Video-Language Entailment for Video Similarity Estimation

Benedetta Liberatori[1], Lorenzo Vaquero[2], Elisa Ricci[1,2], Paolo Rota[1]
[1] University of Trento, [2] Fondazione Bruno Kessler

Project website: https://benedettaliberatori.github.io/vice

Abstract

Comparing videos to assess their semantic similarity remains a challenging task for current Large Multimodal Models (LMMs), which are primarily trained with single-video objectives. While most prior works have focused on computing global similarity, recent studies have introduced the more challenging problem of *concept-based similarity*, which measures how similar two videos are with respect to specific concepts (e.g., activity, location). However, no dedicated approaches have yet leveraged LMMs to explicitly address this task. In this paper, we aim to fill this gap introducing **Vi**deo **C**aption **E**ntailment for video similarity estimation (VICE): a dual-branch design that leverages LMMs for cross-video reasoning. Each branch jointly processes one video and the caption of the other, measuring their alignment for a specific concept via video-language entailment. The outputs from both branches are fused to produce a final similarity score, along with a textual explanation of the model reasoning. This approach leverages the model's native single-video understanding while enabling fine-grained comparative reasoning between video and text. Experiments on ConViS-Bench and StepDiff show that VICE outperforms models that process both videos simultaneously, highlighting the promise of single-video reasoning for concept-based similarity estimation.

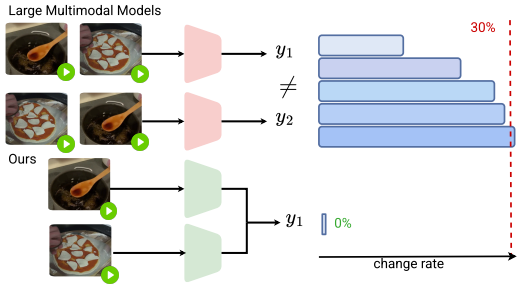

Figure 1: **Order sensitivity of Large Multimodal Models in video-pair understanding.** Large Multimodal Models (LMMs) exhibit strong order dependence: swapping the order of videos in a pair leads to different predictions in approximately 25–30% of cases. In contrast, our method is order-invariant (0% change rate).

1 Introduction

Understanding video similarity lies at the core of numerous computer vision tasks, from video retrieval (Wu et al., 2025) and content recommendation to reasoning over multi-source information (Meng et al., 2025). Yet, defining what makes two videos *similar* remains challenging, as similarity depends on several factors like semantics, context, and purpose. Two clips may depict an athlete kicking a ball in different environments, or people cooking pasta with different ingredients and utensils. Capturing semantic correspondences requires models that go beyond low-level visual cues and single-concept matches, revealing a deeper understanding of the events depicted in videos. This observation has motivated research on *concept-based video similarity estimation* (ConViS) (Liberatori et al., 2025): estimating how similar two videos are with respect to a specific concept (e.g., location, actions). Unlike traditional global video similarity estimation (Liu et al., 2024; Kordopatis-Zilos et al., 2023; Jo et al., 2024), ConViS requires models to selectively attend to one concept while ignoring irrelevant variations, making it more challenging.

Recent Large Multimodal Models (LMMs) (Li et al., 2025; Zhu et al., 2025a; Bai et al., 2025b; Zhang et al., 2025) have achieved remarkable progress in video understanding. Currently, LMMs can accurately describe videos, answer questions about their content, and even make comparisons using natural language (Nagarajan and Torresani, 2024; Burgess et al., 2025). Their generative nature allows them to be easily prompted with natural language, making them particularly suitable for ConViS: by prompting them with a concept-specific instruction, users can query an LMM to estimate a video similarity score along the desired di-

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 35690–35706
October 24-29, 2026 ©2026 Association for Computational Linguistics

mension, without the overhead of concept-specific fine-tuning. This aspect was recently investigated in (Liberatori et al., 2025), where the capabilities of different LMMs were analyzed and compared across a variety of concepts.

Previous works targeting video-pair understanding with LMMs feed both videos at once (Zhu et al., 2025b; Liberatori et al., 2025; Burgess et al., 2025). We empirically find that this approach suffers from a significant limitation: LMMs are highly sensitive to the ordering of the videos. As illustrated in Fig. 1, simply reversing the order of videos in input can result in inconsistent judgments, with variations in responses ranging from 20% to 30% across different models. These findings align with earlier observations in multiple-choice question answering, where the order of possible answers influences model predictions despite constituting semantically equivalent inputs (Zheng et al., 2024; Alzahrani et al., 2024; Pezeshkpour and Hruschka, 2024). Such instability undermines the reliability of current LMMs for tasks that require robust similarity estimation between video pairs. We hypothesize that this brittleness stems from a mismatch between their pretraining objectives and the requirements of the intended multi-video reasoning tasks. Current LMMs are pretrained on single-video objectives (e.g., captioning or question answering) (Bai et al., 2025b), limiting their ability to reason across multiple visual streams. Consequently, multi-video inputs push these models outside their training regime, often yielding inconsistent outputs (Meng et al., 2025; Qiu et al., 2025; Burgess et al., 2025).

We introduce **Vi**deo **C**aption **E**ntailment for video similarity estimation (VICE), a novel test-time inference approach that harnesses the inherent capabilities of LMMs while avoiding their weaknesses. Rather than feeding both videos simultaneously, we decompose the task into a dual-branch single-video pipeline: each branch evaluates entailment between one video and the textual description of the other with respect to a target concept, and the outputs are fused into a final similarity score with a textual explanation. This design ensures LMMs operate strictly within their native single-video setting, avoiding degradation from multi-video prompts while enabling interpretable, language-grounded similarity judgments.

We evaluate our approach on ConViS-Bench (Liberatori et al., 2025), a benchmark explicitly designed for concept-conditioned video similarity. Additionally, we repurpose the StepDiff (Nagarajan and Torresani, 2024) dataset, originally proposed for difference description in video pairs, by providing human-curated annotations following the same protocol as ConViS-Bench. Our evaluation, conducted on multiple LMMs, shows that VICE consistently produces more semantically-faithful similarity estimates than directly inputting the video pairs. This shows that framing ConViS within a caption-then-entail paradigm not only addresses the limitations of models trained with single-video objectives, but also improves interpretability and robustness in concept-based video comparison.

In summary, our contributions are threefold: *i)* We introduce VICE, a dual-branch framework that leverages LMMs to compare videos using video-language entailment. To the best of our knowledge, it is the first method specifically designed for this task. *ii)* VICE operates within the model's native single-video setting, yielding more semantically faithful similarity estimates and interpretable textual explanations. *iii)* We validate VICE on ConViS-Bench and repurpose the StepDiff dataset with concept-level human similarity judgments, highlighting the effectiveness of caption-then-entail reasoning on multiple LMMs.

