MORSE: LEARNING COMPONENT-WISE RESIDUAL TRANSFORMATIONS FOR COMPOSED VIDEO RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Composed Video Retrieval (CoVR) retrieves a target video given a reference video and a text instruction that describes the change. Existing methods learn composed representations through target alignment, but global target alignment alone is insufficient to distinguish the intended target from visually similar candidates that fail to satisfy the requested modification. How to coordinate changes across the components of the reference representation to capture the instruction remains underexplored. We propose MORSE, which formulates target-feature prediction as context-conditioned corrections over an ordered residual decomposition. This decomposition organizes a reference feature into additive components with cumulative dependencies. Residual Transformation predicts differentiated corrections conditioned on the instruction and each component's residual context, then aggregates them with the reference feature to form a target-feature prediction. Target-supervised Refinement constrains the cumulative composition of the transformed components using the target decomposition under shared codebooks, connecting overall target discrimination with structural supervision without intermediate change annotations. Compared with recent methods on each benchmark, MORSE achieves relative improvements of 17.81% in R@1 on FineCVR-1M and 10.25% in mAP@50 on TF-CoVR. Further analyses support the effectiveness of context-conditioned corrections and target-side structural supervision in distinguishing instruction-specific changes. Code will be available upon acceptance.

1 INTRODUCTION

Composed Video Retrieval (CoVR) retrieves a target video from a gallery given a reference video and a text instruction that specifies the desired change (Ventura et al., 2024b;a; Gupta et al., 2025). Unlike conventional video retrieval, the instruction does not describe the target in isolation; its meaning is defined relative to the reference video. An effective composed feature must retain relevant reference information, incorporate the instruction, and distinguish the requested target from otherwise similar candidates (Wu et al., 2025; Gupta et al., 2025). Consequently, retrieval requires more than broad visual relevance: it must resolve the particular attribute, action, or relation specified by the text instruction.

Existing approaches improve instruction understanding and target alignment by enriching the context for the query (Thawakar et al., 2024), disentangling the requirements expressed in text instructions (Wu et al., 2025), and calibrating visual and textual contributions to the composed feature (Li et al., 2026a). However, identifying feature contributions and modeling feature interactions do not, by themselves, specify how the components of a reference representation should jointly realize the requested change. A single instruction may require different corrections across components, and the effectiveness of the resulting target representation depends on how these corrections are coordinated within the reference context. Multiple features can provide rich matching evidence without necessarily defining the compositional relationships needed for context-conditioned correction and cumulative supervision. Beyond aligning the composed feature as a whole with the target, we therefore ask how to establish explicit dependencies among reference components so that the requested change can be realized through coordinated corrections conditioned on their respective contexts.

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.

To address this challenge, we propose MORSE, which explicitly structures the dependencies among reference components and learns context-conditioned corrections that jointly realize the change specified by the text instruction. MORSE is a target-feature correction framework based on an ordered residual decomposition. Its central idea is to organize a single reference video feature into additive components linked by cumulative dependencies, learn instruction-conditioned corrections within their residual contexts, and supervise their composition using the target video's decomposition in the same representation space. The reference decomposition supplies the increments and context for componentwise correction, while the target decomposition provides cumulative representations that supervise how these corrections combine. MORSE thus connects overall target discrimination with the composition of the underlying feature changes.

Specifically, to ground each correction in its reference context, we propose a Residual Transformation module, which decomposes a single video feature into a sequence of full-dimensional codeword increments through residual quantization (RQ), where each level approximates the information left unrepresented by preceding levels (Chen et al., 2010). The accumulated prefix and remaining residual establish a hierarchy from current approximations to remaining details (Huijben et al., 2024). Thus, our Residual Transformation module predicts continuous corrections for each level's increment, conditioned on the current reference codeword, textual instruction, and prefix–residual context. The module generates differentiated updates across levels and all corrections are added to the reference feature to produce a target-feature prediction.

To constrain both the overall target prediction and the composition of its componentwise corrections, we design target-supervised Refinement. MORSE projects the transformed increments onto the shared codebooks, accumulates the selected codewords into predicted prefixes, and aligns them with target prefixes obtained from the same codebooks. Contrastive learning on the continuous target prediction guides the overall transformation toward the retrieval target; cumulative prefix alignment constrains how the corrections are composed; and learning on the final quantized prefix supervises target discrimination after projection. By incorporating residual structure into composition learning, MORSE learns both the overall change required by the instruction and how that change is realized across reference components. Our main contributions are as follows:

- We propose MORSE, an ordered residual correction framework for CoVR. By organizing a single video feature into additive components with cumulative dependencies, MORSE connects reference-conditioned transformation with target-side structural supervision within a shared representation.
- We introduce a Transformation module that learns differentiated corrections from the current codeword, the text instruction, and the original prefix–residual context, then aggregates these componentwise corrections into a target-feature prediction.
- We design a Refinement mechanism that constrains the transformation and its structured composition by prefix alignment, providing supervision for target-feature prediction without intermediate change annotations.
- We demonstrate the effectiveness of MORSE across multiple public benchmarks. MORSE outperforms the compared methods on WebVid-CoVR, FineCVR-1M, and TF-CoVR. Ablation studies and qualitative analyses further substantiate the effectiveness of residual-conditioned transformation and target-supervised Refinement.

