Vote-in-Context: VLMs as Explainable Zero-Shot Rank Fusers

Mohamed Eltahir[1] Ali Habibullah[1] Lama Ayash[1,2] Tanveer Hussain[3*†] Naeemullah Khan[1†]

[1] King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
[2] Department of Computer Science, King Khalid University (KKU), Abha, Saudi Arabia
[3] Department of Computer Science, Edge Hill University, Ormskirk, England
{mohamed.hamid, ali.habibullah, lama.ayash}@kaust.edu.sa
hussaint@edgehill.ac.uk, naeemullah.khan@kaust.edu.sa

Abstract

*In retrieval domain, fusing candidates from heterogeneous retrievers (R), especially for multi-modal data like videos is challenging. Typical training-free fusion methods lack content-awareness, relying either on rank or score signals. We introduce **Vote-in-Context (ViC)**, a generalized, training-free framework that re-thinks list-wise reranking and fusion as a zero-shot reasoning task for a Vision-Language Model (VLM). ViC serializes content evidence and retriever metadata into the VLM's prompt, allowing it to adaptively weigh retriever consensus against visual–linguistic content. This generalized framework naturally operates as content-aware rank fuser ($R > 1$) and single-list reranker ($R = 1$). We demonstrate ViC's potentials in video retrieval, where we serialize video contents into the VLM via our efficient $S - Grid$ representation. Across video retrieval benchmarks, ViC sets new zero-shot SOTA, outperforming strong fusion baselines ($R > 1$) and boosting individual retrievers ($R = 1$), demonstrating its effectiveness in handling complex visual and temporal signals alongside text. ViC achieves massive gains of up to +40 Recall@1 over SOTA on all benchmarks, proving it as a simple, reproducible, and highly effective recipe for turning modern VLMs into powerful zero-shot rerankers and fusers capable of yielding grounded natural-language rationales for their top-K decisions. Code and resources are publicly available at https://github.com/mohammad2012191/ViC.*

1. Introduction

Modern applications increasingly rely on large repositories of unstructured text and complex multimodal data, such as video, which integrates visual, auditory, and temporal signals [17]. The scale and heterogeneity of these collections make efficient retrieval and analysis challenging [29].

Retrieval systems tackle this by aligning natural-language queries with semantically relevant content, but performance still degrades in the presence of high-dimensional, temporally structured data and sparse or ambiguous queries.

Considering this complexity, a two-stage retrieval paradigm is commonly adopted [14, 18]. A fast dual-encoder first retrieves a high-recall candidate set, which is then refined by a more powerful but expensive reranker [6,

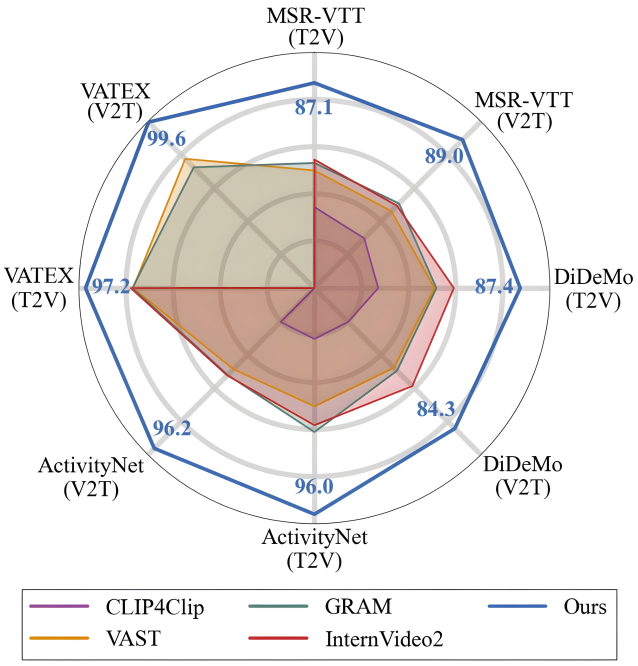

Figure 1. R@1 for T2V/V2T on MSR-VTT, DiDeMo, VATEX, and ActivityNet versus strong baselines.

* Corresponding author
† Principal Investigator (PI)

19]. Two-stage designs also enable using multiple diverse retrievers as a first stage. Fusing their outputs based on ranks or scores, using Reciprocal Rank Fusion (RRF) [9] or CombSUM/CombMNZ [10], respectively, typically yields significant performance gains [5].

However, applying two-stage template to complex multimodal data exposes limitations in the second stage. First-stage retrievers typically rely on global embeddings and may rank irrelevant candidates highly when they miss query-specific details. A reranker is therefore essential, yet conventional approaches are often costly, require in-domain fine-tuning, or are tied to a specific retriever's features [21]. Moreover, when ensembling multiple retrievers, standard fusion methods are *content-blind*, operating only on rank or score signals while ignoring the candidates' rich content. This motivates a universal, training-free framework that acts as both a content-aware reranker and fuser.

Recent advances in instruction-following Large Language Models (LLMs) offer a promising alternative. In text retrieval, LLMs already act as strong zero-shot list-wise rerankers [19]. This paradigm extends to Vision-Language Models (VLMs) which demonstrate strong zero-shot reasoning and cross-modal alignment capabilities. By adapting videos into a format interpretable by VLMs, these models can themselves serve as powerful zero-shot relevance estimators.

