

Text-Conditioned Static-Dynamic Composition for Composed Video Retrieval

Yulin Zhang^{1*}, Haitian Zheng^{1*}, Zhengwei Yang^{1,2*}, Hao Li¹, Zheng Wang^{1†}

¹National Engineering Research Center for Multimedia Software,
Institute of Artificial Intelligence,
School of Computer Science, Wuhan University

²School of Information Engineering, Zhongnan University of Economics and Law
{zyl-252, skyzheng, sc.lihao, wangzwhu}@whu.edu.cn
yzw@zuel.edu.cn

Abstract

Composed Video Retrieval (CVR) aims to retrieve a target video given a reference video and a modification text. The central challenge is to identify and use video content that is relevant to the retrieval need specified by the modification text. Existing methods typically encode videos before considering this textual need, causing visual evidence to be formed in a largely text-agnostic manner. We propose StaDy-CLIP, a compact dual-pathway framework for text-conditioned evidence composition. It preserves the frozen CLIP visual encoder as a static pathway for CLIP-aligned static evidence, introduces a lightweight side pathway for dynamic evidence, and uses Text-Conditioned Composition Blocks to perform patch-wise dynamic evidence selection and textual modulation. A learnable Gallery Conditioning Prompt further enables gallery videos to be encoded within the same evidence composition framework. StaDy-CLIP achieves state-of-the-art performance on FineCVR-1M and TF-CoVR with only 7.78M trainable parameters, improving prior best results by +4.05 in Recall@1 and +1.49 in mAP@5. Ablations and encoder-transfer results validate its effectiveness and encoder-level transferability for existing CVR methods.¹

1 Introduction

Composed retrieval addresses a flexible search scenario where a target instance is specified relative to a reference example through a natural-language modification. The modification text not only describes how the target differs from the reference, but also determines which reference content should be preserved for retrieval. While this paradigm has been widely studied in image-based settings such as Composed Image Retrieval (CIR) (Liu et al.,

2021), Composed Video Retrieval (CVR) (Ventura et al., 2024b) extends it to temporally structured videos, where retrieval depends on both a reference video and a modification text.

Compared with composed retrieval over static images, CVR involves a richer and more heterogeneous space of visual evidence. A reference video may contain static cues, such as scenes, objects, attributes, and appearances, as well as dynamic cues, such as actions, interactions, and motion patterns. The retrieval relevance of these cues is determined not only by the video itself, but also by the modification text, which specifies how the reference video should be interpreted for retrieval. As illustrated in Figure 1a, the same reference video may require different evidence under different textual modifications: object-level modifications may require preserving the ongoing action while changing static semantics, whereas action-level modifications may preserve the scene context while suppressing the original actions. Therefore, effective CVR models must go beyond generic spatio-temporal modeling and organize video evidence according to the language-specified retrieval need.

Recent CVR benchmarks make this language-dependent evidence demand explicit through diverse modification types (Yue et al., 2025). Most existing CVR methods have focused primarily on how to compose the modification text with video features extracted by an upstream visual encoder. Compared with earlier methods that rely on image-based encoders (Ventura et al., 2024b; Chen et al., 2025), recent approaches introduce stronger video encoders to provide richer temporal evidence for downstream composition (Hummel et al., 2024; Gupta et al., 2025). However, as shown in Figure 1b, replacing the visual encoder can bring noticeable gains, but the improvements are often biased toward particular types. Such gains arise because the modification text is typically introduced only after video features have been extracted, leav-

*Equal contribution.

†Corresponding author.

¹Code: <https://github.com/Skye-ye/StaDy-CLIP>

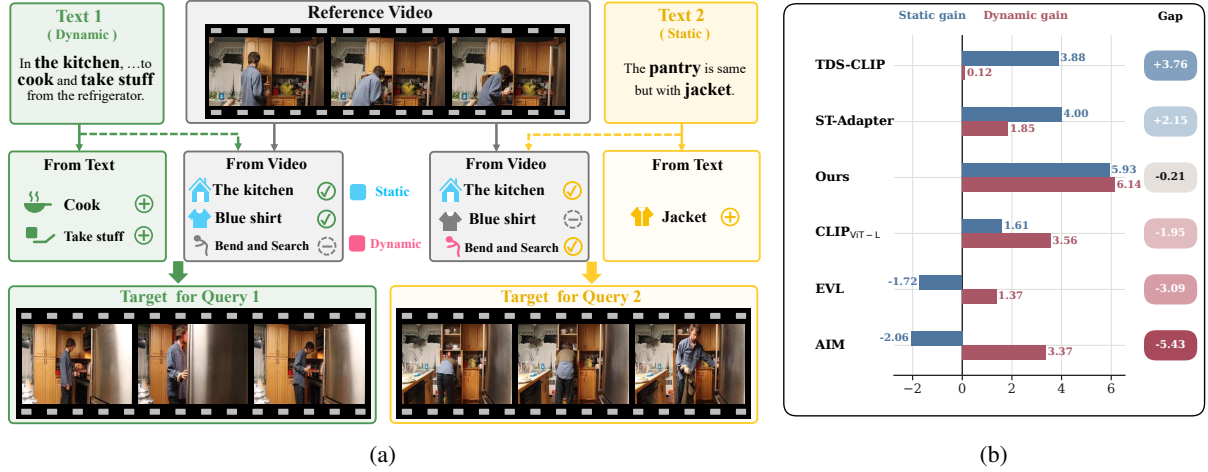


Figure 1: Motivation. *Left*: ✓ means preserving the evidence, while ✗ means ignoring the evidence. ⊕ means adding new evidence. The same reference video requires different evidence under different text. When only the object changes, dynamic information should be preserved for retrieval; when the modification changes the action, retrieval should rely more on static semantics. *Right*: Bars show the gains on static and dynamic query subsets, while the right column reports the gap $\Delta_{\text{Static}} - \Delta_{\text{Dynamic}}$, when CLIP is replaced by video encoders.

ing the formation of video evidence largely unaffected by the current textual requirement. As a result, the encoded video evidence is shaped primarily by the encoder’s architecture-dependent priors, such as appearance modeling or motion sensitivity, and thus tends to better serve specific textual requirements that match these priors.

To address these limitations, we propose **StaDy-CLIP**, a CLIP-based dual-pathway framework for text-conditioned video modeling in composed video retrieval. Rather than treating the reference video as a fixed backbone output, StaDy-CLIP allows the modification text to condition how visual evidence is selected and organized. Specifically, it preserves the original CLIP visual encoder as a static pathway to provide static evidence, while introducing a lightweight dynamic pathway to construct temporal evidence. The two pathways interact through a sequence of **Text-Conditioned Composition Blocks**, where the modification text is grounded onto patch-level visual evidence to guide dynamic evidence selection and textual modulation. In this way, StaDy-CLIP can emphasize static cues for appearance-, object-, or scene-related modifications, while incorporating temporal evidence for action- and motion-related modifications. To maintain a unified conditioning framework for both query and gallery videos, we further introduce a learnable **Gallery Conditioning Prompt**, which enables gallery video representations to be encoded under the same composition mechanism.

Overall, our main contributions are as follows.

- We revisit composed video retrieval as text-conditioned video modeling, where the modification text specifies the retrieval need and determines how static and dynamic evidence should be selected from the reference video.
- We propose StaDy-CLIP, a compact dual-pathway framework that performs text-conditioned composition for both query and gallery video encoding.
- Experiments on FineCVR-1M and TF-CoVR show that StaDy-CLIP achieves state-of-the-art performance with only 7.78M trainable parameters, while ablations and analysis validate its effectiveness and encoder-level generalizability for existing CVR methods.

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 learn a shared embedding space for visual and textual inputs (Baldrati et al., 2022; Chen et al., 2020; Vo et al., 2019; Lee et al., 2021; Anwaar et al., 2021), where the fused representation is used to match the target image. 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., 2024a; Suo et al., 2024; Tang et al., 2024), which removes the

reliance on pre-defined triplet training data. Despite their effectiveness, CIR methods are fundamentally designed for static visual composition. Their modifications are typically expressed at the level of objects, attributes, or scenes in a single image, without reasoning over temporal evolution.