2 Related Work

2.1 Composed Video Retrieval

CoVR inherits the reference-plus-modification formulation from Composed Image Retrieval (CIR), whose methods learn image–text fusion, textual inversion, or explicit semantic shifts Baldrati et al. (2022a;b; 2023); Wen et al. (2023). WebVid-CoVR and CoVR-2 established the task, while EgoCVR, FineCVR-1M, and TF-CoVR broadened its evaluation to diverse and fine-grained modifications Ventura et al. (2024a;b); Hummel et al. (2024); Wu et al. (2025); Gupta et al. (2025). Existing CoVR methods enrich modification context, bridge modalities, disentangle preserved and changed semantics, or calibrate composed features Thawakar et al. (2024); Uesugi et al. (2024); Thawakar et al. (2025); Wu et al. (2025); Chen et al. (2025); Wang et al. (2025); Ding et al. (2025).

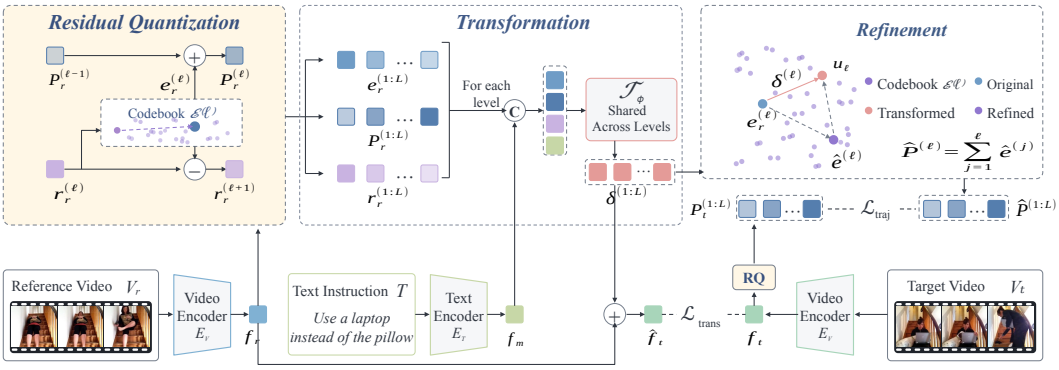

Figure 1: Overview of MORSE. *Residual Quantization* (left) decomposes the reference feature into codewords with prefix and residual context. *Transformation* (center) concatenates each codeword and its context with the instruction feature and uses a shared predictor to produce corrections in parallel. *Refinement* (right) maps each transformed vector $\mathbf{u}_\ell$ to a codeword and accumulates the selected codewords into predicted prefixes. The lower branches supervise the target prediction $\widehat{\mathbf{f}}_t$ and align predicted prefixes with target prefixes. Circled C denotes concatenation; circled $+$ and $-$ denote vector addition and subtraction, respectively.

Recent studies further explore evidence-driven matching, shared and differential semantics, transformation reasoning, and reason-aware recall and reranking Li et al. (2026a); Hu et al. (2026); Huang et al. (2026); Thawakar et al. (2026); Li et al. (2026b), while interactive and audio-visual extensions broaden the task setting Zhang et al. (2026); Han et al. (2026); Ji et al. (2026). MORSE focuses on how the requested modification is realized through changes to reference representation components. It learns instruction-dependent corrections conditioned on the residual decomposition and supervises their composition toward the target.

2.2 Quantized Representations for Retrieval

Multi-codebook quantization provides structured approximations of dense representations. Additive quantization (AQ) and composite quantization (CQ) represent vectors as sums of codewords for vector approximation and approximate nearest-neighbor search (Babenko & Lempitsky, 2014; Zhang et al., 2014). Residual quantization (RQ) constructs this approximation successively, assigning each codeword to the residual left by preceding levels (Chen et al., 2010). QINCo and QINCo2 further condition implicit codebooks on partial reconstructions to improve vector encoding (Huijben et al., 2024; Vallaeys et al., 2025). Quantization also participates in retrieval representation learning: HCQ aligns quantized video and text representations at multiple granularities (Wang et al., 2022), while GENIUS learns discrete semantic identifiers for generative multimodal retrieval (Kim et al., 2025).

MORSE uses residual decomposition to learn instruction-dependent transformations of a reference representation. The accumulated prefix and remaining residual contextualize each component, providing the inputs for predicting its modification and the intermediate representations for target-prefix supervision. Unlike reconstruction-conditioned codebook adaptation, MORSE keeps the prepared codebooks fixed during modification learning and predicts continuous corrections to reference components from the instruction and their decomposition context. Refinement expresses the transformed components through the shared codebooks, while query discrimination and prefix alignment coordinate their composition toward the target. The distinction is therefore in learning the reference-to-target transformation over a residual representation, rather than improving the quantization of a given embedding.

3 Method

Given a reference video V_r and a text instruction T, composed video retrieval aims to retrieve a target video V_t that realizes the requested modification while preserving relevant reference content (Ventura et al., 2024a). MORSE formulates target-feature prediction as coordinated corrections to an

ordered residual decomposition of the reference representation. As illustrated in Figure 1, residual quantization provides additive components and their prefix–residual contexts. *Residual Transformation* uses these contexts and the instruction to predict componentwise corrections, whose sum updates the reference feature into a continuous target prediction for retrieval. *Refinement* projects the transformed components onto shared codebooks and uses the target decomposition to supervise their cumulative composition.