To this end, we introduce **Vote-in-Context (ViC)**, a general, training-free framework that turns a frozen VLM into a universal list-wise reranker and rank fuser. Instead of using a fixed content-blind fusion formula, **ViC** adaptively weighs all the available signals and treats fusion as a zero-shot reasoning problem over a serialized view of the candidate lists.

In this paper, we apply **ViC** to video retrieval. We introduce the **S-Grid**, a compact video serialization map that represents each clip as a single image grid of uniformly sampled frames, optionally paired with subtitles, providing the VLM-readable content evidence for each candidate.

The framework operates in two modes. First, as a powerful single-list reranker ($R = 1$), where **ViC** uses S-Grids to re-evaluate the top-K items from one retriever. Second, as a novel rank fuser ($R > 1$), where **ViC** constructs a candidate list by interleaving multiple retrievers. This assembly explicitly encodes rank and consensus metadata in the list order and item multiplicity, allowing the VLM to weigh these signals jointly with the S-Grid content evidence. The experiments show this combination yields massive gains, saturating several benchmarks in a zero-shot settings, as illustrated in 1. Beyond this high performance, the reasoning-based nature of the **ViC** framework provides a crucial advantage: inherent interpretability. The VLM can be prompted to generate natural-language rationales for its decisions, offering a fully transparent fusion process, as exemplified in 2.

The main contributions of this work are summarized as follows:

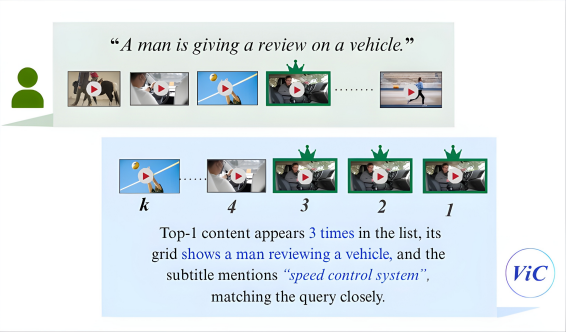

Figure 2. A qualitative example of **ViC** as an interpretable fuser.

- We propose **Vote-in-Context (ViC)**, a generalized, training-free framework that turns a frozen VLM into a powerful list-wise reranker and fuser by serializing both content and retriever metadata into its prompt.
- We introduce the **S-Grid**, a compact and effective video representation that serves as the content serialization map for **ViC**, enabling VLM-based reasoning over video without costly sequence processing.
- We comprehensively evaluate **ViC** in both its $R = 1$ (single-list) and $R > 1$ (multi-list) modes, analyze its scaling and context-size sensitivity, and release code and evaluation protocols for reproducibility.
- We show that **ViC** can be prompted to generate evidence-grounded rationales for its top-K decisions, making the fusion step more interpretable without any extra training.

This paper is organized as follows. Section 2 reviews related work. Section 3 details the proposed **ViC** framework and its application for video retrieval. Section 4 presents the experimental results and ablation studies, followed by a discussion of the framework's limitations and future directions.

2. Related Work

Modern video retrieval has progressed from early temporal-attention architectures [27, 30] to large-scale unified pretraining [4, 25]. CLIP-style adaptations such as CLIP4Clip [15] and X-CLIP [16] transfer powerful image-text encoders to video, while recent foundation-scale systems like VAST [7] and InternVideo2 [24] further improve recall by incorporating audio, subtitles, and larger video-specific backbones. These models typically rely on coarse global representations: efficient for high-recall candidate generation, but often insufficient to capture fine-grained, query-specific details in the top ranks.

When multiple first-stage lists are available, they are typically fused using classical score- or rank-level methods such as CombSUM/CombMNZ [10] and Reciprocal Rank Fusion (RRF) [9]. These approaches are simple and robust

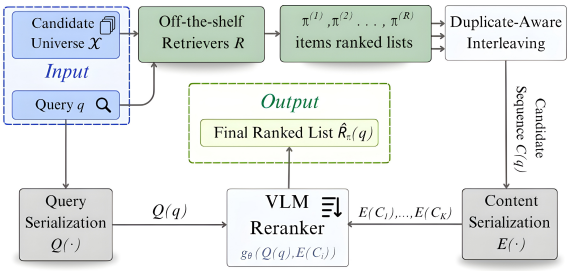

Figure 3. The **Vote-in-Context (ViC)** framework. A `VLM Reranker` takes the serialized query $Q(q)$ and candidate content $E(C_1), \ldots, E(C_K)$, and jointly weighs them against retriever metadata (rank, multiplicity) encoded in the `Candidate Sequence` $C(q)$ via the `Duplicate-Aware Interleaving` step to produce the final ranked list $\widehat{R}(q)$.

but rely on fixed weighting formulas and hyperparameters, and operate solely on rank or score signals, leaving other modalities unexploited.

The emergence of LLMs and VLMs has introduced a new paradigm for re-ranking in retrieval systems. In text retrieval, LLMs have demonstrated strong zero-shot list-wise reranking performance [1, 19, 28], while recent work shows that serializing video as a grid of sampled frames allows image-centric VLMs to reason over temporal content [12]. At the same time, modern instruction-following VLMs, such as InternVL [22] and Qwen-VL [3], provide the robust zero-shot, multimodal alignment required to make such designs highly-performing.