2 Related Work

Large Multimodal Models for Video. Recent advances in foundation models have led to the emergence of architectures capable of processing videos natively (Zhang et al., 2025). Meanwhile, several works have extended image-centric models to the video domain, allowing unified understanding across static and dynamic visual data (Ye et al., 2025; Zhu et al., 2025a; Wang et al., 2024; Bai et al., 2025b,a). Despite these developments, most existing LMMs remain limited in their ability to jointly process and reason over multiple distinct videos. Recent efforts have started to investigate large-scale video foundation models in multi-video contexts (Meng et al., 2025; Zhu et al., 2025b; Qiu et al., 2025), uncovering significant limitations in the current architectures. These models are neither architecturally designed nor pretrained to jointly reason over multiple videos, limiting their effectiveness in tasks that require integrated understanding across distinct visual streams. Differently, this work specifically focuses on reasoning over two videos, as required by video similarity estimation.

Concept-based Video Similarity Estimation.

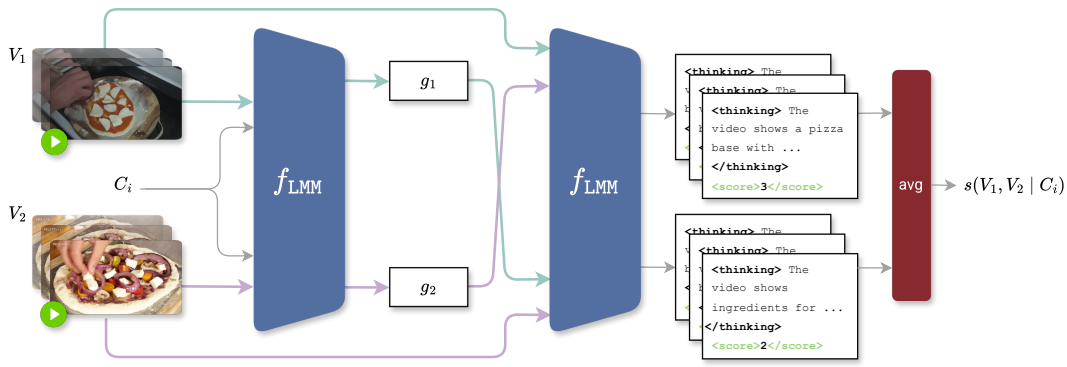

Figure 2: **Overview of VICE:** Given a video pair (V_1, V_2) and a concept C_i, e.g., **`Ingredients`**, VICE processes the videos through two parallel branches. Each branch generates a concept-conditioned caption (g_1, g_2), then performs video–caption entailment by comparing one video with the caption of its counterpart. The branches produce reasoning traces inside **`<thinking>. . . </thinking>`** and similarity scores inside **`<score>. . . </score>`**. The final score $s(V_1, V_2 \mid C_i)$ is averaged across both branches. For clarity, we omit visualizations of the reasoning trace summarization (Sec. 3.3), as it does not affect the predicted score and is used solely for interpretability.

Motivated by prior work on concept-based image similarity (Achille et al., 2024), recent research has explored how to compare videos with respect to specific semantic concepts. VidDiffBench (Burgess et al., 2025) advances this direction by prompting models to generate textual explanations of subtle variations in how actions are performed, focusing on fine-grained differences. StepDiff (Nagarajan and Torresani, 2024), in contrast, investigates difference-centric descriptions of procedural videos and extends this paradigm by comparing videos along multiple dimensions relevant to the domain, such as ingredients, tools, or techniques. However, its scope remains largely limited to cooking-related content, emphasizing difference captioning within this narrow procedural domain. To address this, ConViS-Bench (Liberatori et al., 2025) introduces quantitative, concept-specific similarity measures across a broader spectrum of domains and semantic categories (e.g., objects, locations, actions). While ConViS-Bench focuses on benchmarking LMMs, our work introduces the first dedicated approach for concept-based video similarity estimation.

Video-Language Entailment. Video-language entailment examines whether a video semantically supports a given textual hypothesis. Originally posed as a binary classification task (Liu et al., 2020), it has evolved into a framework for fine-grained reasoning and compositional inference (Saravanan et al., 2025), with early works exploring multiple pretraining objectives and explicit visual grounding (Li et al., 2020; Chen and Golisano, 2021; Sanders et al., 2024). Recent work leverages LMMs for entailment directly through prompting (Zhao et al., 2025). Advances along this line include task-specific fine-tuning with synthetic videos (Zanella et al., 2025) and the use of textual explanations to guide model decisions and enhance interpretability (Bansal et al., 2024). In this work, we extend video-language entailment to multi-video reasoning.

3 Method

We introduce **Vi**deo **C**aption **E**ntailment for video similarity estimation (VICE), a test-time inference, dual-branch method that uses LMMs to estimate concept-conditioned similarities between video pairs. In the following, we first formally define the task (Sec. 3.1), then describe our proposed VICE (Sec. 3.2), and show how we can synthesize human-readable explanations from it (Sec. 3.3).

3.1 Problem Formulation

Given a video pair and a concept of interest, *concept-based video similarity estimation* (ConViS) aims to quantify how similar the two videos are with respect to that specific concept. Formally, let $\mathcal{C} = \{C_1, \ldots, C_K\}$ denote a set of concepts in natural language, and (V_1, V_2) a video pair. ConViS aims to predict a concept-conditioned similarity score $s(V_1, V_2 \mid C_i) \in [1, 5]$, for each $C_i \in \mathcal{C}$.

For example, consider the two cooking videos shown in Fig. 2. Conditioned on $C_i =$ *"action"*, the similarity should be high, as both videos depict adding toppings to a pizza. Conditioned on $C_i =$ *"ingredients"*, it should be lower, as the videos differ in toppings and cheese type.

3.2 Video Caption Entailment for ConViS

Large multimodal models have demonstrated impressive zero-shot capabilities on many downstream video tasks, primarily driven by large-scale pretraining on video-text data (Zhang et al., 2025; Zhu et al., 2025a; Bai et al., 2025b). However, these models remain much stronger at processing individual videos than reasoning across multiple video inputs simultaneously (Meng et al., 2025; Zhu et al., 2025b; Qiu et al., 2025).

Building upon these findings, we propose VICE, a zero-shot framework that leverages LMMs to estimate concept-based similarity between video pairs. As illustrated in Fig. 2, VICE follows a two-stage pipeline with two parallel branches. The first stage, *concept-conditioned captioning*, generates textual descriptions for each video that explicitly emphasize the target concept. In the second stage, *video-caption entailment*, each video is paired with the other video's caption and fed to an LMM, which evaluates how well the video content aligns with the corresponding concept-conditioned description.

