MODELING RETRIEVAL INTENT FROM VIDEO-TEXT CONTEXT FOR COMPOSED VIDEO RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Composed Video Retrieval (CVR) retrieves a target video by composing a reference video with a text modification describing desired changes. Effective CVR requires resolving the user's retrieval intent from joint video-text context: what to modify and what to preserve. However, existing methods merely perform task-specific refinement within each modality, underexploring how video-text interaction determines the user's retrieval intent. This leads to two critical weaknesses: 1) **imprecise modification**, where changes are not confined to relevant visual regions, and 2) **indiscriminate preservation**, where insignificant video content is fully retained. To address this, we propose **IC-CoVR**, an intent-centric framework that resolves the retrieval intent within the joint video-text context. Specifically, we characterize retrieval intent along two complementary dimensions: explicit intent for targeted modification and implicit intent for selective preservation. Based on this formulation, IC-CoVR first constructs structured video representations to support video-text interaction and intent reasoning, then uses explicit intent to locate and guide targeted modification, and finally leverages implicit intent to selectively preserve unmodified content by suppressing irrelevant details. Extensive experiments on FineCVR-1M and EgoCVR show that IC-CoVR achieves state-of-the-art performance and strong generalization, with a substantial gain of +9.22% R@1 on FineCVR-1M. Comprehensive ablation studies further validate the effectiveness of our intent-centric formulation.

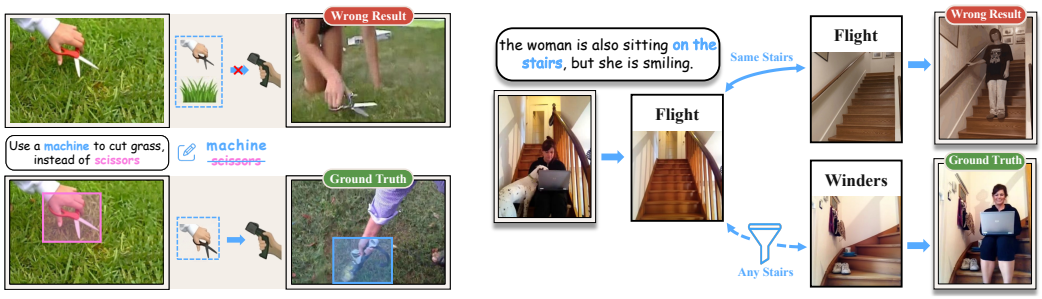
1 INTRODUCTION

As video content continues to grow at scale, effective video retrieval becomes increasingly important. Beyond conventional text-video retrieval (Wang et al., 2021; Gorti et al., 2022), many practical search scenarios are reference-driven: users begin with an existing example that is close to the desired target and then describe the additional conditions it should satisfy. Composed Video Retrieval (CVR) was proposed (Ventura et al., 2024a) for this setting, where the query is formulated as a reference video paired with a textual modification. Its goal is to retrieve a target video that satisfies the specified change while remaining consistent with the relevant content in the reference.

While CVR provides a flexible query formulation, it also introduces a distinctive characteristic: neither the modification text nor the reference video can independently define the target semantics. The reference video contains more information than needed, while the modification text describes only a partial transformation. As a result, the central difficulty of CVR lies not merely in fusing a reference video and a modification text, but in determining retrieval relevance under video-text context. Therefore, the model must reason over the joint input, determining what to modify and what to preserve for retrieval. We refer to this relevance structure as **retrieval intent**, a relevance criterion induced by the joint context of the reference video and the modification text.

Existing CVR methods mainly advance in two directions. One line of work tends to explore the modification text (Thawakar et al., 2025; Wu et al., 2025) to express the desired change more precisely. The other line focuses on video modeling (Ventura et al., 2024b; Hummel et al., 2024; Thawakar et al., 2024), leveraging advanced video encoders or captions to incorporate richer information. Despite their differences, these methods mainly perform refinement within each modality but lack an explicit mechanism for determining retrieval relevance under a composed query. Specifically, two major issues persist. **1)** Modification is imprecisely grounded in the visual context. As illustrated

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Imprecise Modification (b) Indiscriminate Preservation

Figure 1: Limitations. *Left*: Applying the replace command holistically to the video fails to change the scissors to a machine. *Right*: Indiscriminately preserving all background (straight wooden stairs) details results in a wrong video with a similar background.

in Figure 1a, without fine-grained interaction, modification is applied to the entire video, failing to change the scissors to a machine. **2)** Preservation is indiscriminately applied to the reference content. As shown in Figure 1b, current methods keep all unmodified content, even if some is irrelevant to the target: faithfully preserving straight stairs from the reference conflicts with the target's winders, making retrieval sensitive to irrelevant differences.

To operationalize this retrieval intent, we decompose it into two complementary forms: explicit intent for modification and implicit intent for preservation. **Explicit intent** captures the retrieval-critical modification specified by the query. Derived from direct video-text interaction, it determines where the modification should be grounded and what new semantics should be introduced. **Implicit intent** governs the content not specified for modification. Inferred from the joint context, it regulates which unmodified content should remain relevant for retrieval and at what granularity it should be preserved, suppressing details irrelevant to the retrieval objective.

Based on this formulation, we propose **IC-CoVR**, an intent-centric framework with a three-stage architecture that progressively models intent for precise modification and selective preservation. **Firstly**, in the Representation Structuring Stage (RSS), we construct a Component-Structured Profile for each video, as a component-wise, multi-granularity representation for subsequent intent reasoning. **Secondly**, in the Explicit-intent Modification Stage (EMS), a Bidirectional Intent Parser locates the target region in the video and defines operations by parsing the explicit intent from the joint context. Then an Enhanced-Intent Modifier performs the operations on the corresponding regions. **Thirdly**, in the Implicit-intent Preservation Stage (IPS), an Intent-guided Granularity Alignment Module captures implicit intent through video-text context at different granularities, and accordingly adjusts the weight of each granularity to suppress irrelevant details, achieving selective preservation. In this way, IC-CoVR turns intent into an operational principle that governs modification and preservation.

Overall, our main contributions are as follows.

- We formulate CVR around retrieval intent, decomposing joint video-text relevance into explicit intent for targeted modification and implicit intent for selective preservation.
- We propose IC-CoVR, an intent-centric framework that addresses both aspects in a unified architecture through a three-stage pipeline of representation structuring, explicit-intent modification, and implicit-intent preservation.
- We conduct extensive experiments on FineCVR-1M and EgoCVR. IC-CoVR achieves state-of-the-art performance with an R@1 improvement of +9.22% on FineCVR-1M. Comprehensive ablations verify the effectiveness of each intent-modeling component.

2 Related Work

2.1 Composed Image Retrieval

CIR aims to retrieve a target image given a reference image and a modification text (Vo et al., 2019). Early CIR methods typically project visual and textual inputs into a shared, joint embedding

space (Baldrati et al., 2022; Chen et al., 2020; Vo et al., 2019; Lee et al., 2021; Anwaar et al., 2021), forming a compositional representation for target image matching. More recently, with the emergence of large vision-language models (Radford et al., 2021), several studies have explored zero-shot CIR (Saito et al., 2023; Baldrati et al., 2023; Gu et al., 2024; Suo et al., 2024; Tang et al., 2024), which removes the reliance on pre-defined triplet training data. Nevertheless, these methods are designed for static image retrieval and do not account for the temporal dynamics and action-level modifications introduced in CVR.

2.2 Composed Video Retrieval

CVR aims to retrieve a target video based on a reference video and modification text. Some methods (Ventura et al., 2024a;b; Wang et al., 2025; Chen et al., 2025) adapt CIR techniques by leveraging the image-grounded text encoder of BLIP (Li et al., 2022; 2023) to fuse the reference video with modification text. Other studies improve text representation by generating richer descriptions of reference videos (Thawakar et al., 2024) and denser modification texts (Thawakar et al., 2025). Some approaches (Gupta et al., 2025) model temporal dynamics for better video understanding. The FDCA framework (Wu et al., 2025) disentangles text at sequence and token levels to obtain fine-grained features for alignment. In contrast, EgoCVR (Hummel et al., 2024) proposes a training-free reranking method using cross-modal models. Supporting these efforts, datasets such as WebVid-CoVR (Ventura et al., 2024a) focus on static, object-centric content, facilitating CIR adaptation, while FineCVR-1M (Wu et al., 2025), EgoCVR (Hummel et al., 2024), and TF-CoVR (Gupta et al., 2025) include richer human actions and temporal patterns requiring deeper video understanding. However, existing methods rely on explicit textual semantics, overlooking the user's implicit intent.

2.3 Modeling User Intents

Modeling user intents focuses on how user goals can be accurately represented and incorporated into generation or retrieval systems, which has emerged as a central design challenge. Some studies (Ci et al., 2025; Mondal et al., 2024; Shukla et al., 2019) investigate intent modeling by shifting from instruction-based commands to description-based goal specification, which provides richer semantic information and better aligns with the priors of pre-trained text-to-image models. Other works (Cheng et al., 2024; Chen et al., 2022) address the "intent gap" in visual generation through knowledge-enhanced iterative refinement, progressively aligning system outputs with user intents by integrating diverse knowledge sources across multiple rounds of feedback.

3 Method

3.1 Method Overview

Task Formulation. The core task of a CVR model is to learn a function f that fuses the reference video v_{r} and modification text t_{m} into a composed query feature $\boldsymbol{q}_{\mathrm{c}} = f(v_{\mathrm{r}}, t_{\mathrm{m}})$ to retrieve the target video v_{t}. In this work, we instead view this composition process as governed by retrieval intent. The retrieval intent is modeled from the video-text context: $\boldsymbol{I} = M(v_{\mathrm{r}}, t_{\mathrm{m}})$. The composed query is then constructed by applying this intent to the reference video: $\boldsymbol{q}_{\mathrm{c}} = g(\boldsymbol{I}, v_{\mathrm{r}})$.