Beyond performance, explainable vision–language modeling has explored natural-language and multimodal rationales: e-ViL [11] benchmarks joint prediction and explanation across Video Question-Answering (VQA), retrieval, and captioning, and NLX-GPT [20] generates free-form textual explanations conditioned on visual features. These trends point toward VLM-based rerankers that can both improve retrieval and provide human-readable justifications.

3. Methodology

This paper introduces **Vote-in-Context (ViC)**, a general, training-free, and multimodal framework that utilizes the VLM reasoning capabilities to solve the ranked-list fusion problem. **ViC** provides a uniform candidate prompt to the VLM containing both: (a) *content evidence* (such as images/text), and (b) *retriever metadata*, including ranks and cross-list multiplicity encoded directly in the prompt. This approach stands in contrast to classical, non-content-aware fusion methods (such as RRF or CombSUM), which operate only on rank/score signals and ignore candidate content.

The VLM receives this metadata alongside the candidates' content and implicitly weighs retriever metadata versus content evidence on a per-query basis in a zero-shot setting. A candidate's rank is conveyed by each list's order, while cross-list consensus is represented by allowing duplicates to appear in the candidate set. Compared to the traditional fusion methods, **ViC** is hyperparameter-free and modality-aware, yielding per-query decisions that adaptively weight all available signals. The idea is modality-agnostic, requiring only that candidates can be serialized into a VLM-readable prompt (such as passages for text search, images with metadata, tables, or audio transcripts). See Fig. 3 for a high-level overview.

Beyond retrieval quality, this joint encoding of content and metadata also makes the fusion process naturally *explainable*, as the same prompt can be used to elicit natural-language rationales for the final ranking.

To demonstrate **ViC** generality, we apply it to video retrieval. We introduce the **S-Grid**, an efficient uniform visual–linguistic representation that serializes each video into a single grid of frames, optionally paired with subtitles, and show how **ViC** uses this representation to operate both as a powerful single-list reranker and as a rank fuser over multiple heterogeneous retrievers.

3.1. Problem Setup and Notation

The **ViC** fusion framework is formalized as follows. Let $\mathcal{X}$ denote the universe of candidate items (such as videos or text passages). For a given query q, assume access to R retrievers, $\mathcal{R} = \{1, \ldots, R\}$. Each retriever $r \in \mathcal{R}$ returns a ranked list of items $L_r(q)$ drawn from $\mathcal{X}$:

$$L_r(q) = \big(x_{r,1}, x_{r,2}, \ldots, x_{r,n_r}\big), \quad \text{where } x_{r,j} \in \mathcal{X}. \quad (1)$$

Candidate Assembly and Metadata Encoding. The process begins by constructing a single *candidate sequence* $C(q)$ of length K. This sequence retains both the rank and multiplicity metadata from the initial retrieval lists. Define a per-list depth as $k_{\max} = \lceil K/R \rceil$, and truncate each list accordingly before assembling the final candidate sequence.

$$\mathrm{Top}_{k_{\max}}(L_r) = \big(x_{r,1}, \ldots, x_{r,\min(k_{\max}, n_r)}\big). \quad (2)$$

The candidate sequence $C(q)$ is formed by a round-robin (RR) interleaving of these truncated lists, preserving duplicates:

$$C(q) = \mathrm{RR}_K\big(\mathrm{Top}_{k_{\max}}(L_1), \ldots, \mathrm{Top}_{k_{\max}}(L_R)\big). \quad (3)$$

The $\mathrm{RR}_K(\cdot)$ operator appends items in the order $\big(x_{1,1}, x_{2,1}, \ldots, x_{R,1}, x_{1,2}, \ldots\big)$, skipping any exhausted lists, and truncates the final sequence to length K. This sequence $C(q) = (C_1, \ldots, C_K)$ inherently encodes retriever metadata: per-list rank is signaled by position, and cross-list consensus is signaled by an item's multiplicity, $\mu_C(x) = \sum_{i=1}^{K} \mathbf{1}\{C_i = x\}$.

