88	-11.86	-15.22	-12.36
LLaVA-Video-7B [51]	-13.56	-11.96	-11.09	-9.65	-1.96	-2.26	-10.65	-9.45	-15.81	-13.23
Qwen2.5-VL-3B [3]	-26.80	-22.22	-10.96	-8.64	-17.98	-14.92	-13.03	-10.80	–	–
Qwen2.5-VL-7B [3]	-29.57	-24.39	-16.45	-13.53	-17.92	-14.75	-21.96	-17.80	-22.81	-18.75
InternVL2.5-4B [5]	+14.88	+11.55	+5.03	+3.87	+17.84	+13.88	+5.82	+4.54	+15.65	+11.93
InternVL2.5-8B [5]	-8.91	-7.38	-1.59	-1.71	-1.73	-1.43	+8.70	+6.65	-2.33	-2.06
InternVL3-8B [52]	-0.59	-0.09	-3.56	-2.26	+2.93	+3.01	+2.12	+2.10	+1.67	+1.56
Gemini-2.0-Flash [39]	-3.57	-2.06	-0.93	-0.62	+9.85	+8.62	+7.01	+5.10	-6.86	-4.59

F Qualitative examples

In this section, we present qualitative examples of video pairs from ConViS-Bench. As shown in Figures 14 to 18, each example includes two videos with four representative frames per video, together with concept scores and annotated similarity and difference tags. ConViS-Bench contains similar video pairs that differ in specific aspects such as *location*. For example, a young blonde woman performing the same workout exercise at the seaside and in a park (see 14c), or in an indoor gym setting (see 14e). Another example features a person crocheting outdoors in a park versus indoors on a sofa (see 17e). The dataset also includes video pairs with similar *location* and *main subjects*, where the *main action* differs, *e.g.*, 14d, 16e, and 18e where only the action performed changes. Other examples show low similarity in the *order of actions*, even when the overall activity or context is related. For instance, both videos may show a person interacting with plant leaves, but in one case a leaf is deliberately detached (see 15e). Finally, some video pairs are nearly identical in setting and action but differ in the *main subjects*, such as a male versus female athlete competing in a track and field event (see 18d).

Additionally, this Appendix includes qualitative examples used for the retrieval evaluation discussed in Sec. 5.3. These examples are provided in video format to better demonstrate the full temporal dynamics and allow simultaneous comparison of entire videos.

Table 7: **Alignment performance of LMMs on ConViS-Bench at varying number of frames.** We report Spearman's ρ and Kendall's τ correlations with human judgment ($\times 100$). *Frames* indicates the number of frames per every video in the pair. A dash (–) indicates that the correlation could not be computed because the model outputs constant predictions.

