RECONTEXT: CONTEXT-COUPLED MATCHING FOR MULTI-EVENT VIDEO–TEXT RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

In multi-event video–text retrieval, a description refers to selected parts of one or several activities within a video. Local actions and objects help locate relevant content and distinguish similar activities, while the surrounding activity provides context for judging their relevance to the description. Pooling the strongest local and contextual matches independently discards their pairing, making different local–activity associations indistinguishable when their pooled maxima coincide. For descriptions that span several activities, relevance also depends on combining support from different contexts. To this end, we introduce ReContext, which supplements each context's compatibility with the description using matches from its own constituent spans, then selects the highest-scoring context. We combine this context score with an aggregation of each description fragment's best match across the video, yielding a shared bidirectional score using frozen representations. Our analysis shows that coupled selection can retain distinctions between local–context configurations that independent pooling assigns identical local and contextual maxima. Experiments on Charades-Event and ActivityNet-Event show improved retrieval in both directions, with controlled comparisons examining the contributions of membership coupling and cross-activity support.

1 INTRODUCTION

Videos in the multi-event setting contain several activities, while descriptions concern one activity or selected parts of several activities. Multi-event video–text retrieval ranks videos for descriptions and descriptions for videos in this partially relevant setting (Zhang et al., 2023). In instructional video collections, these directions support finding videos of a specified activity and retrieving descriptions of activities present in a given video for semantic indexing. In either direction, relevance depends on whether the described content occurs within the video, even when other activities are unrelated to the description. Comparing text fragments with short video spans produces local matches based on actions and objects (Jun et al., 2025). A matching action or object, however, does not by itself establish that the local detail occurs in an activity compatible with the description.

Consider the example in Fig. 1, with the description "A woman puts white paint on a table, and uses a small toothpick to make dots on the nail." The rank-1 candidate contains a strong local match to "white paint on a table," but its surrounding activity provides limited support for the complete description. The annotated positive appears at rank 11 under the same element-matching ranking. In that video, the nail-art activity clarifies that "paint" refers to nail polish, while "dots on the nail" identifies the decoration described by the second fragment. The example motivates assessing local details together with the activity context that gives them meaning.

This example points to a broader issue for retrieval in multi-event videos. Existing retrieval methods incorporate context through local temporal aggregation (Fragomeni et al., 2022) and broader semantic representations (Moon et al., 2025a; Li et al., 2026b). Yet a local match is informative partly because of the activity context that contains it. Matching strength alone does not specify that relation. As illustrated in Fig. 1(b), independent pooling merges fragment matches into a single description-level response and removes the context that supplied each match. Two pairs can therefore receive the same pooled response while associating their strongest matches with different activities. The resulting score cannot tell those associations apart. The limitation lies in score composition. Representing

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.

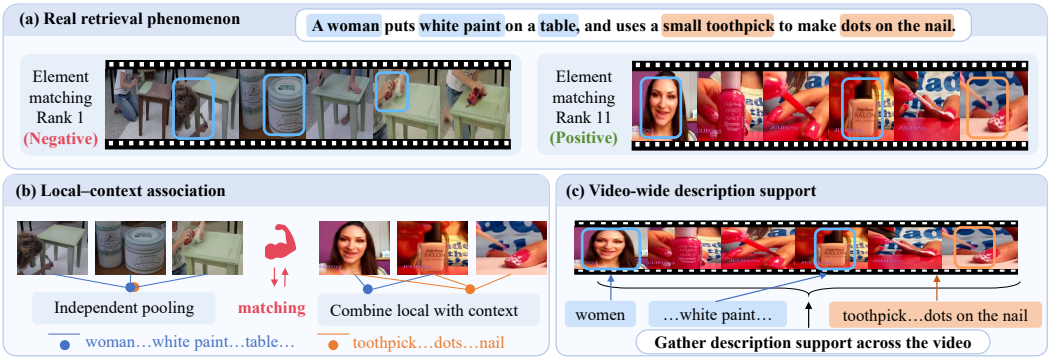

Figure 1: Motivation example for ReContext. (a) Element matching ranks a distractor first and the annotated positive at rank 11 for the same description; red and green labels denote them, respectively. (b) Independent pooling merges fragment matches and discards their context membership. (c) Different video spans can support different description fragments, motivating video-wide support.

local details and activity context is insufficient when their membership relation is discarded. This motivates retaining membership when local matches contribute to a context score.

The role of membership changes with the scope of the judgment. For a context score, membership determines which context receives a local match. For the complete description, support can be distributed across several activities or video spans, as illustrated in Fig. 1(c). A context that explains one fragment need not account for the remaining fragments, and their absence from that context says nothing about their absence from the video. Restricting every fragment to the selected context would therefore turn a relation used to identify the supporting activity into an unintended restriction on description support. Context scoring and description support consequently require different scopes.

These two roles therefore operate at different scopes in the scoring process. For context scoring, local matches and context scores are combined before a context is selected, so each local contribution remains tied to the context containing its visual element. For description support, the strongest match for each fragment is retained over the video, allowing different fragments to draw support from different activities. To this end, we propose **ReContext**, which combines these two operations in a shared bidirectional score.

ReContext represents each video with local elements and contexts that retain their constituent elements and membership links. It enhances each context's compatibility with the complete description using matches from its own members, selects the highest-scoring context after this coupling, and separately aggregates the strongest match for each description fragment over the full video. The score uses these fixed elements, contexts, and pretrained representations, keeping the analysis centered on score composition rather than task-specific encoder adaptation. Our analysis characterizes when membership coupling distinguishes local–context pairings that independent pooling maps to the same pooled value.

The contributions of this work are as follows:

- We identify how independent score aggregation can discard local–context association. We characterize when membership coupling distinguishes pairings that receive the same independently aggregated local and contextual scores.
- We propose ReContext, which couples each context with matches from its own members and aggregates each description fragment's best support over the full video in a bidirectional score.
- We evaluate ReContext on Charades-Event and ActivityNet-Event in both retrieval directions, which demonstrates that it outperforms competitors in both text-to-video and video-to-text retrieval.

2 Problem Formulation

Let $\mathcal{V}$ and $\mathcal{T}$ denote collections of videos and descriptions. We use a shared compatibility score $s(V, T; \mathcal{R})$ to rank videos for a description and descriptions for a video. Here, $\mathcal{R}$ is a fixed reference

video collection used for score calibration. The top-ranked candidates satisfy:

$$\widehat{V}(T) \in \arg\max_{V \in \mathcal{V}} s(V, T; \mathcal{R}), \quad \widehat{T}(V) \in \arg\max_{T \in \mathcal{T}} s(V, T; \mathcal{R}). \tag{1}$$

The full rankings order candidates by descending compatibility. For benchmark evaluation, we use $\mathcal{R} = \mathcal{V}$ and keep this reference unchanged across retrieval directions. We subsequently write $s(V, T)$ with $\mathcal{R}$ fixed. A description concerns one activity or selected parts of several activities in a video. Its supporting content can span several activities and be interleaved with unrelated material. We represent local details as semantic elements and group related visual elements into activity contexts.

2.1 Semantic Elements and Activity Contexts

We model a video V and a description T through visual and linguistic elements:

$$\mathcal{E}^v(V) = \{e_i^v\}_{i=1}^{N_v}, \qquad \mathcal{E}^t(T) = \{e_j^t\}_{j=1}^{N_t}, \tag{2}$$

where N_v and N_t are the numbers of visual and linguistic elements, respectively. A visual element captures a localized semantic component, such as an action, an object interaction, or scene information, while retaining its temporal location. A linguistic element is a description fragment that retains the participants and relations needed to interpret its content.

Related visual elements form candidate activity contexts:

$$\mathcal{C}^v(V) = \{c_k^v\}_{k=1}^{K_v}, \quad c_k^v \equiv \{e_i^v : i \in \mathcal{I}_k\}, \quad \varnothing \neq \mathcal{I}_k \subseteq \{1, \ldots, N_v\}, \qquad c^t = T. \tag{3}$$

Here, K_v is the number of contexts, and $\mathcal{I}_k$ records the indices of the elements in context c_k^v. The complete description serves as the textual context c^t. A visual context can contain content unmentioned in the description and support only some of its fragments.

Context memberships need not form a disjoint temporal partition. The groups can share elements, connect elements separated in time, or contain smaller groups associated with parts of an activity. Memberships are proposed by semantic parsing and specify which local elements can contribute to a context's score. The element and context sets are nonempty.

2.2 Matching at Complementary Granularities

Let $\mathbf{e}_i^v$, $\mathbf{e}_j^t$, $\mathbf{c}_k^v$, and $\mathbf{c}^t$ denote unit-normalized representations in a shared embedding space, constructed as described in Section 3.1. The elementary matching signals are:

$$s_{ij}^e(V, T) = \langle \mathbf{e}_i^v, \mathbf{e}_j^t \rangle, \qquad s_k^c(V, T) = \langle \mathbf{c}_k^v, \mathbf{c}^t \rangle. \tag{4}$$

The local score compares a visual element with a description fragment, capturing matches in specific actions and objects. The context score compares an activity context with the complete description. A fragment can have a weak match inside one context and a strong match elsewhere in the video. This motivates separate aggregation scopes for context scoring and video-wide description support.

3 Method

Fig. 2 shows how ReContext uses local matches at two scopes. The context-coupled branch uses matches from each context's members to enhance its score before selection. The video-wide branch aggregates support for description fragments across activities. Both branches use the same calibrated local matches and contribute to a shared bidirectional score. We first describe semantic structure construction and calibration, then the two scoring operations.