2.2 Multimodal Grounding

Multimodal grounding establishes correspondences between language and visual content, covering tasks such as referring expression comprehension, phrase grounding, moment retrieval, and spatio-temporal video grounding. Early image-level work treated frame referring expression and phrase grounding as text-conditioned region prediction (Mao et al., 2015; Deng et al., 2021), later scaled up by grounded pretraining that learns region-phrase alignment from large-scale image-text data (Kamath et al., 2021; Li et al., 2021; Liu et al., 2023). In the video domain, grounding extends to temporal sentence grounding and moment retrieval (Gao et al., 2017; Hendricks et al., 2017; Lei et al., 2021; Cao et al., 2024) and spatio-temporal video grounding (Yang et al., 2022; Gu et al., 2024b; Zhao et al., 2024), with recent video-LLMs sharpening temporal localization via specialized temporal tokens (Ren et al., 2023; Huang et al., 2023; Wang et al., 2024b; Zeng et al., 2024). A complementary line pursues fine-grained alignment between visual and text tokens within the encoder (Yao et al., 2021; Ma et al., 2022; Wang et al., 2023; Asokan et al., 2025). However, these methods focus on text describing content present in the visual input, unlike the modification-style language in composed retrieval.

2.3 Composed Video Retrieval

CVR aims to retrieve the target video based on a reference video and modification text. Many existing CVR methods remain centered on static semantic alignment, either through CIR-style fusion of reference videos and modification text (Ventura et al., 2024b,a; Wang et al., 2025), or through enhanced fine-grained video-text alignment via richer textual representations (Thawakar et al., 2024, 2025; Yue et al., 2025), without explicitly modeling temporal dynamics. Some recent works further extend CVR beyond static semantic alignment by introducing simple video-level modeling, such as incorporating video encoders or leveraging MLLMs to generate video descriptions (Gupta et al., 2025; Hummel et al., 2024). Although these methods incorporate

temporal information to some extent, they mainly capture generic temporal patterns and lack task-specific temporal modeling for CVR. Several CVR benchmarks have also been introduced. WebVid-CoVR (Ventura et al., 2024b) mainly covers static content. In contrast, FineCVR-1M (Yue et al., 2025) focuses on fine-grained static and dynamic modifications, EgoCVR (Hummel et al., 2024) targets egocentric actions, and TF-CoVR (Gupta et al., 2025) emphasizes fine-grained actions.

3 Method

3.1 Method Overview

In composed video retrieval (CVR), each query consists of a reference video v_r and a modification text t_m . The goal is to retrieve the target video v_t from a gallery $\mathcal{V}$ according to the composed query (v_r, t_m) . t_m specifies the information need under which visual evidence in v_r should be selected and organized. We formulate the video representation as $v_c = f(v_r | t_m)$, emphasizing that the modification text conditions the composition of reference-video evidence. Then compose the video and text to form the final query representation.

We propose **Stady-CLIP**, a CLIP-based dual-pathway framework for text-conditioned video modeling in CVR. As shown in Figure 2, Stady-CLIP preserves the original CLIP visual encoder as a static pathway to provide static evidence, while attaching a lightweight dynamic pathway to construct temporal evidence. The two pathways interact through a sequence of **Text-Conditioned Composition Blocks (TCC Blocks)**, where static and dynamic evidence are composed at the patch level under the modification text. Within each block, the modification text is first grounded onto visual evidence to provide text-grounded guidance, which is then used to select dynamic evidence and modulate visual tokens according to the specified modification. For gallery encoding, we replace the modification text with a learnable **Gallery Conditioning Prompt**, enabling gallery videos to be represented within the same evidence composition framework.

3.2 Static-Dynamic Evidence Decomposition

To make video evidence selectable under different modification texts, each reference video is represented as complementary evidence sources capturing different types of visual information. Stady-CLIP decomposes video evidence into these two sources through a dual-pathway design.

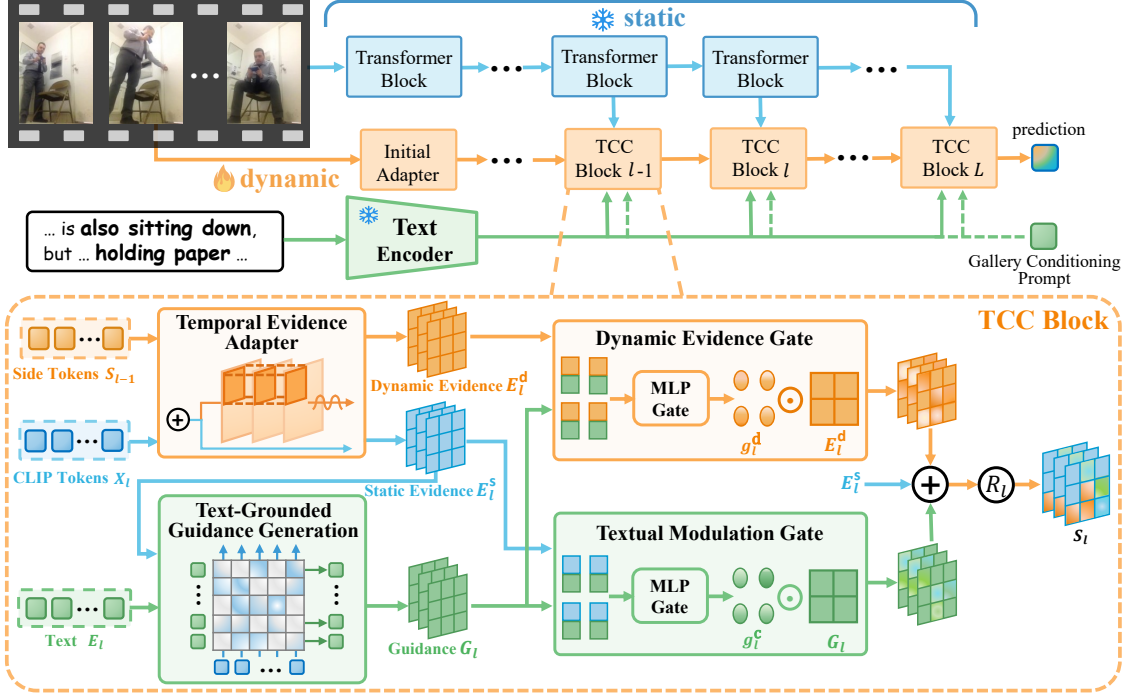


Figure 2: Architecture of StaDy-CLIP. A frozen CLIP visual encoder provides static evidence, while a lightweight dynamic pathway constructs temporal evidence. Text-Conditioned Composition Blocks ground the modification text onto visual evidence and perform patch-wise gated selection of dynamic evidence and textual modulation. Gallery videos are encoded with a learnable Gallery Conditioning Prompt within the same evidence composition framework.

Dual-Pathway Structure. We adopt frozen CLIP (Radford et al., 2021) as the static pathway, extracting frame-level static cues such as objects, scenes, and visual attributes from raw frames. The dynamic pathway is a smaller trainable network with dimension d_s , in which lightweight adapters (Wang et al., 2024a) are inserted to capture temporal information. It is initialized by an SME Adapter (Wang et al., 2024a) with frame difference, denoted as $S_0 = \text{SME}(v_r)$. Starting from this initial side representation, the dynamic pathway interacts with the CLIP visual encoder through L TCC Blocks attached to its last L layers.

Layer-Wise Evidence Construction. At the l -th layer, the previous side representation S_{l-1} is combined with the CLIP tokens X_l from current layer to form the current static evidence tokens $E_l^s \in \mathbb{R}^{T \times N \times d_s}$. Then, a lightweight temporal evidence adapter $T_l(\cdot)$ is applied to produce dynamic evidence $E_l^d \in \mathbb{R}^{T \times N \times d_s}$. we construct the two evidence sources as:

$$E_l^s = \alpha \cdot P_l^v(X_l) + S_{l-1}, \quad E_l^d = T_l(E_l^s), \quad (1)$$

where $P_l^v(\cdot)$ projects CLIP tokens in the side-pathway dimension d_s , and α is a fixed weight

that balances the initial static contribution. These static (E_l^s) and dynamic (E_l^d) tokens form complementary evidence sources, which are passed to subsequent modules that perform text-conditioned grounding and selection.

3.3 Text-Conditioned Evidence Composition

Given the static evidence E_l^s and dynamic evidence E_l^d , CVR requires deciding how each type of evidence should contribute under the modification text. Rather than uniformly fusing static and dynamic tokens, StaDy-CLIP performs text-conditioned evidence composition at the patch level. Specifically, each composition block first grounds the modification text onto patch-level visual evidence, and then uses the resulting guidance to decide where dynamic evidence should be injected and where visual tokens should be textually modulated.