Overall Architecture. As shown in Figure 2, IC-CoVR involves three stages: Representation Structuring Stage (RSS), Explicit-intent Modification Stage (EMS) and Implicit-intent Preservation Stage (IPS). In RSS, we adopt an MLLM to generate a Component-Structured Profile (CSP) for each video, and leverage a frozen encoder to construct representations. During EMS, the Bidirectional Intent Parser (BIP) disentangles the textual and visual representations, thereby locating the region to be modified and extracting operation intents. Later, in the Enhanced-Intent Modifier, we incorporate the modifications into the target regions and then combine the modified parts with the unmodified parts, finally constructing the modified visual and CSP representations. At the beginning of IPS, we retrieve an initial candidate set based on the visual features, whereupon we leverage the Intent-guided Granularity Alignment Module (IGAM) to synchronously align the granularities between the query CSP and the candidate CSPs. Ultimately, we determine the final retrieval results by computing the similarities between the respective CSP representations.

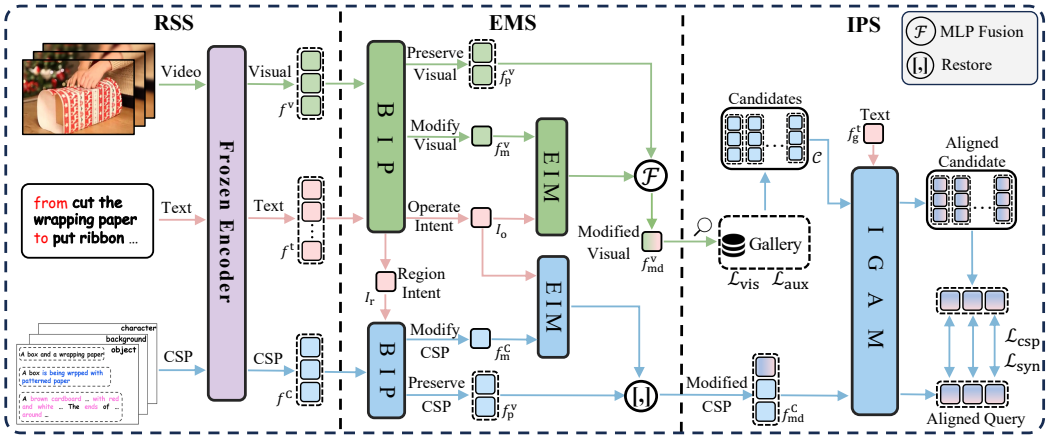


Figure 2: The pipeline of the proposed IC-CoVR framework. (1) RSS: Generates CSPs and extracts initial multimodal features. (2) EMS: Implements targeted modification according to explicit intent. (3) IPS: Preserves information in modified features selectively based on implicit intent.

3.2 Representation Structuring Stage

RSS organizes and encodes video content to lay the foundation for fine-grained video-text interaction and selective preservation.

As illustrated in Figure 3a, the CSP is designed as a component-wise, multi-granularity representation. On the one hand, it decomposes video content into distinct components (character, object and background) for fine-grained video-text interaction. On the other hand, it organizes each component in a three-level coarse-to-fine hierarchy to support selective preservation. We employ an MLLM (Bai et al., 2025) to generate a CSP for each video, which is represented as $\{s_{i,j} \mid i, j \in \{1, 2, 3\}\}$, where i indexes the component and j indexes the granularity level.

To construct the representations, we adopt frozen CLIP (Radford et al., 2021) as the backbone. Both CSP descriptions and modification text are encoded by the text encoder, yielding CSP features $\boldsymbol{f}^{\mathrm{C}} \in \mathbb{R}^{3\times3\times d}$ and textual sequence features $\boldsymbol{f}^{\mathrm{t}} \in \mathbb{R}^{L\times d}$. For videos, to enable interaction between visual and textual features, we apply a parameter-free cross-frame aggregation on patch-level features. Specifically, we group and aggregate patch tokens by measuring their similarity to CSP components, producing visual features $\boldsymbol{f}^{\mathrm{v}} \in \mathbb{R}^{3\times d}$. Additionally, a global visual feature $\boldsymbol{f}_{\mathrm{g}}^{\mathrm{v}}$ is obtained via mean pooling. Refer to Appendix C for more details.

3.3 Explicit-intent Modification Stage

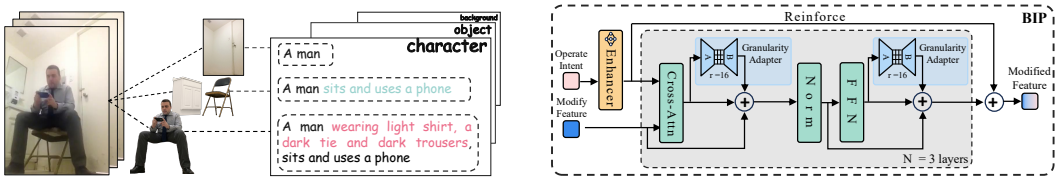
In this stage, the model aims to apply the editing operation to specific video regions, conditioned on the explicit intent.

Bidirectional Intent Parser. BIP captures the explicit intent to build correspondence between text and video. It finally yields a precise modification region and operation for downstream editing. The texts usually expresses two explicit decisions: the target region and the modification operation. Typically, target regions relate to content already present in the video, whereas the operation introduces novel concepts, which are usually different from the original video. Motivated by this distinction, we adopt the disentanglement method of Wu et al. (2025) to obtain two groups of disentangled text features, $\mathbb{T}_1$ and $\mathbb{T}_2$. We assign their intent roles according to their similarity to the global visual feature $\boldsymbol{f}_{\mathrm{g}}^{\mathrm{v}}$. The group more similar to the reference video represents the region intent, while the other represents the operation intent. The process could be formulated as:

$$(\boldsymbol{I}_{\mathrm{r}}, \boldsymbol{I}_{\mathrm{o}}) = \mathrm{Sort}_{\downarrow \mathrm{sim}(\cdot, \boldsymbol{f}_{\mathrm{g}}^{\mathrm{v}})} \left(\mathrm{Pool}(\mathbb{T}_1), \mathrm{Pool}(\mathbb{T}_2)\right). \tag{1}$$

Here, $\mathrm{Pool}(\cdot)$ aggregates the features within each group.

Using region-specific features, we partition the video into regions to be modified and regions to be preserved. Specifically, we select the top-k components that are most similar to $\boldsymbol{I}_{\mathrm{r}}$ as the target for



(a) CSP Illustration (b) Enhanced-Intent Modifier

Figure 3: *Left*: CSP incorporates multi-granularity descriptions for video components. Taking character as an example, the description is progressively refined: basic identification, followed by action descriptions, and finally, detailed attributes. *Right*: The details of EIM implementation.

modification: $\boldsymbol{f}_{\mathrm{m}}^{\mathrm{v}} \in \mathbb{R}^{k\times d}$ from visual features and $\boldsymbol{f}_{\mathrm{m}}^{\mathrm{C}} \in \mathbb{R}^{k\times 3\times d}$ from CSP. The remaining visual and CSP features are defined as preserved features, $\boldsymbol{f}_{\mathrm{p}}^{\mathrm{v}}$ and $\boldsymbol{f}_{\mathrm{p}}^{\mathrm{C}}$.

Enhanced-Intent Modifier. The operation intent $\boldsymbol{I}_{\mathrm{o}}$ determines the final modification effect and acts as the dominant editing signal. We propose the Enhanced-Intent Modifier (EIM) to strengthen its role. As shown in Figure 3b, we first apply a non-linear transformation (Vaswani et al., 2017) to obtain $\tilde{\boldsymbol{I}}_{\mathrm{o}}$, enhancing its expressivity before interacting with the video. A multi-head cross-attention (MHCA) network (Vaswani et al., 2017) is then used for intent-guided modification. To address heterogeneity across granularities, we introduce granularity adapters (Hu et al., 2022) on the CSP branch. Later, $\tilde{\boldsymbol{I}}_{\mathrm{o}}$ is injected into the fused features via residual connections to reinforce consistency with the intent, controlled by α_{I}. This process can be formulated as:

$$\tilde{\boldsymbol{f}}_{\mathrm{m}}^{\mathrm{v}} = \mathrm{MHCA}_{\mathrm{v}}(\boldsymbol{f}_{\mathrm{m}}^{\mathrm{v}}, \tilde{\boldsymbol{I}}_{\mathrm{o}}, \tilde{\boldsymbol{I}}_{\mathrm{o}}) + \alpha_{\mathrm{I}} \cdot \tilde{\boldsymbol{I}}_{\mathrm{o}}, \quad \tilde{\boldsymbol{f}}_{\mathrm{m}}^{\mathrm{C}} = \mathrm{MHCA}_{\mathrm{C}}(\boldsymbol{f}_{\mathrm{m}}^{\mathrm{C}}, \tilde{\boldsymbol{I}}_{\mathrm{o}}, \tilde{\boldsymbol{I}}_{\mathrm{o}}) + \alpha_{\mathrm{I}} \cdot \tilde{\boldsymbol{I}}_{\mathrm{o}}. \tag{2}$$

Finally, we feed the modified visual features $\tilde{\boldsymbol{f}}_{\mathrm{m}}^{\mathrm{v}}$ and preserved visual features into an MLP for a final fusion, yielding modified visual features $\boldsymbol{f}_{\mathrm{md}}^{\mathrm{v}} \in \mathbb{R}^{d}$. Meanwhile, the modified CSP features are restored into their original positions, forming locally modified CSP features $\boldsymbol{f}_{\mathrm{md}}^{\mathrm{C}} \in \mathbb{R}^{3\times 3\times d}$.

3.4 Implicit-intent Preservation Stage

In this stage, the model adjusts the granularity structure according to implicit intent, enabling selective preservation while suppressing irrelevant details.