3.1 Semantic Structure Construction and Encoding

We instantiate the elements and context memberships defined in Eqs. (2) and (3) through semantic parsing with frozen models. Visual parsing organizes observed actions, objects, interactions, and scene transitions into local units with temporal boundaries. Linguistic parsing extracts description fragments while retaining their participants, actions, and explicit relations. Related visual elements are grouped by activity, with each candidate group retaining the indices and temporal locations of

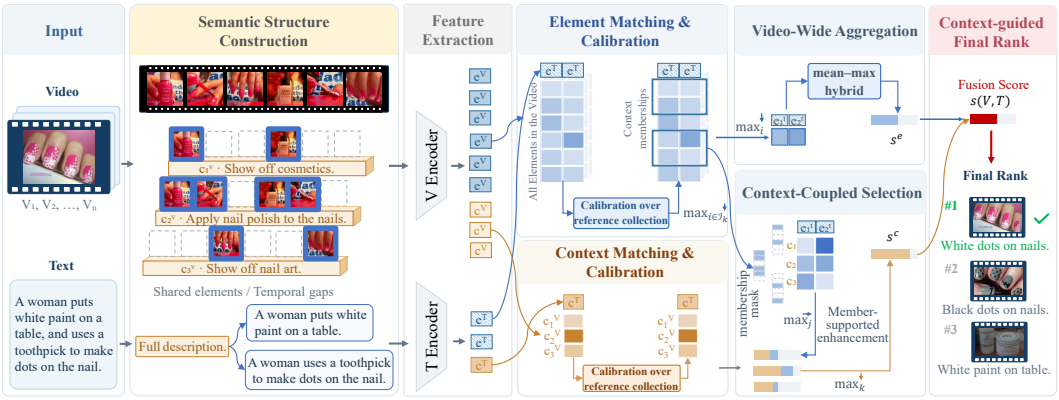

Figure 2: Overview of ReContext. Local elements and activity contexts are encoded with frozen models, and their matching scores are calibrated over a fixed reference video collection. The upper branch aggregates each fragment's best match across the video into s^e. The lower branch gathers local support within each context's membership, enhances the context scores, and selects the highest as s^c. The branches are fused into the shared bidirectional score $s(V,T)$. Small arrows beside the maxima indicate aggregation axes; connecting arrows show data flow. The rightmost ranking illustrates text-to-video retrieval schematically. Green marks the positive match.

its members. These groups define the activity contexts used for scoring and can overlap or connect temporally separated spans.

For each visual element, we uniformly sample frames within its temporal support. For each visual context, we sample frames from its constituent elements and assemble them into a context-level input. Using frozen visual and textual encoders f^v and f^t, we obtain unit-normalized representations:

$$\mathbf{e}_i^v = \frac{f^v(e_i^v)}{\|f^v(e_i^v)\|_2}, \mathbf{e}_j^t = \frac{f^t(e_j^t)}{\|f^t(e_j^t)\|_2}, \quad \mathbf{c}_k^v = \frac{f^v(c_k^v)}{\|f^v(c_k^v)\|_2}, \mathbf{c}^t = \frac{f^t(T)}{\|f^t(T)\|_2}. \tag{5}$$

Here, $f^v(c_k^v)$ directly encodes the context's constituent video content. Element inputs expose local details, while context inputs provide the broader visual scope of their members. The representations and membership relations remain fixed throughout retrieval scoring.

3.2 Candidate-Relative Similarity Calibration

Different description fragments can produce systematically different similarity levels, affecting their contributions to an aggregate score. We compare each similarity with the other visual responses to the same text. Starting from Eq. (4), we normalize the similarities for each linguistic element over all visual elements in $\mathcal{R}$, and those for the complete description over all contexts in that collection:

$$\bar{s}_{ij}^e(V,T) = \frac{\exp\left(s_{ij}^e(V,T)/\tau^e\right)}{\sum_{V'\in\mathcal{R}}\sum_{i'=1}^{N_{v'}}\exp\left(s_{i'j}^e(V',T)/\tau^e\right)}, \quad \bar{s}_k^c(V,T) = \frac{\exp(s_k^c(V,T)/\tau^c)}{\sum_{V'\in\mathcal{R}}\sum_{k'=1}^{K_{v'}}\exp\left(s_{k'}^c(V',T)/\tau^c\right)}. \tag{6}$$

Here, $\tau^e, \tau^c > 0$ are temperature parameters, and $N_{v'} = |\mathcal{E}^v(V')|$ and $K_{v'} = |\mathcal{C}^v(V')|$ denote the numbers of elements and contexts in a reference video V'. Both denominators use the full reference collection and remain the same for a given video–description pair in either retrieval direction.

Calibration preserves the ordering of visual candidates for each textual component while removing a shared additive baseline from their similarities. The resulting weights are used in both scoring branches. Because the element and context pools have different sizes, their weights are not directly comparable in semantic strength; we therefore keep the reference collection fixed across retrieval directions. Appendix C details this property and its scope.

3.3 CONTEXT-COUPLED SELECTION

A context representation summarizes several spans, so a brief action or object can have limited influence on its similarity to a description. The lower scoring branch in Fig. 2 makes those details directly available to context selection through the local matches of its members. The membership restriction preserves which activity supplies the details contributing to each context score.

The membership masks in Fig. 2 represent the sets $\mathcal{I}_k$. For each linguistic element, we first find its strongest match among the members of context c_k^v. We then retain the strongest support across linguistic elements:

$$s_{kj}^{e|c}(V,T) = \max_{i \in \mathcal{I}_k} \bar{s}_{ij}^{e}(V,T), \quad s_k^{e|c}(V,T) = \max_{1 \le j \le N_t} s_{kj}^{e|c}(V,T). \tag{7}$$

The maximum over $i \in \mathcal{I}_k$ finds support for each fragment within a context, producing the context-by-fragment matrix shown in the figure. The maximum over j reduces each row to one local support score for that context. This permits a context to contribute through its strongest fragment match, even when other fragments are supported by different activities.

The member-supported enhancement combines this local support with the score of the same context through a one-sided update:

$$\widetilde{s}_k^c(V,T) = \bar{s}_k^c(V,T) + \beta \max\Big\{0,\, s_k^{e|c}(V,T) - \bar{s}_k^c(V,T)\Big\}, \tag{8}$$

where $0 \le \beta \le 1$ controls the enhancement. The update interpolates between the context score and a larger member-level weight whenever one is available. Otherwise, it retains the context score. Local details can thus compensate for a lower context score, while fragments with weak matches in this context contribute no subtraction to the enhancement. Their support is collected separately across the video. The base context score still compares the context with the complete description and can change when that description changes.

After enhancing all contexts in the candidate video, we select the highest-scoring context and retain its value:

$$k^\star(V,T) \in \arg\max_{1 \le k \le K_v} \widetilde{s}_k^c(V,T), \quad s^c(V,T) = \max_{1 \le k \le K_v} \widetilde{s}_k^c(V,T). \tag{9}$$

The selected context combines compatibility with the complete description and support from its constituents. Its selection leaves matches elsewhere in the video available to video-wide aggregation.

Coupling before selection. The order of combination and selection determines which associations the score retains. We compare context-coupled selection with independent pooling of local support and context scores. For a fixed pair (V,T), write $C_k = \bar{s}_k^c$, $L_k = s_k^{e|c}$, $C_\star = \max_k C_k$, and $L_\star = \max_k L_k$. The context score has the equivalent form

$$\begin{aligned} s^c &= \max\Big\{C_\star,\ \max_k \left[(1-\beta)C_k + \beta L_k\right]\Big\} \\ &\le \max\{C_\star,\ (1-\beta)C_\star + \beta L_\star\}\,. \end{aligned} \tag{10}$$

The right-hand side applies the same enhancement after independent max pooling. The coupled term retains a shared context index k, and the inequality can be strict when the two maxima belong to different contexts. Consequently, identical pooled maxima can conceal different local–context associations that the coupled score can still distinguish. This result concerns the use of the supplied memberships; their semantic quality depends on structure construction. Appendix C provides the proof, an example, and the parameter settings in which the distinction disappears.

3.4 VIDEO-WIDE AGGREGATION AND BIDIRECTIONAL SCORING

Selecting one context leaves open how the other parts of the description are supported. The upper scoring branch in Fig. 2 therefore finds the best match for each linguistic element over the full video, retaining access to spans outside the selected context:

$$s_j^e(V,T) = \max_{1 \le i \le N_v} \bar{s}_{ij}^e(V,T). \tag{11}$$

Table 1: Text-to-video retrieval on Charades-Event and ActivityNet-Event. Recall is in percent (higher is better); mRank is lower-is-better. The upper block lists task-supervised methods; the final two rows use frozen models without task-specific training. $*$ marks I3D features, – means unavailable, † means training-free, and red X marks pending results.