Text-Grounded Guidance Generation. To obtain patch-level modeling cues, we ground the modification text onto visual evidence through cross-attention. Instead of reducing the text to a single pooled sentence embedding, we keep the token-level CLIP text representation as the conditioning context, so that different words in the modification can guide different visual regions. Specifically,

the static evidence tokens E_l^s are used as queries, while the projected text-token features $\tilde{E}_l$ are used as keys and values:

$$G_l = \text{LN} \left(\text{CrossAttn}_l \left(E_l^s, \tilde{E}_l, \tilde{E}_l \right) \right). \quad (2)$$

The resulting text-grounded guidance $G_l \in \mathbb{R}^{T \times N \times d_s}$ preserves the patch-wise structure of the visual evidence. Each token in G_l encodes how the corresponding visual region should be interpreted under the modification text, establishing localized correspondences between visual evidence and textual semantics. This guidance provides the linguistic basis for subsequent evidence modulation, indicating which regions require stronger textual reinterpretation and where dynamic evidence should be incorporated.

Text-Specified Evidence Selection. Given the static evidence E_l^s , dynamic evidence E_l^d , and text-grounded guidance G_l , the composition block selects and modulates visual evidence at the patch level. Rather than treating dynamic evidence and textual guidance as uniformly useful for all visual tokens, we predict two gates that reflect the evidence demand specified by the modification text. The **Dynamic Evidence Gate** determines where motion-sensitive evidence should be injected, while the **Textual Modulation Gate** determines where visual evidence should be reinterpreted or emphasized by the text-grounded guidance:

$$\begin{aligned} g_l^d &= \sigma(\text{MLP}_l^d([E_l^d; G_l])), \\ g_l^t &= \sigma(\text{MLP}_l^t([E_l^s; G_l])). \end{aligned} \quad (3)$$

The patch-wise scalar weights g_l^d and g_l^t are broadcast along the channel dimension during modulation. Specifically, g_l^d controls the contribution of dynamic evidence, allowing motion-sensitive cues to be incorporated only for regions where they are needed by the modification text. In parallel, g_l^t controls the contribution of text-grounded guidance, allowing each visual token to be reinterpreted or emphasized according to the specified modification. The two gates therefore separate dynamic evidence selection from textual modulation, so that each patch can form its representation from the evidence required by the current text.

We then use a lightweight self-attention layer $\mathcal{R}_l$ to propagate the selected evidence across patches and produce the updated side representation S_l , which is passed to the next composition block:

$$S_l = \mathcal{R}_l(E_l^s + g_l^d \odot E_l^d + g_l^t \odot G_l), \quad (4)$$

where $\odot$ denotes element-wise multiplication with broadcasted gate weights.

3.4 Prompt-Conditioned Gallery Encoding

The proposed composition blocks require a conditioning input to guide evidence composition. For the query branch, this conditioning input is naturally provided by the modification text. However, gallery videos are not paired with modification texts, and directly encoding them without any conditioning would make the gallery branch fall back to unconditioned video representation learning. To keep the query and gallery branches within the same evidence composition framework, we introduce a learnable **Gallery Conditioning Prompt** as the conditioning input for the gallery branch.

The gallery prompt is a short sequence of learnable task tokens shared across TCC Blocks. During gallery encoding, it replaces the modification-text tokens as the conditioning context, allowing gallery videos to pass through the same text-conditioned composition mechanism used by the query branch. Unlike the modification text, the gallery prompt does not specify an instance-level retrieval intent. Instead, it provides a task-level prior learned from the composed retrieval objective, encouraging gallery videos to organize static and dynamic evidence in a retrieval-oriented form.

This design keeps query and gallery encoding symmetric at the module level while respecting their different roles. The query branch uses the modification text to select evidence for a specific retrieval need, whereas the gallery branch uses the learned prompt to produce query-independent gallery representations that can be precomputed for retrieval.

3.5 Objective and Inference

Training Objective. During training, the query branch is conditioned on the modification text, while the gallery branch is conditioned on the Gallery Conditioning Prompt. Given a batch of B triplets, we denote the composed query representation as q_i and the target video representation as k_i . StaDy-CLIP is optimized with a symmetric in-batch InfoNCE loss. We first compute the query-to-target loss $\mathcal{L}_{q \rightarrow k}$:

$$\mathcal{L}_{q \rightarrow k} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(q_i^\top k_i / \tau)}{\sum_{j=1}^B \exp(q_i^\top k_j / \tau)}. \quad (5)$$

We then compute the symmetric target-to-query loss $\mathcal{L}_{k \rightarrow q}$ by swapping the roles of q and k . The

Table 1: Performance on FineCVR-1M and TF-CoVR compared with baseline and existing methods. #Params denotes trainable parameters.

Method	Backbone	#Params	FineCVR-1M				TF-CoVR			
			R1	R5	R10	R50	mAP5	mAP10	mAP25	mAP50
CA	CLIP	–	11.94	37.97	55.76	88.43	3.74	4.53	5.78	6.79
MLP	CLIP	–	16.19	46.74	64.62	93.13	6.40	7.46	9.21	10.40
CIR	Uncertainty-R	CLIP	–	5.23	18.45	28.28	60.78	–	–	–
	Pic2word	CLIP	179M	8.10	25.79	38.32	71.01	–	–	–
	Combiner	CLIP	237M	19.57	49.46	65.38	92.11	–	–	–
	FreestyleRet	CLIP	–	20.39	52.98	68.37	93.03	–	–	–
CVR	TFR-CVR	CLIP	–	15.21	40.12	52.78	81.75	–	–	–
	CoVR	BLIP	142M	17.05	41.57	56.60	85.56	11.07	13.94	16.07
	ECDE-CoVR	BLIP	142M	22.81	47.63	60.12	90.32	13.03	15.90	18.62
	BSE-CoVR	BLIP	143M	24.40	51.01	63.42	91.21	–	–	–
	FDCA _{AIM}	AIM	289M	25.72	53.10	66.62	90.67	–	–	–
	FDCA _{CLIP}	CLIP	214M	25.84	55.84	70.23	<u>94.33</u>	14.43	17.02	19.54
	TF-CoVR-Base	AIM	85M	19.02	43.57	57.42	84.94	<u>21.85</u>	<u>24.23</u>	<u>26.47</u>
	HUD	BLIP-2	191M	26.85	55.92	70.28	94.28	17.13	21.81	25.45
	StaDy-CLIP	CLIP	7.78M	30.90	64.05	78.02	96.52	23.34	26.97	29.89
	Δ (vs. second best)		$\downarrow$ 183M / 77M	$\uparrow$ 4.05	$\uparrow$ 8.13	$\uparrow$ 7.74	$\uparrow$ 2.19	$\uparrow$ 1.49	$\uparrow$ 2.74	$\uparrow$ 3.50

final contrastive objective is:

$$\mathcal{L}_{con} = \frac{1}{2} (\mathcal{L}_{q \rightarrow k} + \mathcal{L}_{k \rightarrow q}), \quad (6)$$

where τ is a temperature parameter.

Inference. During inference, StaDy-CLIP can be used in two modes. In the end-to-end composition mode, the query branch is conditioned on the modification text, while gallery videos are conditioned on the Gallery Conditioning Prompt. The composed query representation is then matched against gallery representations in the shared embedding space. In another mode, StaDy-CLIP can also serve as a video encoder for downstream CVR frameworks. In this mode, videos are encoded through the gallery branch, and the resulting representations can be used by an external composition or retrieval module. Because they are learned through evidence composition under the composed retrieval objective, these representations encode retrieval-oriented evidence rather than generic video features.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on two benchmarks that reflect different textual retrieval needs. **FineCVR-1M** (Yue et al., 2025) (1,000,028 training / 10,043 test triplets) covers fine-grained static modifications, such as scene shifts and object changes, as well as dynamic modifications, such as similar- and distinct-action changes. **TF-CoVR** (Gupta et al.,

2025) (177,789 training / 473 test triplets) contains temporally fine-grained sports triplets and evaluates transfer to an action-centric domain. Results on an additional dataset are reported in Appendix B.1.

Evaluation Metrics. We report Recall@ K on FineCVR-1M (Yue et al., 2025) and mAP@ K on TF-CoVR (Gupta et al., 2025), consistent with their respective baselines. We further report static and dynamic scores on FineCVR-1M based on its query-type annotations, where the score is computed as the average of Recall@5 and Recall@10, ensuring a stable measure of the performance.

Implementation Details. We adopt CLIP ViT-B/16 as the frozen visual-textual backbone and attach a parallel trainable side pathway of width $d_s=320$ over its last 4 layers (layers 9–12); gradients are detached at the backbone interface so no CLIP parameters are updated, yielding a total trainable budget of 7.78 M parameters. More details are provided in the Appendix A.1.