Concept-conditioned captioning. Let $f_{\texttt{LMM}}$ denote the LMM and $j \in \{1, 2\}$ the index of the branch. In the first stage, each branch produces a *concept-conditioned* textual description of each video in the pair (V_1, V_2), explicitly reflecting the conditioning concept:

$$f_{\texttt{LMM}}(V_j, C_i) = g_j, \quad j \in \{1, 2\},\ C_i \in \mathcal{C}, \quad (1)$$

where g_j represents the concept-conditioned caption of video V_j produced by branch j under conditioning C_i.

In this phase, we propose to leverage the zero-shot capabilities of the LMM, which was pretrained on single-video captioning tasks (Zhang et al., 2025; Zhu et al., 2025a; Bai et al., 2025b), allowing it to generate meaningful concept-conditioned captions from a single video in input. Back to our example in Fig. 2, conditioned on the concept C_i = *"ingredients"*, the model produces captions such as *"tomato sauce, mozzarella cheese, and basil"* for V_1 and *"mushrooms, tomatoes, red onion, cheese, and herbs"* for V_2. In contrast, for C_i = *"action"*, both videos are described as *"adding toppings to the pizza"*, highlighting the shared procedural aspect.

Video-caption entailment. In the second stage, VICE measures semantic alignment between the two videos by evaluating bidirectional entailment (i.e., how well each video supports the other's concept-conditioned caption). For videos V_1 and V_2, this is performed in both directions:

$$f_{\texttt{LMM}}(V_j, g_k, C_i) = \tau_j, j, k \in \{1, 2\},\ j \neq k, \quad (2)$$

where τ_j is the entailment output of the model.

Judging the degree to which a video entails a description requires complex reasoning with at least two steps: understanding the content of the video and assessing the relevance of the description. To stimulate this reasoning capability, we generate the output τ_j as a sequence of intermediate steps, leading to the final entailment score a_j as:

$$\tau_j = [z_{j_1}, \ldots, z_{j_N}, a_j] \sim p(\cdot \mid V_j, g_k, C_i), \quad (3)$$

where $p(\cdot)$ is the probability distribution over the vocabulary of $f_{\texttt{LMM}}$ and each intermediate step is generated autoregressively as:

$$z_{j_n} \sim p(\cdot \mid V_j, g_k, C_i, z_{j_1}, \ldots, z_{j_{n-1}}), \quad (4)$$

for $n = 1, \ldots, N,\ j, k \in \{1, 2\},\ j \neq k$.

With reference to Fig. 2, intermediate steps z_{j_n} and final score a_j are wrapped in `<thinking>...</thinking>` and `<score>...</score>` tags, constituting a reasoning trace. This syntax not only provides free-form rationales describing the evidence and comparisons relevant to the concept C_i, but also actively guides the model's decision process (Kojima et al., 2022).

Returning to our example in Fig. 2, when evaluating whether V_1 entails V_2's caption for C_i = *"ingredients"*, the reasoning trace might focus first on observing *mushrooms* in V_1, then comparing to V_2's mention of *tomato sauce, mozzarella, and basil*, and finally output a lower entailment score for the *ingredients* concept. For the C_i = *"action"* concept, the two directions observe the videos adding toppings and output a high entailment score, reflecting the common procedure.

To improve robustness, we generate reasoning traces τ_j^m through repeated sampling from $f_{\texttt{LMM}}$:

$$\{\tau_j^1, \ldots, \tau_j^M\} \quad \text{for } m = 1, \ldots, M. \quad (5)$$

These reasoning traces may capture a diverse set of scores and mitigate the impact of outlier predictions on the semantic alignment evaluation (Wang et al., 2023), as sampled traces can lead to different final entailment scores.

Lastly, the final concept-conditioned similarity score between V_1 and V_2 under concept C_i is obtained by averaging the entailment scores across all

Model	Main Action	Main Subjects	Main Objects	Location	Order of Actions	Avg.
LLaVA-OV-7B (Li et al., 2025)						
zero-shot	51.8	48.4	**58.6**	58.9	**41.0**	**51.7**
zero-shot CoT	39.9	28.6	28.3	44.1	28.0	33.8
VICE	**54.4**	**55.9**	50.0	**64.8**	19.2	48.9$_{\downarrow 2.8}$
LLaVA-Video-7B (Zhang et al., 2025)						
zero-shot	44.2	39.8	45.8	56.0	41.2	45.4
zero-shot CoT	25.1	25.1	31.3	30.2	14.7	25.3
VICE	**63.8**	**56.0**	**60.1**	**64.7**	**57.9**	**60.5**$_{\uparrow 15.1}$
Qwen3-VL-2B (Bai et al., 2025a)						
zero-shot	29.2	31.3	35.9	40.9	16.6	30.8
zero-shot CoT	31.5	37.5	38.2	59.1	9.8	35.2
VICE	**58.4**	**50.7**	**60.0**	**61.8**	**52.3**	**56.3**$_{\uparrow 25.5}$
Qwen3-VL-8B (Bai et al., 2025a)						
zero-shot	43.0	43.3	47.1	52.8	35.7	44.4
zero-shot CoT	51.0	46.9	50.2	64.5	41.1	50.7
VICE	**64.6**	**49.6**	**61.4**	**69.8**	**62.7**	**61.6**$_{\uparrow 17.2}$
Qwen3-VL-32B (Bai et al., 2025a)						
zero-shot	60.4	50.2	51.8	58.8	52.7	54.8
zero-shot CoT	65.3	52.8	57.7	70.1	64.5	62.1
VICE	**69.4**	**61.7**	**61.4**	**70.2**	**65.3**	**65.6**$_{\uparrow 10.8}$

Table 1: **Correlation performance on ConViS-Bench** (Liberatori et al., 2025). We apply VICE to multiple LMMs, reporting Spearman's ρ correlations with human judgment ($\times 100$).

generated traces, as in eq. (3), and both branches:

$$s(V_1, V_2 \mid C_i) = \frac{1}{2M} \sum_{j=1}^{2} \sum_{m=1}^{M} a_j^m, \qquad (6)$$

where a_j^m is the score from the m-th trace τ_j.

Asymmetry between branches. Importantly, we do not require the two branches to produce mutually consistent scores. Because entailment is inherently asymmetric, it is possible that the caption g_1 distilled from V_1 describes only a subset of the ingredients that appear in V_2, while g_2 distilled from V_2 lists a richer set of toppings. In this case, the entailment score for (V_2, g_1, C_i) can be high, as V_2 easily satisfies the simpler description, whereas the reverse direction (V_1, g_2, C_i) may yield a lower score, since V_1 does not contain all the ingredients referenced in g_2. We deliberately preserve this asymmetry, as it captures whether one video conceptually subsumes the other under C_i; by combining the scores from both branches, VICE arrives at a more reliable and balanced similarity estimate.