	Charades-Event					ActivityNet-Event				
Method	R@1	R@5	R@10	R@50	mRank	R@1	R@5	R@10	R@50	mRank
CenterCLIP (Zhao et al., 2022)	1.93	6.61	10.67	30.13	131.00	15.53	37.21	49.78	77.19	11.00
CLIP4Clip (Luo et al., 2022)	1.53	6.48	10.54	29.38	136.00	15.25	36.43	49.11	76.64	11.00
GMMFormer* (Wang et al., 2024c)	2.00	7.80	12.50	34.80	98.50	8.30	24.95	36.69	66.77	20.00
MeVTR (avg) (Zhang et al., 2023)	1.88	6.69	10.59	29.97	126.00	15.99	37.13	49.60	76.66	11.00
MeVTR (max) (Zhang et al., 2023)	2.02	7.61	12.74	34.49	112.00	14.93	34.44	46.36	74.65	13.00
MS-SL++ (Chen et al., 2026c)	1.50	5.20	9.30	–	–	12.70	33.10	46.00	–	–
T-MASS (Wang et al., 2024a)	0.90	3.80	6.60	–	–	12.80	31.60	44.10	–	–
TC-MGC (Jing et al., 2025)	0.70	3.40	6.60	–	–	10.10	28.30	40.60	–	–
TempMe (Shen et al., 2025a)	1.48	4.79	7.71	23.70	201.00	–	–	–	–	–
X-Pool (Gorti et al., 2022)	1.34	5.26	5.46	23.74	205.00	13.10	33.80	45.50	–	–
InternVideo2† (Wang et al., 2024b)	2.87	9.46	14.65	37.96	98.00	17.40	37.72	49.36	73.78	11.00
ReContext†	**3.17**	**10.08**	**15.81**	**41.02**	**77.00**	**24.20**	**47.07**	**59.03**	**81.38**	**6.00**

To retain a prominent match while accounting for the remaining fragments, we combine the maximum and mean of these scores:

$$s^e(V,T) = (1-\alpha) \max_{1\le j\le N_t} s_j^e(V,T) + \frac{\alpha}{N_t}\sum_{j=1}^{N_t} s_j^e(V,T), \tag{12}$$

where $0 \le \alpha \le 1$. Setting $\alpha = 0$ retains only the strongest element match, whereas $\alpha = 1$ averages the best matches of all description elements. For videos with equally strong best cues, a positive mean weight favors stronger average support for the remaining fragments. These matches can originate from different activities, so the aggregation assesses description support over the full video.

The context branch combines context compatibility with support from its constituent details, and the video-wide branch collects support for the description across activities. We combine them as:

$$s(V,T) = (1-\lambda)s^e(V,T) + \lambda s^c(V,T), \tag{13}$$

where $0 \le \lambda \le 1$ controls the contribution of context compatibility relative to aggregated local support. The fusion directly reuses the candidate-relative scores from both branches without an additional normalization step. Context coupling influences the retrieval score through member-supported context selection. The local matrix is reused throughout, and the same visual match can serve multiple linguistic elements. For text-to-video retrieval, we rank $V \in \mathcal{V}$ for each description; for video-to-text retrieval, we rank $T \in \mathcal{T}$ for each video.

4 Experiments

4.1 Experimental Setup

We evaluate on ActivityNet-Event and Charades-Event following MeVTR (Zhang et al., 2023), without ground-truth temporal boundaries as inputs. We report Recall@K and Median Rank (mRank) for text-to-video (T2V) retrieval, and Average, One-Hit, and All-Hit Recall@K for video-to-text (V2T) retrieval. ReContext uses frozen InternVideo2 representations (Wang et al., 2024b) without task-specific training. Dataset statistics, metric definitions, and implementation settings are provided in Appendix B.

4.2 Main Bidirectional Retrieval Results

Tabs. 1–3 report bidirectional retrieval results. The flat InternVideo2 baseline shares our encoder family; comparisons with task-supervised systems span different backbones. ReContext achieves the highest T2V recall among the listed methods at all reported cutoffs on both datasets. Relative to flat InternVideo2, R@1 increases by 0.30% on Charades-Event and 6.80% on ActivityNet-Event, with R@50 gains of 3.06% and 7.60%. For V2T, ReContext leads in Average and One-Hit recall at every reported cutoff on both datasets. R@10-Average exceeds the strongest listed baseline

Table 2: Video-to-text retrieval on Charades-Event. Each cell gives Average / One-Hit / All-Hit Recall (%). R@1-All-Hit is unreported for the comparison methods.

Method	R@1	R@5	R@10	R@50
CenterCLIP	1.32 / 3.00 / –	3.71 / 7.87 / 1.57	6.33 / 13.42 / 2.62	17.37 / 33.43 / 7.72
CLIP4Clip	1.32 / 2.92 / –	4.07 / 8.32 / 1.80	5.65 / 11.84 / 2.55	16.84 / 32.01 / 7.65
GMMFormer*	0.61 / 1.27 / –	2.74 / 5.92 / 1.20	4.34 / 9.82 / 1.57	16.15 / 30.96 / 6.82
MeVTR (avg)	1.39 / 3.22 / –	5.04 / 9.90 / **2.47**	7.02 / 13.72 / 3.37	17.62 / 32.08 / 8.70
MeVTR (max)	1.01 / 2.62 / –	4.03 / 9.00 / 1.72	6.39 / 13.72 / 2.70	19.08 / 36.66 / 8.10
TempMe	1.06 / 2.47 / –	3.96 / 7.75 / 1.91	6.27 / 11.79 / **3.47**	15.76 / 28.14 / 7.60
X-Pool	0.95 / 1.90 / –	3.75 / 7.07 / 2.12	5.11 / 10.22 / 2.65	12.59 / 24.51 / 5.60
InternVideo2†	2.04 / 4.80 / –	5.34 / 12.29 / 1.87	7.61 / 16.41 / 3.00	19.39 / 36.88 / 8.47
ReContext†	**2.19 / 5.40** / –	**5.88 / 13.27** / 2.10	**8.25 / 18.52** / 3.15	**20.80 / 39.96 / 9.15**

Table 3: Video-to-text retrieval on ActivityNet-Event. Each cell gives Average / One-Hit / All-Hit Recall (%). R@1-All-Hit is unreported for the comparison methods.

Method	R@1	R@5	R@10	R@50
CenterCLIP	6.93 / 21.62 / –	20.58 / 47.98 / 3.66	29.90 / 61.72 / 6.96	58.88 / 90.34 / 26.28
CLIP4Clip	6.60 / 20.74 / –	20.32 / 47.02 / 3.46	29.37 / 61.26 / 6.71	58.41 / 89.77 / 26.24
GMMFormer*	2.62 / 8.60 / –	10.11 / 28.71 / 0.47	17.14 / 43.63 / 1.40	45.00 / 80.56 / 13.93
MeVTR (avg)	8.52 / 26.91 / –	23.56 / 54.18 / 3.86	32.95 / 67.13 / 7.79	61.26 / 92.50 / 27.80
MeVTR (max)	6.98 / 22.72 / –	18.72 / 46.37 / 2.30	27.04 / 60.62 / 5.76	54.19 / 88.37 / 20.90
InternVideo2†	8.29 / 26.04 / –	21.54 / 51.59 / 2.75	30.00 / 64.94 / 5.72	56.47 / 89.97 / 23.93
ReContext†	**11.08 / 35.31** / –	**28.05 / 63.60 / 4.20**	**37.46 / 75.23 / 8.49**	**65.41 / 95.59 / 31.90**

by 0.64% on Charades-Event and 4.51% on ActivityNet-Event. The improvements are larger on ActivityNet-Event in both retrieval directions. All-Hit gains are less uniform: ActivityNet-Event R@10-All-Hit improves by 0.70%, whereas Charades-Event reaches 3.15%, below TempMe's 3.47%. Thus, improved retrieval of relevant descriptions does not consistently translate into recovering the complete annotated set within the top ten.

4.3 Ablation Study

Tab. 4 compares six scoring configurations. Baseline is the frozen InternVideo2 similarity without parsing or ReContext branches. **Local** denotes element-level support s^e; **Context** denotes context compatibility $\max_k \bar{s}^c_k$; **Coupled context** uses s^c without video-wide support. **Fusion** denotes independent branch fusion of s^e and context compatibility, whereas ReContext combines s^e with context-coupled selection. Structured rows share parsed inputs and frozen representations.

Table 4: Component ablations on the ActivityNet dataset.

Configuration	T2V R@1	V2T R@1 Avg./One/All
Baseline	17.41	8.30 / 26.04 / –
Local	20.58	9.49 / 31.23 / –
Context	22.69	10.55 / 33.55 / –
Coupled context	23.25	10.72 / 34.08 / –
Fusion	24.02	10.91 / 34.94 / –
ReContext	**24.20**	**11.08 / 35.31 / –**

Element-level support and context compatibility each improve over the baseline in both retrieval directions. Their fusion outperforms either branch alone, reaching 24.02% T2V R@1. Local details and compatibility with activity context therefore provide complementary retrieval cues. Member-supported enhancement further improves the context-only branch by 0.56% in T2V R@1. With video-wide support included, ReContext improves over Fusion by 0.18% in T2V R@1 and 0.17%/0.37% in V2T R@1-Average/One-Hit. These results motivate examining where context compatibility helps and when preserving local–context association adds further benefit.

4.4 Local Ambiguity and Activity Context

We first examine whether similar local matches also imply similar activity contexts. We compare visual elements from the annotated video and a competing video that strongly match the same description fragment with similar scores, obtaining 36,848 pairs across 5,667 queries.

In Fig. 3, Each point pairs visual elements matching the same fragment; L is their cosine similarity and C that of their most similar containing contexts. Colored dashed lines mark means; the diagonal is $C = L$. Local and context similarities increase together overall, but the vertical spread shows

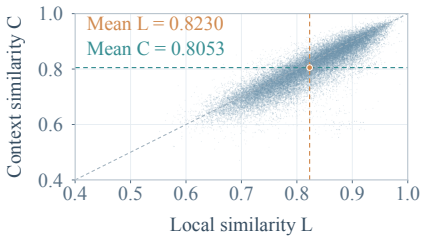

Figure 3: Local ambiguity analysis.

Table 5: ActivityNet-Event T2V R@1 (%). Subset columns give query counts.

Configuration	All	8,502	1,981
Local	20.58	19.94	19.08
Context	22.69	25.48	27.81
Fusion	24.02	**26.76**	29.63
ReContext	**24.20**	26.51	**29.73**

that comparable local matches can occur in differently represented contexts. Points below $C = L$, including those at high L, have less similar contexts than local elements; the mean gap is 0.0177. Thus, local resemblance does not fully describe the surrounding activity context. The retrieval comparison below tests whether this additional information helps distinguish locally similar candidates.