Model	Frames	Main Action		Main Subjects		Main Objects		Location		Order of Actions	
		ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
mPLUG-Owl37B	16	30.64	25.12	20.59	16.84	28.53	23.47	21.00	17.00	23.11	18.74
mPLUG-Owl37B	8	31.18	25.62	23.15	18.98	29.05	23.95	22.02	17.98	21.43	17.51
mPLUG-Owl37B	4	32.82	27.09	23.19	19.23	33.67	27.92	14.14	11.64	23.73	19.56
mPLUG-Owl37B	2	29.32	24.20	18.69	15.49	25.04	20.77	8.12	6.67	23.81	19.62
mPLUG-Owl37B	1	24.19	19.99	18.73	15.49	17.64	14.62	9.86	8.11	23.68	19.52
LLaVA-OV-0.5B	16	1.95	1.54	-5.05	-3.91	-4.00	-3.20	5.66	4.59	1.30	0.90
LLaVA-OV-0.5B	8	-4.04	-3.32	-8.69	-6.87	0.09	0.06	-1.06	-0.81	-0.32	-0.30
LLaVA-OV-0.5B	4	3.75	3.09	6.52	5.36	10.60	8.74	4.73	3.83	7.02	5.78
LLaVA-OV-0.5B	2	6.12	5.01	-2.59	-2.14	0.35	0.29	-1.58	-1.33	5.66	4.66
LLaVA-OV-0.5B	1	-1.89	-1.57	-4.43	-3.67	-0.50	-0.41	-0.45	-0.37	5.69	4.71
LLaVA-OV-7B	16	51.76	42.48	48.43	39.14	58.64	48.17	58.94	47.72	41.02	33.31
LLaVA-OV-7B	8	47.61	38.94	47.30	38.64	50.23	41.36	57.00	46.10	43.46	35.49
LLaVA-OV-7B	4	44.39	35.87	39.84	32.28	48.68	40.02	57.39	46.55	37.51	30.54
LLaVA-OV-7B	2	35.81	28.96	40.51	33.04	45.06	36.93	50.33	40.90	32.60	26.72
LLaVA-OV-7B	1	32.36	26.24	37.06	30.09	37.71	30.48	44.06	35.86	25.80	20.95
LLaVA-Video-7B	16	44.17	36.53	39.81	32.81	45.85	37.96	55.96	46.33	41.25	34.05
LLaVA-Video-7B	8	44.86	36.97	35.08	28.85	43.71	36.12	54.84	45.16	39.43	32.59
LLaVA-Video-7B	4	46.50	38.42	35.88	29.57	43.76	36.12	53.98	44.54	39.86	32.95
LLaVA-Video-7B	2	37.79	30.95	32.00	26.33	40.88	33.55	48.22	39.33	33.68	27.81
LLaVA-Video-7B	1	30.61	24.57	28.72	23.16	43.89	35.70	45.31	36.88	25.44	20.82
Qwen2.5-VL-3B	16	21.84	18.10	8.62	6.70	15.20	12.61	13.44	11.14	12.79	10.60
Qwen2.5-VL-3B	8	23.34	19.35	16.62	13.75	20.20	16.75	14.27	11.76	8.37	6.91
Qwen2.5-VL-3B	4	20.56	17.06	12.65	10.48	7.09	5.88	13.28	11.00	5.86	4.85
Qwen2.5-VL-3B	2	9.76	8.09	7.51	6.22	8.11	6.73	4.32	3.57	-	-
Qwen2.5-VL-3B	1	-4.96	-4.12	-2.34	-1.94	-2.78	-2.31	0.41	0.34	-	-
Qwen2.5-VL-7B	16	37.88	31.28	17.53	14.43	26.97	22.26	23.63	19.17	23.85	19.61
Qwen2.5-VL-7B	8	29.13	23.92	16.02	13.21	24.23	20.01	19.59	16.02	20.24	16.64
Qwen2.5-VL-7B	4	26.47	21.83	19.53	16.08	19.75	16.35	18.29	15.03	22.88	18.89
Qwen2.5-VL-7B	2	17.84	14.76	13.99	11.58	10.85	9.00	7.34	6.08	16.01	13.23
Qwen2.5-VL-7B	1	8.31	6.89	1.08	0.90	9.05	7.51	1.67	1.37	1.04	0.86
InternVL2_5-4B	16	13.23	10.13	16.71	12.89	14.99	11.71	13.93	10.71	9.88	7.54
InternVL2_5-4B	8	17.31	13.30	19.95	15.51	21.42	17.05	22.18	17.13	9.98	7.68
InternVL2_5-4B	4	27.58	21.07	28.47	21.70	31.76	24.63	21.47	16.12	19.77	14.99
InternVL2_5-4B	2	27.98	21.34	22.97	17.44	29.00	22.22	22.37	17.10	22.63	17.16
InternVL2_5-4B	1	28.11	21.68	21.74	16.76	32.83	25.59	19.75	15.25	25.53	19.47
InternVL2_5-8B	16	28.70	22.42	28.60	22.08	25.06	19.40	19.64	15.18	18.15	14.07
InternVL2_5-8B	8	22.93	17.68	27.35	20.99	26.67	20.65	32.21	24.96	25.20	19.31
InternVL2_5-8B	4	28.53	21.95	30.41	23.20	32.27	24.89	28.17	21.80	22.98	17.46
InternVL2_5-8B	2	25.30	19.43	27.19	20.53	27.87	21.46	31.65	24.36	21.87	16.46
InternVL2_5-8B	1	19.79	15.04	27.01	20.37	23.33	17.97	28.34	21.83	15.82	12.01
InternVL3-8B	16	40.69	31.31	36.54	27.80	42.50	32.88	45.47	34.98	32.74	24.88
InternVL3-8B	8	45.68	35.26	41.74	31.86	47.25	36.79	49.04	38.08	31.98	24.25
InternVL3-8B	4	48.24	37.23	42.11	31.91	49.65	38.65	53.78	41.75	32.51	24.69
InternVL3-8B	2	41.84	32.43	34.02	26.07	41.49	32.33	52.05	40.44	30.64	23.52
InternVL3-8B	1	40.10	31.22	32.98	25.54	45.43	35.89	47.59	37.08	34.41	26.44