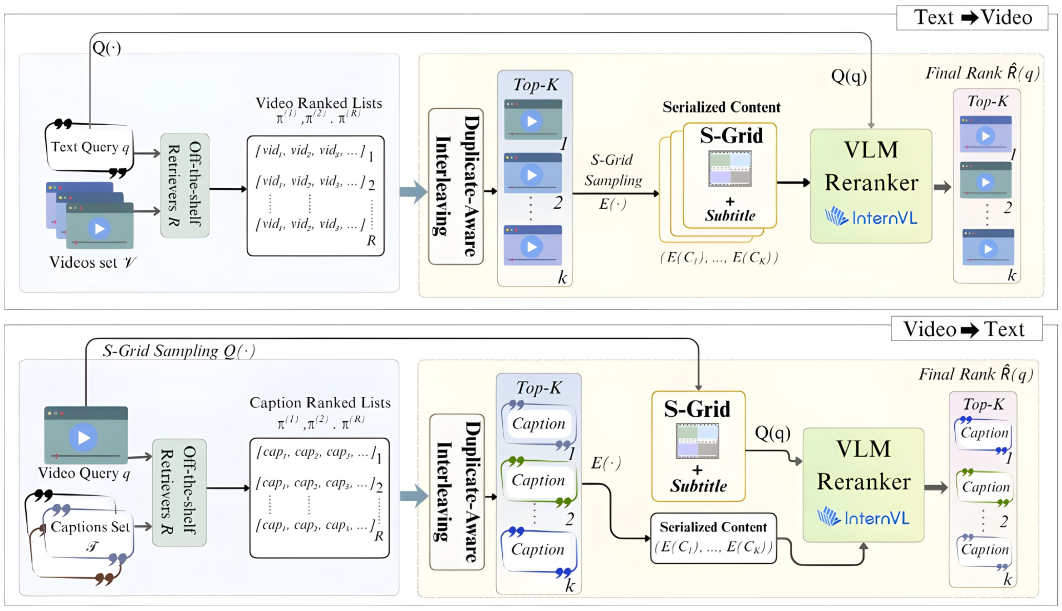

Figure 4. The **Vote-in-Context (ViC)** framework applied for Text-to-Video (T2V, top) and Video-to-Text (V2T, bottom). The left block shows the initial retrieval stage. The right block (green) shows our **ViC** framework. The serialization maps ($Q(\cdot)$, $E(\cdot)$) are modality-dependent: `S-Grid Sampling` is applied to video inputs, while text inputs use the identity.

This serialization process also provides practical control mechanisms. A round-robin assembly based on $k_{\max}$ ensures balanced coverage across all retrievers, while the candidate sequence $C(q)$ can be optionally reordered to bias the VLM's early context, by prioritizing items from stronger backbones, for instance. Such flexibility is inherently absent from fixed-formula fusion methods.

VLM Reranking. The sequence is passed to a frozen, list-wise VLM g_Θ, where Θ denotes the frozen model parameters, for reranking. Let $E(\cdot)$ be the *content serialization map* that converts an item $x \in \mathcal{X}$ into its VLM-readable format (i.e., the content evidence), and let $Q(q)$ be the serialized query. The VLM computes a permutation $\pi \in \mathfrak{S}_K$, where $\mathfrak{S}_K$ is the set of all permutations of the indices $\{1, \ldots, K\}$:

$$\pi = g_\Theta\Big(Q(q),\ \big(E(C_1), E(C_2), \ldots, E(C_K)\big)\Big). \tag{4}$$

The final fused and reranked output $\widehat{R}(q)$ is the sequence C reordered by this permutation:

$$\widehat{R}(q) = \big(C_{\pi(1)},\ C_{\pi(2)},\ \ldots,\ C_{\pi(K)}\big). \tag{5}$$

Special Case: Single-List Reranking ($R = 1$). When $R = 1$, the candidate sequence reduces to the usual top-K list from the single retriever,

$$C(q) = \mathrm{Top}_K(L_1(q)) = (x_{1,1}, \ldots, x_{1,K}). \tag{6}$$

In this setting, **ViC** reduces to a pure list-wise reranker where the VLM's decision depends only on the content evidence $E(\cdot)$, since cross-list metadata (multiplicity, rank-of-ranks) is absent.

3.2. Applying ViC to Video Retrieval.

Applying **ViC** to video retrieval requires a method to serialize video candidates into a VLM-readable format. This section first defines our video representation, the S-Grid, and then maps it to the **ViC** framework.

S-Grid: A Uniform Video Prompt. A video v is represented as a regular grid of uniformly sampled frames composited into a single $H \times W$ image, optionally paired with a subtitle or Automated Speech Recognition (ASR) string a_v (if available). Let s denote the grid dimension (i.e., the grid has $s \times s$ cells). Given video length F frames, let $f_i \in F$ be the i-th selected frame. s^2 frame indices $\{f_i\}_{i=1}^{s^2}$ are selected uniformly via $f_i = \lfloor (i-1)\frac{F}{s^2-1} \rfloor$. These frames are extracted, resized to $\lfloor H/s \rfloor \times \lfloor W/s \rfloor$, and tiled in row-major order to form an $H \times W$ canvas, denoted as $\mathrm{Grid}(v; s)$. When a subtitle a_v is available, it is concatenated to the textual prompt as an auxiliary input. This representation is denoted as:

$$\text{S-Grid}(v) \ = \ \big(\mathrm{Grid}(v; s),\ a_v\big),$$

This design provides the VLM with both visual snapshots and audio transcripts within a single prompt. Such a uni-

form interface enables a single VLM to process candidates retrieved from *any* upstream model.

Formalizing the Video Retrieval Tasks. The **ViC** framework is applied to cross-modal video retrieval, where the candidate universe consists of videos $\mathcal{V}$ and text captions $\mathcal{T}$. In the Text-to-Video (T2V) retrieval task, the query $q \in \mathcal{T}$ is text ($Q(q) = q$), and the candidates $C_i \in \mathcal{V}$ are videos, which are serialized as $E(C_i) = \big(\text{S-Grid}(C_i), a_{C_i}\big)$, where a_{C_i} denotes optional subtitle/ASR text aligned with C_i. In the Video-to-Text (V2T) retrieval task, the query $q \in \mathcal{V}$ is a video serialized as $Q(q) = \big(\text{S-Grid}(q), a_q\big)$ (with a_q the optional subtitle/ASR aligned with q), and the candidates $C_i \in \mathcal{T}$ are text captions, so the content map is the identity ($E(C_i) = C_i$). This bidirectional retrieval process is illustrated in Fig. 4.