3.1 PRELIMINARY

We first introduce the backbone encoders used to obtain the base video and text representations, followed by residual quantization and codebook preparation.

Video and Text Representations. Given a training triplet (V_r, T, V_t), we encode the videos and the text instruction as

$$\mathbf{f}_s = E_V(V_s), \quad s \in \{r, t\}, \qquad \mathbf{f}_m = E_T(T), \tag{1}$$

where E_V and E_T denote the video and text encoders, respectively, and $\mathbf{f}_s, \mathbf{f}_m \in \mathbb{R}^D$. The video encoder is shared by reference, target, and gallery videos and produces unit-normalized embeddings. The video encoder uses a Combiner to fuse multi-frame features extracted by a pretrained visual backbone into a single video embedding. The text encoder builds on a pretrained text backbone. Architecture details are provided in Appendix A.3.

Residual Decomposition. Residual quantization (RQ) (Chen et al., 2010; Huijben et al., 2024) approximates a video feature by an ordered sum of full-dimensional codewords, each selected to represent the residual left by preceding levels. Let $\mathcal{E}^{(\ell)} = \{\mathbf{e}_k^{(\ell)}\}_{k=1}^K$ be the codebook at level ℓ, where $\mathbf{e}_k^{(\ell)} \in \mathbb{R}^D$, K is the number of codewords, and L is the number of levels. Starting from $\mathbf{P}_s^{(0)} = \mathbf{0}$ and $\mathbf{r}_s^{(1)} = \mathbf{f}_s$, we compute

$$\begin{aligned} c_s^{(\ell)} &= \arg\min_{k \in \{1,\dots,K\}} \|\mathbf{r}_s^{(\ell)} - \mathbf{e}_k^{(\ell)}\|_2^2, \qquad \mathbf{e}_s^{(\ell)} = \mathbf{e}_{c_s^{(\ell)}}^{(\ell)}, \\ \mathbf{P}_s^{(\ell)} &= \mathbf{P}_s^{(\ell-1)} + \mathbf{e}_s^{(\ell)}, \qquad\qquad \mathbf{r}_s^{(\ell+1)} = \mathbf{f}_s - \mathbf{P}_s^{(\ell)}. \end{aligned} \tag{2}$$

Here, $c_s^{(\ell)}$ indexes the selected codeword $\mathbf{e}_s^{(\ell)}$. The prefix $\mathbf{P}_s^{(\ell-1)}$ represents the accumulated approximation, while $\mathbf{r}_s^{(\ell)}$ captures the remaining residual; together, they define the reference context at level ℓ. Prefixes are accumulated without intermediate normalization, preserving $\mathbf{f}_s = \mathbf{P}_s^{(L)} + \mathbf{r}_s^{(L+1)}$.

The codebooks are pretrained by feature reconstruction and remain fixed during modification learning; preparation details are provided in Appendix A.4. This shared additive structure supplies reference components and their contexts for Residual Transformation, and target prefixes for supervising the transformed components. The levels represent successive approximations rather than predefined semantic categories.

3.2 RESIDUAL TRANSFORMATION

Residual Transformation learns how each reference component should change under the instruction. The central block of Figure 1 conditions this correction on what preceding levels already represent (the prefix) and what remains before the current assignment (the residual). This context distinguishes a component's role in the reference decomposition, allowing its update to depend on more than the codeword and instruction alone.

Level-Wise Transformation. For each level, a shared predictor maps the instruction and reference decomposition to a continuous codeword correction:

$$\begin{aligned} \boldsymbol{\delta}^{(\ell)} &= \mathcal{T}_\phi\Big([\mathbf{P}_r^{(\ell-1)}; \mathbf{e}_r^{(\ell)}; \mathbf{f}_m; \mathbf{r}_r^{(\ell)}]\Big), \\ \mathbf{u}_\ell &= \mathbf{e}_r^{(\ell)} + \boldsymbol{\delta}^{(\ell)}. \end{aligned} \tag{3}$$

where $[\,;\,]$ denotes concatenation and $\mathcal{T}_\phi$ is a learned predictor with parameters shared across levels. Its output $\boldsymbol{\delta}^{(\ell)} \in \mathbb{R}^D$ is an unnormalized correction to the reference codeword; the prefix and residual condition its prediction. The resulting $\mathbf{u}_\ell$ is a continuous vector awaiting quantization, not a codeword selected from the codebook. All corrections are conditioned on the original reference decomposition and can therefore be predicted in parallel in a single batched evaluation. The RQ decomposition itself remains recursive.

Query Transformation. The corrections are added to the reference feature to form the target-feature prediction used for retrieval:

$$\Delta\mathbf{f} = \sum_{\ell=1}^{L} \boldsymbol{\delta}^{(\ell)}, \qquad \widehat{\mathbf{f}}_t = \mathbf{f}_r + \Delta\mathbf{f}, \tag{4}$$

$$\widehat{\mathbf{f}}_t = \sum_{\ell=1}^{L} \mathbf{u}_\ell + \mathbf{r}_r^{(L+1)}. \tag{5}$$

The second equality follows from the additive RQ decomposition: the prediction combines the transformed components with the terminal reference residual. Here, $\mathbf{f}_t$ denotes the feature encoded from the actual target video. Refinement uses the same component corrections to provide structural supervision during training.