Candidate Set. The candidate set serves as the interface between explicit modification and implicit preservation: the visual feature provides modification-aware recall, while the CSP features of candidates offer multi-granularity structures for subsequent preservation reasoning. Specifically, we use the modified visual feature $\boldsymbol{f}_{\mathrm{md}}^{\mathrm{v}}$ to retrieve a candidate set of K videos for each query, denoted as $\mathbb{C} = \{\boldsymbol{f}^{\mathrm{C}}\}$. During training, if the ground-truth video is absent from $\mathbb{C}$, we replace the lowest-ranked candidate with it at probability p, which decays to maintain training stability.

Intent-guided Granularity Alignment Module. During preservation, implicit intent serves as a granularity preference, which IGAM infers from the video-text context for both the query and candidate CSPs. For each CSP component, the global text feature $\boldsymbol{f}_{\mathrm{g}}^{\mathrm{t}}$ interacts with its representations through a cross-attention network, producing adaptive weights over the three granularity levels. These weights determine what detail should be preserved for retrieval, emphasizing intent-relevant semantics while suppressing incidental details. The weighted features are then fused within each component, forming the final CSP representation $\tilde{\boldsymbol{f}}^{\mathrm{C}} \in \mathbb{R}^{3\times d}$. The process is formulated as:

$$\tilde{\boldsymbol{f}}^{\mathrm{C}} = \mathrm{cat}(\{\mathrm{MHCA}_i(\boldsymbol{f}_{\mathrm{g}}^{\mathrm{t}}, \boldsymbol{f}_i^{\mathrm{C}}) \mid i \in \{1,2,3\}\}). \tag{3}$$

Pairwise Comparison. To enhance component-wise correspondence between the query and candidate, we propose a pairwise comparison method. Besides holistic features (fused across components), we compute similarity between corresponding CSP components and combine both scores to obtain the final CSP similarity as:

$$\mathrm{sim}_{\mathrm{C}} = \beta \cdot \mathrm{sim}(\tilde{\boldsymbol{f}}^{\mathrm{C_q}}, \tilde{\boldsymbol{f}}^{\mathrm{C_t}}) + (1-\beta) \cdot \sum_{i=1}^{3} \mathrm{sim}(\tilde{\boldsymbol{f}}_i^{\mathrm{C_q}}, \tilde{\boldsymbol{f}}_i^{\mathrm{C_t}}). \tag{4}$$

The final similarity score for a video pair is a weighted combination of the visual similarity and the CSP similarity: $\text{sim} = \text{sim}_\text{v} + \gamma \cdot \text{sim}_\text{C}$.

3.5 Training Objective

IC-CoVR is trained with two objectives: explicit-intent modification and implicit-intent preservation. Accordingly, the training objective includes visual losses for precise modification, a CSP loss for selective preservation, and a joint loss for their synergy in retrieval. Key settings include batch size B, candidate set $\mathbb{C}_i$ with $|\mathbb{C}_i| = K$, scale τ, cosine similarities to ground-truth visual and CSP features ($\boldsymbol{S}^\text{v}$ and $\boldsymbol{S}^\text{C}$), and CSP similarity weight γ.

Visual Losses. The visual losses supervise the EMS by aligning the modified visual representation with the intended semantics. Specifically, $\mathcal{L}_\text{vis}$ applies contrastive learning between the modified and target visual features, promoting precise modification. As the Bidirectional Intent Parser relies on preliminary textual decomposition, we retain the auxiliary loss $\mathcal{L}_\text{aux}$ from Wu et al. (2025).

CSP Loss. To supervise CSP-based matching, we introduce a candidate-level cross-entropy loss $\mathcal{L}_\text{csp}$. By optimizing the similarity between the query CSP and its target against the candidate CSPs, this loss guides the model to preserve retrieval-relevant information. This objective is defined as:

$$\mathcal{L}_\text{csp} = -\frac{1}{B}\sum_{i=1}^{B} \mathbf{1}_{i\in\mathbb{C}_i} \cdot \log \frac{\exp(S^\text{C}_{i,i}/\tau)}{\sum_{j\in\mathbb{C}_i}\exp(S^\text{C}_{i,j}/\tau)}. \tag{5}$$

Synergy Loss. To couple explicit modification with selective preservation in final retrieval, we introduce an InfoNCE-style synergy loss $\mathcal{L}_\text{syn}$. It jointly optimizes visual and CSP similarities, encouraging consistent target retrieval. This objective is defined as:

$$\mathcal{L}_\text{syn} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp((S^\text{v}_{i,i} + \mathbf{1}_{i\in\mathbb{C}_i}\cdot\gamma\cdot S^\text{C}_{i,i})/\tau)}{\sum_{j\notin\mathbb{C}_i}\exp(S^\text{v}_{i,j}/\tau) + \sum_{j\in\mathbb{C}_i}\exp((S^\text{v}_{i,j} + \gamma\cdot S^\text{C}_{i,j})/\tau)}. \tag{6}$$

Overall Loss. The model is trained in an end-to-end manner where the overall loss function is:

$$\mathcal{L} = \mathcal{L}_\text{vis} + \mathcal{L}_\text{aux} + \mathcal{L}_\text{csp} + \mathcal{L}_\text{syn}. \tag{7}$$

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our approach on two benchmarks with distinct modification characteristics. For action-centric scenarios with strong temporal dynamics, we use EgoCVR (Hummel et al., 2024) (a test-only set of 2,295 triplets). To assess fine-grained retrieval, we employ FineCVR-1M (Wu et al., 2025) (1,000,028 training and 10,043 test triplets), where roughly half involve static modifications and the rest temporal changes. This setup enables robust evaluation in complex, high-precision video retrieval scenarios. Results on additional datasets are provided in Appendix A.

Implementation Details. In line with previous work (Thawakar et al., 2025), we adopt the standard Recall@K metric to evaluate the retrieval performance. In our framework, we adopt Qwen3-VL-8B (Bai et al., 2025) to generate CSPs. Following prior work (Ventura et al., 2024a; Wu et al., 2025), we evaluate IC-CoVR with two frozen vision encoders, CLIP (Radford et al., 2021) and BLIP (Li et al., 2022), while keeping all other hyperparameters identical. The model is trained on a single NVIDIA RTX 3090 GPU with a batch size of 1024. During training, the candidate set size is 10 and the top-k for location is 1. More details are provided in Appendix A.

4.2 Main Results

Performance on FineCVR-1M. As shown in Table 1, IC-CoVR achieves state-of-the-art performance on FineCVR-1M, with substantial gains in R1 and R5. Given that FineCVR-1M requires

Table 1: Performance on FineCVR-1M compared with baseline and existing methods.

Method	Backbone	R1	R5	R10	R50
CoVR (Ventura et al., 2024a)	BLIP	17.05	41.57	56.60	85.56
ECDE-CoVR (Thawakar et al., 2024)	BLIP	22.81	47.63	60.12	90.32
BSE-CoVR (Thawakar et al., 2025)	BLIP	24.40	51.01	63.42	91.21
HUD (Chen et al., 2025)	BLIP-2	<u>26.85</u>	55.92	70.28	94.28
$\text{FDCA}_{\text{BLIP}}$ (Wu et al., 2025)	BLIP	26.79	<u>63.21</u>	**78.65**	**97.25**
$\text{IC-CoVR}_{\text{BLIP}}$	BLIP	**30.48**	**64.33**	<u>77.17</u>	<u>96.43</u>
Pic2Word (Saito et al., 2023)	CLIP	8.10	25.79	38.32	71.01
Combiner (Baldrati et al., 2022)	CLIP	19.57	49.46	65.38	92.11
FreestyleRet (Li et al., 2024)	CLIP	20.39	52.98	68.37	93.03
TFR-CVR (Hummel et al., 2024)	CLIP	15.21	40.12	52.78	81.75
$\text{FDCA}_{\text{CLIP}}$ (Wu et al., 2025)	CLIP	<u>25.84</u>	<u>55.84</u>	<u>70.23</u>	<u>94.33</u>
$\text{IC-CoVR}_{\text{CLIP}}$	CLIP	**35.06**	**59.12**	**72.04**	**94.58**
Δ (vs. SOTA)	CLIP	$\uparrow$ 9.22	$\uparrow$ 3.28	$\uparrow$ 1.81	$\uparrow$ 0.25

Table 2: Performance on EgoCVR compared with baseline and existing methods.

	Global			Local		
Method	R1	R5	R10	R1	R2	R3
$\text{Only Text}_{\text{CLIP}}$	0.7	1.7	2.7	33.5	48.8	61.8
$\text{Only Visual}_{\text{CLIP}}$	7.4	33.2	<u>55.3</u>	26.1	43.4	57.7
CoVR (Ventura et al., 2024a)	5.4	15.2	24.3	33.1	49.5	62.9
ECDE-CoVR (Thawakar et al., 2024)	6.0	16.3	24.8	33.4	49.3	63.0
EgoVLPv2 (Pramanick et al., 2023)	9.5	34.9	52.1	30.7	51.3	66.0
TFR-CVR (Hummel et al., 2024)	14.1	39.5	54.4	44.2	61.0	73.2
BSE-CoVR (Thawakar et al., 2025)	**14.6**	<u>41.3</u>	54.9	<u>44.8</u>	<u>61.7</u>	<u>74.0</u>
$\text{FDCA}_{\text{BLIP}}$ (Wu et al., 2025)	8.7	19.2	27.1	36.3	51.1	63.8
$\text{IC-CoVR}_{\text{CLIP}}$	<u>10.3</u>	**43.9**	**63.7**	**54.5**	**74.6**	**87.5**
Δ (vs. SOTA)	$\downarrow$ 4.3	$\uparrow$ 2.6	$\uparrow$ 8.4	$\uparrow$ 9.7	$\uparrow$ 12.9	$\uparrow$ 13.5

fine-grained discrimination and covers both static modifications and temporal changes, the results suggest that our intent-centric design is effective at handling query-conditioned changes without losing retrieval-relevant context. Moreover, compared with prior methods (Thawakar et al., 2025; 2024) that employ MLLMs for data augmentation, IC-CoVR maintains highly competitive performance.