Computational Efficiency. The reranker processes one image per video candidate and a short text block per item. The complexity per query is $\mathcal{O}(K \cdot C_{\text{VLM}})$ where K is the number of candidates, and C_{VLM} is one forward pass cost. This cost is independent of the raw video length, as each video is represented by a single image, keeping the per-candidate cost effectively constant. The approach is significantly lighter than frame-level cross-attention and permits larger candidate sets to be evaluated within the VLM's context window.

3.3. ViC as a Natural-Language Explainer

A key advantage of the **ViC** framework is its inherent interpretability. Because g_Θ is an instruction-following VLM, it can be prompted to produce both the ranking and a justification in a single forward pass. Concretely, the VLM is given one instruction asking it to both *rank the candidates* and *provide a natural-language rationale* for its top choices. The model executes the listwise reranking and appends its reasoning to the output. This process elicits faithful, evidence-grounded rationales without any additional training, architectural changes, or extra inference cost. As the VLM jointly conditions on the query $Q(q)$, the content evidence $E(C_i)$, and the retriever metadata (rank and multiplicity), its explanation can reference all three signals to justify its final decision.

4. Experiments

4.1. Benchmarks and Protocol

Evaluation is conducted on the MSR-VTT [26], DiDeMo [2], ActivityNet Captions [13], and VATEX [23] benchmarks, following the standard retrieval protocols established in prior work. Notably, only MSR-VTT and VATEX provide subtitles, which are incorporated into the S-Grid representation where applicable. On MSR-VTT, the standard 1k-A split is used. For DiDeMo, evaluation is performed at the video level by pooling the moment annotations into a single retrieval target per video. ActivityNet is evaluated using the official validation split for retrieval. For VATEX, the community 1.5k test subset is adopted. Out of the intended 1,500 videos from prior work, only 1,252 were successfully recovered due to the online unavailability of some videos. To ensure fair comparison, captions were re-indexed to this fixed subset, and all baselines and the proposed method were reproduced on the same 1,252 test videos. All evaluation items correspond to test-only instances, and the final video list is publicly released to facilitate reproducibility.

4.2. Implementation Details

The first-stage retrievers are CLIP4Clip[15], VAST[7], GRAM [8], and InternVideo2-6B [24]. CLIP4Clip is a canonical CLIP-style video retriever. VAST provides omni-modality pretraining. GRAM is a strong global-regional baseline. InternVideo2-6B serves as the strongest recent baseline. Each model is reproduced or re-evaluated using official checkpoints and released evaluation configurations, and all retrievers are kept frozen during experimentation. Tokenization, frame sampling, and text preprocessing strictly follow the original repository implementations to ensure consistency and reproducibility.

InternVL 3.5 38B [22] is employed as the main training-free VLM reranker. It takes S-Grid inputs along with the video/text query and is used in a zero-shot setting without any dataset-specific fine-tuning. Unless otherwise noted, the same candidate counts are used for each comparison. The standard ensemble configuration fuses all backbones except VAST, as this combination yielded the highest performance on average. A notable exception occurs in VATEX, where the ensemble includes only InternVideo2 and VAST, as these were the models successfully reproduced for this benchmark. Backbone-combination ablations are reported in §5.4.4.

4.3. Metrics and Hyperparameters

Results are reported using Recall@1 (R@1), the proportion of queries for which the top-ranked result is correct, for both T2V and V2T directions. For T2V, the **ViC** framework receives $K = 14$ candidate S-Grids per query, while for V2T, it receives $K = 20$ candidate captions, unless stated otherwise, as these settings performed best in our context-window ablation (see §5.4.3). The default S-Grid size is 3×3 frames. For the Soft Voting baseline, similarity scores are min-max normalized per query (row) before aggregation with uniform weights. For **ViC** rank fuser ($R > 1$), candidate lists are assembled by interleaving each retriever's list up to depth $k_{\max}$, preserving duplicates. The VLM output is parsed into a permutation, with the identity mapping used

Backbone	MSRVTT		DiDeMo		ActivityNet		VATEX	
	T2V	V2T	T2V	V2T	T2V	V2T	T2V	V2T
BASELINES (NO RERANK)								
C4C	34.4	29.9	27.1	20.3	21.6	20.3	—	—
VAST	49.9	46.2	51.0	47.8	50.2	48.7	77.0	77.6
GRAM	53.1	50.8	51.8	49.6	61.1	52.1	77.3	72.5
IV2-6B	54.5	49.5	59.2	58.8	58.2	52.4	80.7	—
VIC SINGLE- LIST RERANKER (R=1)								
C4C	62.8	61.3	60.4	53.8	64.6	62.8	—	—
C4C*	64.2	62.5	—	—	—	—	—	—
VAST	67.3	62.2	70.2	63.4	79.7	75.2	91.9	99.4
VAST*	68.7	63.1	—	—	—	—	92.4	**99.6**
GRAM	75.4	72.3	70.9	63.9	82.4	77.2	—	—
GRAM*	**76.2**	73.6	—	—	—	—	—	—
IV2-6B	74.0	74.1	**78.1**	**70.7**	**89.8**	**84.9**	95.5	—
IV2-6B*	75.9	**76.6**	—	—	—	—	**95.8**	—

Table 1. **Zero-shot R@1 for single backbones.** Abbreviations: C4C for CLIP4Clip, IV2-6B for InternVideo2-6B. Rows marked * use S-Grid; unmarked rows use Grid. Bold indicates the best per benchmark.