3.3 Refinement

Supervising the final target prediction alone does not determine how corrections are distributed across components: adding a vector to one correction and subtracting it from another leaves $\widehat{\mathbf{f}}_t$ unchanged. Refinement provides an additional structural constraint by projecting the transformed components onto shared codebooks and supervising their cumulative composition against the target decomposition.

Codeword Projection. Each transformed vector $\mathbf{u}_\ell$ is assigned to its nearest codeword in $\mathcal{E}^{(\ell)}$:

$$\widehat{c}^{(\ell)} = \arg\min_{k\in\{1,\dots,K\}} \|\mathbf{u}_\ell - \mathbf{e}_k^{(\ell)}\|_2^2. \tag{6}$$

To train through this discrete selection, we use a straight-through estimator (STE) (van den Oord et al., 2017):

$$\begin{aligned} \widehat{\mathbf{e}}^{(\ell)} &= \mathbf{u}_\ell + \mathrm{sg}\Big(\mathbf{e}^{(\ell)}_{\widehat{c}^{(\ell)}} - \mathbf{u}_\ell\Big), \\ \widehat{\mathbf{P}}^{(\ell)} &= \sum_{j=1}^{\ell} \widehat{\mathbf{e}}^{(j)}. \end{aligned} \tag{7}$$

Here, sg denotes stop-gradient. The forward pass returns the selected codeword, while the backward pass uses an identity derivative to pass gradients to $\mathbf{u}_\ell$ and the shared predictor; the codebooks remain fixed. The predicted prefixes $\widehat{\mathbf{P}}^{(\ell)}$ are compared with target prefixes obtained from the same codebooks. These constraints act on the quantized composition; they do not uniquely determine the continuous corrections within a fixed code-assignment region.

3.4 Training Objectives and Inference

The lower branches of Figure 1 connect the predicted transformation to target supervision at two scales: the aggregate query and the accumulated residual prefixes. We use query discrimination before and after projection, together with target-prefix alignment, to couple the component updates with their retrieval objective. The pretrained video encoder and prepared codebooks remain fixed during modification learning; reconstruction is used for codebook preparation, not as a trainable task objective.

Query Discrimination. For a batch of B queries, let $\mathcal{J}$ index distinct target videos and $\mathcal{P}_i \subseteq \mathcal{J}$ contain all known relevant targets for query i. With $s(\mathbf{a}, \mathbf{b}) = -\|\mathbf{a} - \mathbf{b}\|_2^2$, we define the multi-positive contrastive loss van den Oord et al. (2018)

$$\mathcal{L}_{\mathrm{NCE}}(\{\mathbf{z}_i\}; \tau) = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\sum_{j \in \mathcal{P}_i} \exp(s(\mathbf{z}_i, \mathrm{sg}(\mathbf{f}_{t,j}))/\tau)}{\sum_{j \in \mathcal{J}} \exp(s(\mathbf{z}_i, \mathrm{sg}(\mathbf{f}_{t,j}))/\tau)}, \tag{8}$$

where $\tau > 0$ is a temperature. We apply this objective to the continuous target prediction and the final refined prefix:

$$\mathcal{L}_{\mathrm{trans}} = \mathcal{L}_{\mathrm{NCE}}(\{\widehat{\mathbf{f}}_{t,i}\}; \tau_{\mathrm{T}}), \qquad \mathcal{L}^{\mathrm{NCE}}_{\mathrm{traj},L} = \mathcal{L}_{\mathrm{NCE}}(\{\widehat{\mathbf{P}}_i^{(L)}\}; \tau_L). \tag{9}$$

The first objective supervises the continuous target prediction, while the second makes the final refined prefix discriminative in the target feature space. For nonzero queries and unit-normalized targets, the distance logits induce the same ranking as cosine similarity, with query norm affecting the contrastive scale (Appendix B).

Prefix Alignment. We also align the refined prefixes with the paired target decomposition:

$$\mathcal{L}^{\mathrm{geo}}_{\mathrm{traj}} = \frac{1}{B} \sum_{i=1}^{B} \sum_{\ell=1}^{L} \omega_\ell \left\| \widehat{\mathbf{P}}_i^{(\ell)} - \mathrm{sg}(\mathbf{P}_{t,i}^{(\ell)}) \right\|_2^2, \tag{10}$$

where $\omega_\ell > 0$ weights depth ℓ. Prefixes retain their original scale, so alignment constrains both direction and magnitude at each depth. This supervision captures intermediate discrepancies that can cancel in the final sum, complementing target discrimination.

Overall Objective. We group prefix alignment and final-prefix discrimination into the Refinement objective and combine it with continuous-target supervision:

$$\begin{aligned} \mathcal{L}_{\mathrm{traj}} &= \lambda_{\mathrm{R}} \mathcal{L}^{\mathrm{geo}}_{\mathrm{traj}} + \lambda_L \mathcal{L}^{\mathrm{NCE}}_{\mathrm{traj},L}, \\ \mathcal{L}_{\mathrm{total}} &= \lambda_{\mathrm{T}} \mathcal{L}_{\mathrm{trans}} + \mathcal{L}_{\mathrm{traj}}, \end{aligned} \tag{11}$$

with nonnegative weights. We train the instruction head and shared predictor while keeping the video encoder and codebooks fixed; target features and prefixes are detached. Through the STE, the Refinement losses update the same corrections that form the continuous target prediction.