G Broader impacts

The introduction of ConViS and ConViS-Bench presents significant potential for advancing the interpretability and alignment of models with human judgment in video understanding. By facilitating more reliable and human-aligned evaluation of video models, our research contributes to the development of responsible AI technologies that serve diverse communities and societal needs. We do not foresee any immediate negative societal impacts resulting from our work. However, there is a risk that ConViS-Bench may not adequately reflect the video understanding capabilities of models in specific tasks or scenarios. We encourage users to carefully consider the intended use cases and limitations of their models when utilizing ConViS-Bench for evaluation, and to supplement it with additional benchmarks or domain-relevant assessments where appropriate.

H Institutional review board approvals

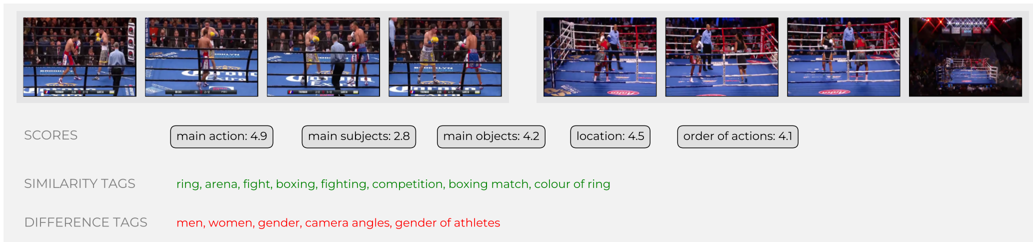
The institutional review board (IRB) reviewed and approved this study, concluding that *"the research in question does not involve risks to the psycho-physical well-being of the subjects involved that could possibly even limit their right to confidentiality, information and autonomy in decision-making."*. The official document is available on request.

I Assets

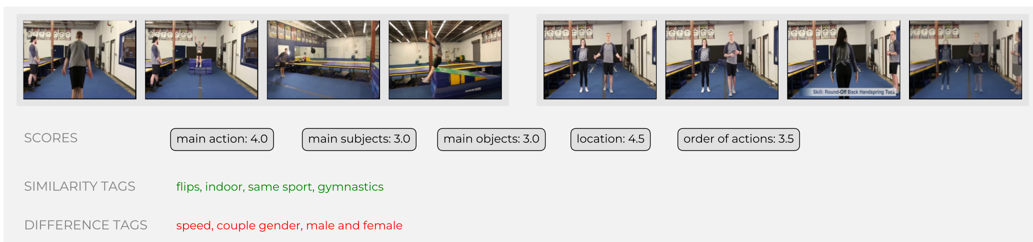
Table 8 lists URLs and licenses for all the assets used in the paper.