3.3 From Reasoning Traces to Explanations

While VICE aggregates entailment scores across directions and reasoning traces, its intermediate `<thinking>...</thinking>` traces also contain rich, concept-specific evidence (e.g., object mentions, temporal cues, counterexamples). To better understand the underlying f_{LMM} model's behavior and make this information useful for real-world applications, we synthesize a single, human-readable explanation by summarizing the reasoning traces generated for both directions. Let:

$$\mathcal{R}_i = \left\{ \mathbf{z}_j^m \;\middle|\; j \in \{1,2\},\, m \in \{1,\dots,M\} \right\} \qquad (7)$$

denote the multiset of reasoning traces associated with concept C_i, where each $\mathbf{z}_j^m = [z_{j_1}^m, \dots, z_{j_N}^m]$ denotes the generated reasoning trace omitting the final score. We employ a large language model f_{LLM} to produce a concept-grounded explanation:

$$e_i \;=\; f_{\text{LLM}}\big(\mathcal{R}_i\big). \qquad (8)$$

This explanation covers consensus evidence supporting similarity under C_i. This step is strictly post hoc and does not modify the numeric similarity $s(V_1, V_2 \mid C_i)$. It yields a holistic, concept-aligned narrative that explains the model's decision, making VICE transparent and practical for downstream use. We validate the quality of these explanations through a human evaluation study; details and results are provided in Appendix C.

4 Experiments

In this section, we present the experimental protocol and evaluate VICE on the ConViS task, showing improvements across multiple LMMs through extensive quantitative and qualitative analyses.

Dataset and Metrics. We use two datasets designed for ConViS: ConViS-Bench (Liberatori et al., 2025) and StepDiff (Nagarajan and Torresani, 2024). ConViS-Bench comprises 610 video pairs, annotated with human judgments across five broad concepts: main action, main subjects, main objects, location, and order of actions. StepDiff contains

5946 test videos, originally annotated on a subset of five cooking-related dimensions: technique, tools/equipment, actions, visuals, and ingredients. Since the public release included only partial annotations, we re-annotated the full set with human annotators to ensure complete and consistent coverage on 1269 video pairs. Details on the annotation process are provided in Appendix A. We use Spearman's ρ (scaled by $\times 100$) to evaluate correlation with humans, following standard protocols (Achille et al., 2024), as it provides a robust measure in the presence of ties (Puth et al., 2015). Implementation details and prompts are provided in Appendix B and G, respectively.

4.1 Main Results

We evaluate our proposed ViCE on multiple strong off-the-shelf LMMs for video understanding: Qwen3-VL-2B, Qwen3-VL-8B, Qwen3-VL-32B (Bai et al., 2025a), LLaVA-OneVision-7B (Li et al., 2025), LLaVA-Video-7B (Zhang et al., 2025). As no prior work has addressed ConViS with a dedicated methodology, we adopt these LMMs as baselines. We employ a *zero-shot* baseline where the two videos are concatenated into a single input, following prior video-pair understanding works (Liberatori et al., 2025). We further extend this baseline with a chain-of-thought variant (i.e., *zero-shot CoT*), prompting the model to produce an explicit reasoning trace before the final output score.

Tab. 1 and Tab. 2 report the results obtained on ConViS-Bench (Liberatori et al., 2025) and StepDiff (Nagarajan and Torresani, 2024) respectively. ViCE consistently outperforms the underlying LMMs across both datasets, yielding substantial gains on the *zero-shot* baseline of up to 25.5% on ConViS-Bench and 23.3% on StepDiff on average. This highlights the effectiveness of ViCE in leveraging LMMs within their preferred single-video input setting, while still enabling reliable cross-video comparison in a zero-shot regime.

ViCE elevates compact models to performance tiers typically reserved for much larger counterparts: a 2B model equipped with our framework consistently outperforms standard 8B baselines (e.g., +11.9% on average, Tab. 1); while an 8B model surpasses even 32B baselines (e.g., +6.8% on average, Tab. 1). This shows that test-time inference is a compelling and resource-efficient alternative to parameter scaling, delivering higher accuracy at a fraction of the GPU memory cost.

Fig. 3 shows the score distribution of predictions

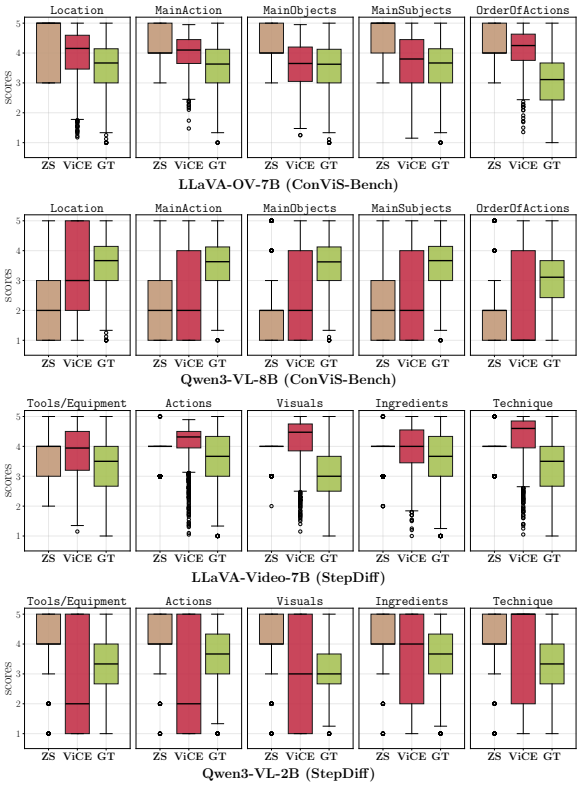

Figure 3: **Distributions of predicted scores across datasets and models.** We compare zero-shot predictions, predictions with ViCE, and ground truths.

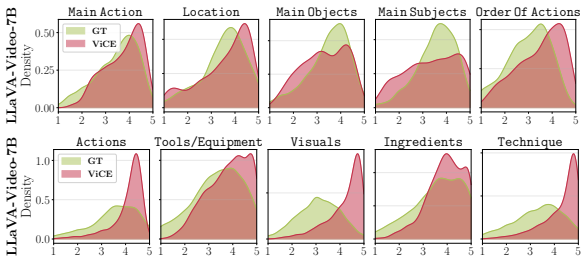

Figure 4: **Distributions of predicted scores.** We compare ViCE's predictions to the ground truth using LLaVA-Video-7B model as the LMM.

made by the *zero-shot* LMM and ViCE, as well as the ground truth scores. LLaVA-OV-7B produces generally high similarity scores across all concepts, with no scores falling below 3 for *location*. Applying ViCE mitigates this bias, yielding a wider distribution that better matches the ground truth. For the Qwen3-VL models, the *zero-shot* baseline consistently predicts low scores on ConViS-Bench, while skewing toward very high scores on StepDiff, whereas ViCE consistently produces distributions that better reflect the ground truth, including both low and high extremes. For LLaVA-Video-7B on StepDiff, the *zero-shot* baseline produces scores heavily peaked at 4; ViCE again yields more diverse, better-calibrated distributions.