Transformation Scoring. We rank candidate videos by the cosine similarity between the continuous target prediction $\widehat{\mathbf{f}}_t$ and each candidate feature $\mathbf{f}_{t,j} = E_V(V_j)$:

$$S_{\mathrm{trans}}(V_j \mid V_r, T) = \mathrm{sim}(\widehat{\mathbf{f}}_t, \mathbf{f}_{t,j}), \tag{12}$$

where $\mathrm{sim}(\cdot, \cdot)$ denotes cosine similarity. This scoring rule, denoted Trans, is used for the main results.

Refinement Reranking. As an optional step, the predicted prefix at a selected level ℓ provides a refinement score for the top-100 candidates returned by Trans:

$$S_{\mathrm{traj}}(V_j \mid V_r, T) = \mathrm{sim}\left(\widehat{\mathbf{P}}^{(\ell)}, \mathbf{P}_{t,j}^{(\ell)}\right), \tag{13}$$

where $\mathbf{P}_{t,j}^{(\ell)}$ is the candidate's prefix at the same level, obtained using the shared codebooks. This refinement reranking operates within the retrieved candidate set; the main comparisons use Transformation scoring alone.

4 Experiments

4.1 Experimental Setup

Benchmarks and Metrics. WebVid-CoVR Ventura et al. (2024a) contains 1,648,789 automatically generated training triplets over 130,775 web videos; its manually verified test set has 2,556

triplets over 2,444 videos. FineCVR-1M Wu et al. (2025) broadens the edits to attributes, objects, scenes, and actions, with 1,000,028 training triplets, 10,043 test triplets, and 136,547 videos. TF-CoVR Gupta et al. (2025) targets subtle gymnastics and diving transitions and associates each query with 3.94 valid targets on average. Our FineGD artifacts contain 177,787/379/473 train/validation/test queries and a 473-video test gallery. We evaluate MORSE on publicly available CoVR benchmarks that provide both training and test splits for task-specific fine-tuning: WebVid-CoVR Ventura et al. (2024a), FineCVR-1M Wu et al. (2025), and TF-CoVR Gupta et al. (2025). Following their official evaluation protocols, we report Recall@K (R@K) for $K \in \{1, 5, 10, 50\}$ on the first two, and mean Average Precision (mAP@K) for $K \in \{5, 10, 25, 50\}$ on TF-CoVR.

Implementation Details. We follow established backbones: BLIP-2 with four uniformly sampled frames on WebVid-CoVR and FineCVR-1M Ventura et al. (2024b); Li et al. (2026a), and official pre-extracted AIM embeddings with a lightweight Transformer composer on TF-CoVR Gupta et al. (2025). The main results use Transformation scoring with the continuous target prediction $\widehat{\mathbf{f}}_t$. Optional Refinement Reranking uses the predicted prefix at a selected level ℓ to reorder the top-100 candidates returned by Transformation. Additional baselines, hyperparameter optimization, configurations, and inference details are provided in the supplementary material.

Table 1: Comparison on WebVid-CoVR-Test. Recall@K (%) is reported. The best results are bold and the second-best results are underlined.

Method	Recall@K			
	1	5	10	50
Not fine-tuned				
CLIP Radford et al. (2021)	44.37	69.13	77.62	93.00
BLIP Li et al. (2022)	45.46	70.46	79.54	93.27
Fine-tuned				
BLIP_{CA} Ventura et al. (2024a)	53.13	79.93	86.85	97.69
ECDE Thawakar et al. (2024)	60.12	84.32	91.27	98.72
BLIP2_{CA} Ventura et al. (2024b)	59.82	83.84	91.28	98.24
FDCA Wu et al. (2025)	54.80	82.27	89.84	97.70
HUD Chen et al. (2025)	63.38	86.93	92.29	98.76
REFINE Hu et al. (2026)	62.48	86.82	92.41	98.87
ReTrack Li et al. (2026a)	<u>63.85</u>	<u>87.05</u>	<u>92.80</u>	<u>99.10</u>
MORSE	**65.49**	**87.56**	**93.27**	**99.18**

Table 2: Comparison on FineCVR-1M.

Method	Recall@K			
	1	5	10	50
Not fine-tuned				
CLIP_{V} Radford et al. (2021)	8.27	27.89	40.69	70.00
CLIP_{T} Radford et al. (2021)	2.86	9.19	14.61	34.15
CLIP Radford et al. (2021)	8.47	24.85	46.08	65.01
TFR-CVR Hummel et al. (2024)	15.21	40.12	52.78	81.75
Fine-tuned				
CLIP_{MLP} Radford et al. (2021)	14.68	41.42	57.82	88.74
BLIP_{MLP} Li et al. (2022)	19.13	51.79	68.60	93.62
BLIP_{CA} Ventura et al. (2024a)	17.05	41.57	56.60	85.56
Pic2Word Saito et al. (2023)	8.10	25.79	38.32	71.01
Combiner Baldrati et al. (2022b)	19.57	49.46	65.38	92.11
FreestyleRet Li et al. (2024)	20.39	52.98	68.37	93.03
$\text{FDCA}_{\text{CLIP}}$ Wu et al. (2025)	25.84	55.84	70.23	94.33
$\text{FDCA}_{\text{BLIP}}$ Wu et al. (2025)	<u>26.79</u>	<u>63.21</u>	<u>78.65</u>	<u>97.25</u>
MORSE	**31.56**	**68.83**	**81.21**	**98.01**

4.2 Main Results

WebVid-CoVR. As shown in Table 1, MORSE achieves the best result at every cutoff, improving ReTrack by 1.64/0.51/0.47/0.08 points at R@1/5/10/50. The largest gain at R@1 demonstrates more accurate early ranking, indicating that the modification-adaptive composed feature and transition objectives better distinguish subtly modified targets.