Method	MSRVTT		DiDeMo		ActivityNet		VATEX	
	T2V	V2T	T2V	V2T	T2V	V2T	T2V	V2T
BASELINES AND TRADITIONAL FUSION								
IV2-6B†	54.5	49.5	59.2	58.8	58.2	52.4	80.7	—
RRF	78.3	80.2	72.8	73.2	**96.8**	**97.4**	94.7	—
CombSUM	84.4	83.0	80.4	83.1	95.8	95.2	96.1	—
CombMNZ	85.3	86.9	78.0	80.8	95.0	92.2	96.4	—
VIC RANK FUSER (R>1)								
ViC‡	84.2	80.7	85.5	76.1	94.8	91.9	96.1	—
ViC	**87.1**	**88.1**	**87.4**	**84.3**	96.0	96.2	**97.5**	—

Table 2. **Zero-shot R@1 for rank fusion.** IV2–6B† denotes the previous strong backbone baseline. ViC‡ indicates ViC without duplicates in the candidate sequence. Bold indicates the best per benchmark.

as a fallback in very rare cases. The resulting ranked list $\widehat{R}(q)$ may include duplicate candidates; however, only the highest-ranked instance of each is considered during evaluation, consistent with standard practice.

5. Results

5.1. ViC as a Single-List Reranker ($R = 1$)

ViC is first evaluated in its simplest form as a single-list reranker ($R = 1$). In this setting, the VLM reranks the top-K candidates from a single retriever, using only content evidence (S-Grids and subtitles) without any cross-list fusion metadata, as presented in Table 1.

Applying **ViC** reranking to a single backbone yields substantial and consistent R@1 improvements across all datasets and models. For example, on MSR-VTT (T2V), **ViC** lifts the weakest backbone (CLIP4Clip) by 29.8 points (increases from 34.4 to 64.2) and the strongest (InternVideo2) by 21.4 points (increases from 54.5 to 75.9). On ActivityNet (T2V), the gains are even larger, adding 31.6 R@1 to InternVideo2 (increases from 58.2 to 89.8). On VATEX (V2T), **ViC** boosts VAST by 22.0 points (increases from 77.6 to 99.6), achieving near-saturation in R@1 performance.

These results demonstrate that, even without fusion, the VLM performs highly effective list-wise reasoning over S-Grid content evidence. This provides a training-free mechanism to correct the coarse similarity biases of dual-encoder retrievers. Moreover, a comparison between "Grid" (visuals only) and "S-Grid" (visuals and subtitles) configurations shows that incorporating textual evidence consistently enhances performance, confirming that the VLM effectively utilizes all available modalities during reranking.

5.2. ViC as a Rank Fuser ($R > 1$)

The full **ViC** framework is evaluated as a rank fuser ($R > 1$), utilizing both content evidence and retriever metadata (rank, multiplicity). A detailed comparison between **ViC** fusion and traditional fusion baselines (RRF, CombSUM, and CombMNZ) is summarized in Table 2

ViC consistently outperforms all traditional fusion methods across nearly all benchmarks. On MSR-VTT (T2V), **ViC** achieves 87.1 R@1, surpassing the best baseline (CombMNZ) by +1.8 points. On DiDeMo (T2V), the gain is most significant, where **ViC**'s 87.4 R@1 is +7.0 points higher than the next-best baseline (CombSUM). On VATEX (T2V), **ViC** reaches 97.5 R@1, once again setting the highest overall performance.

While RRF remains a strong competitor on ActivityNet, **ViC** demonstrates substantially greater stability across the other datasets, where RRF and other score-level fusion methods exhibit notable performance fluctuations.

Furthermore, Table 2 includes a **ViC** (No Duplicates) ablation. This variant deduplicates the candidate sequence $C(q)$ before passing it to the VLM, thus removing the multiplicity metadata. The resulting performance drop (such as 87.1 to 84.2 on MSR-VTT T2V) confirms that the VLM actively uses cross-list consensus as a strong relevance signal.

Finally, comparing the **ViC** fusion result (87.1 on MSR-VTT, Table 2) with the best single-backbone reranking result (75.9 on MSR-VTT, Table 1) highlights the additive advantage of fusion. Reranking a single model (**ViC**, $R = 1$) yields a +21.4 point improvement, while incorporating fusion (**ViC**, $R > 1$) contributes an additional +11.2 points, underscoring the complementary strengths of the two components within the **ViC** framework. This consistent, state-of-the-art performance across all benchmarks is visualized

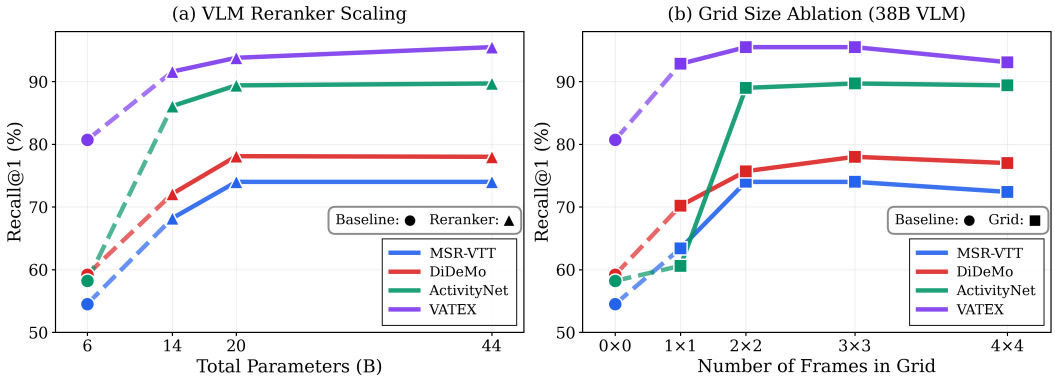

Figure 5. (a) Effect of reranker scale (InternVL 3.5, 3×3 grid) on T2V Recall@1. (b) Impact of grid size on T2V performance, using InternVideo2-6B and InternVL 3.5-38B.

in Figure 1.