Performance on EgoCVR. As shown in Table 2, IC-CoVR also shows strong performance in the zero-shot setting. Compared to the previous state-of-the-art BSE-CoVR (Thawakar et al., 2025) and the training-free method TFR-CVR (Hummel et al., 2024), IC-CoVR trained on FineCVR-1M has a lower R1 in global setting, while showing notable improvement at R5 and R10. In the "Local mode", i.e., retrieving a target clip from multiple candidates within a long video, our model demonstrates a stronger lead, indicating robust discrimination among highly similar candidates.

4.3 Ablation Studies

Ablation on Main Stages. We adopt FDCA (Wu et al., 2025) as the baseline, with only textual disentanglement. As shown in Table 3, adding EMS brings a clear improvement, verifying the benefit of modeling intent for targeted modification. Further introducing IPS boosts R1 by 3.48%, showing that intent-guided selective preservation provides complementary gains beyond modification.

Ablation in EMS. EMS incorporates two modules to achieve targeted modification: BIP and EIM. To evaluate their individual contributions, we conduct ablation studies on both modules. For **BIP**, we vary top-k to control the scope of the modified region. As shown in Table 4, enlarging the modified region leads to a clear performance drop. This indicates that effective modification depends on locating the relevant content according to explicit intents, rather than spreading the edit signal across broader visual regions. Regarding **EIM**, we vary α_{I} to control the strength of the operation

Table 3: Ablation studies on main stages.

Components		R1	R5	R10	R50
A.	Baseline	25.84	55.84	70.23	94.33
B.	A + EMS	31.58	59.42	72.04	94.58
C.	B + IPS	35.06	59.12	72.04	94.58

Table 4: Ablation studies on intent location.

Top-k	R1	R5	R10	R50
3	33.43	57.83	71.03	94.07
2	34.49	**59.17**	**72.13**	94.30
1	**35.06**	59.12	72.04	**94.58**

Table 5: Ablation studies on IPS granularity.

Module	Coarse	Middle	Fine	R1	R5
-	✓			30.34	57.50
-		✓		31.00	58.22
-			✓	32.64	58.65
IGAM	✓	✓		32.43	58.06
IGAM	✓		✓	34.46	58.23
IGAM		✓	✓	33.83	58.65
Mean	✓	✓	✓	30.80	58.44
IGAM	✓	✓	✓	**35.06**	59.12

Table 6: Ablation studies on each loss.

$\mathcal{L}_{\text{vis}}$	$\mathcal{L}_{\text{aux}}$	$\mathcal{L}_{\text{csp}}$	$\mathcal{L}_{\text{syn}}$	R1	R5	R10	R50
✓				28.18	57.77	71.47	94.48
✓		✓		33.05	58.48	71.10	94.40
✓			✓	31.00	56.61	71.59	94.49
	✓	✓	✓	30.94	56.40	68.62	93.53
✓		✓	✓	33.97	57.87	71.34	94.14
✓	✓		✓	31.94	57.31	71.93	**94.78**
✓	✓	✓		34.14	58.99	71.70	94.59
✓	✓	✓	✓	**35.06**	**59.12**	**72.04**	94.58

intent $\tilde{\boldsymbol{I}}_{\text{o}}$ in modification. As shown in Figure 4, increasing this guidance consistently improves performance up to a certain point. This verifies that targeted modification benefits from making the explicit operation intent a dominant signal, so that the resulting representation stays aligned with the intended change, while the gain saturates as α_{I} grows.

Ablation in IPS. The core mechanism of IPS is the intent-guided granularity alignment. We conduct an ablation study on granularity to isolate its specific contribution, demonstrating that the model's efficacy is not merely a byproduct of the MLLM's information augmentation. As shown in Table 5, relying solely on fine-grained information yields limited improvement, while performance improves when multiple granularities are adaptively combined. This result shows that IPS improves retrieval not by introducing more detailed descriptions, but by selectively preserving retrieval-relevant information while suppressing irrelevant details (35.06 vs. 30.80 R1 for mean fusion).

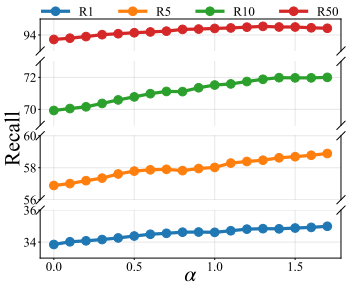


Figure 4: Results on different α_{I}.

Ablation on Loss Functions. Table 6 demonstrates that each loss contributes positively, improving all metrics. $\mathcal{L}_{\text{vis}}$ supervises the modified visual representation for modification-aware retrieval, while $\mathcal{L}_{\text{csp}}$ optimizes CSP representations for preservation-aware matching. Removing either loss weakens one side of the retrieval intent, resulting in a marked decrease. $\mathcal{L}_{\text{aux}}$ stabilizes textual disentanglement, contributing positively to intent parsing. $\mathcal{L}_{\text{syn}}$ aligns the two stages during training, enabling IPS to refine retrieval based on the modification-aware predictions from EMS. Overall, the loss design jointly supports targeted modification and selective preservation, with all four losses together giving the best R1 (35.06) and R5 (59.12).

4.4 In-depth Analysis

Synergy Analysis of the Two Stages. We analyze how EMS and IPS cooperate to model the two constraints in CVR: modification and preservation. The candidate set serves as the interface between the two stages. As shown in Table 7, the best performance is achieved at $K = 10$. Performance drops with a larger candidate set, suggesting that preservation reasoning is most effective when applied after the modification intent has narrowed the search space. We also find that the fusion weight γ performs best within a small range ($0.004 \sim 0.008$), indicating that preservation should complement rather than dominate modification-aware matching in ranking.

Table 7: Studies on candidate set size.

Size	R1	R5	R10	R50
5	35.0	**59.4**	72.0	94.6
10	**35.1**	59.1	**72.0**	**94.6**
15	35.0	58.2	72.0	94.6
20	35.0	58.1	71.2	94.6
25	35.0	58.0	70.9	94.6
30	35.0	58.0	70.6	94.6

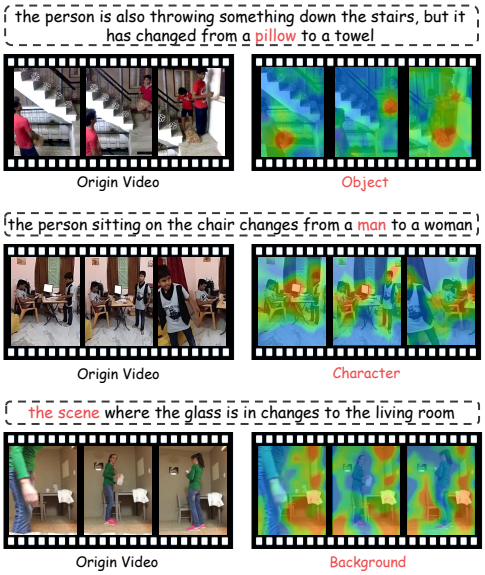
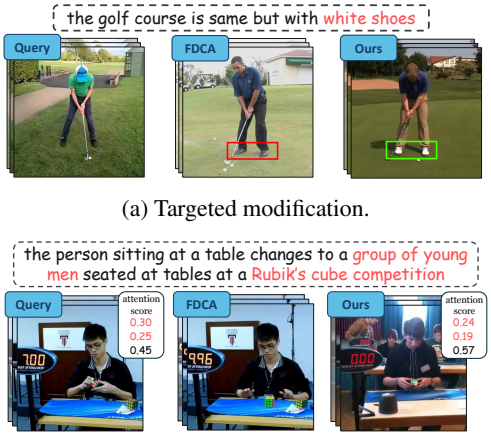


Figure 5: Visualization of the regions located by intent $\boldsymbol{I}_{\mathrm{r}}$.



(a) Targeted modification.

(b) Selective preservation.

Figure 6: IC-CoVR vs. FDCA on modification and preservation. In example (b), background attention scores are attached to the videos with granularity from coarse to fine in the top-to-bottom direction.

Visualization of Bidirectional Intent Parser. To further verify the effectiveness of BIP, we visualize heatmaps of the located target region intent $\boldsymbol{I}_{\mathrm{r}}$. As shown in Figure 5, the responses induced by $\boldsymbol{I}_{\mathrm{r}}$ consistently focus on query-relevant regions rather than diffusing across the video. This shows that BIP can accurately locate the target region, adapt its scope to different semantic types, and maintain stable responses across frames, providing a reliable basis for downstream modification.

Case Studies. In the Introduction, we identify two major limitations of existing methods: imprecise modification and indiscriminate preservation. To illustrate how our method addresses both issues, we present representative case studies on FineCVR-1M (Wu et al., 2025) in Figure 6, comparing IC-CoVR with FDCA (Wu et al., 2025). **Modification:** In Figure 6a, when faced with a fine-grained modification requirement, our method localizes the subtle change to the specific region and successfully retrieves the correct video. **Preservation:** In Figure 6b, the text requires changing the number of people, while the background also slightly differs. According to the context, retrieval only needs high-level scene semantics, e.g., an indoor Rubik's cube competition, rather than fine-grained background details. Accordingly, IC-CoVR assigns higher weights to coarse and middle background representations (red) and downweights the fine-grained one (black). This adaptive reweighting preserves relevant scene context while suppressing unmentioned visual differences, enabling IC-CoVR to retrieve the correct target.