FineCVR-1M. Table 2 shows that MORSE establishes new best results across all recall levels. MORSE improves over $\text{FDCA}_{\text{BLIP}}$ by 4.77/5.62/2.56/0.76 points at R@1/5/10/50, with particularly strong gains at early and intermediate ranks. These improvements demonstrate that the learned composed feature and residual transitions generalize across FineCVR-1M's diverse object, attribute, scene, and action modifications.

Table 3: Comparison on TF-CoVR. mAP@K (%) is reported under the official protocol.

Method	mAP@K			
	5	10	25	50
Not fine-tuned				
BLIP2_{T} Li et al. (2023)	1.34	1.79	2.20	2.50
BLIP2_{V} Li et al. (2023)	1.74	2.20	3.06	3.62
BLIP_{CA} Ventura et al. (2024a)	2.33	2.99	3.90	4.50
ECDE Thawakar et al. (2024)	0.78	0.88	1.16	1.37
TF-CVR Hummel et al. (2024)	0.56	0.76	0.99	1.24
Fine-tuned				
BLIP2_{T} Li et al. (2023)	10.69	13.02	15.35	16.41
BLIP2_{V} Li et al. (2023)	4.86	6.49	8.92	10.06
CLIP_{MLP} Radford et al. (2021)	6.40	7.46	9.21	10.40
$\text{BLIP2}_{\text{MLP}}$ Li et al. (2023)	11.64	14.81	16.74	17.55
BLIP_{CA} Ventura et al. (2024a)	11.07	13.94	16.07	16.88
ECDE Thawakar et al. (2024)	13.03	15.90	18.62	19.83
$\text{TF-CoVR-Base}_{\text{Stage 2}}$	15.08	18.70	21.78	22.61
TF-CoVR-Base Gupta et al. (2025)	<u>21.85</u>	<u>24.23</u>	<u>26.47</u>	<u>27.22</u>
MORSE	**23.38**	**26.19**	**28.89**	**30.01**

TF-CoVR. MORSE improves over TF-CoVR-Base by 1.53/1.96/2.42/2.79 mAP points at $K = 5/10/25/50$ and reaches 30.01 mAP@50, a

10.25% relative gain. The increasing absolute gain toward larger cutoffs shows that MORSE recovers multiple valid targets rather than only improving the first match.

4.3 Ablation Study

Training Objectives. Table 4 examines the three training objectives, with Transformation scoring used throughout. Removing continuous-target discrimination reduces TF-CoVR mAP@50 from 30.01 to 23.93, showing that structural supervision alone is insufficient to match the full objective. With both discrimination losses active, adding prefix alignment raises FineCVR-1M R@1 from 30.92 to 31.56 and TF-CoVR mAP@50 from 29.57 to 30.01. The full objective achieves the highest values across all reported metrics, supporting the complementary roles of direct target discrimination and target-side structural supervision.

Table 4: Training-Objective Ablation on FineCVR-1M and TF-CoVR.

TRANS	REFINEMENT		FineCVR	TF-CoVR	
$\mathcal{L}_{\text{trans}}$	$\mathcal{L}_{\text{traj}}^{\text{geo}}$	$\mathcal{L}_{\text{traj},L}^{\text{NCE}}$	**R@1**	**mAP@5**	**mAP@50**
✓			29.64	20.52	27.35
✓	✓		31.19	22.72	28.42
✓		✓	30.92	22.85	29.57
	✓	✓	31.00	17.35	23.93
✓	✓	✓	**31.56**	**23.38**	**30.01**

Effect of Transformation. Table 5 examines component-wise updates and their conditioning inputs. Global update (GU) uses a shared displacement across levels, whereas global context (GC) replaces the prefix–residual context with $\mathbf{f}_r$; the remaining variants zero the specified inputs. MORSE exceeds GU by 3.84 and 1.18 R@1 points on WebVid-CoVR and FineCVR-1M, respectively, and by 1.19 mAP@50 points on TF-CoVR. It also outperforms GC on all three benchmarks, supporting the use of level-specific reference context for differentiated corrections. Removing the prefix lowers performance on all three datasets, and removing the instruction produces the largest drop. The residual input has a dataset-dependent effect, so its inclusion is not uniformly beneficial.

Table 5: Effect of level-specific transformation and conditioning inputs.

	WebVid	FineCVR-1M	TF-CoVR
Variant	R@1	R@1	mAP@50
MORSE	65.49	31.56	30.01
GU	61.65	30.38	28.82
w/o context	63.38	29.68	29.65
w/o prefix	55.47	28.68	29.14
w/o residual	64.71	32.71	29.86
w/o instruction	45.73	12.53	12.25
w/ GC	59.74	22.39	27.63

In Figure 2, GU retrieves bare-chested people and a person beside horses, whereas MORSE correctly matches the requested sumo costumes and horse-feeding action.