Model	Technique	Tools/Equipment	Actions	Visuals	Ingredients	Avg.
LLaVA-OV-7B (Li et al., 2025)						
zero-shot	31.1	**39.4**	32.8	**32.3**	**28.5**	32.8
zero-shot CoT	29.5	35.3	29.4	25.3	24.1	28.7
ViCE	**42.2**	36.6	**32.8**	31.8	26.5	**34.0**$_{\uparrow 1.2}$
LLaVA-Video-7B (Zhang et al., 2025)						
zero-shot	30.6	37.4	28.9	**29.2**	17.5	28.7
zero-shot CoT	26.3	35.3	27.4	22.7	15.1	25.4
ViCE	**48.4**	**37.8**	**34.4**	28.5	**34.5**	**36.7**$_{\uparrow 8.0}$
Qwen3-VL-2B (Bai et al., 2025a)						
zero-shot	11.7	20.2	12.9	5.5	18.5	13.8
zero-shot CoT	18.5	25.6	20.3	15.3	14.7	18.9
ViCE	**48.1**	**47.1**	**50.2**	**39.8**	**25.9**	**42.2**$_{\uparrow 23.3}$
Qwen3-VL-8B (Bai et al., 2025a)						
zero-shot	42.6	**52.6**	48.2	41.8	35.9	44.2
zero-shot CoT	43.4	47.4	47.0	41.3	36.7	43.2
ViCE	**53.7**	50.1	**49.2**	**43.0**	**39.8**	**47.2**$_{\uparrow 3.0}$
Qwen3-VL-32B (Bai et al., 2025a)						
zero-shot	47.1	51.3	49.2	42.1	43.5	46.6
zero-shot CoT	48.2	50.0	51.6	43.8	45.4	47.8
ViCE	**50.1**	**52.2**	**51.8**	39.8	**45.4**	**47.9**$_{\uparrow 1.3}$

Table 2: **Correlation performance on StepDiff** (Nagarajan and Torresani, 2024). We apply ViCE to multiple LMMs, reporting Spearman's ρ correlations with human judgment ($\times 100$).

Fig. 4 shows the distributions of ViCE's predicted scores against the ground truth using LLaVA-Video-7B across both datasets. On ConViS-Bench, ViCE's predictions are more closely aligned with the ground truth for the concepts of *main action*, *location*, and *main objects*. In contrast, on StepDiff, the model performs better on the more object-centric concepts of *tools/equipment* and *ingredients*. We attribute the weaker action-related performance on StepDiff to its finer granularity and the fact that its *action* concept encompasses not only the action itself but also procedural step ordering.

Qualitative results. Fig. 5 presents a qualitative example, with additional ones in Appendix E. The concept-conditioned descriptions effectively isolate the target concept, and the reasoning traces provide evidence that it influences the final score, yielding assessments that align more closely with ground-truth preferences than the *zero-shot* baseline.

5 Ablation Studies

We present ablation studies on the contribution of each component of ViCE in Tab. 3.

Blind. To isolate the contribution of visual information during the entailment step, we introduce a *Blind* variant that removes the video–caption entailment entirely and replaces it with a purely text-based large language model (LLM). For each video pair, we provide the two concept-conditioned captions generated via Eq. (1) to the LLM and prompt it to predict a similarity score based solely on the text. We adopt LLaMA3-8B-Instruct (Grattafiori et al., 2024) for this purpose. Across all LMMs, the *Blind* variant consistently yields lower performance than ViCE, indicating that visual inputs provide complementary information that is not necessarily captured by caption text alone.

Effect of dual-branch design. The core idea of ViCE lies in its dual-path architecture. For this experiment, we caption one video only and entail it against the other one, having a *single-branch* design. Results show that while each branch contributes to the final prediction, a single branch alone is not sufficient to match ViCE's performance, showing the complementary nature of the branches and the effectiveness of their joint use.

Effect of repeated sampling. ViCE generates multiple outputs within each branch, effectively exploring a richer set of reasoning paths. This produces similarity estimates that better align with human judgments, as different reasoning traces can capture complementary aspects of a concept. We show that a single-CoT approach ($M = 1$) underperforms ViCE, highlighting the benefit of aggregating multiple reasoning processes for robustness.

Effect of reasoning traces. ViCE produces concept-conditioned similarity scores along with corresponding reasoning traces. These traces add interpretability by revealing the model's decision process, while also improving score quality. When we disable reasoning generation and prompt ViCE

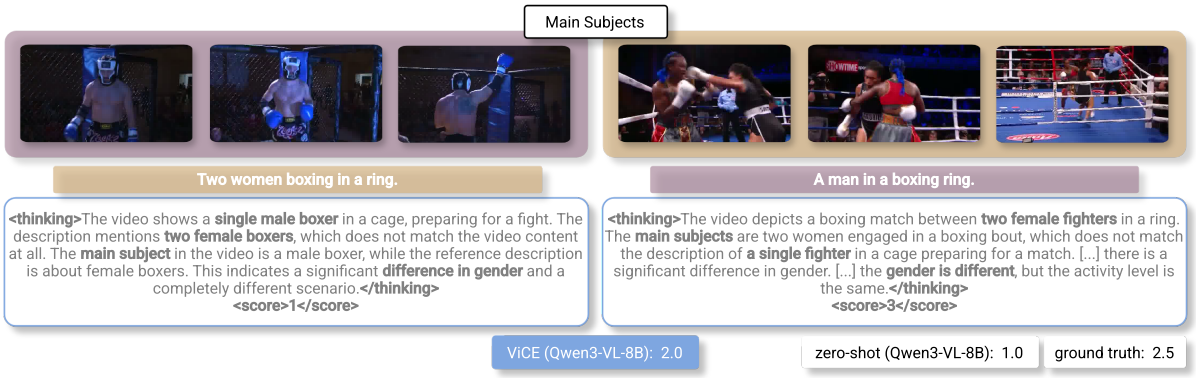

Figure 5: Example of a video pair with the captions and reasoning traces generated when conditioning on the target concept. Compared to the *zero-shot* baseline, ViCE produces scores that more closely align with the ground truth.