5 Conclusion

In this paper, we revisit composed video retrieval from the perspective of query intent, arguing that effective retrieval requires reasoning over video-text context to determine what to modify and what to preserve. We propose an intent-centric model, IC-CoVR, which overcomes key limitations of existing methods. To address imprecise modification and indiscriminate preservation, we model retrieval intent along two dimensions: explicit intent for targeted modification and implicit intent for selective preservation. Explicit intent localizes the required change, while implicit intent suppresses retrieval-irrelevant details. As a result, IC-CoVR retrieves videos that satisfy desired modifications without being distracted by incidental differences. Experiments on FineCVR-1M and EgoCVR demonstrate strong effectiveness, achieving state-of-the-art performance on FineCVR-1M with a 9.22% improvement in R@1; ablations further validate each component. These results highlight intent modeling as a principled solution to CVR, and suggest its potential for related vision-language tasks such as controllable video generation and editing.

AI USE STATEMENT

In this work, we used generative AI tools to assist in implementing our method. We have not used generative AI tools to develop the conceptual framework, formulate hypotheses, design the methodology or experiments, generate or clean datasets, translate text, or interpret results, and formulating mathematical claims or writing proofs is not applicable to this work. Additionally, we used generative AI tools for literature retrieval and discovery (e.g., finding related work), to polish the writing of the paper, and to format references. Using an MLLM to generate Component-Structured Profiles is part of the proposed method itself and is described in Section 3.2. We have reviewed all AI-assisted work: the authors reviewed all AI-assisted code before running the experiments reported in this paper, and checked all AI-edited text against the intended meaning and the experimental records. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses the publicly released benchmarks FineCVR-1M, EgoCVR and WebVid-CoVR in accordance with their licenses and terms of use; we do not collect or release any new videos. These videos contain people, and the CSPs generated by the MLLM describe their visible attributes, including the number and apparent gender of individuals. Such model-generated descriptions may inherit biases of the underlying MLLM, so CSPs should not be treated as reliable personal attributes. The human evaluation of CSP quality in Appendix C.2 was conducted by six researchers who rated model-generated descriptions of benchmark videos; no personal information about the raters was collected. Like other video retrieval systems, IC-CoVR could be misused to search video collections for specific individuals, and any deployment should respect privacy and data-protection regulations.

REPRODUCIBILITY STATEMENT

We provide the anonymized source code of IC-CoVR as supplementary material, including scripts for feature extraction, training and evaluation, together with instructions for reproducing our results on FineCVR-1M. Section 3 describes the full IC-CoVR architecture and training objective. Appendix A.1 lists the backbones, optimizer, learning-rate schedule and all hyperparameters, and Algorithm 1 gives the CSP-guided cross-frame aggregation step by step. The CSP components and granularity levels are specified in Appendix C.1.

REFERENCES

Muhammad Umer Anwaar, Egor Labintcev, and Martin Kleinsteuber. Compositional learning of image-text query for image retrieval. In *Proceedings of the IEEE/CVF Winter conference on Applications of Computer Vision*, pp. 1140–1149, 2021.

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

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining CLIP-based features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21466–21474, 2022.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15338–15347, 2023.

Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3001–3011, 2020.

Yongjun Chen, Zhiwei Liu, Jia Li, Julian McAuley, and Caiming Xiong. Intent contrastive learning for sequential recommendation. In *Proceedings of the ACM web conference 2022*, pp. 2172–2182, 2022.

Zhiwei Chen, Yupeng Hu, Zixu Li, Zhiheng Fu, Haokun Wen, and Weili Guan. HUD: Hierarchical uncertainty-aware disambiguation network for composed video retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 6143–6152, 2025.

Yi Cheng, Ziwei Xu, Dongyun Lin, Harry Cheng, Yongkang Wong, Ying Sun, Joo Hwee Lim, and Mohan Kankanhalli. Bridging the intent gap: Knowledge-enhanced visual generation. *arXiv preprint arXiv:2405.12538*, 2024.

En Ci, Shanyan Guan, Yanhao Ge, Yilin Zhang, Wei Li, Zhenyu Zhang, Jian Yang, and Ying Tai. Describe, don't dictate: Semantic image editing with natural language intent. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19185–19194, 2025.

Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-Pool: Cross-modal language-video attention for text-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5006–5015, 2022.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only training of zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13225–13234, 2024.

Animesh Gupta, Jay Parmar, Ishan Rajendrakumar Dave, and Mubarak Shah. From play to replay: Composed video retrieval for temporally fine-grained videos. *arXiv preprint arXiv:2506.05274*, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Thomas Hummel, Shyamgopal Karthik, Mariana-Iuliana Georgescu, and Zeynep Akata. EgoCVR: An egocentric benchmark for fine-grained composed video retrieval. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2024.

Seungmin Lee, Dongwan Kim, and Bohyung Han. CoSMo: Content-style modulation for image retrieval with text feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 802–812, 2021.

Hao Li, Yanhao Jia, Peng Jin, Zesen Cheng, Kehan Li, Jialu Sui, Chang Liu, and Li Yuan. FreestyleRet: Retrieving images from style-diversified queries. In *European Conference on Computer Vision*, pp. 258–274. Springer, 2024.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Ishani Mondal, Zongxia Li, Yufang Hou, Anandhavelu Natarajan, Aparna Garimella, and Jordan Lee Boyd-Graber. SciDoc2Diagrammer-MAF: Towards generation of scientific diagrams from documents guided by multi-aspect feedback refinement. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13342–13375, 2024.

Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. EgoVLPv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5285–5297, 2023.

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 *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2Word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19305–19314, 2023.

Pushkar Shukla, Carlos Elmadjian, Richika Sharan, Vivek Kulkarni, Matthew Turk, and William Yang Wang. What should I ask? Using conversationally informative rewards for goal-oriented visual dialog. *arXiv preprint arXiv:1907.12021*, 2019.

Yucheng Suo, Fan Ma, Linchao Zhu, and Yi Yang. Knowledge-enhanced dual-stream zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26951–26962, 2024.

Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Yue Hu, and Qi Wu. Context-I2W: Mapping images to context-dependent words for accurate zero-shot composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5180–5188, 2024.

Omkar Thawakar, Muzammal Naseer, Rao Muhammad Anwer, Salman Khan, Michael Felsberg, Mubarak Shah, and Fahad Shahbaz Khan. Composed video retrieval via enriched context and discriminative embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26896–26906, 2024.

Omkar Thawakar, Dmitry Demidov, Ritesh Thawkar, Rao Muhammad Anwer, Mubarak Shah, Fahad Shahbaz Khan, and Salman Khan. Beyond simple edits: Composed video retrieval with dense modifications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20435–20444, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. CoVR: Learning composed video retrieval from web video captions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5270–5279, 2024a.

Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. CoVR-2: Automatic data construction for composed video retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024b.

Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6439–6448, 2019.

Hao Wang, Fang Liu, Licheng Jiao, Jiahao Wang, Shuo Li, Lingling Li, Puhua Chen, and Xu Liu. Vision-by-Prompt: Context-aware dual prompts for composed video retrieval. *Pattern Recognition*, pp. 112378, 2025.

Xiaohan Wang, Linchao Zhu, and Yi Yang. T2VLAD: Global-local sequence alignment for text-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5079–5088, 2021.

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

APPENDIX

In this appendix, we provide:

- **Appendix A: Additional Experiments.** This section supplements the main experiments with implementation details, additional benchmark comparisons, ablation studies, hyperparameter analysis, and further studies of CSP design, frame count, temporal reasoning, and efficiency.
- **Appendix B: Additional Visualization.** This section presents qualitative visualizations of attention distributions, CSP examples, and cross-frame aggregation results, further illustrating how the model captures component-level and granularity-level information.
- **Appendix C: CSP Analysis.** This section provides a detailed analysis of the Component-Structured Profile (CSP), including its construction, quality evaluation, annotation rules, statistical properties, and the CSP-guided cross-frame aggregation strategy.

A ADDITIONAL EXPERIMENTS

In addition to the experiments presented in the main text, we conducted additional ablation studies and hyperparameter analyses.

A.1 DETAILED SETUP

For the MLLM, we employ Qwen3-VL-8B, and we use frozen CLIP (ViT-L/14) and BLIP as our backbones (the results for other backbones can be found in Table 9). We trained the model with the following hyperparameters:

- **Optimizer:** Adam
- **Batch Size:** 1024
- **Epochs:** 100
- **Learning Rate:** Initial `init_lr` set to 0.0001.
- **LR Scheduler:** OneCycleLR, configured as follows:
  - Warm-up phase (`pct_start`) = 0.05 (i.e., 5%)
  - Annealing strategy (`anneal_strategy`) = 'cos'
  - `div_factor` = 1.2
  - `final_div_factor` = 1.5
- **Other Hyperparameters:**
  - $K = 10$
  - top-$k = 1$
  - $\beta = 0.6$
  - $\gamma = 0.006$
  - $\tau = 0.001$

A.2 COMPARISON ON WEBVID-COVR

WebVid-CoVR mainly focuses on static edits and contains 1.6M training samples and 2K test samples. To further validate the generalization ability of our model, we compare it with prior methods evaluated on WebVid-CoVR. Since our approach uses a fully frozen CLIP encoder, we choose to compare against the variants of previous methods that also keep the backbone frozen. All models in Table 8 use a fully frozen backbone and train only the downstream network. Our model demonstrates strong generalization ability and consistently excellent handling of static information.

A.3 ABLATION STUDIES

Table 8: Performance on WebVid-CoVR compared with baseline and existing methods.