(a) the person changes to wear sumo costumes instead of funny clothes and white wigs

(b) the person is also talking, but feeding a horse instead of holding a phone

Figure 2: Each group shows the reference, GU top-1, and MORSE top-1 from left to right, with the instruction below. Throughout this paper, videos outlined in green are target videos for the corresponding queries.

Effect of Refinement Reranking. Table 6 evaluates optional Refinement Reranking with the model and gallery fixed. Using the predicted prefix, it reorders the top-100 candidates returned by Transformation. R@1 improves by 0.99 points and R@5 also increases, while R@10 and R@50 decline slightly, indicating a benefit at early ranks rather than a uniform improvement

Table 6: Effect of refinement reranking on FineCVR-1M.

Inference	R@1	R@5	R@10	R@50
Transformation	31.56	68.83	**81.21**	**98.01**
Refinement Rerank	**32.55**	**69.24**	79.91	97.75

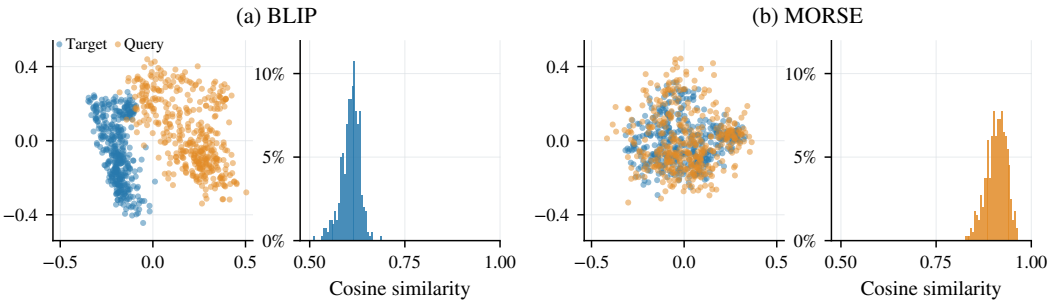

Figure 5: Query–target alignment on FineCVR-1M: per-model PCA projections (left) and paired cosine similarities in the original spaces (right), using unquantized targets.

across cutoffs. In Figure 3, refinement promotes the pedal-installation target from rank 2 to 1 and demotes the brake-repair clip from rank 1 to 3.

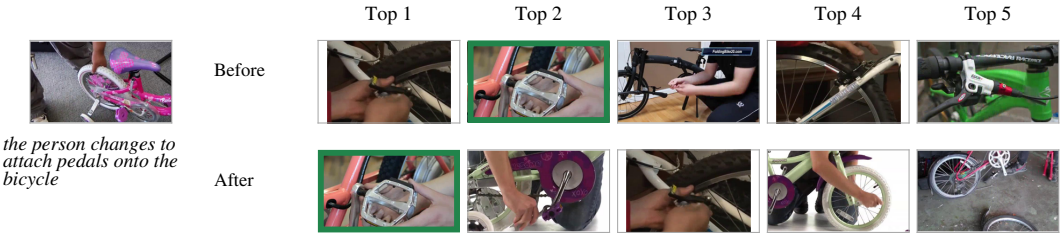

Figure 3: The reference and instruction are on the left; the top-five results from Transformation (before) and Refinement Rerank (after) are arranged on the right.

4.4 Further Analysis

Modification Discrimination. Figure 4 holds the reference video fixed and varies the instruction between measuring shims and sealing a roof. BLIP returns the same upholstery-related clip for both requests, whereas MORSE ranks the corresponding target first in each case. This contrast illustrates instruction-specific retrieval responses: the same reference leads to different targets when the requested change differs. The BLIP checkpoint was selected by test R@1, so these examples illustrate behavior rather than establish a controlled aggregate advantage.

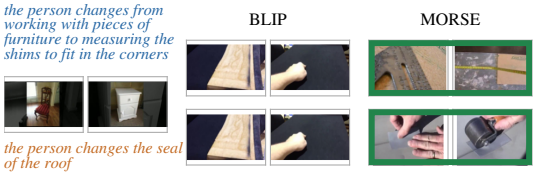

Figure 4: Two instructions, above and below the reference, with the corresponding top-1 results in the upper and lower rows. Each video is represented by two chronological frames.

Query–Target Alignment. Figure 5 compares BLIP queries and MORSE's eight-level refined prefixes with unquantized targets on FineCVR-1M, yielding mean paired cosine similarities of 0.6072 and 0.9083 in their respective feature spaces. These samples, selected to favor MORSE, illustrate refinement-branch alignment rather than test-wide performance. Separately fitted PCA plots provide within-model views, not directly comparable coordinates.

5 Conclusion

We proposed MORSE, which predicts target features through reference-conditioned corrections over an ordered residual decomposition. Target-side structural supervision coordinates these corrections to realize the requested modification, improving modification discrimination and composed video retrieval.

AI USE STATEMENT

Generative AI tools assisted with research ideas and experimental design, code development and review, analysis of results, and mathematical arguments. They were also used for literature search, figure preparation, and manuscript editing. The authors checked AI-assisted results and claims against the underlying experiments, code, and cited sources and take responsibility for the final paper and its artifacts.

REPRODUCIBILITY STATEMENT

The main text specifies the model, training objectives, retrieval procedure, benchmark protocols, and evaluation metrics. Appendix A documents the datasets and implementation, including the encoders; Appendix B provides supporting representation and optimization analyses.