4.2 Main Results

Results on FineCVR-1M and TF-CoVR. Table 1 compares StaDy-CLIP against CIR methods: Uncertainty-R (Chen et al., 2022), Pic2word (Saito et al., 2023), Combiner (Baldrati et al., 2022), FreestyleRet (Li et al., 2024), and CVR methods: CoVR (Ventura et al., 2024b), TF-CoVR-Base (Gupta et al., 2025), ECDE-CoVR (Thawakar et al., 2024), BSE-CoVR (Thawakar et al., 2025), FDCA (Yue et al., 2025), HUD (Chen et al., 2025),

Table 2: Ablation studies on key modules.

Configuration				Retrieval					
Insert Pos.	TA	g^t	g^d	R1	R5	R10	R50	Static	Dynamic
Backbone				16.90	48.92	66.55	94.42	67.38	51.42
Backbone	✓			19.96	54.92	72.58	96.09	72.94	57.75
Side Path	✓			28.35	61.01	74.78	94.61	74.78	64.66
Side Path	✓	✓		31.26	63.10	76.44	95.35	75.36	66.16
Side Path	✓		✓	31.52	61.95	75.24	95.24	75.05	64.41
Side Path	✓	✓	✓	30.90	64.05	78.02	96.52	76.47	67.49

Table 3: Effect of text-conditioned video modeling in the Side Network.

Input Pos.	R1	R5	R10	R50
12	29.88	62.17	77.01	96.51
11-12	31.01	63.38	77.25	96.38
10, 12	31.14	63.39	77.16	96.35
10-12	30.85	62.82	77.66	96.24
9-12	30.90	64.05	78.02	96.52

Table 4: Absolute gains on FineCVR-1M static/dynamic query scores from swapping in our encoder across existing methods.

Method	Static			Dynamic		
	Scene	Object	Overall	Sim. Act	Dist. Act	Overall
CA	+3.52	+5.18	+4.38	+5.06	+4.48	+4.63
MLP	+6.27	+6.43	+6.35	+9.50	+9.52	+9.51
CoVR	+5.23	+2.86	+4.01	+5.52	+9.20	+8.22
TF-CoVR-Base	+16.88	+14.90	+15.85	+12.91	+13.55	+13.38
FDCA	+3.85	+3.91	+3.88	+4.42	+1.78	+2.48

and TFR-CVR (Hummel et al., 2024).

On FineCVR-1M, StaDy-CLIP achieves the best performance across all Recall@K metrics, improving over the previous best method HUD (Chen et al., 2025) by about 4–8 absolute points on major retrieval metrics. These gains show that composing video evidence under the text is effective for fine-grained CVR, where different textual retrieval needs require different use of static evidence and dynamic evidence; On TF-CoVR, a temporally fine-grained benchmark, StaDy-CLIP further obtains the best results across all mAP@K metrics. Since TF-CoVR requires retrieval under subtle action-level modifications, the improvements indicate that the proposed text-conditioned composition can effectively incorporate dynamic evidence according to the textual retrieval need.

Notably, StaDy-CLIP achieves these improvements with only 7.78M trainable parameters, substantially fewer than competitive methods such as FDCA, HUD, and TF-CoVR-Base. The consistent gains on both benchmarks suggest that StaDy-CLIP is effective on both fine-grained static-dynamic and action-centric benchmark settings, providing an efficient way for CVR.

4.3 Ablations

Ablation on model architecture. Table 2 analyzes the architectural choices of StaDy-CLIP. Adding temporal modeling to the backbone improves over the CLIP baseline, especially on dy-

namic queries, showing that CVR benefits from motion-sensitive evidence beyond static evidence. Side-path insertion further outperforms backbone insertion, suggesting that dynamic evidence is more effective when constructed as a complementary source, while the frozen CLIP preserves CLIP-aligned static evidence as a stable anchor for text-conditioned evidence composition. We also ablate the components of the Text-Conditioned Composition Block. The Textual Modulation Gate improves the use of text-grounded guidance, while the Dynamic Evidence Gate controls where motion-sensitive evidence should be incorporated according to the textual retrieval need. Combining both gates yields the most balanced performance and the strongest static and dynamic query scores, indicating that effective CVR requires not only extracting static and dynamic evidence, but also selecting and modulating them under the modification text.

Effect of Text-Conditioned Video Modeling.

We examine where the modification text should be introduced in the side network. In this experiment, the selected TCC Blocks use the modification text as conditioning input, while the remaining blocks use a shared learnable prompt, following the gallery-side conditioning strategy. As shown in Table 3, conditioning only the last layer yields weaker performance, suggesting that the modification text should not be used only at the final representation stage. Introducing text into multiple late TCC Blocks generally improves retrieval, and the

Table 5: Analysis of Gallery Conditioning Prompt.

Gallery-side Conditioning	R1	R5	R10	R50	Static	Dynamic
<i>Non-learned (frozen, $K=4$)</i>						
Zero	30.86	63.66	77.00	95.84	75.00	65.99
Random Gaussian	26.39	58.50	73.04	95.01	72.49	62.05
Dummy Caption	25.51	60.74	75.52	96.26	75.86	63.07
<i>Learned prompt (length sweep)</i>						
$K = 2$	30.24	63.74	77.17	96.49	76.47	66.55
$K = 4$ (default)	30.90	64.05	78.02	96.52	76.47	67.49
$K = 8$	31.55	64.33	78.06	96.48	78.05	66.75

Table 6: Controlled inference cost comparison using CLIP ViT-B/16 dimensions.

Model	Query GFLOPs	Latency (ms)
CLIP ViT-B/16	275.6	34.1
TF-CoVR-Base	275.2	36.4
FDCA-CLIP	282.1	54.9
StaDy-CLIP	296.9	37.6

default 9–12 setting achieves the best R@5, R@10, and R@50 results. Although conditioning layers 10 and 12 gives the best R@1, the full 9–12 setting remains close and provides the most balanced overall performance. This indicates that multi-layer text conditioning helps StaDy-CLIP progressively select and modulate static and dynamic evidence according to the language-specified retrieval need.

Encoder transfer to existing CVR methods. Table 4 evaluates whether the gallery-side encoder learned by StaDy-CLIP can serve as a transferable video encoder for existing CVR methods. We replace the original visual encoder in several prior methods with our gallery-side video encoder, while keeping their composition modules unchanged. Here, Overall reports performance over all queries, Static and Dynamic correspond to the two major query types, and the other columns are their fine-grained subsets. Across different baselines, the substituted encoder improves the Overall score and most category-level and subset-level results. This shows that the learned representations are not specific to the StaDy-CLIP composition head, but provide retrieval-oriented static and dynamic evidence that benefits downstream composed modules.

4.4 Analysis

Analysis of Gallery Conditioning Prompt. Table 5 analyzes how gallery videos should be conditioned when modification text is unavailable. Fixed conditioning signals consistently underperform learnable prompts, indicating that gallery-side



Figure 3: Gate Visualization: The left heatmaps show the Dynamic Evidence Gate, while the right heatmaps show the Textual Modulation Gate.

conditioning should not be treated as a placeholder or manually specified prior. Instead, the conditioning input should be optimized with the retrieval objective, so that videos can organize static evidence and dynamic evidence within the same text-conditioned evidence composition framework. Increasing the learnable prompt length partially improves retrieval within the tested range, suggesting that additional conditioning capacity helps form more effective retrieval-oriented representations.

Gate Visualization Analysis. To understand how StaDy-CLIP performs text-conditioned evidence composition, we visualize gate responses in Figure 3. The results show that TCC Blocks do not uniformly enhance all visual or temporal information. The Dynamic Evidence Gate responds more strongly to motion-sensitive regions relevant to the modification text, while appearing to assign lower responses to motions and background changes. Meanwhile, the Textual Modulation Gate highlights regions requiring textual reinterpretation or emphasis, such as objects, interaction, or scene elements specified by the text. These complementary responses suggest that StaDy-CLIP constructs representations by jointly selecting dynamic evidence and modulating visual evidence according to the language-specified retrieval need. More visualizations are provided in Appendix C.5.

Efficiency Analysis. Table 6 reports the computational cost of StaDy-CLIP under a controlled protocol using CLIP ViT-B/16, 8 frames at 224² resolution, batch size 1, fp32 inference, and one NVIDIA RTX 3090 GPU. The TCC cross-attention is lightweight because it is applied only in the last four CLIP layers and uses the side-pathway width $d_s = 320$, rather than the backbone width 768. Across all four TCC Blocks, cross-attention adds approximately 3.3 GFLOPs, about 1.1% of one full query encoding. The entire trainable side pathway adds 27.2 GFLOPs, increasing the frozen CLIP video encode from 269.7 to 296.9 GFLOPs.