Model	Main Action	Main Subjects	Main Objects	Location	Order of Actions	Avg.
LLaVA-OV-7B (Li et al., 2025)						
Blind 🙈	53.4	38.5	42.9	47.0	8.1	38.0
ViCE w/o *CoT*	48.1	48.9	46.7	57.2	20.2	44.2
ViCE single branch	47.6	48.4	43.0	58.4	**25.0**	44.5
ViCE ($M = 1$)	48.2	49.2	47.1	60.8	40.3	**49.1**
ViCE	**54.4**	**55.9**	**50.0**	**64.8**	19.2	48.9
LLaVA-Video-7B (Zhang et al., 2025)						
Blind 🙈	53.3	40.2	44.4	48.7	42.2	45.8
ViCE w/o *CoT*	57.4	50.1	54.2	56.7	49.7	53.6
ViCE single branch	56.2	48.6	52.0	59.1	50.4	53.3
ViCE ($M = 1$)	56.6	49.4	57.3	61.3	52.0	55.3
ViCE	**63.8**	**56.0**	**60.1**	**64.7**	**57.9**	**60.5**
Qwen3-VL-2B (Bai et al., 2025a)						
Blind 🙈	47.5	36.2	42.6	41.3	44.3	42.4
ViCE w/o *CoT*	50.0	41.5	53.6	61.5	50.3	51.4
ViCE single branch	52.8	41.3	50.5	55.4	44.5	48.9
ViCE ($M = 1$)	51.3	46.4	54.9	58.5	43.7	51.0
ViCE	**58.4**	**50.7**	**60.0**	**61.8**	**52.3**	**56.3**
Qwen3-VL-8B (Bai et al., 2025a)						
Blind 🙈	54.8	39.3	51.0	45.8	45.1	47.2
ViCE w/o *CoT*	59.0	46.2	58.3	68.7	59.2	58.3
ViCE single branch	55.8	42.8	49.9	62.0	52.3	52.6
ViCE ($M = 1$)	61.6	45.2	57.9	67.5	58.7	58.2
ViCE	**64.6**	**49.6**	**61.4**	**69.8**	**62.7**	**61.6**
Qwen3-VL-32B (Bai et al., 2025a)						
Blind 🙈	56.1	42.8	52.3	53.7	54.3	51.8
ViCE w/o *CoT*	62.6	52.1	58.2	68.9	61.3	60.6
ViCE single branch	63.3	55.4	55.1	63.6	62.5	60.0
ViCE ($M = 1$)	66.2	56.3	60.9	69.4	60.5	62.7
ViCE	**69.4**	**61.7**	**61.4**	**70.2**	**65.3**	**65.6**

Table 3: **Ablation Studies on ConViS-Bench** (Liberatori et al., 2025). We report Spearman's ρ ($\times 100$).

to output the final similarity score directly in the caption entailment step, performance drops by up to 6.9% on average. This demonstrates that generating explicit reasoning traces encourages more structured and coherent representations, leading to more accurate similarity estimation.

6 Conclusions

In this work, we addressed concept-based video similarity, a task where existing LMMs struggle due to their sensitivity to video ordering. We introduced ViCE, a novel test-time inference method based on a dual-branch, caption-then-entail approach, which additionally produces human-readable explanations that summarize the reasoning behind its predictions. Experiments on two benchmarks show consistent improvements in both accuracy and interpretability over baselines that process video pairs jointly. As an additional contribution to the community, we release human-curated concept-level similarity annotations for StepDiff video pairs, providing a new benchmark resource for concept-based video understanding.

7 Limitations

Being a test-time inference method, VICE's performance depends on the capabilities of the underlying LMM, and may suffer if the model cannot properly explain the concept of interest. However, this same property makes VICE flexible: without requiring concept-specific fine-tuning, it can be directly applied to any concept expressible in natural language. As LMMs advance, we expect VICE's performance to improve correspondingly.

Additionally, VICE requires more forward passes than a standard zero-shot baseline: two captioning operations and $2M$ entailment operations per video pair. However, the M reasoning traces within each branch are generated in parallel via batched sampling, and the two branches can be parallelized as well, reducing the number of sequential forward passes to just two. This test-time compute overhead enables smaller models to outperform larger ones, as discussed in Sec. 4.

Finally, VICE is not designed for long videos, as concept-conditioned captioning may fail to capture all relevant moments under sparse frame sampling. We note, however, that cross-video understanding remains an open challenge even for short videos, and we see long-video comparison as a natural direction for future work as the field matures.

Acknowledgments

This work was supported by the EU Horizon projects IAMI (No. 101168272), and ELLIOT (No. 101214398). We also thank the Deep Learning Lab of the ProM Facility for the GPU time.

References

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

Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M. Saiful Bari, and Haidar Khan. 2024. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *ACL*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *CoRR*.

Hritik Bansal, Yonatan Bitton, Idan Szpektor, Kai-Wei Chang, and Aditya Grover. 2024. Videocon: Robust video-language alignment via contrast captions. In *CVPR*.

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

Junwen Chen and Yu Kong Golisano. 2021. Explainable video entailment with grounded visual evidence. In *ICCV*.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *NeurIPS*.

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

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

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. HERO: hierarchical encoder for video+language omni-representation pre-training. In *EMNLP*.

Benedetta Liberatori, Alessandro Conti, Lorenzo Vaquero, Yiming Wang, Elisa Ricci, and Paolo Rota. 2025. Convis-bench: Estimating video similarity through semantic concepts. In *NeurIPS*.

Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. 2020. Violin: A large-scale dataset for video-and-language inference. In *CVPR*.

Yang Liu, Qianqian Xu, Peisong Wen, Siran Dai, and Qingming Huang. 2024. Not all pairs are equal: Hierarchical learning for average-precision-oriented video retrieval. In *ACMMM*.

Yiran Meng, Junhong Ye, Wei Zhou, Guanghui Yue, Xudong Mao, Ruomei Wang, and Baoquan Zhao. 2025. Videoforest: Person-anchored hierarchical reasoning for cross-video question answering. In *ACMMM*.

Tushar Nagarajan and Lorenzo Torresani. 2024. Step differences in instructional video. In *CVPR*.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *NAACL-HLT (Findings)*.

Marie-Therese Puth, Markus Neuhäuser, and Graeme D Ruxton. 2015. Effective use of spearman's and kendall's correlation coefficients for association between two measured traits. *Animal Behaviour*.

Yue Qiu, Yanjun Sun, Takuma Yagi, Shusaku Egami, Natsuki Miyata, Ken Fukuda, Kensho Hara, and Ryusuke Sagawa. 2025. Videosetdiff: Identifying and reasoning similarities and differences in similar videos. In *ICCV*.