Table 8: **List of URLs and licenses for all assets used.**

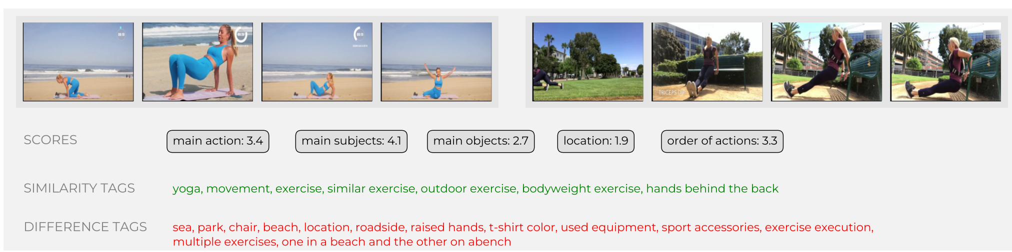
Name	URL	License
mPLUG-Owl3-7B	https://huggingface.co/mPLUG/mPLUG-Owl3-7B-240728	Apache 2.0
LLaVA-OV-7B	https://huggingface.co/lmms-lab/llava-onevision-qwen2-7b-ov	Apache 2.0
LLaVA-OV-0.5B	https://huggingface.co/lmms-lab/llava-onevision-qwen2-0.5b-ov	Apache 2.0
LLaVA-Video-7B	https://huggingface.co/lmms-lab/LLaVA-Video-7B-Qwen2	Apache 2.0
Qwen2.5-VL-7B	https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct	Apache 2.0
Qwen2.5-VL-3B	https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct	Apache 2.0
InternVL2.5-4B	https://huggingface.co/OpenGVLab/InternVL2_5-4B	MIT
InternVL2.5-8B	https://huggingface.co/OpenGVLab/InternVL2_5-8B	MIT
InternVL3-8B	https://huggingface.co/OpenGVLab/InternVL3-8B	MIT
Gemini-2.0-Flash	https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash	Apache 2.0
VideoMAE	https://huggingface.co/MCG-NJU/videomae-large	CC BY-NC 4.0
DINOv2	https://huggingface.co/facebook/dinov2-large	Apache 2.0
CLIP	https://huggingface.co/openai/clip-vit-large-patch14	MIT
InternVideo	https://huggingface.co/OpenGVLab/InternVideo1.0	Apache 2
FineVideo	https://huggingface.co/datasets/HuggingFaceFV/finevideo	CC-BY 4.0



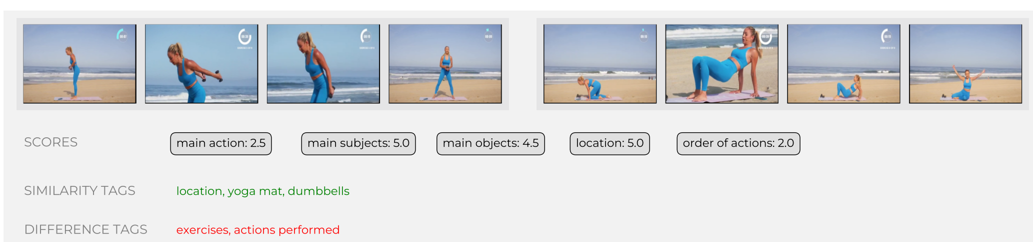
(a)



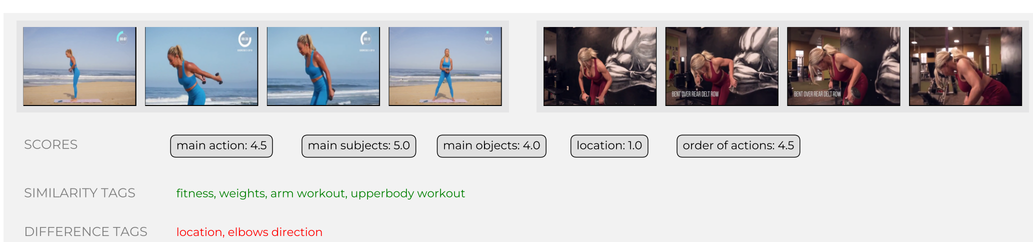
(b)



(c)