REFERENCES

Artem Babenko and Victor Lempitsky. Additive quantization for extreme vector compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 931–938, 2014.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4959–4968, 2022a.

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, 2022b.

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.

Yongjian Chen, Tao Guan, and Cheng Wang. Approximate nearest neighbor search by residual vector quantization. *Sensors*, 10(12):11259–11273, 2010. doi: 10.3390/s101211259.

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.

Zhixiang Ding, Lilong Liu, Zhenyu Yang, and Shengsheng Qian. Multi-view captioning with semantic delta re-ranking for zero-shot composed video retrieval. In *Image and Graphics: 13th International Conference, ICIG 2025, Xuzhou, China, October 31–November 2, 2025, Proceedings, Part II*, volume 16162 of *Lecture Notes in Computer Science*, pp. 80–91. Springer, 2025. doi: 10.1007/978-981-95-3393-0_7.

Animesh Gupta, Jay Parmar, Ishan Rajendrakumar Dave, and Mubarak Shah. From play to replay: Composed video retrieval for temporally fine-grained videos. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/c192d0aefb044150948577bf6e7b3f62-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Gyuwon Han, Young Kyun Jang, and Chanho Eom. Cova: Text-guided composed video retrieval for audio-visual content. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2026.

Yupeng Hu, Zixu Li, Zhiwei Chen, Qinlei Huang, Zhiheng Fu, Mingzhu Xu, and Liqiang Nie. Refine: Composed video retrieval via shared and differential semantics enhancement. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 22(7):1–24, 2026. doi: 10.1145/3796712.

Sihong Huang, Jiaxin Wu, Dongmei Jiang, Yi Cai, Yaowei Wang, and Xiaoyong Wei. Compositional transformation reasoning for composed video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Iris A. M. Huijben, Matthijs Douze, Matthew J. Muckley, Ruud Van Sloun, and Jakob Verbeek. Residual quantization with implicit neural codebooks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20682–20699, 2024.

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 Nature Switzerland Cham, 2024.

Junyang Ji, Shengjun Zhang, Da Li, Yuxiao Luo, Yan Wang, Di Xu, Biao Yang, Wei Yuan, Fan Yang, Zhihai He, and Wenming Yang. OmniCVR: A benchmark for omni-composed video retrieval with vision, audio, and text. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=KxxR7emO5K.

Sungyeon Kim, Xinliang Zhu, Xiaofan Lin, Muhammet Bastan, Douglas Gray, and Suha Kwak. GENIUS: A generative framework for universal multimodal search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19659–19669, 2025.

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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742, 2023.

Zixu Li, Yupeng Hu, Zhiwei Chen, Qinlei Huang, Guozhi Qiu, Zhiheng Fu, and Meng Liu. Retrack: Evidence-driven dual-stream directional anchor calibration network for composed video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 23373–23381, 2026a. doi: 10.1609/aaai.v40i28.39507.

Zixu Li, Yupeng Hu, Zhiheng Fu, Zhiwei Chen, Weili Guan, and Liqiang Nie. R3: Composed video retrieval via reasoning-guided recalling and re-ranking. *arXiv preprint arXiv:2606.01113*, 2026b.

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.

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.

Omkar Thawakar, Dmitry Demidov, Vaishnav Potlapalli, Sai Prasanna Teja Reddy Bogireddy, Viswanatha Reddy Gajjala, Alaa Mostafa Lasheen, Rao Muhammad Anwer, and Fahad Khan. Covr-r: Reason-aware composed video retrieval. *arXiv preprint arXiv:2603.20190*, 2026.

Kenta Uesugi, Naoki Saito, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama. Zero-shot composed video retrieval with projection module bridging modality gap. In *2024 IEEE 13th Global Conference on Consumer Electronics (GCCE)*, pp. 6–7. IEEE, 2024.

Théophane Vallaeys, Matthew J. Muckley, Jakob Verbeek, and Matthijs Douze. QINCo2: Vector compression and search with improved implicit neural codebooks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=2zMHHZ569S.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 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*, pp. 5270–5279, 2024a. doi: 10.1609/aaai.v38i6.28334.

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*, 46(12):11409–11421, 2024b. doi: 10.1109/TPAMI.2024.3463799.

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.

Jinpeng Wang, Bin Chen, Dongliang Liao, Ziyun Zeng, Gongfu Li, Shu-Tao Xia, and Jin Xu. Hybrid contrastive quantization for efficient cross-view video retrieval. In *Proceedings of the ACM Web Conference 2022*, pp. 3020–3030, 2022. doi: 10.1145/3485447.3512022.

Haokun Wen, Xian Zhang, Xuemeng Song, Yinwei Wei, and Liqiang Nie. Target-guided composed image retrieval. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 915–923, 2023.

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 *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2dae7d1ccf1edf76f8ce7c282bdf4730-Abstract-Conference.html.

Bingqing Zhang, Yi Zhang, Zhuo Cao, Yang Li, Xue Li, Jiajun Liu, and Sen Wang. Recovr: Closing the loop in interactive composed video retrieval. *arXiv preprint arXiv:2605.09836*, 2026.

Ting Zhang, Chao Du, and Jingdong Wang. Composite quantization for approximate nearest neighbor search. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 838–846, 2014. URL https://proceedings.mlr.press/v32/zhangd14.html.