5.3. Qualitative Analysis: Interpreting ViC's Decisions

The reasoning-based nature of the **ViC** framework allows it to provide faithful, natural-language explanations for its decisions, as shown in Figure 2. For the query, "A man is giving a review on a vehicle," the VLM generates a rationale that explicitly references the three key evidence types. In this example, the VLM notes that the top candidate appears multiple times across lists, then ties this consensus to the visual grid (a man reviewing a vehicle) and the subtitle ("speed control system"), confirming that the VLM is not just counting votes like RRF or just matching content like a simple reranker but is adaptively weighing all three information sources to make its final, interpretable decision.

5.4. Ablation Studies

5.4.1. Grid size

As **ViC** relies on content-derived evidence, the S-Grid constitutes the key visual representation driving its performance. Figure 5 (b) studies 1×1 to 4×4 grids. 2×2 and 3×3 are the sweet spots. 1×1 undercovers the video. 4×4 begins to compress each frame too aggressively and can introduce redundant visual tokens. This trend holds across the benchmarks that have been tested. Small grids are well matched to the evaluated datasets: MSR-VTT uses 10-30 s clips, DiDeMo videos are about 25-30 s, and VATEX clips are around 10 s. ActivityNet Captions contains longer, untrimmed videos with average durations on the order of minutes, though, the reranker performs strongly.

5.4.2. Reranker scale

Scaling the VLM in **ViC** from 8B to 38B at a fixed 3×3 grid steadily improves R@1, then plateaus as model size grows, as shown in Figure 5a. Even the 8B model delivers strong zero shot reranking, while smaller models fail to produce stable permutations, which makes 8B the minimum effective scale for list wise reranking in the proposed **ViC** pipeline. The strength of the 8B model without any training also suggests that light fine tuning could turn much smaller VLMs into efficient rerankers.

5.4.3. VLM type and context size

Varying the number of candidates given to the VLM in **ViC** has a clear effect on retrieval, as shown in Table 3. Preliminary analysis confirmed that R@30 is already near 100% on all benchmarks, so the correct item is almost always present in the top 30 and larger context mainly affects how well the VLM reranks. However, most VLMs do not exploit the extra coverage at $K = 30$ and their discrimination gets worse when the list is too long. For T2V, increasing K from 10 to 14 improves R@1, but pushing K to 30 lowers R@1 and gives almost no gain in R@10. Qwen3 VL is strong on T2V yet drops sharply on V2T when K grows, while Gemma 3 stays stable at $K = 30$ and reaches the best R@10 in both directions. We still use InternVL 3.5 in the main experiments because it gives strong overall results and comes in multiple scales that enable our scaling analysis. For V2T, $K = 20$ captions is the best operating point, since performance flattens or declines beyond that. Overall, these trends show that current VLMs have a practical limit on how much list context they can use for relevance judgment, even when the ground truth is already present in the candidate pool.

5.4.4. Backbone combinations

Table 4a shows that the strongest three-model mix uses CLIP4Clip, GRAM, and InternVideo2. The choice follows directly from the overlap patterns in Table 4b. GRAM and VAST share very similar top k neighbors, so combining them adds little. CLIP4Clip and InternVideo2 also overlap more than we would like, so we begin with the most diverse high-performers. Among two-backbone ensembles, GRAM with InternVideo2 gives the best averages, and it is also a strong single backbone.

Reranker	T2V — grids per query			V2T — captions per query		
	10	14	30	10	20	30
	R@1/R@10	R@1/R@10	R@1/R@10	R@1/R@10	R@1/R@10	R@1/R@10
InternVL 3.5 38B	73.8 / 82.7	74.0/ 83.8	71.3 / 84.5	**75.8 / 85.3**	**74.1** / 89.0	**70.0** / 90.8
Qwen3-VL 30B (A3B)	**76.5 / 82.7**	**77.0** / 84.7	**77.0** / 86.5	59.5 / 85.3	55.2 / 84.2	51.8 / 85.4
Gemma-3 27B IT	76.2 / 82.7	76.7 / **84.8**	73.3 / **88.1**	**75.8 / 85.3**	71.2 / **90.5**	69.3 / **91.5**

Table 3. Reranker type and context size in one view. Left: T2V vs. grids per query. Right: V2T vs. captions per query.

Backbones	MSRVTT		DiDeMo		ActivityNet		Avg	
	T2V	V2T	T2V	V2T	T2V	V2T	T2V	V2T
(i) Two-backbone ensembles								
V + IV2	77.5	75.9	77.8	78.5	86.6	87.7	80.6	80.7
G + IV2	83.6	**84.7**	78.8	80.0	90.9	90.0	84.4	84.9
(ii) Three-backbone ensembles								
C4C + G + IV2	**84.4**	82.9	80.4	**83.1**	**95.8**	**95.2**	**86.9**	**87.1**
V + G + IV2	73.0	73.4	74.7	74.7	84.8	82.5	77.5	76.9
C4C + V + IV2	81.0	79.4	79.2	81.7	94.8	94.9	85.0	85.3
C4C + V + G	65.3	64.3	64.9	63.1	75.3	71.6	68.5	66.3
(iii) All four backbones								
All	81.5	80.8	**81.4**	80.8	92.9	90.2	85.3	83.9

(a) Zero-shot R@1 for CombSUM ensembles without VLM reranking.