We then evaluate retrieval on a stricter 1,981-query subset of the 8,502 queries with multiple fragments and contexts. Each selected query has a fixed pair with high local similarity, lower context similarity, and nonempty, disjoint video-level activity labels. Selection uses matching scores and labels without retrieval success or gain.

In Tab. 5, element-level support performs worse on the local-confusion subset, while the benefit of incorporating context is larger. ReContext improves R@1 over element-level support by 10.65% on this subset, compared with 6.57% among all 8,502 queries with multiple fragments and contexts and 3.63% overall, using unrounded recalls. The larger gain therefore persists when the comparison is restricted to queries with multiple fragments and contexts. Independent branch fusion achieves a similar 10.55% gain on the subset without member-supported enhancement, supporting context compatibility as the main source of improvement under local confusion. A label-based control also finds larger local–context similarity gaps without similarity-threshold filtering (Appendix D.1).

4.5 Context-Coupling Analysis

We next test whether retrieval depends on which context receives member support. Shuffle randomly reassigns that support among contexts within each video while preserving response values and scoring parameters; its results are averaged over ten seeds. We analyze 1,758 queries where high context compatibility aligns with strong member support and member-supported enhancement is active. Queries are ordered from weak to strong enhancement and grouped using fixed development-input quintiles. The horizontal labels give each group's cumulative query-position interval in this ordering, not retrieval ranks.

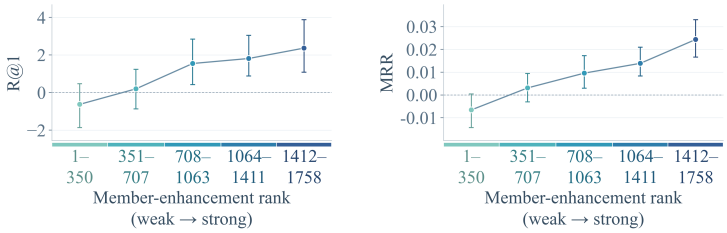

Figure 4: Local–context association experiment on ActivityNet-Event. Panels show ReContext minus Shuffle in T2V R@1 (%, left) and MRR (right). Band widths indicate group sizes; darker colors indicate stronger member-supported enhancement. Error bars show 95% video-cluster bootstrap intervals.

Fig. 4 shows an increasing advantage over Shuffle across the five enhancement groups. The R@1 difference rises from -0.63% to $+2.36\%$ in the left panel, while the MRR difference rises from -0.0066 to $+0.0244$ in the right. For both metrics, confidence intervals span zero in the two weakest groups and lie above zero in the three stronger groups. Since Shuffle changes only which context receives member support, the growing difference indicates that preserving local–context association matters more when that support provides stronger enhancement. The weakest group has no clear benefit despite active enhancement. Additional controls show no overall improvement over Shuffle and a decrease under low response alignment (Appendix D.3).

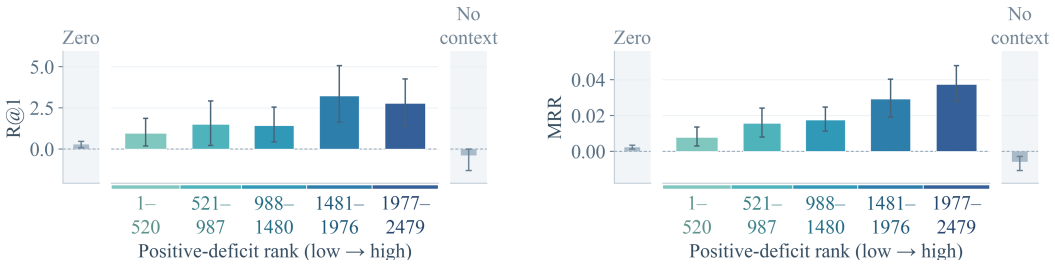

Figure 5: Description-support scope experiment on ActivityNet-Event. Left and right panels show ReContext minus Restricted in T2V R@1 (%) and MRR; positive values favor video-wide support. The central bars cover 2,479 queries with $D > 0$, ordered by D and grouped using fixed development-input quintiles. Labels give cumulative query positions in this ordering, not retrieval ranks; bar widths indicate group sizes and darker colors indicate larger deficits. Gray side panels show $D = 0$ (14,613 queries) and undefined D (258 queries without a nonempty context); their widths do not encode counts. Error bars show 95% video-cluster bootstrap intervals.

4.6 Support Across Activity Contexts

Different description fragments can find their best visual matches in different activity contexts. We test whether restricting these matches to individual contexts loses support available elsewhere in the video. ReContext searches the full video for each fragment, whereas Restricted scores all fragments within each selected context before combining the resulting local scores with the context weights. Both variants share features, context scoring, selection, and fusion weights.

We group queries by how well a single context in the annotated video retains the fragments' best video-wide matches. The support deficit D measures the normalized shortfall in the best such context. Larger D means that evaluating all fragments within one context leaves more of their available support unaccounted for.

Fig. 5 shows greater benefit from video-wide matching as a single context becomes less able to retain support for the description. Across the positive-deficit groups, MRR gains increase from 0.0079 to 0.0374. R@1 gains are positive throughout, increasing from 0.96% in the lowest group to 2.78% in the highest, although they peak at 3.23% in the preceding group. The MRR trend and nonmonotonic R@1 gains indicate a growing ranking benefit without a uniform increase in first-place retrieval. With context scoring fixed, this supports retaining each fragment's best match across the video. Membership can guide context scoring without requiring that context to explain every fragment of the description.

The zero-deficit group has a smaller R@1 gain of 0.27%; a context retaining the fragments' near-best matches exists but need not be selected. Queries without a nonempty context have a negative difference because competing videos' scores still change. A supplementary comparison directly examines descriptions whose fragment matches require multiple contexts and finds larger gains than when one context suffices (Appendix D.4).

5 Conclusion

This work studies how local details and their activity context jointly contribute to relevance in multi-event video–text retrieval. ReContext evaluates contexts together with their constituent matches while gathering description support across the video. The resulting score supports both retrieval directions, with context membership determining which details contribute to each context score. Our analysis shows that this coupling can distinguish score configurations with identical independently pooled local and contextual maxima. More broadly, the formulation shows how aggregation can retain or discard the association between local details and their activities even with fixed representations. For video–language retrieval, it makes the associations retained during scoring an explicit design consideration alongside the content captured by the representations.

AI USE STATEMENT

AI-based tools were used solely to polish the language of this manuscript, including grammar, phrasing, and readability. They were not used to develop the method, design the experiments, generate results, or make research decisions. The authors reviewed and verified the final text.

REFERENCES

Chuanshen Chen, Kai Zhou, Feiqi Wang, Yutao Ning, Zhendong Xiong, Yirui Li, Zhiquan Wen, and Mingkui Tan. CaptAin: Caption-driven alignment for bridging modality gaps in partially relevant video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 6208–6217, 2026a.

Chuanshen Chen, Kai Zhou, Zhiquan Wen, Zeng You, Yirui Li, Tianhang Xiang, and Mingkui Tan. Action-and-object aware alignment for partially relevant video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2814–2822, 2026b.

Xianke Chen, Daizong Liu, Xun Yang, Xirong Li, Jianfeng Dong, Meng Wang, and Xun Wang. PRVR: Partially relevant video retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(2):1262–1277, 2026c.

Shuaiqi Cheng, Siyu You, Yanbi Wu, Yuxi Chen, Jiahao Zhang, and Xuming Hu. Concentrate after imagination: Text-conditioned evidence grounding for partially relevant video retrieval. *arXiv preprint arXiv:2609.08999*, 2026.

Cheol-Ho Cho, WonJun Moon, WooJin Jun, MinSeok Jung, and Jae-Pil Heo. Ambiguity-restrained text-video representation learning for partially relevant video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2500–2508, 2025.

Jianfeng Dong, Xianke Chen, Minsong Zhang, Xun Yang, Shujie Chen, Xirong Li, and Xun Wang. Partially relevant video retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 246–257, 2022.

Adriano Fragomeni, Michael Wray, and Dima Damen. ConTra: (con)text (tra)nsformer for cross-modal video retrieval. In *Proceedings of the Asian Conference on Computer Vision*, pp. 3481–3499, 2022.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5267–5275, 2017.

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

Donghoon Han, Eunhwan Park, Gisang Lee, Adam Lee, and Nojun Kwak. MERLIN: Multimodal embedding refinement via LLM-based iterative navigation for text-video retrieval-rerank pipeline. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 547–562, Miami, Florida, US, November 2024. Association for Computational Linguistics.

Xiaolun Jing, Genke Yang, and Jian Chu. TC-MGC: Text-conditioned multi-grained contrastive learning for text–video retrieval. *Information Fusion*, 121:103151, 2025.

WooJin Jun, WonJun Moon, Cheol-Ho Cho, MinSeok Jung, and Jae-Pil Heo. Bridging the semantic granularity gap between text and frame representations for partially relevant video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4166–4174, 2025.

Dohwan Ko, Ji Soo Lee, Minhyuk Choi, Zihang Meng, and Hyunwoo J. Kim. Bidirectional likelihood estimation with multi-modal large language models for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22263–22273, 2025.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 706–715, 2017.

Jun Li, Jinpeng Wang, Chaolei Tan, Niu Lian, Long Chen, Yaowei Wang, Min Zhang, Shu-Tao Xia, and Bin Chen. Enhancing partially relevant video retrieval with hyperbolic learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23074–23084, 2025.

Jun Li, Peifeng Lai, Xuhang Lou, Jinpeng Wang, Yuting Wang, Ke Chen, Yaowei Wang, and Shu-Tao Xia. Revisiting uncertainty: On evidential learning for partially relevant video retrieval. In *Forty-Third International Conference on Machine Learning*, 2026a.