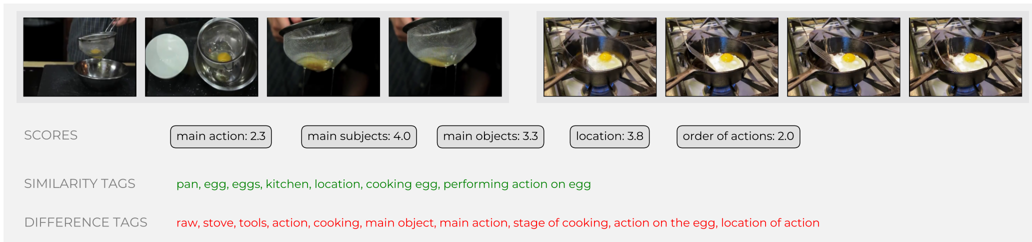


(d)

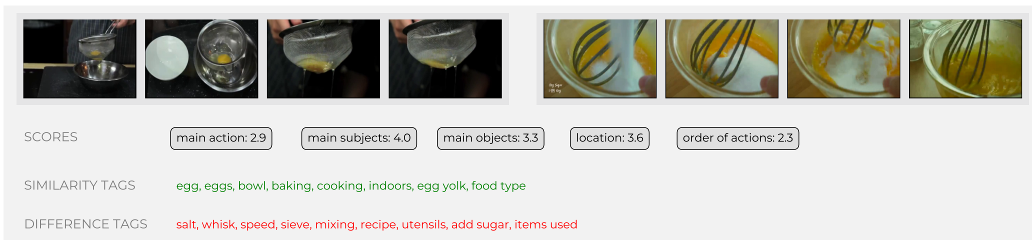


(e)

Figure 14: **Samples from ConViS-Bench.**

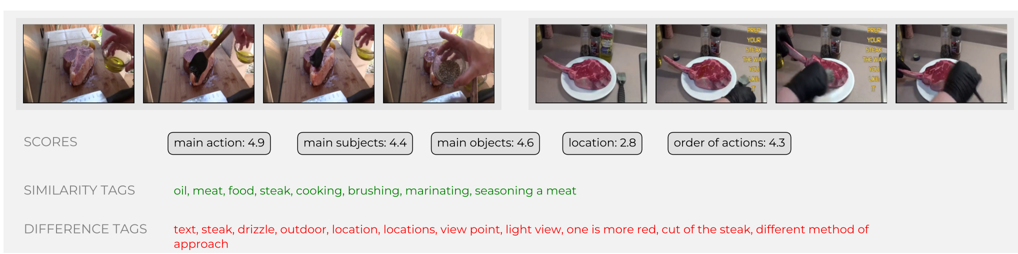


(a)

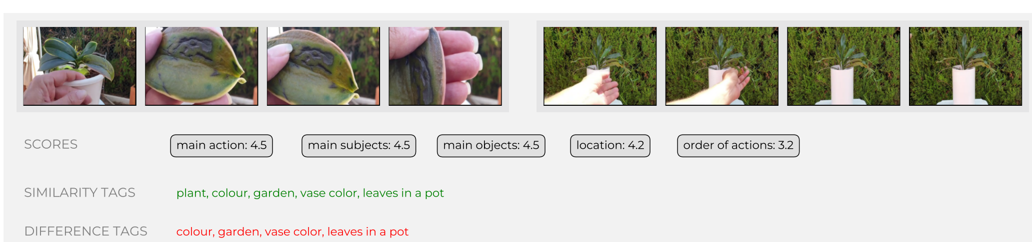


(b)

(c)