Kate Sanders, Nathaniel Weir, and Benjamin Van Durme. 2024. TV-TREES: multimodal entailment trees for neuro-symbolic video reasoning. In *EMNLP*.

Darshana Saravanan, Varun Gupta, Darshan Singh S, Zeeshan Khan, Vineet Gandhi, and Makarand Tapaswi. 2025. VELOCITI: benchmarking video-language compositional reasoning with strict entailment. In *CVPR*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *ICLR*.

Yue Wu, Zhaobo Qi, Yiling Wu, Junshu Sun, Yaowei Wang, and Shuhui Wang. 2025. Learning fine-grained representations through textual token disentanglement in composed video retrieval. In *ICLR*.

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

Luca Zanella, Massimiliano Mancini, Willi Menapace, Sergey Tulyakov, Yiming Wang, and Elisa Ricci. 2025. Can text-to-video generation help video-language alignment? In *CVPR*.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2025. Llava-video: Video instruction tuning with synthetic data. *TMLR*.

Lingjun Zhao, Mingyang Xie, Paola Cascante-Bonilla, Hal Daumé III, and Kwonjoon Lee. 2025. Can hallucination correction improve video-language alignment? In *ACL (Findings)*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *ICLR*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025a. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *CoRR*.

Nannan Zhu, Yonghao Dong, Teng Wang, Xueqian Li, Shengjun Deng, Yijia Wang, Zheng Hong, Tiantian Geng, Guo Niu, Hanyan Huang, Xiongfei Yao, and Shuaiwei Jiao. 2025b. Cvbench: Evaluating cross-video synergies for complex multimodal understanding and reasoning. *CoRR*.

A StepDiff Annotation

The original StepDiff (Nagarajan and Torresani, 2024) dataset features video pairs annotated with captions that describe differences across five main categories, primarily focusing on textual descriptions of these differences. It also includes similarity scores for some categories. However, the annotations are only partially complete: most video pairs have captions for fewer than five categories, and many similarity scores are either missing or reported as zero, which is outside the valid range. To enable more consistent and comprehensive evaluation for the ConViS task, we annotated all video pairs using a clear scoring rubric. Each pair received similarity scores from 1 to 5 in all categories, ensuring complete and uniform coverage.

Human annotation was performed via Prolific (Douglas et al., 2023) and annotators were compensated at a rate of £8.50 per hour. Since the videos in StepDiff are originally identified via YouTube IDs, some videos have been taken down or made private. Consequently, our reannotated dataset consists of the subset of videos that remain publicly available. In order to have video pairs with sufficient temporal structure, we kept those with videos longer than 10 seconds. The corresponding video IDs for this subset will be disclosed to facilitate reproducibility.

B Implementation Details

All models are evaluated using 16 frames per video. The number of reasoning traces per branch is set to $M = 10$. We use LLaMA3-8B-Instruct (Grattafiori et al., 2024) as f_{LLM} to generate explanations from reasoning traces in Sec. 3.3. When sampling multiple reasoning traces, we employ stochastic decoding with a temperature of 0.7 and nucleus sampling (top-p=0.9). For all other generations, we default to greedy decoding. All experiments were conducted on a single NVIDIA A100 GPU.

C User Study

We conducted a user study to assess the quality of the explanations generated by VICE (section 3.3). We randomly selected a total of 30 video pairs, drawn in equal proportion from ConViS-Bench (Liberatori et al., 2025) and StepDiff (Nagarajan and Torresani, 2024). For each video pair, we gathered explanations from three models (Zhang et al., 2025; Li et al., 2025; Bai et al., 2025b), generated as defined in section 3.3. We recruited five independent human raters, who were asked to rank the three explanations for each video pair from best to worst based on how well they described information relevant to the video pair content. The three explanations for a given video pair were shown in randomized order to mitigate position bias. Raters were instructed to focus solely on relevance and clarity, without considering model identity or stylistic preferences. The interface used for the study is shown in Fig. 6.

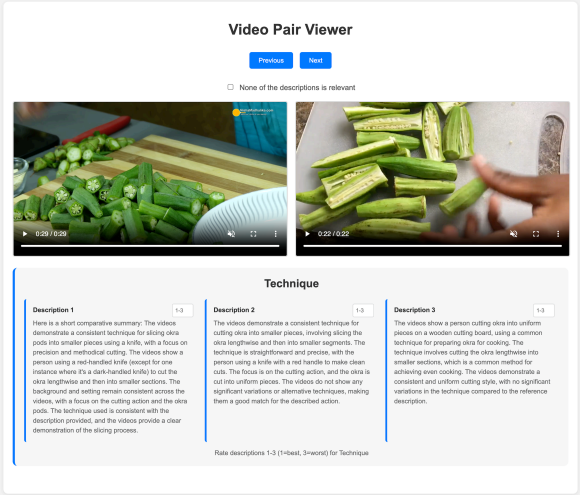

Figure 6: **User Study Interface:** Annotators are asked to rank the descriptions or select the checkbox if none of them apply.

Model	1st(%)	2nd(%)	3rd(%)
Qwen3-VL-8B (Bai et al., 2025a)	38.34	24.87	36.79
LLaVA-Video-7B (Zhang et al., 2025)	31.61	42.49	25.90
LLaVA-OV-7B (Li et al., 2025)	29.53	33.16	37.31

Table 4: **User Study Results:** We report the percentage of times each model's explanation generated by VICE with different LMMs (see Sec. 3.3) was ranked 1st, 2nd, or 3rd.

We find that in 97% of cases, at least one explanation was considered relevant by the raters, highlighting the robustness of VICE across different underlying models. Tab. 4 reports how often each model's explanation was ranked 1st, 2nd, or 3rd among the provided options, showing that models are broadly comparable, with Qwen3-VL-8B marginally outperforming the others, a trend consistent with the quantitative results in Tab. 1 and Tab. 2.

D Runtime and Memory Analysis

VICE requires more forward passes than a direct video-pair comparison. However, most of these

Figure 7: **Qualitative results of ViCE:** We present three example video pairs for two distinct concepts and datasets (Liberatori et al., 2025; Nagarajan and Torresani, 2024), along with the captions and reasoning traces generated when conditioning on those concepts. Compared to the *zero-shot* baseline, ViCE produces scores that more closely align with the ground truth ones.

Method	Peak GPU Memory (GiB)	Time / Prediction (s)
Baseline (zero-shot)	17.5	2.5
VICE	22.4	5.8

Table 5: **Computational Cost:** Peak GPU memory and wall-clock time per prediction for the zero-shot baseline and VICE using Qwen3-VL-8B on a single NVIDIA A100 GPU.

forwards (the M sampled entailment traces per branch) can be batched rather than run sequentially, substantially limiting the added latency in practice.