Jun Li, Xuhang Lou, Jinpeng Wang, Yuting Wang, Yaowei Wang, Shu-Tao Xia, and Bin Chen. Imagine before concentration: Diffusion-guided registers enhance partially relevant video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9710–9721, 2026b.

Xuye Liu, Yimu Wang, and Jian Zhao. ELIOT: Zero-shot video-text retrieval through relevance-boosted captioning and structural information extraction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pp. 381–391, Albuquerque, USA, April 2025. Association for Computational Linguistics.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing*, 508: 293–304, 2022.

WonJun Moon, Cheol-Ho Cho, Woojin Jun, Taeoh Kim, Inwoong Lee, Dongyoon Wee, Minho Shim, and Jae-Pil Heo. Prototypes are balanced units for efficient and effective partially relevant video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21789–21799, October 2025a.

WonJun Moon, MinSeok Jung, Gilhan Park, Tae-Young Kim, Cheol-Ho Cho, Woojin Jun, and Jae-Pil Heo. Mitigating semantic collapse in partially relevant video retrieval. In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

Prasanna Reddy Pulakurthi, Jiamian Wang, Majid Rabbani, Sohail Dianat, Raghuveer Rao, and Zhiqiang Tao. X-CoT: Explainable text-to-video retrieval via LLM-based chain-of-thought reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 31184–31195. Association for Computational Linguistics, 2025.

Leqi Shen, Tianxiang Hao, Tao He, Sicheng Zhao, Yifeng Zhang, Pengzhang Liu, Yongjun Bao, and Guiguang Ding. TempMe: Video temporal token merging for efficient text-video retrieval. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. LongVU: Spatiotemporal adaptive compression for long video-language understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54582–54599. PMLR, 2025b.

Jiamian Wang, Guohao Sun, Pichao Wang, Dongfang Liu, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. Text is MASS: Modeling as stochastic embedding for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16551–16560, 2024a.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2: Scaling foundation models for multimodal video understanding. In *Computer Vision – ECCV 2024*, pp. 396–416. Springer, 2024b.

Yuting Wang, Jinpeng Wang, Bin Chen, Ziyun Zeng, and Shu-Tao Xia. GMMFormer: Gaussian-mixture-model based transformer for efficient partially relevant video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5767–5775, 2024c.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. VideoTree: Adaptive tree-based video representation for LLM reasoning on long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3272–3283, June 2025.

Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. Cap4Video: What can auxiliary captions do for text-video retrieval? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10704–10713, 2023.

Bingqing Zhang, Zhuo Cao, Heming Du, Yang Li, Xue Li, Jiajun Liu, and Sen Wang. Quantifying and narrowing the unknown: Interactive text-to-video retrieval via uncertainty minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22120–22130, 2025a.

Gengyuan Zhang, Jisen Ren, Jindong Gu, and Volker Tresp. Multi-event video-text retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22113–22123, October 2023.

Long Zhang, Peipei Song, Jianfeng Dong, Kun Li, and Xun Yang. Enhancing partially relevant video retrieval with robust alignment learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 4615–4629. Association for Computational Linguistics, 2025b.

Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. CenterCLIP: Token clustering for efficient text-video retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 970–981, 2022.

ABSTRACT

This appendix provides background and analysis for ReContext. We review related work on multi-event retrieval, context modeling, and pretrained video–text models, and record the implementation details needed to reproduce the evaluation. We then analyze how context-coupled scoring preserves local–context associations that independent pooling discards. The analysis establishes the independent-pooling bound, illustrates when it is strict, and examines parameter limits, partial description support, and calibration invariance. Additional diagnostics examine local confusion, fragment support, enhancement strength, parameter sensitivity, and remaining retrieval failures.

Contents

A Related Work **13**

A.1 Multi-Event Video–Text Retrieval 13

A.2 Context Modeling in Video Understanding 14

A.3 Pretrained Models for Video–Text Retrieval 14

B Implementation Details **14**

C Analysis of Context-Coupled Scoring **16**

D Additional Experiments **18**

D.1 Controls for the Local-Confusion Subset 18

D.2 Variation in Fragment Support 18

D.3 Context Maxima and Pairing Controls 19

D.4 Controls for Description-Support Scope 20

D.5 Failure Analysis 20

D.6 Efficiency Analysis 21

D.7 Parameter and Robustness Checks 21

A Related Work

A.1 Multi-Event Video–Text Retrieval

Video–text retrieval methods improve cross-modal compatibility through pretrained representations (Luo et al., 2022), text-conditioned aggregation and interaction (Gorti et al., 2022; Jing et al., 2025), and stochastic text representations that accommodate diverse video semantics (Wang et al., 2024a). Efficiency-oriented methods reduce video encoding costs through token clustering and temporal merging (Zhao et al., 2022; Shen et al., 2025a). When a video contains several activities, relevance also depends on which parts of its content support a description. Partially relevant video retrieval studies this setting for text-to-video search (Dong et al., 2022), while multi-event video–text retrieval explicitly supports both retrieval directions and descriptions with different event scopes (Zhang et al., 2023).

Research on partial relevance has developed multi-scale video representations (Chen et al., 2026c; Wang et al., 2024c), alignment across semantic granularities (Jun et al., 2025; Li et al., 2025), and objectives that preserve distinctions among events (Moon et al., 2025b). Related approaches address ambiguous training correspondences (Cho et al., 2025) and uncertainty in cross-modal alignment (Zhang et al., 2025b; Li et al., 2026a). These advances address how to represent and align the relevant parts of a video. We study which associations between local details and their activity contexts remain available when these matching responses are reduced to a retrieval score.

A.2 Context Modeling in Video Understanding

Context modeling connects local observations to the surrounding events that help explain them. In video retrieval, temporal neighborhood aggregation (Fragomeni et al., 2022) and action/object-aware alignment (Chen et al., 2026b) connect local evidence to broader video semantics, while prototype and diffusion-based representations (Moon et al., 2025a; Li et al., 2026b) provide complementary context for local matching. Long-video understanding also studies how hierarchical selection and adaptive compression expose query-relevant content under limited input budgets (Wang et al., 2025; Shen et al., 2025b).

Context also influences retrieval through score refinement, as in concurrent work TRACE (Cheng et al., 2026), which uses query-conditioned global information to calibrate local responses before temporal selection. ReContext combines each context score with matches from its own constituent spans before selecting the highest-scoring context, preserving membership in this aggregation.

A.3 Pretrained Models for Video–Text Retrieval

Pretrained models support retrieval refinement by enriching the information available for matching and revisiting initial ranking decisions. Caption-based methods supply video-level or temporally localized descriptions for alignment and similarity estimation (Wu et al., 2023; Chen et al., 2026a), while structured video textualization supports zero-shot matching (Liu et al., 2025). Query-focused approaches use language-model feedback and user interaction to clarify retrieval intent (Han et al., 2024; Zhang et al., 2025a). At ranking time, language-model reasoning enables comparisons between candidate video descriptions (Pulakurthi et al., 2025), and multimodal likelihood estimation provides another basis for scoring and calibration (Ko et al., 2025). ReContext uses frozen pretrained models to construct semantic elements, activity contexts, and their representations. Keeping these outputs fixed during scoring allows us to examine the contribution of local–context association separately from encoder adaptation.

B Implementation Details

Datasets and metrics. The released MeVTR benchmark (Zhang et al., 2023) provides 4,917 videos and 17,505 descriptions in ActivityNet-Event's *val1* split, derived from ActivityNet Captions (Krishna et al., 2017), and 1,334 videos and 3,720 descriptions in the Charades-Event test split, derived from Charades-STA (Gao et al., 2017).[1] Each retrieval item is a complete video or an individual description; ground-truth temporal boundaries are excluded from the matching inputs. Recall is reported as a percentage at $K \in \{1, 5, 10, 50\}$. T2V Recall@K measures the fraction of queries with a relevant video in the top K, and mRank is the median rank of the relevant video. For V2T, Recall@K-Average measures the fraction of relevant descriptions recovered per video, One-Hit measures whether any relevant description is found, and All-Hit measures whether the full annotated set is recovered. All three V2T metrics are averaged over query videos.

Models and scoring. ReContext uses the frozen InternVideo2-CLIP-1B-224p-f8 checkpoint (Wang et al., 2024b), while Qwen3.5-Plus performs text decomposition and video-span organization. Each visual element or context is encoded using at most four frames. Element frames are uniformly sampled within their spans, and context frames are drawn from constituent elements. The whole-video baseline uses 32 uniformly sampled frames. Decompositions, memberships, and encoder features are cached, and parsing and encoding costs are reported separately from reference normalization and pairwise scoring. Both retrieval directions use a shared video–description score with the full evaluation video collection as the fixed normalization reference; controlled scoring comparisons reuse identical representations. We set $\tau^e = 0.0125$ and $\tau^c = 0.015$ for element and context calibration, respectively. The cached experiments use $\alpha = 0.15$, $\beta = 0.6$, and $\lambda = 0.8$, with the same settings on Charades-Event. Member support averages up to three highest responses per fragment before maximizing over fragments; the context branch combines its maximum and top-three mean with weights 0.25 and 0.75.

[1] The cached ActivityNet results use 4,877 videos, 17,497 V2T candidate descriptions, and 17,350 valid T2V queries. Eight descriptions without valid parsed actions are excluded; 147 retained descriptions have no positive video in the gallery and serve only as V2T candidates.

Figs. 7 and 8 examine the scoring weights and aggregation counts, respectively. Code and processed data will be publicly released.