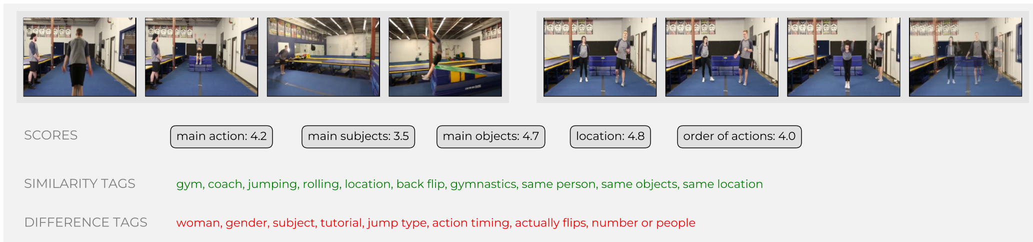


(d)

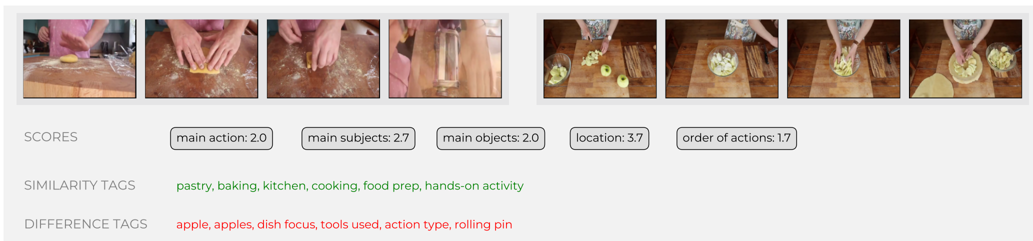


(e)

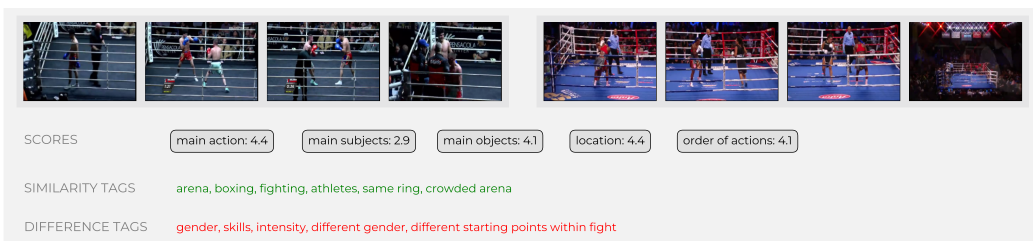
Figure 15: **Samples from ConViS-Bench.**



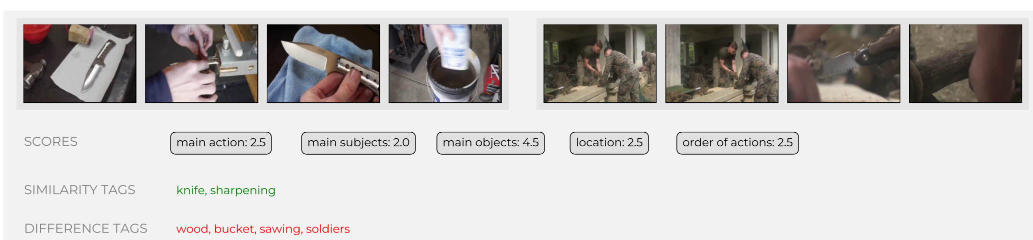
(a)



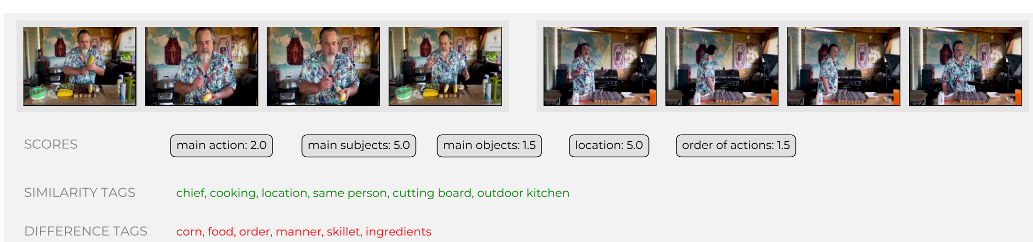
(b)



(c)