Method	Backbone	R1	R5	R10	R50
Only Text	BLIP	19.68	37.09	45.85	65.14
Only Visual	BLIP	34.90	59.23	68.04	85.95
CoVR (Ventura et al., 2024a)	BLIP	45.46	70.46	79.54	93.27
CoVR-2 (Ventura et al., 2024b)	BLIP-2	45.66	71.71	81.30	94.80
ECDE-CoVR (Thawakar et al., 2024)	BLIP	47.52	72.18	**82.37**	**95.06**
BSE-CoVR (Thawakar et al., 2025)	BLIP	<u>48.08</u>	73.36	81.06	93.78
Vision-by-Prompt (Wang et al., 2025)	BLIP-2	46.56	72.81	81.81	94.21
CoVR (Ventura et al., 2024a)	CLIP	44.37	69.13	77.62	93.00
$\text{FDCA}_{\text{CLIP}}$ (Wu et al., 2025)	CLIP	41.23	71.36	80.87	94.35
IC-CoVR$_{\text{CLIP}}$	**CLIP**	**49.07**	70.06	79.08	94.35

Table 10: Effect of the MLLM used for CSP generation on FineCVR-1M. EMS denotes the first-stage result; EMS+IPS denotes the final retrieval result. ECDE-CoVR (Thawakar et al., 2024) and BSE-CoVR (Thawakar et al., 2025) also use MLLMs.

		EMS				EMS+IPS	
Method	MLLM	R1	R5	R10	R50	R1	R5
IC-CoVR	Qwen2.5-VL-7B	31.58	59.46	71.97	94.55	34.89	58.91
IC-CoVR	Qwen3-VL-4B	31.56	59.39	71.96	94.56	34.47	58.71
IC-CoVR	Qwen3-VL-8B	31.58	59.42	72.04	94.58	35.06	59.12
ECDE-CoVR	MiniGPT-4	22.81	47.63	60.12	90.32	–	–
BSE-CoVR	Gemini-Pro/GPT-4o	24.40	51.01	63.42	91.21	–	–

Ablation on Backbone. There is a lack of a unified standard for visual encoders in the CVR task, with most papers using backbones that vary slightly. For example, regarding CLIP, Ventura et al. (2024a) utilize the ViT-L/14 version, while Wu et al. (2025) employ the RN50x4 version. Furthermore, some studies (Gupta et al., 2025) have employed adapted versions of CLIP, such as AIM. Similarly, regarding BLIP, Thawakar et al. (2025) use the original BLIP, whereas Ventura et al. (2024b) use BLIP-2. Therefore, we experimented with different backbones. As shown in Table 9, our model maintains superior performance across different backbones, consistently outperforming the previous SOTA model (Wu et al., 2025).

Table 9: Ablation studies on different CLIP backbones.

Method	Backbone	R1	R5	R10	R50
FDCA	RN50x4	25.84	55.84	70.23	94.33
IC-CoVR	RN50x4	33.62	59.36	72.16	94.69
IC-CoVR	ViT-L/14	35.06	59.12	72.04	94.58

Impact of the MLLM Backbone. To examine whether the gains of IC-CoVR primarily arise from the capability of the MLLM used to generate CSPs, we evaluate several Qwen2.5-VL and Qwen3-VL variants while keeping the retrieval backbone and the remaining settings unchanged. Table 10 reports retrieval performance after the Explicit-intent Modification Stage (EMS) and after the full model (EMS+IPS). Changing the MLLM barely affects EMS performance and causes only modest variation after IPS. These results suggest that the retrieval gains are not primarily explained by the choice of a stronger MLLM; rather, the proposed modification and preservation mechanisms effectively exploit the structured CSP.

We also compare IC-CoVR with ECDE-CoVR and BSE-CoVR, two methods that employ MLLMs. As shown in Table 10, IC-CoVR achieves better final-stage retrieval performance than both methods, further supporting the effectiveness of our design.

Ablation on Intent Enhancer. In EIM, we introduce a non-linear transformation to enhance the expressivity of $\boldsymbol{I}_{\text{o}}$. To validate its efficacy, we conduct ablation on it with multiple structural choices. Table 11 shows the model's performance with different intent enhancers. The current architecture gives the best R1 while remaining competitive at R5, R10, and R50.

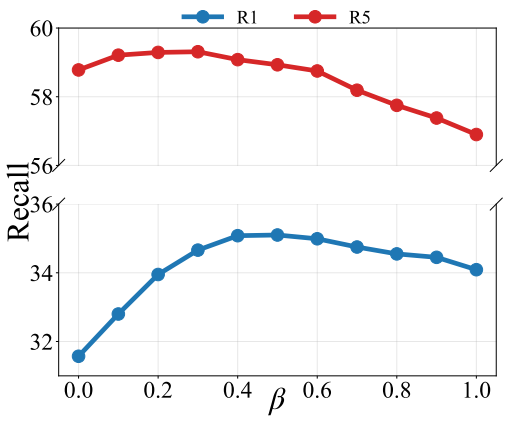


Figure 7: Results on different β.

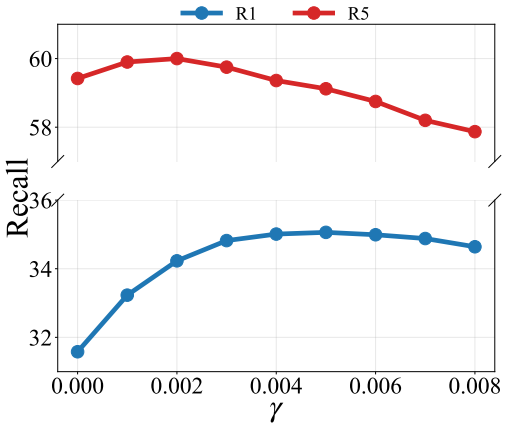


Figure 8: Results on different γ.

Table 11: Ablation studies on intent enhancer.

Method	R1	R5	R10	R50
None	34.66	**59.29**	71.68	94.71
Linear	34.43	58.72	**72.23**	94.82
No Expansion	33.98	58.50	71.89	**94.84**
4× Expansion	34.29	59.19	72.05	94.38
IC-CoVR	**35.06**	59.12	72.04	94.58

Table 12: Studies on enhanced CSP.

Coarse	Middle	Fine	R1	R5	R10	R50
0.0	0.0	0.5	34.22	58.86	**72.31**	**95.13**
0.3	0.3	0.3	33.85	58.91	72.07	94.59
0.1	0.3	0.5	33.48	**59.27**	72.06	94.68
0.2	0.5	0.8	34.37	59.14	71.53	94.53
0.0	0.0	0.0	**35.06**	59.12	72.04	94.58

A.4 Hyperparameter Studies

Studies on the pairwise comparison weight β. We further analyze the fusion weight β between the component-wise score and the global score. As shown in Figure 7, the performance is more sensitive when β is small, where a larger emphasis on local matching leads to noticeable changes in R@1. As β increases, the results become relatively more stable, suggesting that incorporating more global information improves robustness. R@1 achieves the best performance around $\beta = 0.5$–0.6, while R@5 remains competitive over a broad range.

Studies on the final similarity weight γ. We further study the similarity fusion weight γ to examine the relative roles of the two stages in final scoring. As shown in Figure 8, performance peaks at small γ values ($0.004 \sim 0.008$), indicating that IPS acts as a lightweight but important refinement over the modification-aware predictions established by EMS, rather than replacing them. In summary, the two stages are complementary rather than competitive.

A.5 Further Analysis

Studies on Enhanced CSP. To further investigate the role of CSP, we attempted to enhance it with visual features. Table 12 shows the model's performance when injecting visual features into CSP features of different granularities at varying degrees. The results demonstrate that while this enhancement slightly benefits R10 and R50, it is detrimental to precise retrieval. This further indicates that the significance of CSP lies in its ability to provide a mechanism for selective information filtering, rather than offering excessively detailed descriptions that perfectly match the visual content. This helps the model ignore irrelevant disturbances during the retrieval process.

Studies on Frame Count. We evaluate how the number of sampled frames affects performance. As shown in Table 13, the model's performance suffers a significant degradation as the number of sampled frames decreases. We attribute this to the fact that reducing the number of frames severely undermines the cross-frame aggregation process. This not only leads to a substantial loss of effective information but also increases the risk—due to the stochasticity of sparse sampling—of capturing

Table 13: Studies on frame count.

Count	R1	R5	R10	R50
1	28.00	44.65	51.89	76.86
2	31.17	52.26	62.89	87.68
4	33.32	57.57	69.50	93.30
8	**35.06**	**59.12**	**72.04**	**94.58**

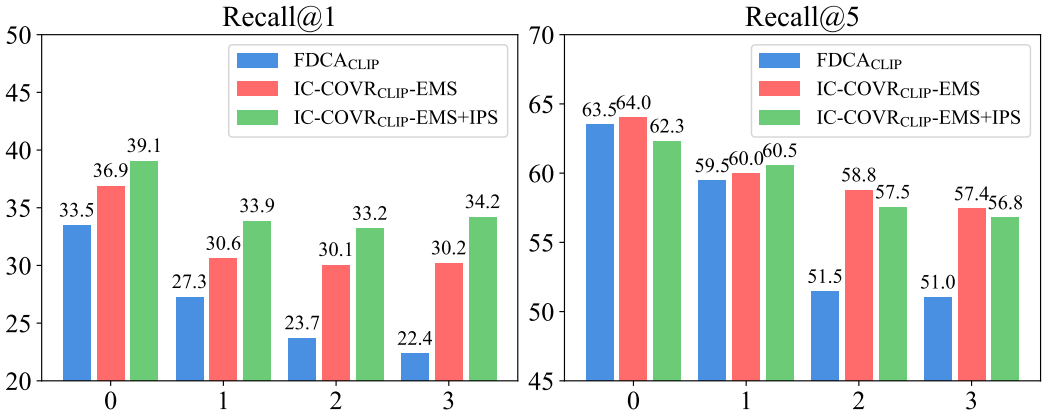


Figure 9: Performance on different query types, where 0-background change, 1-object change, 2-similar action change, 3-distinct action change. IC-CoVR achieves higher performance gains on action-related query types.

non-informative frames. Furthermore, this also suggests that the CVR task possesses a more complex information structure than the CIR task, further underscoring the unique nature of this task.