Local-confusion selection. We construct pairs that are difficult to distinguish from their local matches but differ in the contexts containing those matches. We first retain 8,502 queries with at least two distinct description fragments and two activity contexts with distinct, nonempty member sets in the annotated video. For each fragment, we compare its strongest element match in the annotated video with matches from its ten highest-ranked distinct videos, excluding the annotated video. Each matched visual element belongs to at least one activity context. Their responses must reach the fragment's 90th percentile over 2,048 fixed reference elements and differ by at most $0.1(p_{95} - p_{05})$, where p_r is the corresponding rth percentile. These conditions yield the 36,848 pairs across 5,667 queries in Fig. 3. For the retrieval subset, their visual cosine similarity must additionally reach the 95th percentile of fixed random cross-video element pairs, approximately 0.732464.

For each matched visual element, we retain its containing contexts except those whose member sets strictly contain that of another context containing the same element. Features of contexts with identical member sets are averaged and normalized before comparison. Even the most similar retained context pair must have cosine similarity at least 0.03 below that of the elements. A fixed product of fragment weight and this similarity gap selects one qualifying pair per query. Fragment weights combine dominant-fragment frequencies over the Local top-20 candidates with uniform weights in a 0.85/0.15 ratio. We then require nonempty, disjoint official ActivityNet v1.3 activity-label sets, without replacing a pair that fails this check, yielding 1,981 queries. These exploratory rules use annotated video identity and video-level labels, but no temporal annotations or per-query retrieval outcomes. All scoring variants share frozen features, corresponding component parameters, and the candidate gallery.

Description-support scope. The comparison in Section 4.6 tests whether evaluating all fragments within each selected context misses matches available elsewhere in the video. Restricted computes each fragment's best response within each selected context, aggregates these responses as $0.15\,\mathrm{mean} + 0.85\,\mathrm{max}$, and then weights the resulting local scores. ReContext applies the same fragment aggregation to video-wide best responses. Both use the cached context branch $0.25\,\mathrm{max} + 0.75\,\mathrm{top3mean}$, giving selected-context weights of $0.5/0.25/0.25$, $0.625/0.375$, or 1 for three, two, or one context, respectively. The final weights are 0.8 for the context branch and 0.2 for the local branch. Restricted does not pool the union of selected contexts' members. Both variants fall back to video-wide local support for videos without nonempty contexts.

For grouping, let a_{ij} be the base response of visual element i to fragment j, $u_j = \max_i a_{ij}$ its video-wide best support, and $u_{kj} = \max_{i \in \mathcal{I}_k} a_{ij}$ its best support within context k. Percentiles $p_{r,j}$ use 2,048 reference visual elements sampled with seed 20260925. Set $d_j = p_{95,j} - p_{05,j}$ and $\epsilon_j = \max\{0.05 d_j, 2 \times 10^{-6} \max_{i \in \mathcal{R}_e} |a_{ij}|\}$, where $\mathcal{R}_e$ is this reference set. The support deficit is

$$D(V,T) = \min_{k:\mathcal{I}_k \neq \varnothing} \frac{1}{N_t} \sum_{j=1}^{N_t} \frac{[u_j - u_{kj} - \epsilon_j]_+}{\max(d_j, 10^{-8})}. \tag{14}$$

Taking the minimum asks whether even the most supportive single context loses fragment matches retained by video-wide aggregation. We use the annotated video, deduplicate fragments and member sets for grouping, and consider all context levels. The statistic does not depend on context selection or retrieval outcomes and does not count activities.

All 17,350 valid queries enter Fig. 5. Zero deficits remain one group; absent nonempty contexts make D undefined. Positive deficits use development-input quintile boundaries of approximately 0.017983, 0.042375, 0.087476, and 0.233019, fixed across partitions. Lower bounds are included and upper bounds excluded; identical values are not split. The displayed positions run only over the 2,479 positive-deficit queries, without indicating retrieval rank or numerical distances in D.

For the supplementary support groups, a context covers fragment j if $u_{kj} \geq p_{90,j}$ and $u_{kj} \geq u_j - \epsilon_j$. Of the 8,502 queries with multiple fragments and distinct nonempty context member sets, 708 lack strong video-wide support for at least one fragment and 765 more lack joint context coverage, leaving 7,029. Cross-context support requires joint coverage but no single context covering every fragment (817 queries). Compact support requires one minimal context to cover every fragment (3,910 queries).

A context is minimal if no other nonempty context has a member set strictly contained in its own. The remaining 2,302 require a broader parent context. All remain in the global analysis. These groups describe response and membership structures, not annotated activity categories.

Groups use annotated video identity, base responses, and memberships without filtering on retrieval gains. Both variants rank the same 4,877-video gallery; intervals use 2,000 video-cluster bootstrap resamples. These are diagnostics on historical data, not independent confirmation experiments.

Context–member response pairing. For the cached diagnostics in Section 4.5, C_k denotes context compatibility and E_k the support from context k's own members. The latter averages the top three member responses for each fragment, using all members if fewer than three are available, then takes the maximum over fragments. Contexts with identical member sets are deduplicated during extraction. Shuffle permutes complete member-response bundles among the resulting contexts within every gallery video using seeds 0–9. Each video's permutation is shared across queries. Context scores, member-response values, video-wide support, and scoring parameters remain fixed; metrics are averaged after retrieval is run separately for each seed.

For the maximum-response statistics, we use 8,502 queries with multiple distinct fragments and context member sets, and 4,383 candidate videos with at least two distinct, nonempty member sets. This gives 37,264,266 pairs. A context belongs to a maximizing set if its response differs from that component's maximum by at most 10^{-6}. We count pairs whose two maximizing sets do not intersect, retaining ties and overlapping memberships. The candidate comparison filters each query's fixed top 20 unannotated videos ranked by element-level support to the same video requirements. No response-strength or retrieval-gain filter enters these statistics.

We test pairing on queries with strong context and member responses that vary across contexts. Reference percentiles for C_k, E_k use 2,048 context rows sampled with seed 20260925. Let d_C, d_E be their 95th–5th percentile spreads, floored at 10^{-8}. We require both maxima in the annotated video to reach their reference 90th percentiles, both response ranges to reach 0.1 reference spreads, and mean Jaccard distance between distinct member sets to reach 0.5. These conditions retain 6,408 of the 8,502 queries. For each query, enhancement strength is the largest normalized excess $(E_k - C_k)/d_E$ among contexts in the annotated video satisfying $\max_r C_r - C_k \leq 0.1 d_C$. This measures how strongly members can enhance contexts with near-maximal compatibility. We retain 1,758 queries with positive enhancement strength, Pearson correlation at least 0.5, and a context within 0.1 reference spreads of both maxima. The low-alignment control instead requires nonpositive correlation and positive enhancement strength, yielding 264 queries. It does not require a context near both maxima, so the comparison does not isolate correlation as the sole factor.

Development-input quintiles of enhancement strength give approximate boundaries 0.046525, 0.098641, 0.166230, and 0.283539, applied unchanged across video partitions. Bins include their lower bounds and exclude their upper bounds; the lowest bin contains only positive values, and identical values are not split. Fig. 4 labels each group by its cumulative query-position interval in ascending enhancement order; each point summarizes all queries in that interval. The horizontal spacing reflects group sizes, not numerical differences in enhancement strength. Rules and the stronger-enhancement threshold of 0.25 were developed using historical results, while per-query selection and binning use only memberships and input responses. The reused partitions provide internal checks.

C Analysis of Context-Coupled Scoring

The key distinction is whether local support is combined with its context before or after the maxima are taken. We analyze this ordering with the calibrated local matches, context scores, and reference collection fixed; membership relations determine which local entries are pooled into each context-local score L_k. As in Eq. (10), let $C_k = \bar{s}^c_k$ and $L_k = s^{e|c}_k$, with maxima $C_\star$ and $L_\star$ over contexts. We first establish what is retained by coupling local matches to their contexts, then examine how the rule accommodates partial description support. All statements concern numerical matching weights under the specified structures and reference collection; membership changes below are diagnostic interventions on the scoring operator, whereas regrouping video content would also change the encoded context responses.

Table 6: Constructed score configurations illustrating the role of local–context association at $\beta = 0.5$. S_{ind} denotes the right-hand side of Eq. (10). The configurations share the same component maxima and differ in which context supplies the stronger local match.

Configuration	C_1	C_2	L_1	L_2	s^c	S_{ind}
Same-context maxima	0.06	0.02	0.09	0.01	0.075	0.075
Different-context maxima	0.06	0.02	0.01	0.09	0.060	0.075

Coupled selection and independent pooling. For $0 \leq \beta \leq 1$, each enhanced context score satisfies:

$$C_k + \beta[L_k - C_k]_+ = \max\{C_k, (1-\beta)C_k + \beta L_k\}. \tag{15}$$

Here, $[x]_+ = \max\{0, x\}$. If $L_k \leq C_k$, both sides equal C_k; otherwise, both equal $(1-\beta)C_k + \beta L_k$. Taking the maximum over k gives the first line of Eq. (10). The second line follows from $C_k \leq C_\star$ and $L_k \leq L_\star$ with nonnegative coefficients. Thus, independently maximizing the two components provides an upper bound on their coupled combination. The bound can be viewed as a relaxation that allows the two maxima to come from different contexts; the gap therefore quantifies the score difference caused by discarding membership at the scoring-operator level. Because $\lambda \geq 0$, replacing s^c with this pooled quantity also gives an upper bound on the fused score for the same s^e. Retrieval accuracy still depends on the quality of the representations and memberships.

The second configuration in Tab. 6 makes the inequality strict, whereas the first attains the bound. Both configurations have the same context maximum and local-support maximum. A score computed solely from those two extrema assigns the same value to both configurations. Member-restricted combination retains their different pairings and yields different scores. This establishes that the coupled result cannot, in general, be recovered from the independently pooled extrema alone.