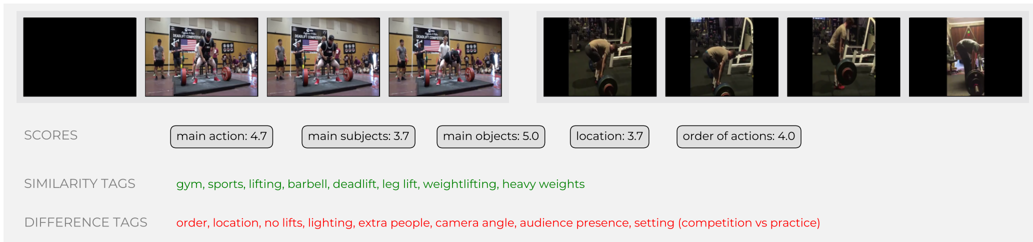


(d)

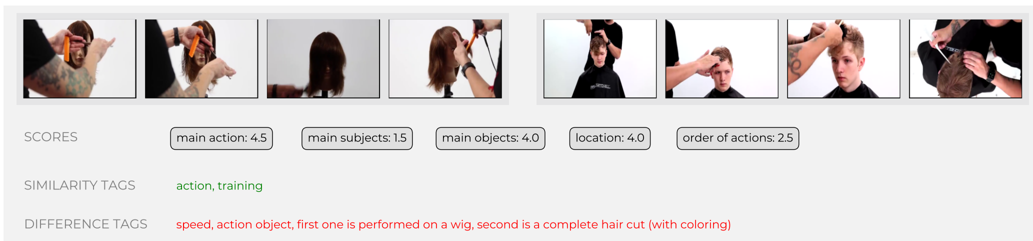


(e)

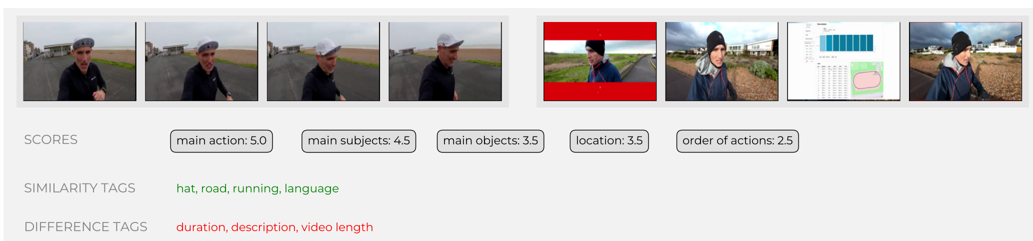
Figure 16: **Samples from ConViS-Bench.**



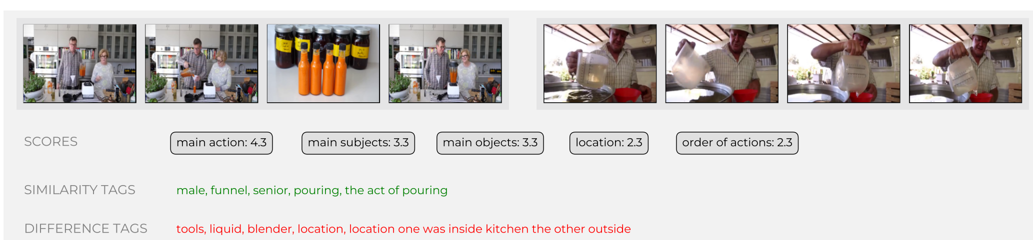
(a)



(b)



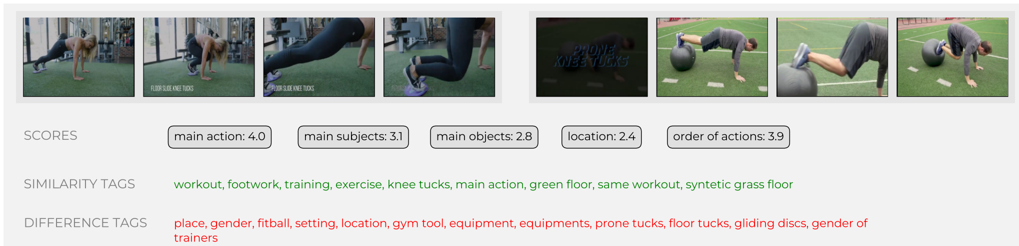
(c)