Although StaDy-CLIP performs text-conditioned video encoding, gallery embeddings do not depend on the query text. Gallery videos are conditioned by the fixed learned prompt after training and can therefore be precomputed offline, as in text-agnostic retrieval systems. At query time, text-conditioned computation is applied only to the reference video. Peak activation memory increases only slightly, from 847.4 MiB for the frozen CLIP visual encoder to 851.6 MiB for StaDy-CLIP under the same protocol.

5 Conclusion

In this paper, we studied composed video retrieval as a problem of text-conditioned video modeling, where the modification text determines how static and dynamic evidence should be organized in the reference video. Based on this view, we proposed StaDy-CLIP, a dual-pathway framework that preserves CLIP-aligned static evidence, constructs dynamic evidence through a lightweight side pathway, and performs text-conditioned evidence composition via Text-Conditioned Composition Blocks. A learnable Gallery Conditioning Prompt further enables gallery videos to be encoded within the same framework. Experiments on FineCVR-1M and TF-CoVR show that StaDy-CLIP achieves state-of-the-art performance with only 7.78M trainable parameters, while ablations and encoder-transfer results support the effectiveness of the proposed design. These results highlight text-conditioned evidence composition as an effective and parameter-efficient direction for composed video retrieval.

Limitations

This work has two main limitations. First, although StaDy-CLIP achieves strong performance with a compact trainable budget, its zero-shot domain gen-

eralization remains limited under large distribution shifts, which is reflected by EgoCVR, an egocentric human activity test dataset. Since the CLIP backbone is frozen and only a lightweight side pathway is trained, the model may have limited adaptation capacity for domain shift. Improving robustness beyond the current parameter-efficient setting remains an important future direction.

Second, the current conditioning mechanism mainly captures semantic correspondence between the modification text and visual evidence. Although fine-grained results show that StaDy-CLIP handles both static and dynamic query types, dynamic/action-centered queries remain more difficult. These cases often require joint reasoning over actions, environments, interacted objects, and interaction-induced state changes, which are not explicitly modeled in the current framework.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62571379). We thank the anonymous reviewers for their constructive comments and suggestions, which helped improve the quality of this paper.

References

- Muhammad Umer Anwaar, Egor Labintcev, and Martin Kleinstueber. 2021. Compositional learning of image-text query for image retrieval. In *Proceedings of the IEEE/CVF Winter conference on Applications of Computer Vision*, pages 1140–1149.
- Mothilal Asokan, Kebin Wu, and Fatima Albreiki. 2025. [Finelip: Extending clip’s reach via fine-grained alignment with longer text inputs](#). *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14495–14504.
- Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. 2023. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15338–15347.
- Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21466–21474.
- Zhuo Cao, Bingqing Zhang, Heming Du, Xin Yu, Xue Li, and Sen Wang. 2024. [Flashvtg: Feature layering and adaptive score handling network for video temporal grounding](#). *2025 IEEE/CVF Winter Conference*

- on *Applications of Computer Vision (WACV)*, pages 9226–9236.
- Yanbei Chen, Shaogang Gong, and Loris Bazzani. 2020. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3001–3011.
- Yiyang Chen, Zhedong Zheng, Wei Ji, Leigang Qu, and Tat-Seng Chua. 2022. Composed image retrieval with text feedback via multi-grained uncertainty regularization. *arXiv preprint arXiv:2211.07394*.
- Zhiwei Chen, Yupeng Hu, Zixu Li, Zhiheng Fu, Haokun Wen, and Weili Guan. 2025. Hud: Hierarchical uncertainty-aware disambiguation network for composed video retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 6143–6152.
- Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. 2021. [Transvg: End-to-end visual grounding with transformers](#). 2021 *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1749–1759.
- J. Gao, Chen Sun, Zhenheng Yang, and Ramakant Nevatia. 2017. [Tall: Temporal activity localization via language query](#). 2017 *IEEE International Conference on Computer Vision (ICCV)*, pages 5277–5285.
- Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. 2024a. Language-only training of zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13225–13234.
- Xin Gu, Hengrui Fan, Yan Huang, Tiejian Luo, and Libo Zhang. 2024b. [Context-guided spatio-temporal video grounding](#). 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18330–18339.
- Animesh Gupta, Jay Parmar, Ishan Rajendrakumar Dave, and Mubarak Shah. 2025. From play to replay: Composed video retrieval for temporally fine-grained videos. *arXiv preprint arXiv:2506.05274*.
- Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. 2017. [Localizing moments in video with natural language](#). 2017 *IEEE International Conference on Computer Vision (ICCV)*, pages 5804–5813.
- Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. 2023. [Vtimellm: Empower llm to grasp video moments](#). 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14271–14280.
- Thomas Hummel, Shyamgopal Karthik, Mariana-Iuliana Georgescu, and Zeynep Akata. 2024. Egocvr: An egocentric benchmark for fine-grained composed video retrieval. In *European Conference on Computer Vision*, pages 1–17. Springer.
- Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. 2021. [Mdetr - modulated detection for end-to-end multi-modal understanding](#). 2021 *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1760–1770.
- Seungmin Lee, Dongwan Kim, and Bohyung Han. 2021. Cosmo: Content-style modulation for image retrieval with text feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 802–812.
- Jie Lei, Tamara L. Berg, and Mohit Bansal. 2021. [Detecting moments and highlights in videos via natural language queries](#). In *Neural Information Processing Systems*.
- Hao Li, Yanhao Jia, Peng Jin, Zesen Cheng, Kehan Li, Jialu Sui, Chang Liu, and Li Yuan. 2024. Freestyleret: retrieving images from style-diversified queries. In *European Conference on Computer Vision*, pages 258–274. Springer.
- Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. 2021. [Grounded language-image pre-training](#). 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10955–10965.
- Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chun yue Li, Jianwei Yang, Hang Su, Jun-Juan Zhu, and Lei Zhang. 2023. [Grounding dino: Marrying dino with grounded pre-training for open-set object detection](#). *ArXiv*, abs/2303.05499.
- Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2125–2134.
- Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. 2022. [X-clip: End-to-end multi-grained contrastive learning for video-text retrieval](#). *Proceedings of the 30th ACM International Conference on Multimedia*.
- Junhua Mao, Jonathan Huang, Alexander Toshev, Oana-Maria Camburu, Alan Loddon Yuille, and Kevin P. Murphy. 2015. [Generation and comprehension of unambiguous object descriptions](#). 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20.
- Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. 2022. St-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35:26462–26477.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models

- from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.
- Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2023. [Timechat: A time-sensitive multi-modal large language model for long video understanding](#). *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14313–14323.
- Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. 2023. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19305–19314.
- Yucheng Suo, Fan Ma, Linchao Zhu, and Yi Yang. 2024. Knowledge-enhanced dual-stream zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26951–26962.
- Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Yue Hu, and Qi Wu. 2024. 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, pages 5180–5188.
- Omkar Thawakar, Dmitry Demidov, Ritesh Thawkar, Rao Muhammad Anwer, Mubarak Shah, Fahad Shahbaz Khan, and Salman Khan. 2025. Beyond simple edits: Composed video retrieval with dense modifications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20435–20444.
- Omkar Thawakar, Muzammal Naseer, Rao Muhammad Anwer, Salman Khan, Michael Felsberg, Mubarak Shah, and Fahad Shahbaz Khan. 2024. Composed video retrieval via enriched context and discriminative embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26896–26906.
- Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. 2024a. Covr-2: Automatic data construction for composed video retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. 2024b. Covr: Learning composed video retrieval from web video captions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5270–5279.
- Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6439–6448.
- Bin Wang, Wentong Li, Wenqian Wang, Mingliang Gao, Runmin Cong, and Wei Zhang. 2024a. Tds-clip: Temporal difference side network for efficient videoaction recognition. *arXiv preprint arXiv:2408.10688*.
- Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. 2024b. [Grouned-videollm: Sharpening fine-grained temporal grounding in video large language models](#). *ArXiv*, abs/2410.03290.
- Hao Wang, Fang Liu, Licheng Jiao, Jiahao Wang, Shuo Li, Lingling Li, Puhua Chen, and Xu Liu. 2025. Vision-by-prompt: Context-aware dual prompts for composed video retrieval. *Pattern Recognition*, page 112378.
- Mengmeng Wang, Jiazheng Xing, Boyuan Jiang, Jun Chen, Jianbiao Mei, Xingxing Zuo, Guang Dai, Jingdong Wang, and Yong Liu. 2024c. A multimodal, multi-task adapting framework for video action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5517–5525.
- Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2023. [Unified coarse-to-fine alignment for video-text retrieval](#). *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2804–2815.
- Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022. [Tubedetr: Spatio-temporal video grounding with transformers](#). *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16421–16432.
- Lewei Yao, Runhu Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. [Filip: Fine-grained interactive language-image pre-training](#). *ArXiv*, abs/2111.07783.
- WU Yue, 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 *The Thirteenth International Conference on Learning Representations*, volume 3, page 4.
- Xiangyun Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, Yali Wang, Yu Qiao, and Limin Wang. 2024. [Timesuite: Improving mllms for long video understanding via grounded tuning](#). *ArXiv*, abs/2410.19702.
- Heng Zhao, Yinjie Zhao, Bihan Wen, Yew Soon Ong, and Joey Tianyi Zhou. 2024. [Video-text prompting for weakly supervised spatio-temporal video grounding](#). In *Conference on Empirical Methods in Natural Language Processing*.