This distinction can carry through to the final retrieval score. For the two configurations in Tab. 6, suppose the visual elements are reassigned between contexts while the calibrated local matrix and context scores remain fixed. The video-wide score s^e is then identical. By Eq. (13), the coupled final scores differ by 0.015λ, while replacing the context branch with the independently pooled bound yields equal final scores. This construction shows how a change in membership can affect the full score for $\lambda > 0$ with matching responses fixed. Regrouping and re-encoding video content changes the representations as well and requires a separate control.

Parameter limits. Under this fixed-response comparison, at $\beta = 0$ the context branch reduces to $C_\star$, while at $\beta = 1$ it becomes $\max(C_\star, L_\star)$. Both limits depend only on independently pooled extrema. If the union of context members is held fixed, reassigning those members among contexts while holding the context scores fixed cannot change $L_\star$. Thus, in this comparison, the explicit pairing constraint can affect the context score only for $0 < \beta < 1$. An intermediate value is necessary but insufficient for a strict gap between the coupled score and its independent-pooling upper bound. More precisely, for $0 < \beta < 1$, the gap is strict exactly when $L_\star > C_\star$ and no single context attains both maxima; otherwise the bound is tight. At $\lambda = 0$, the final score uses only video-wide local support, irrespective of β. These properties motivate testing the contribution of explicit context coupling separately from the quality of the context representations, including enhancement activation and parameter endpoints.

Partial description support. Consider exposing one more fragment to the video-wide aggregation while holding the complete description, context scores, and existing calibrated fragment scores fixed. If its context-local score is no larger than the current L_k, both L_k and the enhanced score of context k remain unchanged. A detail supported in another activity can therefore enter the video-wide aggregation without introducing a missing-detail penalty in this context branch. For the mean term, adding the fragment increases the mean when its best video-wide match exceeds the previous mean, decreases it when the match falls below that mean, and leaves it unchanged when the two are equal; the maximum term can also change if the new match exceeds the previous maximum. The coefficient α controls the contribution of this mean to the video-wide score. This property assumes a fixed complete description and fixed context scores; editing the description falls outside that assumption. Under this fixed-score analysis, the enhancement does not subtract weak fragment responses, which motivates evaluating both valid partial support and spurious local matches.

Calibration invariance. For any fragment j, add a constant b_j to its similarities with every visual element, including all elements in the reference collection. In Eq. (6), the numerator and denominator are both multiplied by $\exp(b_j/\tau^e)$, so the calibrated weights are unchanged. The same argument applies to a shared shift of all description–context similarities. This invariance removes a component's shared additive baseline. It leaves sensitivity to score concentration, temperature, and reference-pool composition, and does not equalize the element and context distributions.

D Additional Experiments

D.1 Controls for the Local-Confusion Subset

We check whether the context-similarity difference persists when further selection uses only activity labels. Fig. 6 applies this condition to the 36,848 pairs in Fig. 3, retaining 18,601 pairs across 4,299 queries. It uses neither the visual-similarity thresholds nor the representative-pair selection used for the 1,981-query retrieval subset. We compare mean $L - C$ within fixed L intervals of width 0.05, retaining seven intervals with at least 20 pairs in each group and weighting them by the original pair counts. The resulting increase is 0.0052, with a 95% interval of $[0.0041, 0.0062]$ from 2,000 annotated-video cluster bootstrap resamples. The orange diamonds lie to the right of the blue circles in every interval, indicating larger local–context gaps after activity-label filtering at comparable local similarity. The largest separation occurs in the sparsely populated final interval and does not establish a monotonic trend. This supports examining context beyond local resemblance; video-level labels do not establish context-level semantic differences.

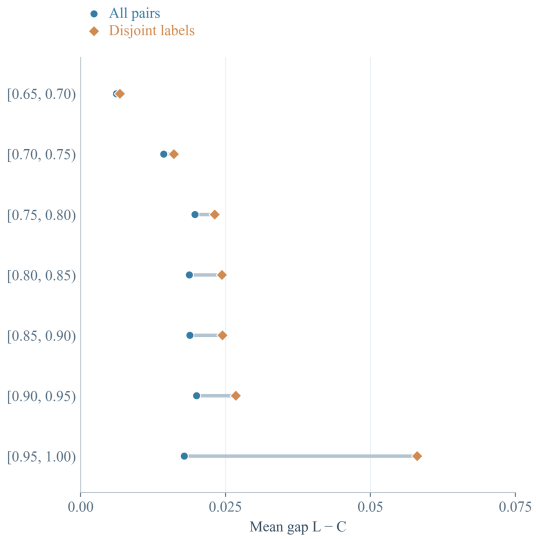

Figure 6: Activity-label control on ActivityNet-Event. Rows give local-similarity intervals; points show mean $L - C$ for all pairs (blue circles) and pairs with nonempty, disjoint activity labels (orange diamonds). L and C are defined in Fig. 3. Connectors join means, not confidence limits; the $[0.95, 1.00)$ interval contains only 32 retained pairs.

We also check whether the larger context benefit arises simply from requiring multiple fragments and contexts. Compared with all 8,502 queries satisfying these requirements, the local-confusion subset has a larger ReContext gain over element-level support, with an R@1 gain difference of 4.08%. The corresponding difference for independent branch fusion is 3.73%. Their 95% intervals, estimated from 2,000 video-cluster bootstrap resamples, are $[2.29, 5.85]\%$ and $[1.90, 5.63]\%$. Thus, the larger context benefit is not explained solely by requiring multiple fragments and contexts. Under the same local and context similarity conditions, comparison with queries whose matched videos share at least one official activity label gives a ReContext gain-difference interval of $[-0.97, 4.89]\%$. This interval includes zero, so the comparison does not establish a larger benefit specifically from differing activity categories.

D.2 Variation in Fragment Support

We examine whether context is more helpful when local support is concentrated in a few description fragments. For each of the 8,502 queries with multiple fragments and contexts, we measure fragment-support variation over the 20 videos ranked highest by element-level support. Within each video, fragment support is its best cached element response. We average the maximum–mean support gap over these videos and normalize by the standard deviation of element-level retrieval scores across the gallery. Queries are divided into five ascending bands using cutoffs of 0.5190, 0.8547, 1.1813, and 1.5721 on this normalized gap. This label- and gain-independent statistic measures support unevenness; it does not identify semantic ambiguity or count supporting activities.

Table 7: ActivityNet-Event T2V R@1 (%) by fragment-support score band, with the same video gallery across bands. Local and Context denote element-level support and context compatibility; Fusion denotes independent branch fusion. Differences use the underlying hit counts.

		Configurations				Differences	
Score band	Queries	Local	Context	Fusion	ReContext	Fusion − Local (%)	ReContext − Fusion (%)
Lowest	1,688	19.37	19.67	20.79	21.80	+1.42	+1.01
Lower	1,702	19.21	23.91	25.38	25.85	+6.17	+0.47
Middle	1,694	21.25	26.33	27.10	27.21	+5.84	+0.12
Higher	1,688	19.02	25.47	27.37	26.60	+8.35	−0.77
Highest	1,730	20.81	31.85	33.01	30.98	+12.20	−2.02

Table 9: ActivityNet-Event T2V R@1 (%). Shuffle averages ten seeds. Differences use unrounded recalls; brackets give 95% video-cluster bootstrap intervals.

Group	Queries	Shuffle	ReContext	ΔR@1 (%) [95% CI]
All queries		24.24	24.20	-0.04 $[-0.18, 0.10]$
Multiple fragments/contexts	8,502	26.60	26.51	-0.09 $[-0.29, 0.10]$
Aligned	1,758	23.29	24.35	$+1.05$ $[0.53, 1.58]$
Low alignment	264	21.48	19.70	-1.78 $[-3.28, -0.51]$
Aligned, stronger enhancement	442	23.19	25.57	$+2.38$ $[1.25, 3.66]$

In Tab. 7, the R@1 gain from independent branch fusion over element-level support generally increases as support becomes less evenly distributed across fragments, from 1.42% to 12.20%. ReContext further improves the lower three bands but trails independent branch fusion by 0.77% and 2.02% in R@1 in the highest two. Context compatibility is thus more helpful when local support is uneven, but member-supported enhancement does not add a gain in every group. This statistic measures concentration of support, not whether the matched content is semantically ambiguous.

D.3 Context Maxima and Pairing Controls

We first check whether independently taking the strongest context compatibility and member support would combine responses from different contexts. Tab. 8 compares all qualifying pairs, annotated pairs, and pairs from fixed local rankings. More than half have no context attaining both maxima, so the two strongest responses cannot be attributed to one shared context. These overlapping groups describe the source of the two maxima, not retrieval errors. Contexts providing different maxima can still share visual elements; their separation does not imply distinct activity categories or a gain from coupling.

Table 8: ActivityNet-Event pairs with no common maximizing context, allowing ties within 10^{-6}.

Pairs	Count	Rate (%)
All	37,264,266	59.83
Annotated	8,502	56.60
Local top 20	163,551	58.26

Tab. 9 reports the corresponding retrieval intervention. Aligned denotes the 1,758 queries with correlated responses, a context near both maxima, and positive enhancement strength; Low alignment denotes the 264-query control. The stronger-enhancement group retains aligned queries with enhancement strength at least 0.25.

ReContext improves R@1 over Shuffle by 1.05% on the aligned group but reduces it by 1.78% on the low-alignment control; the overall difference is -0.04%, with an interval spanning zero. Preserving membership is beneficial here when strong context compatibility is supported by its own members; retaining the original pairing is not uniformly helpful across queries. All ten seeds favor ReContext on the 442-query group.