Temporal Reasoning Analysis. Temporal dynamics distinguish CVR from CIR, yet prior works often sacrifice this rich temporal structure via mean pooling (Hummel et al., 2024; Wu et al., 2025; Ventura et al., 2024a; Thawakar et al., 2025). In contrast, our model effectively leverages temporal cues through cross-frame aggregation and the action descriptions within the CSP. As shown in Figure 9, compared to previous SOTA (Wu et al., 2025), our model achieves the most significant gains on dynamic queries (types 2 and 3), drastically narrowing the performance gap with static ones.

Nevertheless, dynamic queries still lag behind static ones, indicating that temporal reasoning remains an open challenge in CVR. Although our model narrows this gap, more effective mechanisms for modeling fine-grained temporal dependencies and action evolution are still needed.

A.6 Inference Time and Storage Cost

To evaluate the real-world deployment feasibility of our proposed framework, we analyze its computational efficiency with respect to both inference latency and storage overhead. All benchmarking experiments were conducted on a single NVIDIA RTX 3090 GPU. First, generating a CSP with the MLLM takes approximately 7,600 ms per video. The RSS stage requires an average of 33ms per video while maintaining a highly lightweight storage footprint of merely 53KB. Second, the joint computation of EMS and IPS is exceptionally efficient, demanding only 3ms per query when searching against a substantial gallery size of 13,046 samples.

From a practical deployment perspective, generating CSPs with the MLLM introduces an additional one-time cost of approximately 7,600 ms per video. For a new, unindexed video, this brings the total end-to-end latency to approximately 7,636 ms. However, once generated, CSPs and video features can be cached and reused across queries. In a standard retrieval setting with a pre-processed gallery, the online retrieval latency is therefore only 3 ms per query.

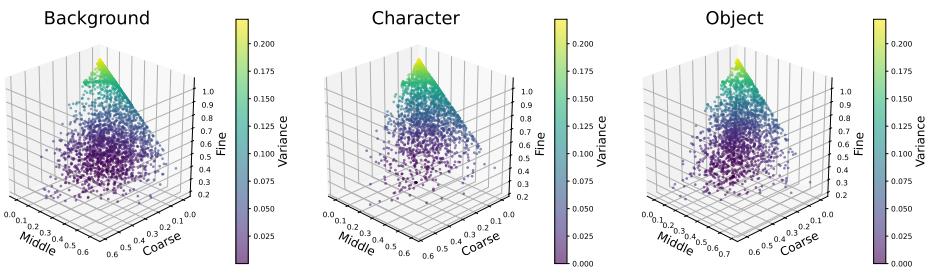


Figure 10: Attention score distribution.

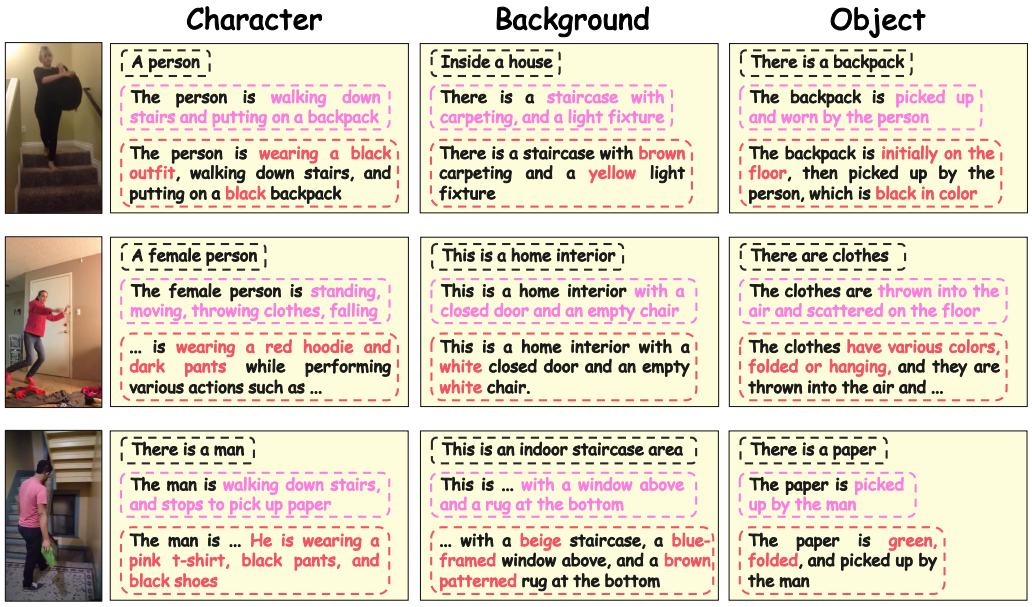


Figure 11: Visualization of CSP.

B Additional Visualization

B.1 Visualization of Attention Distribution

To investigate how the model utilizes information of different granularities during inference, we visualized the attention distribution across these granularity levels. In Figure 10, we map the attention weights of the three granularities to 3D spatial coordinates to generate a 3D scatter plot. The color temperature of each point encodes the variance of its three coordinate values. We observe that for the "Character" component, attention is predominantly concentrated on fine-grained information, indicating a high demand for detailed character attributes in relevant queries. In contrast, the attention distribution for "Object" and "Background" appears relatively uniform, suggesting that detailed information for these components is less critical during the retrieval process.

B.2 Visualization of CSP

Correspondence of CSP. As illustrated in Figure 11, the corresponding CSP effectively partitions each video, maintaining relative independence among components while ensuring that each component fully covers its relevant content. Simultaneously, within each component, we establish a hierarchical structure with progressively increasing granularity, which effectively caters to the information needs of diverse scenarios.

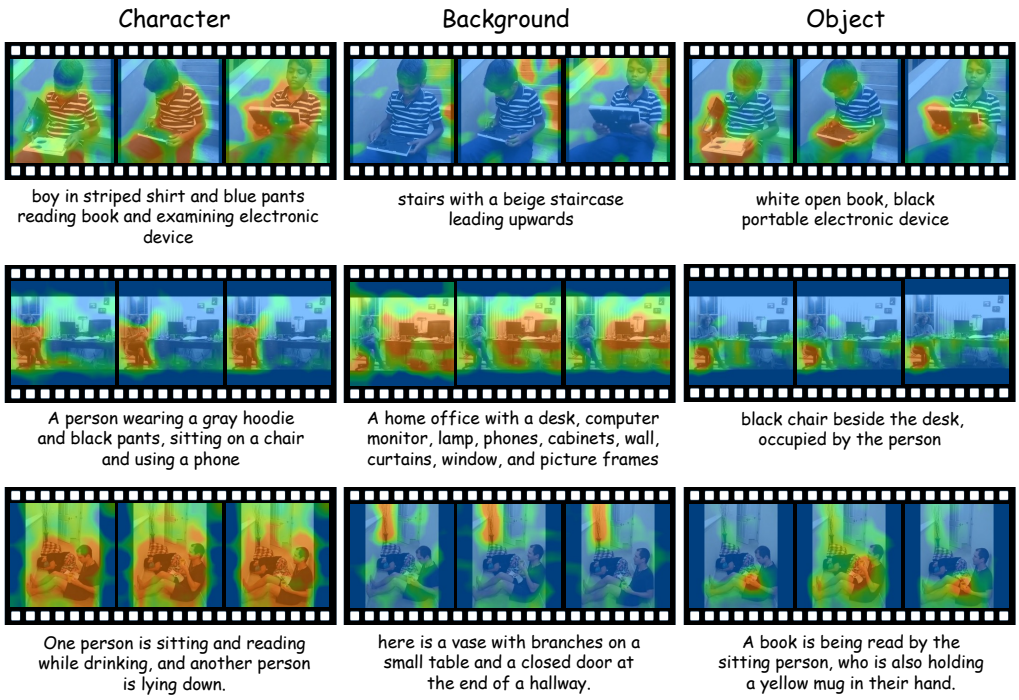


Figure 12: Visualization of cross-frame aggregation.

Visualization of Cross-Frame Aggregation. To validate the effectiveness of the cross-frame aggregation, we visualized the aggregated results. As shown in Figure 12, the content in each frame consistently aligns with its corresponding CSP component, effectively achieving cross-frame tracking of the same object. Ultimately, guided by the CSP, the visual component is extracted to represent specific video regions.

C CSP Analysis

C.1 CSP construction

CSP Structure. Component-Structured Profile (CSP) is a structure that provides multi-granularity descriptions for different regions of a video. To ensure that CSP effectively decomposes the video, its design adheres to two key criteria: (1) the common objective content found in videos, and (2) the entities typically targeted by the modification text. We thus divide the video into three parts:

- **Character:** Individuals present in the video. Its associated attributes include actions, clothing, and quantity.
- **Object:** Objects that either interact with people or undergo significant changes in movement or state. Its attributes include color, position, and state, among others.
- **Background:** This component comprises an abstract location (the scene setting) and static objects. These static objects share similar attributes with dynamic objects.

For each component, we designed the three coarse-to-fine granularity levels based on their attribute characteristics, with each level building upon the previous one by adding the following details:

- **Character:**
  1. Coarse: Quantity and gender of the individuals.
  2. Middle: Activities or actions.
  3. Fine: Static attributes (e.g., clothing, appearance).

- **Object:**
  1. Coarse: General identification of the items.
  2. Middle: Movement or state changes.
  3. Fine: Static attributes (e.g., color, position).
- **Background:**
  1. Coarse: Definition of the location or environment.
  2. Middle: Static objects within the scene.
  3. Fine: Static attributes of static objects.

C.2 CSP Quality Evaluation

CSP Quality Analysis. We evaluate the MLLM-generated CSP from two perspectives: its intrinsic quality and the model's robustness to CSP imperfections.