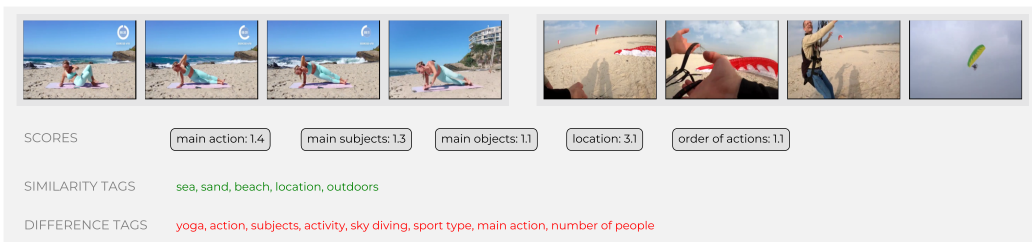
(d)

(e)

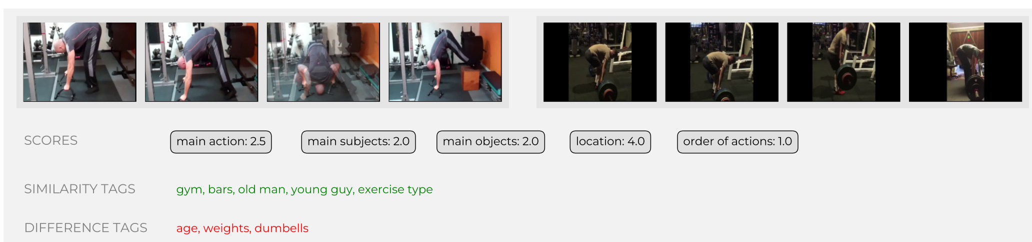
Figure 17: **Samples from ConViS-Bench.**



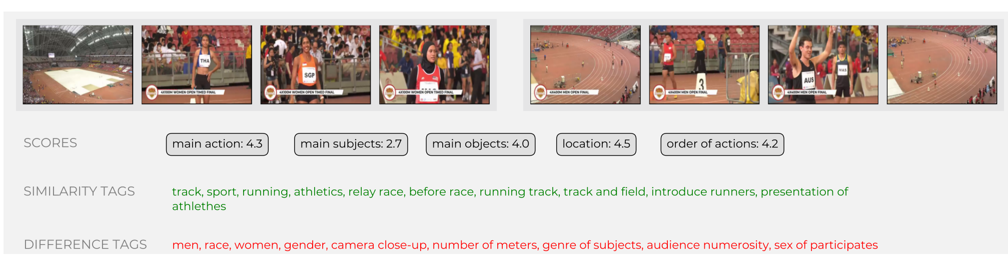
(a)



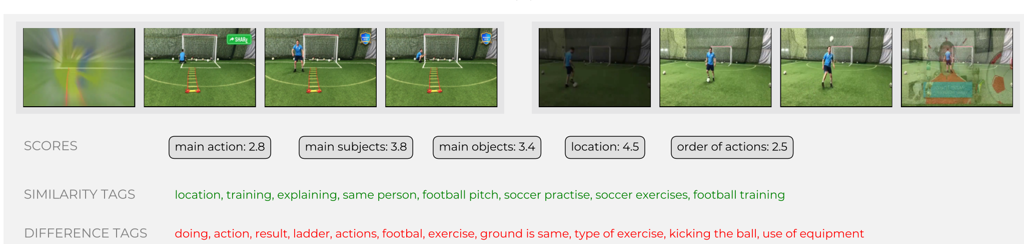
(b)



(c)



(d)



(e)

Figure 18: **Samples from ConViS-Bench.**