Appendix Overview. The appendix is organized as follows:

- Appendix A provides implementation and evaluation details, including training settings, dataset statistics, evaluation protocols, licenses, and intended use.
- Appendix B reports additional main results, including cross-domain evaluation on EgoCVR, fine-grained results on FineCVR-1M, and results on the multi-text query subset.
- Appendix C presents further analysis and visualization, including frame-shuffle and text-perturbation probes, the effect of the static strength coefficient, and additional gate visualizations.
- Appendix D provides supplementary ablation studies on the dynamic pathway width, TCC Block attachment position, temporal evidence adapters, and the number of input frames.

A Implementation and Evaluation Details

A.1 Training Details

Our encoder is trained separately on each benchmark using a symmetric InfoNCE objective with temperature $\tau=0.01$. We optimize the trainable parameters with AdamW, using a weight decay of 0.05, a batch size of 128, an initial learning rate of 2×10^{-4} , and a cosine decay schedule for 30 epochs. The CLIP backbone is kept frozen throughout training, and only the proposed lightweight components are updated. All experiments are conducted on 4 NVIDIA RTX 4090 GPUs with 24 GB memory each. A full training run on FineCVR-1M takes approximately 40 hours end-to-end.

For the downstream encoder-swap evaluation in Table 4, we use the gallery-side StaDy-CLIP encoder as a frozen video encoder for existing CVR frameworks. Specifically, we replace the original visual encoder of each baseline, including MLP, CoVR, TF-CoVR-Base, and FDCA, with our pre-trained encoder, while keeping the corresponding composition module unchanged. We then retrain only the composition module following each baseline’s released optimizer, learning-rate schedule, and loss configuration. This protocol isolates the effect of the video encoder and ensures that the comparison does not benefit from additional changes to the downstream retrieval framework.

Optimization hyperparameters (learning rate, weight decay, batch size, training epochs) were chosen based on FineCVR-1M validation R@1 without extensive grid search; architectural hyperparameters are studied in the ablations of Appendix C.1, C.4, D.1 and D.2.

A.2 Parameter Accounting

To make the efficiency comparison transparent, Table 7 reports total, frozen, and trainable parameters for representative CVR methods. In the main result table, #Params denotes trainable parameters for all methods.

The parameter counts are obtained from released implementations, official checkpoints, or our reproductions following the original training settings. StaDy-CLIP keeps the CLIP backbone frozen and trains only the side pathway, TCC Blocks, Gallery Conditioning Prompt, and projection layers, resulting in 7.78M trainable parameters.

A.3 Dataset and Evaluation Details

FineCVR-1M provides fine-grained annotations of textual retrieval needs, allowing us to evaluate StaDy-CLIP beyond overall retrieval metrics. Following the dataset definition, we group test queries into static and dynamic categories. Static queries require the model to preserve or modify appearance-related evidence, including object changes and scene changes, which account for 19.4% and 20.0% of the test set, respectively. Dynamic queries require reasoning over motion-sensitive evidence and include similar-action changes and distinct-action changes. Similar-action changes account for 16.2% of the test set and refer to cases where only part of the action semantics changes, such as the interacted object, action direction, or execution manner, while distinct-action changes account for 44.5% and involve substantially different action types. Beyond the results provided in the main paper, we further provide per-category Recall@K results to examine how StaDy-CLIP handles different language-specified retrieval needs in Appendix B.2.

Licenses. CLIP (Radford et al., 2021) is released by OpenAI under the MIT License. spaCy (used in the POS-based text perturbation probe in Appendix C.3, with the en_core_web_sm model, v3.7) is released under the MIT License. FineCVR-1M (Yue et al., 2025), TF-CoVR (Gupta et al., 2025), and EgoCVR (Hummel et al., 2024) are released for non-commercial research use under their

Table 7: Parameter breakdown of representative methods.

Method	Backbone	Total (M)	Frozen (M)	Trainable (M)
CoVR	BLIP ViT-L/16@384	446.1	303.9	142.2
FDCA _{CLIP}	CLIP RN50x4	392.3	178.3	214.0
TF-CoVR-Base	BERT-base + offline features	194.4	108.9	85.5
HUD	BLIP-2 EVA ViT-g/14@364	1201.8	1010.6	191.2
StaNy-CLIP	CLIP ViT-B/16	157.4	149.6	7.78

respective dataset licenses; we use each dataset strictly in accordance with the terms stated by its original creators.

Intended Use. All datasets are used as research benchmarks for composed video retrieval, consistent with their stated research-only access conditions. StaDy-CLIP model checkpoints, if released, will be released under a non-commercial research license; we do not intend the model for deployment in surveillance, person identification, or any production retrieval system over non-consensual video data.

B Additional Main Results

B.1 Cross-domain Evaluation on EgoCVR

We further evaluate StaDy-CLIP on EgoCVR to examine its cross-domain generalization ability. EgoCVR is a test-only benchmark containing 2,295 triplets collected from egocentric human activities. Compared with the third-person and web-style videos in FineCVR-1M, EgoCVR exhibits a clear domain shift, including stronger camera motion, frequent hand-object interactions, partial observations, and viewpoint-dependent temporal patterns. It therefore provides a challenging setting for evaluating whether the text-conditioned evidence composition learned on FineCVR-1M can transfer to a distinct first-person video domain.

We directly evaluate the StaDy-CLIP model trained on FineCVR-1M without further training on EgoCVR. The results show that StaDy-CLIP remains comparable to other training-based CVR methods, but still underperforms TFR-CVR, which is specifically designed for training-free composed video retrieval. This gap suggests that although StaDy-CLIP effectively organizes static and dynamic evidence within the training distribution, transferring such evidence composition to egocentric videos remains challenging. In addition, StaDy-CLIP keeps the CLIP backbone frozen and updates only a lightweight dynamic pathway with 7.78M trainable parameters, which supports parameter-

efficient in-domain retrieval but may limit adaptation to large cross-domain variations. These results indicate that improving domain generalization to egocentric videos remains an important direction for future CVR research.

B.2 Fine-grained Results on FineCVR-1M

To complement the static and dynamic scores in Table 1, we report fine-grained retrieval results on FineCVR-1M across all Recall@K metrics. Table 11 provides the detailed results. StaDy-CLIP achieves strong performance across all query types, while also reflecting their different retrieval difficulties. For static queries, the model obtains 39.74 R@1 on scene changes and 32.32 R@1 on object changes. The strong scene-change result suggests that StaDy-CLIP effectively preserves and adjusts global scene semantics, while the high R@10 and R@50 on object-change queries indicate that fine-grained object cues can be reliably captured when broader candidate rankings are considered.

For dynamic queries, StaDy-CLIP obtains 27.85 R@1 on similar-action modifications and 27.32 R@1 on distinct-action modifications. These action-related queries are more challenging because they require reasoning over temporal evolution, action phase, and motion semantics rather than relying mainly on appearance-level correspondence. StaDy-CLIP maintains stable performance across both dynamic categories, showing that the dynamic pathway and TCC Blocks provide effective temporal evidence for action-sensitive composed retrieval. The comparable results on similar-action and distinct-action queries further suggest that the model can handle fine-grained temporal variations conditioned on the modification text.

B.3 Results on Multi-text Query Subset.

We further evaluate on a text-dependent subset of FineCVR-1M: reference videos paired with multiple modification texts, where the same video forms different queries under different textual conditions. This subset is harder than single-query samples —

Table 8: Results on EgoCVR test dataset.