Across the five groups in Fig. 4, ΔR@5 and ΔR@10 also increase, although ReContext's R@1 itself is not monotonic. The trend therefore concerns the relative benefit of preserving the original pairing.

The R@1 gain differs by 2.99% between the strongest and weakest enhancement groups, with a 95% video-cluster interval of $[1.27, 4.93]\%$. Intervals for differences between adjacent groups include zero, and the historical validation partition is not monotonic. These results support an overall trend in this T2V subset, without establishing the same trend for V2T or the independent-pooling comparison.

These intervals use 2,000 video-cluster bootstrap resamples, preserving the dependence between queries from the same video across groups. They are conditional on the selected rules and do not correct for earlier rule exploration.

D.4 Controls for Description-Support Scope

We check whether the benefit of video-wide matching extends beyond descriptions selected for distributed support. Across all valid ActivityNet-Event queries, it raises T2V R@1 from 23.69% to 24.20%, a gain of 0.51% with a 95% video-cluster interval of $[0.33, 0.68]\%$. In Fig. 5, the highest-minus-lowest positive-deficit group difference is 1.82% in R@1 ($[0.22, 3.50]\%$) and 0.0295 in MRR ($[0.0188, 0.0409]$). Across the 17,092 queries with defined D, its Spearman correlation with the per-query reciprocal-rank change is 0.228, with interval $[0.213, 0.244]$. These statistics concern gains over Restricted; ReContext's absolute R@1 does not increase with D.

We check whether the deficit trend also holds when multiple fragments and contexts are present. This matters because 350 of 503 queries in the highest-deficit group lack one of these structures; high deficit alone does not identify support across activities. Repeating the analysis on the 8,502 queries with both structures using the same boundaries retains the increasing MRR-gain trend, with a highest-minus-lowest difference of 0.0333 ($[0.0148, 0.0539]$); the corresponding R@1 interval spans zero.

The separate support-group comparison more directly tests distributed fragment support. Video-wide support improves R@1 by 1.59% on 817 cross-context queries, compared with 0.18% on 3,910 compact-support queries. The difference in gains is 1.41%, with interval $[0.41, 2.46]\%$; the minimum number of contexts needed for coverage is two for 799 queries and three for the remaining 18. Matching queries within each historical partition by fragment count and bins of visual-element count, context count, and the weakest fragment's support score yields 726 pairs without replacement. The matched gain-difference interval is $[-0.04, 2.59]\%$, and trends are not consistent across all historical partitions. The larger gain for distributed fragment matches supports retaining access to the full video when scoring the description, while each context is scored using its own members. These T2V results do not establish a gain for every query or for the untested union-of-members variant.

D.5 Failure Analysis

We examine the limits of score refinement on all 13,151 ActivityNet-Event T2V R@1 failures among 17,350 queries.

We give the annotated video more favorable scoring conditions to test whether changes to member assignment, context selection, and final fusion could recover it. Representations, video-wide support, enhancement strength $\beta = 0.6$, and the competing videos' base context aggregation remain fixed. Let $C_\star$ be the annotated video's highest context compatibility and $m_\star$ its strongest visual-element–fragment response in the cached scores. Allowing any context to use this strongest local response gives the context upper bound $U_c = \max\{C_\star, 0.4C_\star + 0.6m_\star\}$. Every member average is bounded by $m_\star$, and any convex aggregation of enhanced contexts is bounded by U_c.

For annotated video g and competing video v, the resulting final-score upper and lower bounds are

$$U_g(\lambda) = \lambda U_c + (1-\lambda)s^e(g,T), \qquad D_v(\lambda) = \lambda C_0(v,T) + (1-\lambda)s^e(v,T), \tag{16}$$

where C_0 is the cached context aggregation without member-supported enhancement. Removing enhancement gives a lower bound because enhancement only increases context scores. Videos without contexts retain the local-score fallback. We first test $\lambda = 0.8$, then solve the linear inequalities $U_g(\lambda) \geq D_v(\lambda) - 10^{-6}$ for all competing videos in the 4,877-video gallery. The latter test allows one weight per query, shared across its candidates. An empty feasible interval in $[0, 1]$ certifies failure even under these favorable bounds. All original failures are included, irrespective of fragment or context count.

Table 10: ActivityNet-Event T2V failures that remain unrecoverable under optimistic member assignment and context selection. Both shares use all 13,151 original failures as the denominator.

Context weight	Failures	Share (%)
Fixed $\lambda = 0.8$	10,736	81.64
Per-query $\lambda \in [0, 1]$	9,889	75.20

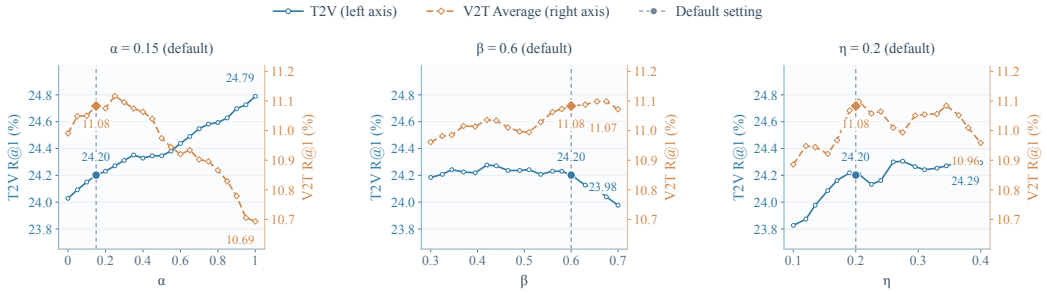

Figure 7: Scoring-weight sensitivity on ActivityNet-Event. Panels vary the fragment-mean weight α, enhancement coefficient β, and local-branch weight $\eta = 1 - \lambda$. Blue circles use the left axis for T2V R@1; orange diamonds use the right axis for V2T R@1-Average (both %). Filled markers and vertical dashed lines indicate defaults.

Tab. 10 shows that allowing per-query fusion reduces the unrecoverable share from 81.64% to 75.20%. Among failures whose annotated video has contexts, the latter share remains 74.86% (9,692/12,946), so missing contexts do not explain the result. Most failures therefore exceed what these association, selection, and fusion changes can repair with the current scores. The remaining 3,262 cases are not certified unrecoverable; the bounds do not demonstrate actual recovery. This analysis neither attributes failures specifically to the encoder nor covers regrouping with re-encoding or changes to the scoring formula. Relevance follows benchmark annotations, so unannotated candidates are not necessarily semantically irrelevant.

D.6 Efficiency Analysis

API-based structure parsing takes an average of 8.71 seconds per video and 7.23 seconds per description. Local feature encoding and retrieval computations use NVIDIA RTX 3090 hardware with 24 GB of GPU memory. Frozen feature encoding takes 1.27 seconds per video and 0.030 seconds per description. With structures and features cached, ReContext completes scoring and ranking for the full ActivityNet-Event evaluation set in 34.76 seconds. Gallery-side parsing and encoding can be performed offline and reused across queries.

D.7 Parameter and Robustness Checks

Fig. 7 examines the balance between fragment support, member-supported enhancement, and branch fusion. Each sweep varies one parameter while keeping all others at their default values. The default weights $\alpha = 0.15$, $\beta = 0.6$, and $\eta = 0.2$ yield 24.20% T2V R@1 and 11.08% V2T R@1-Average. Increasing α shifts video-wide support from the strongest fragment toward the average of all fragments' best matches. T2V generally improves across the sweep, reaching 24.79% at $\alpha = 1$, whereas V2T peaks near $\alpha = 0.25$ and then declines to 10.69%. Accounting more strongly for all fragments thus favors T2V in this sweep, while V2T benefits from retaining a larger contribution from the strongest local cue.

T2V R@1 remains near 24.2% over much of the β range, then falls to 23.98% at $\beta = 0.7$. V2T varies nonmonotonically and favors stronger enhancement than T2V. Thus, giving more weight to member support does not consistently improve both directions. Increasing η initially improves both metrics, but their responses diverge beyond the default region. At $\eta = 0.4$, T2V reaches 24.29% while V2T

falls to 10.96%. The local and context branches therefore require a balance that accounts for both retrieval directions; the default weights do not maximize every individual curve.

Fig. 8 compares the two aggregation counts. Member k counts the strongest member responses averaged for each fragment within a context; Context k counts the strongest enhanced contexts entering the context-branch mean. Both use all available entries when fewer than k exist, with a default of three.

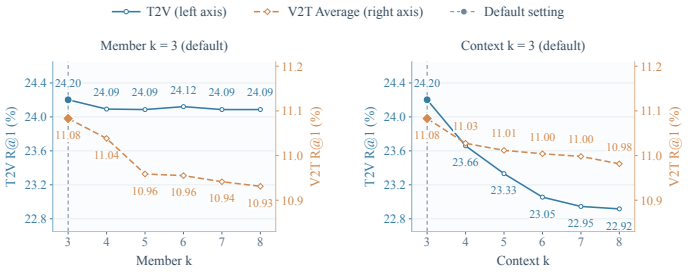

Figure 8: Aggregation-count sensitivity on ActivityNet-Event. Blue circles show T2V R@1 (left axis); orange diamonds show V2T R@1-Average (right axis), both in %. Filled markers and vertical dashed lines mark $k = 3$.

Increasing Member k from 3 to 8 leaves T2V within 24.09–24.20%, while V2T declines from 11.08% to 10.93%. Increasing Context k produces a larger T2V decrease, from 24.20% to 22.92%, together with a smaller V2T decrease to 10.98%. T2V is therefore less sensitive to averaging additional members within a context than to including lower-scoring contexts. This supports selective context aggregation while retaining video-wide fragment support in the separate local branch. Both scans favor $k = 3$ among the tested values; values below three are not evaluated.