To assess CSP quality, we combine human evaluation with feature-level analysis. We recruit six researchers to review 400 samples. Each annotator evaluates 150 videos, including a shared subset of 100 samples for inter-annotator agreement, along two dimensions: *Visual Consistency*, which measures faithful video-text alignment without hallucination, and *Granularity Progression*, which examines whether the CSP follows the intended coarse-to-fine hierarchy. CSP achieves 4.02/5 on Visual Consistency (Fleiss' $\kappa = 0.61$) and 4.14/5 on Granularity Progression (Fleiss' $\kappa = 0.65$), indicating that it is both faithful to the video content and well aligned with our hierarchical design. At the feature level, CSP attains a higher video-text similarity than the original captions (0.717 vs. 0.692), supporting its low-hallucination property, while an inter-component distance of 0.471 further suggests clear semantic separation across components.

To further examine the dependency on CSP quality, we perturb CSP features with Gaussian noise. At a noise level of 0.8, where the perturbation magnitude is comparable to that of the original CSP features, R1 and R5 decrease by only 1.98 and 0.89 points, respectively. This suggests that the model is relatively robust to substantial imperfections in CSP features.

Table 14: Performance with noisy CSPs.

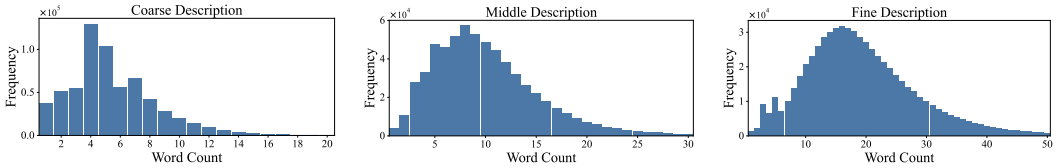
Noise Level	0.2	0.4	0.6	0.8	1.0
R1	34.81	34.39	34.05	33.08	32.52
R5	58.89	58.31	58.29	58.23	58.10

C.3 Evaluation Rules

The purpose of this section is to clarify the annotation protocol and overall evaluation configuration underlying the results reported above. The detailed evaluation criteria for each dimension are presented as follows:

1. Visual Consistency. This dimension measures whether the generated text descriptions, at all granularity levels, faithfully reflect the visual content of the video, without hallucinating non-existent entities, attributes, or actions.

- **5 – Perfect Match.** The descriptions are fully consistent with the video. Core content is accurately captured, without factual errors or hallucinations.
- **4 – Minor Flaws.** The descriptions are largely accurate and cover the core visual content, with only negligible inaccuracies that do not affect the overall semantics.
- **3 – Partial Match.** The descriptions correctly capture the main entities and overall scene, but contain noticeable inaccuracies in motion details or fine-grained attributes.
- **2 – Major Errors.** The descriptions contain substantial mismatches, such as hallucinated objects or actions, or serious misidentification of key entities, actions, or attributes.
- **1 – Irrelevant.** The descriptions are largely unrelated to the video content or are nonsensical.



(a) Coarse (b) Middle (c) Fine

Figure 13: Word count for each granularity in CSP.

2. Granularity Progression. This dimension measures whether the three CSP levels exhibit a clear coarse-to-fine structure, such that fine-grained levels provide progressively richer and more informative details than coarse levels.

- **5 – Strict Coarse-to-Fine.** The progression is clear and well-structured, where each level incorporates a significant incremental increase as designed.
- **4 – Good Progression.** The hierarchy shows a clear increase in detail, although the difference between coarse and middle levels is not significant.
- **3 – Marginal Addition.** Fine-grained levels add only limited or weakly informative details, failing to capture detailed attribute information.
- **2 – Flat or Redundant.** The three levels are largely similar, with minimal increase and substantial redundancy.
- **1 – Contradictory.** The hierarchy fails to progress properly: different levels conflict with each other, or the fine-grained levels become less informative than the coarse.

C.4 CSP Statistics

As mentioned in the main paper, CSP contains three coarse-to-fine granularity levels for each component, which allows for a flexible preservation process. Figure 13 demonstrates that the average word count scales with granularity level, increasing from coarse to fine representations.

For each component, we also provide a word cloud in Figure 14. Specifically, for the Character class (Figure 14a), the CSP emphasizes human-related descriptors. Prominent terms include “person”, “man”, “woman”, and “child”, along with appearance descriptions such as “wearing”, “blue”, “white”, and “shirt”. The vocabulary also captures dynamic states like “playing”, “performing”, and “standing”. For the Background class (Figure 14b), the CSP concentrates on spatial settings with terms like “room”, “kitchen”, “park”, and “indoor”, along with color descriptors “white”, “green”, and “brown” that characterize environmental contexts. For the Object class (Figure 14c), the focus centers on tangible items such as “objects”, “ball”, “table”, with emphasis on color attributes and states including “black”, “used”, “placed”, and “held”.

Furthermore, Figure 15 presents word clouds categorized by granularity levels. At the Coarse level (Figure 15a), the CSP captures fundamental elements with terms like “person”, “object”, and “indoor”, providing a broad overview of the video. At the Middle level (Figure 15b), the vocabulary expands to include actions and movement such as “standing”, “holding”, and “moving”, offering a more detailed depiction of the scene. Finally, at the Fine level (Figure 15c), the CSP incorporates intricate details with words like “black”, “brown”, and “wearing”, reflecting a comprehensive and nuanced understanding of the video’s content.

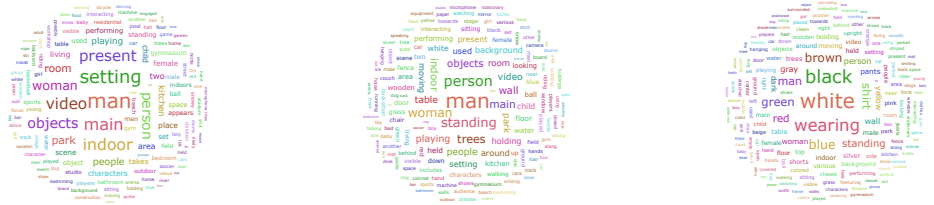
C.5 CSP-Guided Cross-Frame Aggregation

After obtaining the CSP, we designed a CSP-guided cross-frame aggregation algorithm. In this algorithm, we first leverage each CSP component to interact with the patches from each frame, thereby computing attention weights. Subsequently, patches exhibiting high similarity to the component are identified as its visual correspondence. Finally, these corresponding patches are aggregated across frames. The detailed procedure is given in Algorithm 1.



(a) Character (b) Background (c) Object

Figure 14: Word cloud for each component in CSP.



(a) Coarse (b) Middle (c) Fine

Figure 15: Word cloud for each granularity in CSP.

Algorithm 1 CSP-Guided Cross-Frame Aggregation

Require: Image tensors $\mathbf{I} \in \mathbb{R}^{B\times F\times C\times H\times W}$
Require: Multi-granularity descriptions $\mathcal{D} = \{d_b\}_{b=1}^{B}$, where $d_b = \{s : \{g : \text{text}\}\}$
Require: Temperature parameter $\tau > 0$
Require: component $\mathcal{S} = \{\text{background}, \text{character}, \text{object}\}$
Require: Granularity levels $\mathcal{G} = \{\text{coarse}, \text{middle}, \text{fine}\}$
Ensure: Visual Component $\mathbf{E}_{\text{vc}} \in \mathbb{R}^{B\times|\mathcal{S}|\times|\mathcal{G}|\times D}$

// **Step 1: Feature Extraction**
$\mathbf{P} \leftarrow \text{CLIP}_{\text{patch}}(\mathbf{I})$ {Patch embeddings: $[B, F, P, D]$}
$\mathbf{E}_{\text{cls}} \leftarrow \mathbf{P}[:, :, 0, :]$ {Extract CLS tokens: $[B, F, D]$}
// **Step 2: Text Encoding**
Flatten descriptions: $\mathcal{T}_{\text{flat}} \leftarrow \{d_b[s][g] \mid b \in [1, B], s \in \mathcal{S}, g \in \mathcal{G}\}$
$\mathbf{T} \leftarrow \text{CLIP}_{\text{text}}(\mathcal{T}_{\text{flat}})$ $\{[B \cdot |\mathcal{S}| \cdot |\mathcal{G}|, D]\}$
$\mathbf{E}_{\text{text}} \leftarrow \text{reshape}(\mathbf{T}, [B, |\mathcal{S}|, |\mathcal{G}|, D])$
// **Step 3: Compute Text-Patch Similarity**
$\hat{\mathbf{P}} \leftarrow \frac{\mathbf{P}}{\|\mathbf{P}\|_2}$, $\hat{\mathbf{E}}_{\text{text}} \leftarrow \frac{\mathbf{E}_{\text{text}}}{\|\mathbf{E}_{\text{text}}\|_2}$ {L2 normalization}
$\mathbf{Sim} \leftarrow \langle \hat{\mathbf{E}}_{\text{text}}, \hat{\mathbf{P}} \rangle$ {Cosine similarity: $[B, |\mathcal{S}|, |\mathcal{G}|, F \cdot P]$}
// **Step 4: Attention-Weighted Pooling**
$\mathbf{W} \leftarrow \text{softmax}(\mathbf{Sim}/\tau)$ {Attention weights over patches}
$\hat{\mathbf{P}}_{\text{flat}} \leftarrow \text{reshape}(\hat{\mathbf{P}}, [B, F \cdot P, D])$
$\mathbf{E}_{\text{vc}} \leftarrow \mathbf{W} \cdot \hat{\mathbf{P}}_{\text{flat}}$ {Weighted aggregation: $[B, |\mathcal{S}|, |\mathcal{G}|, D]$}
return $\mathbf{E}_{\text{vc}}, \mathbf{E}_{\text{text}}, \mathbf{E}_{\text{cls}}$