Method	Global			Local		
	R1	R5	R10	R1	R2	R3
CoVR (Ventura et al., 2024b)	5.4	15.2	24.3	33.1	49.5	62.9
FDCA _{BLIP} (Yue et al., 2025)	8.7	19.2	27.1	<u>36.3</u>	<u>51.1</u>	<u>63.8</u>
TFR-CVR (Hummel et al., 2024)	14.1	39.5	54.4	44.2	61.0	73.2
StaDy-CLIP	8.1	<u>28.3</u>	<u>46.9</u>	28.6	45.8	59.7

Table 9: Results on the multi-modification subset of FineCVR-1M.

Method	R1	R5	R10	R50
CA	6.72	30.18	49.69	87.64
MLP	10.38	39.31	58.29	92.36
CoVR	11.93	34.83	49.29	83.16
TF-CoVR-Base	12.69	35.26	51.24	83.62
FDCA _{CLIP}	16.93	46.02	62.86	92.92
HUD	18.22	51.57	69.95	95.55
StaDy-CLIP	22.27	53.15	68.73	94.17

the reference video alone is insufficient to determine the target, so the model must rely more on the modification text to decide which visual evidence to preserve, modify, or suppress. As shown in Table 9, all methods perform worse on this subset than on the full test set, confirming that multi-text samples pose a harder retrieval setting. StaDy-CLIP achieves the best R@1 and R@5 on this subset, while remaining competitive at larger cutoffs, supporting our core design: injecting modification text into evidence composition lets the model distinguish different retrieval intents from the same reference video and construct more discriminative query representations.

C Analysis and Visualization

C.1 Static-Dynamic Pathway Zero-Out Analysis

To directly examine whether the two pathways provide complementary static and dynamic evidence, we conduct pathway zero-out analysis on the FineCVR-1M test set. For the *backbone only* variant, we remove the dynamic pathway by setting $\alpha = 0$ and $g_t^d = 0$, so retrieval relies on the frozen backbone-side evidence. For the *side-path only* variant, frozen ViT features are used only to feed the temporal adapters, and only the side-path temporal evidence enters the final representation.

Removing the dynamic pathway causes a larger drop on dynamic queries than on static queries,

Table 10: Pathway zero-out analysis on FineCVR-1M. Static and Dynamic report query-type R@1.

Variant	Overall R@1	Static R@1	Dynamic R@1
Full model	30.90	35.88	27.46
Backbone only	23.61	30.64 (-5.24)	19.04 (-8.87)
Side-path only	15.27	18.28 (-17.60)	13.32 (-14.14)

indicating that the side pathway captures motion-sensitive evidence that the frozen backbone alone cannot fully provide. Conversely, removing the static pathway causes a larger drop on static queries, showing that the frozen CLIP pathway remains important for appearance-, object-, and scene-related evidence. These results support the dual-pathway design as a complementary decomposition of static and dynamic evidence rather than a generic increase in model capacity.

C.2 Frame-shuffle Analysis.

To further examine whether StaDy-CLIP uses temporal evidence, we perturb the temporal order of video frames during evaluation. This diagnostic setting disrupts temporal structure while largely preserving frame-level visual content. As shown in Table 12, frame shuffling causes a larger drop on dynamic queries than on static queries. This is expected because dynamic modifications depend more on motion patterns, action progression, and temporal transitions, whereas static queries rely more on appearance, object, and scene semantics that can remain visible after temporal permutation.

We also observe that perturbing gallery videos causes a larger degradation than perturbing query videos. This suggests that the query branch is more robust under textual anchoring, since the modification text provides an explicit semantic condition for organizing reference-video evidence even when temporal order is partially disrupted. In contrast, the gallery branch is conditioned by the shared Gallery Conditioning Prompt, which provides a

Table 11: StaDy-CLIP’s Fine-grained performance on FineCVR-1M under different query types.

Query Type	R1	R5	R10	R50
Scene Change	39.74	70.26	80.56	95.28
Object Change	32.32	67.63	82.34	97.91
Static	35.88	68.89	81.49	96.65
Similar Action	27.85	61.80	76.46	96.80
Distinct Action	27.32	59.69	73.28	94.89
Dynamic	27.46	60.25	74.13	95.40

Table 12: Frame-shuffle probe.

Perturb.	R1	R5	R10	R50	Static	Dynamic
None	30.90	64.05	78.02	96.52	76.47	67.49
Q-shuffle	30.79	63.78	78.16	96.43	75.87	67.64
T-shuffle	27.37	61.27	75.91	95.67	74.02	64.22
Both-shuffle	27.78	61.32	75.93	95.68	74.10	63.90
Q-reverse	30.81	63.72	77.79	96.29	75.67	67.56
T-reverse	26.36	59.71	74.38	95.18	73.81	62.66

task-level prior but not instance-specific semantic guidance. Gallery representations are therefore more sensitive to disruptions in temporal structure. Overall, this analysis shows that StaDy-CLIP exploits temporal ordering especially for dynamic retrieval, while the asymmetric effect between query and gallery branches highlights the different roles of modification text and the Gallery Conditioning Prompt in text-conditioned evidence composition.

C.3 Text Perturbation Analysis

We conduct an inference-only text perturbation analysis to examine whether StaDy-CLIP relies on the linguistic structure and semantic content of the modification text. We first encode all gallery videos with the original Gallery Conditioning Prompt and keep the gallery embeddings fixed. Then, we perturb only the modification text, re-encode the query branch with the same trained model, and compute Recall@K against the fixed gallery embeddings.

We consider two types of perturbations. For word-order perturbation, we shuffle the words in the modification text or reverse their order, which largely preserves lexical content while disrupting linguistic structure. For POS-based deletion, we either drop function words and keep content words, or drop content words and keep function words (POS tagging is performed with spaCy v3.7 using the `en_core_web_sm` model). As shown in Table 13, all perturbations degrade retrieval performance. Shuffling and reversing reduce R@1 from

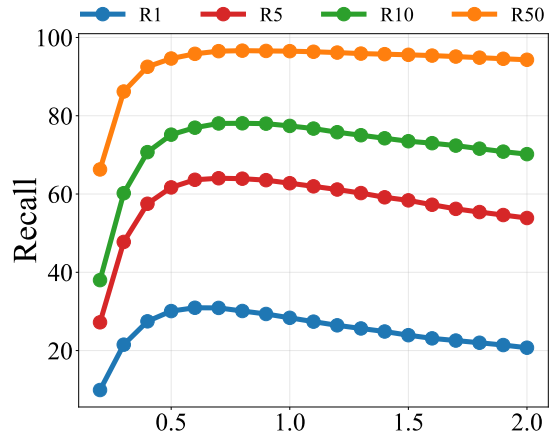


Figure 4: Effect of the static strength coefficient α on retrieval results, emphasizing the importance of balance.

30.90 to 23.83 and 23.11, respectively, indicating that StaDy-CLIP does not rely only on unordered lexical cues. Dropping content words causes the largest decline, reducing R@1 to 11.07, while dropping function words leads to a milder but still clear drop. These results show that content words provide the main semantic evidence, while word order and functional cues also help specify the retrieval need. Overall, the probe supports that StaDy-CLIP uses the modification text as a linguistic conditioning signal for evidence composition, rather than as a shallow bag of keywords.

C.4 Static Strength Coefficient

Figure 4 analyzes the effect of the static strength coefficient α in Eq. 1. This coefficient controls how much CLIP-aligned static evidence is injected into the side representation before temporal evidence construction. As α increases from a small value, retrieval performance first improves, indicating that pretrained static semantics provide an important anchor for text-conditioned evidence composition. These CLIP-aligned cues help preserve object, scene, and attribute information when the model selects and modulates evidence under the modification text.

However, performance decreases when α becomes overly large. This suggests that excessive dependence on backbone tokens can dominate the side representation and suppress the contribution of learned dynamic evidence. Therefore, effective CVR requires a proper balance between pretrained static evidence and side-path dynamic evidence. This observation supports the design of TCC Blocks, which compose static and dynamic evidence under text-grounded guidance rather than

Table 13: Text-perturbation probe on FineCVR-1M.

Variant	R1	R5	R10	R50
<i>Baseline</i>				
Original text	30.90	64.05	78.02	96.52
<i>Order perturbation</i>				
Shuffle	23.83	55.75	71.79	94.61
Reverse	23.11	54.79	70.95	94.35
<i>POS-based deletion</i>				
Drop Function	26.46	59.30	74.34	95.60
Drop Content	11.07	34.96	51.62	84.55

Table 14: Studies on side pathway width d_s .

d_s	#Params	R1	R5	R10	R50
128	1.49M	25.90	60.25	75.59	96.28
320	7.78M	30.90	64.05	78.02	96.52
512	19.83M	32.15	64.12	78.26	96.49

Table 15: Studies on TCC Block attachment positions.