To quantify this, we profile Qwen3-VL-8B on a single NVIDIA A100 GPU, measuring peak GPU memory and wall-clock time per prediction. Considering an 8B model, whose weights alone require 16.97 GiB (a fixed cost common to both methods), we report the average peak memory usage and inference time in Tab. 5. Despite the additional steps, VICE incurs only a moderate computational overhead.

E Qualitative Results

Fig. 7 presents additional qualitative results complementing the example shown in Fig. 5. This appendix is further accompanied by qualitative results in video format, accessible by opening `index.html` in the provided zipped folder.

F Licenses

The datasets (Liberatori et al., 2025; Nagarajan and Torresani, 2024) used in our experiments are released under Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license. The LMMs used in our experiments are all publicly available, released under the Apache 2.0 license.

G Prompts

Our proposed VICE consists of two main steps, each using a dedicated prompt.

In the first concept-conditioned captioning step in Eq. (1), we prompt the LMM to generate descriptions of the input video that are explicitly focused on a given concept. In doing so, we provide guiding examples for each concept.

In the second video-caption entailment step in Eq. (2), we prompt the LMM to assess the semantic alignment between a video and a textual description with respect to a given concept. We guide this estimation using a rubric, identical to the one given to human raters.

Concept-conditioned captioning prompt

You are given a video. Generate a caption that focuses specifically on the **{concept}** of the video. Describe ONLY the **{concept}** and DO NOT include details about other aspects unless they are directly relevant to the **{concept}**.
Keep the caption natural, concise, and directly related to the **{concept}**.

Examples:
{examples[concept]}

Here is the video. Generate a caption that describes its **{concept}** clearly and naturally.

Video-caption entailment prompt

You are given a text description and a video.
Your task is to evaluate how well the video matches the description with respect to the **{concept}**.

Use the following rubric to determine your score:

{rubric[concept]}

Follow the format strictly:
- Inside **<thinking> </thinking>**, explain how the **{concept}** appears in the video and how it relates to the provided description, referencing the rubric levels where relevant.
- Then, output a single similarity score from 1 (least similar) to 5 (most similar) inside **<score> </score>**, reflecting how closely the video aligns with the description regarding the **{concept}**.

Format example:

<thinking>
Analysis of how the video matches the described **{concept}**.
</thinking>
<score> Predicted score **</score>**

Here is a video and a text description. Evaluate how well the video matches the description based on its **{concept}**, following the specified format with **<thinking>** and **<score>**.

Examples

Main Subjects:
- A young boy with curly hair.
- An elderly man with glasses and a beard.

Main Objects:
- A red bicycle leaning against a wall.
- A stack of books on a wooden table.

Location:
- A bustling city street lined with shops.
- A quiet park with tall trees.

Main Action:
- People are tossing a ball back and forth.
- A cat jumps onto the counter.

Order Of Actions:
- The chef chops vegetables, then sautés them in a pan.
- She opens the door, walks outside, and waves goodbye.

Tools/Equipment:
- A sharp chef's knife beside a wooden cutting board.
- A stainless steel saucepan simmering on the stove.

Actions:
- Whisking the batter until smooth.
- Searing the steak on a hot pan.

Visuals:
- A dark brown sauce coating the meat.
- Crispy golden-brown eggs on the plate.

Ingredients:
- Fresh tomatoes, basil, and mozzarella on the counter.
- Flour, sugar, and eggs arranged in small bowls.

Technique:
- Using gentle folding motions to mix the batter.
- Spreading the sauce evenly across the plate with the back of a spoon.

Rubric ConViS-Bench

Main Action:
Score | Description
1 | Completely different types of action (e.g., cooking and sports domains)
2 | Actions are different but related (e.g., both sports)
3 | Actions are similar with significant differences in how they are carried out
4 | Actions are similar with slight differences in how they are carried out
5 | No difference in the main action performed

Main Subjects:
Score | Description
1 | Totally different cardinality, gender, and age (e.g., a man and a group of women)
2 | Different cardinality but some similar in gender or age (e.g., a man and a group of men)
3 | Same cardinality but significant difference in age or gender (e.g., a man and a boy)
4 | Same cardinality but minor difference in age (e.g., a young male and a boy)
5 | Extremely similar (e.g., two young women in both)

Location:
Score | Description
1 | Completely different places
2 | Different but similar type of place (e.g., indoor and outdoor soccer field)
3 | Similar locations still with significant differences
4 | Similar locations with slight differences
5 | Extremely similar (e.g., rooms with a similar size and layout)

Order Of Actions:
Score | Description
1 | Completely different, or no sub-actions are in common
2 | The order of actions has significant changes, and some missing steps
3 | Some actions are reordered, with a few steps missing, but key steps remain intact
4 | Minor changes in the order of actions
5 | Actions occur in the same order, without missing steps

Main Objects:
Score | Description
1 | Main objects are entirely different
2 | A few main objects are shared, but most are different
3 | Most main objects are similar, with some not in common
4 | Minor differences in main objects (e.g., color or shape)
5 | Main objects are the same or extremely similar

Rubric StepDiff

Ingredients:
Score | Description
1 | Completely different main ingredients (e.g., pork vs fish)
2 | Major ingredients differ, but some elements remain the same
3 | Several minor ingredients are added or skipped (3 or more instances)
4 | Only one or two minor ingredients differ (added, skipped, or substituted)
5 | No difference in ingredients used

Tools/Equipment:
Score | Description
1 | Completely different tools or appliances are used
2 | Multiple substitutions or material changes (4 or more) that serve the same purpose
3 | A few tools (2–3) are substituted or differ in material but serve the same function
4 | Minor differences such as material (e.g., metal vs wood), or type within the same category (e.g., non-stick vs ordinary pan)
5 | No difference in tools or equipment used except for minor aspects like color or size

Actions:
Score | Description
1 | Sequence and content of steps are entirely different
2 | Sequence differs with two to three steps added or skipped
3 | Some steps are added, skipped, or slightly changed but the main process remains
4 | Order of steps is slightly altered but all actions are present
5 | All steps occur in the same sequence with no differences

Technique:
Score | Description
1 | Actions are different and techniques are not comparable
2 | Techniques are different (e.g., boiling vs steaming)
3 | Two to three noticeable variations in cooking or preparation techniques
4 | Slight procedural variation (e.g., slicing vs dicing) with similar outcome
5 | No difference in techniques used

Visuals:
Score | Description
1 | Completely different appearance, with major differences in color or consistency (e.g., brown vs white)
2 | Significant difference in size or portioning of dishes
3 | Minor visual differences such as slightly darker or crisper elements
4 | Very small variation in color or texture affecting only a small portion or detail
5 | No visible difference in final appearance