Layers	R1	R5	R10	R50
1–4 (early)	18.41	39.71	51.35	78.19
5–8 (mid)	21.71	51.46	67.02	92.90
9–12 (late)	30.90	64.05	78.02	96.52

relying on either source alone.

C.5 Additional Gate Visualization Results

We provide additional gate visualization results in Figure 5 to complement the qualitative analysis in the main paper. Across diverse retrieval cases, the Dynamic Evidence Gate consistently responds more strongly to motion-sensitive regions related to the modification text, such as action-relevant body parts, interaction areas, or moving objects, while suppressing irrelevant background changes. The Textual Modulation Gate instead highlights visual regions whose semantics need to be emphasized or reinterpreted under the textual condition, including objects, scenes, and interaction targets. These examples further show that the two gates play complementary roles: the Dynamic Evidence Gate selects temporal evidence, while the Textual Modulation Gate adjusts visual evidence according to the language-specified retrieval need. This supports that StaDy-CLIP performs selective text-conditioned evidence composition rather than uniformly enhancing video features.

D Ablation Studies

D.1 Dynamic Pathway Width

We study the effect of the dynamic pathway width by varying the side-path feature dimension while keeping the CLIP backbone frozen and all other settings unchanged. This experiment examines how the capacity of the trainable dynamic pathway influences text-conditioned evidence composition. As shown in Table 14, retrieval performance improves

as the pathway width increases. A larger dimension provides stronger capacity for constructing dynamic evidence and composing it with CLIP-aligned static evidence under text-grounded guidance. This is especially beneficial for dynamic queries, where motion-sensitive cues need to be selected according to the modification text. The improvement on static queries also suggests that a wider pathway helps preserve and reorganize static semantics during composition, rather than only enhancing temporal modeling. Overall, this ablation reflects a capacity-efficiency trade-off: increasing the width improves evidence composition, while the default setting maintains strong retrieval performance with a compact trainable budget.

D.2 TCC Block Attachment Position

We further study where the TCC Blocks should be attached to the frozen CLIP visual encoder. We compare three variants that connect the dynamic pathway to early, middle, and late CLIP layers, while keeping the number of TCC Blocks and all other settings unchanged. As shown in Table 15, the attachment position substantially affects retrieval performance. Early-layer attachment gives the weakest results, middle-layer attachment improves performance, and the best results are obtained when TCC Blocks are attached to the last four layers, reaching 30.90, 64.05, 78.02, and 96.52 on R@1, R@5, R@10, and R@50, respectively.

This result is consistent with the design motivation of StaDy-CLIP. Early CLIP layers mainly encode low-level visual patterns, which are less

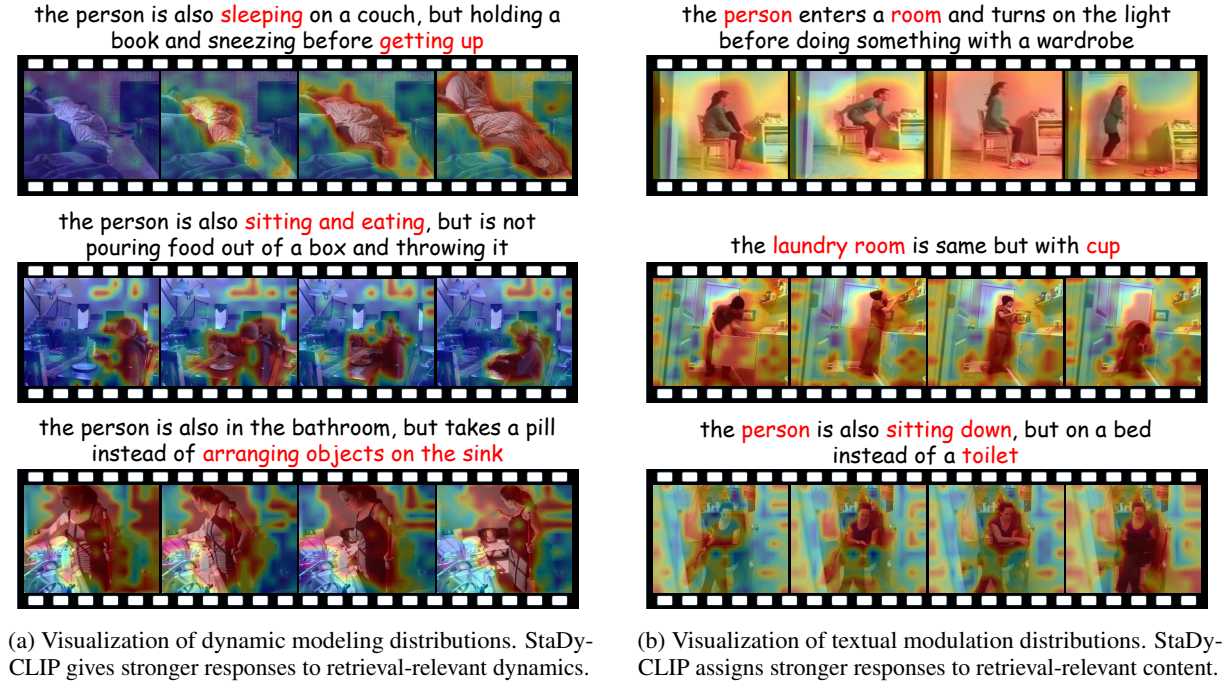


Figure 5: Additional Gate Visualization Results

Table 16: Generalization across Temporal Adapters.

Temporal Adapter	R1	R5	R10	R50
Parameter Free	29.18	61.13	75.84	95.32
ST-Adapter	30.34	64.02	77.36	95.71
TED-Adapter	30.13	63.91	77.12	95.72
TD-Adapter	30.90	64.05	78.02	96.52

Table 17: Studies on the number of input frames T .

#Num	R1	R5	R10	R50	Static	Dynamic
1	10.88	28.43	38.73	62.74	45.81	25.64
2	11.52	30.58	42.28	69.96	46.12	29.94
4	19.93	47.63	62.06	88.02	54.75	48.91
8	30.90	64.05	78.02	96.52	76.47	67.49

directly aligned with the retrieval need specified by the modification text. In contrast, late CLIP layers contain richer object-, scene-, and action-related semantics, providing stronger static anchors for text-conditioned evidence composition. Attaching TCC Blocks to later layers thus allows the dynamic pathway to introduce temporal evidence on top of semantically meaningful visual tokens, while avoiding unnecessary disturbance to lower-level CLIP representations.

D.3 Temporal Evidence Adapter

Table 16 examines whether StaDy-CLIP depends on a specific temporal evidence adapter. We replace the adapter in the TCC Block with both parameter-free and learnable alternatives. In the parameter-free variant, dynamic evidence is directly derived from adjacent-frame differences, while the other variants use lightweight learnable temporal adapters, including ST-Adapter (Pan et al., 2022), TED-Adapter (Wang et al., 2024c), and TD-Adapter (Wang et al., 2024a). StaDy-CLIP remains

consistently effective across all choices, with each variant achieving strong retrieval performance. The TD-Adapter obtains the best results, reaching 30.90 R@1 and 78.02 R@10, but the performance gap among learnable adapters is relatively small.

These results suggest that the gains of StaDy-CLIP do not come from a particular temporal module alone. Instead, the key factor is the text-conditioned evidence composition framework, where TCC Blocks ground the modification text onto visual evidence and selectively incorporate dynamic evidence according to the textual retrieval need. A stronger adapter can provide better temporal evidence, but the composition mechanism remains effective across different forms of dynamic evidence.

D.4 Number of Frames

We study the effect of the number of input frames on FineCVR-1M while keeping the model architecture and training settings unchanged. As illustrated in Table 17, using fewer sampled frames

leads to a clear performance drop on both static and dynamic queries, with a larger degradation on dynamic queries. For static queries, this suggests that relevant objects, attributes, or scene details may appear only in limited temporal segments, and sparse sampling can miss the visual cues needed to preserve or modify the correct static evidence. For dynamic queries, the effect is more direct, since motion patterns, action transitions, and fine-grained temporal changes require sufficient temporal observations. Reducing frame coverage therefore weakens the dynamic evidence available to the dynamic pathway and limits text-conditioned evidence composition.

These results indicate that StaDy-CLIP effectively uses multi-frame video context rather than relying only on isolated frame-level semantics. Richer temporal coverage helps the model compose static and dynamic evidence under the modification text, and the stronger sensitivity of dynamic queries further supports the role of temporal evidence in composed video retrieval.