Backbones	MSRVTT	DiDeMo	ActivityNet	Avg	Overlap
G & V	47.3	47.7	47.9	47.6	High
C4C + IV2	36.4	34.2	4.6	25.1	Medium
G & IV2	6.4	6.3	5.9	6.2	**Low**
IV2 & V	5.7	6.2	5.6	5.9	**Low**
C4C & G	5.4	**5.1**	4.3	4.9	**Low**
C4C & V	**5.0**	**5.1**	**4.1**	**4.7**	**Low**

(b) Top-14 intersection percentages (%) between backbone pairs on T2V retrieval.

Table 4. **Backbone ensembles and diversity.** V for VAST, G for GRAM, C4C for CLIP4Clip, and IV2 for InternVideo2-6B. Bold values mark the best score per dataset and retrieval direction; and lower overlap in (b) indicates higher diversity.

We then add a third model with complementary signals. Adding CLIP4Clip to GRAM and InternVideo2 improves results, with clear gains on ActivityNet. Replacing CLIP4Clip with VAST hurts, and removing InternVideo2 degrades performance. Using all four backbones also underperforms the best triple, which points to redundancy rather than synergy.

The pattern aligns with model roles. CLIP4Clip brings broad frame-level CLIP semantics. GRAM contributes fine regional cues through its global–regional objective. InternVideo2 adds large-scale video pretraining and stronger temporal reasoning. VAST has high overlap with GRAM yet weaker single-backbone recall, which limits its marginal gain in larger ensembles.

5.5. Discussion and Conclusion

The results support **ViC** as a new fusion paradigm for retrieval. Instead of fixed formulas such as RRF or a trained fuser, **ViC** treats fusion as a zero shot list wise reasoning task for a vision language model. The VLM weighs retriever metadata such as rank and multiplicity against content evidence for each query, which yields more context aware and robust fusion.

In video retrieval, S-Grids make it feasible for a VLM to process and rerank whole candidate lists with cost proportional to the number of items rather than video length, while still preserving temporal coverage. This yields large R@1 gains and SOTA zero-shot performance.

ViC introduces several trade offs at the framework level. First, it replaces the near zero arithmetic cost of RRF or CombSUM with a full VLM forward pass over the candidate list, so higher accuracy with a single backbone comes with extra latency. Second, the method is bounded by the VLM context window, and our ablations show that performance can drop as K grows. Third, the framework inherits the reliability issues of the underlying VLM: it depends on instruction following, can exhibit positional bias, and smaller models below 8B sometimes fail to parse the list format.

In video retrieval the method is also recall bound, since **ViC** cannot recover a relevant video that never appears in the top K from the first stage retriever. S-Grid serialization is computationally efficient but lossy, because uniform sampling on long untrimmed videos can miss short events that are crucial for matching a query.

These limitations suggest two main directions for future work. At the framework level, prompt engineering and lightweight VLM fine-tuning may let smaller and cheaper models act as reliable list wise fusers. At the application level, query aware or adaptive keyframe selection could build more informative S-Grids under a fixed token budget.

Acknowledgment

We are grateful to the KAUST Academy for its generous support, and especially to Prof. Sultan Albarakati who made this work possible. For computer time, this research used Ibex managed by the Supercomputing Core Laboratory at King Abdullah University of Science & Technology (KAUST) in Thuwal, Saudi Arabia.

References

[1] Mofetoluwa Adeyemi, Akintunde Oladipo, Ronak Pradeep, and Jimmy Lin. Zero-shot cross-lingual reranking with large language models for low-resource languages. *arXiv preprint arXiv:2312.16159*, 2023. 3

[2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017. 5

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 3

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1728–1738, 2021. 2

[5] Michał Bałchanowski and Urszula Boryczka. A comparative study of rank aggregation methods in recommendation systems. *Entropy*, 25(1), 2023. 2

[6] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*, 2017. 1

[7] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems*, 36:72842–72866, 2023. 2, 5

[8] Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment. *arXiv preprint arXiv:2412.11959*, 2024. 5

[9] Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 758–759, New York, NY, USA, 2009. Association for Computing Machinery. 2

[10] Edward A Fox and Joseph A Shaw. Combination of multiple searches. *NIST special publication SP*, 243, 1994. 2

[11] Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1244–1254, 2021. 3

[12] Wonkyun Kim, Changin Choi, Wonseok Lee, and Wonjong Rhee. An image grid can be worth a video: Zero-shot video question answering using a vlm. *IEEE Access*, 2024. 3

[13] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017. 5

[14] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020. 1

[15] 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. *arXiv preprint arXiv:2104.08860*, 2021. 2, 5

[16] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM international conference on multimedia*, pages 638–647, 2022. 2

[17] Thong Nguyen, Yi Bin, Junbin Xiao, Leigang Qu, Yicong Li, Jay Zhangjie Wu, Cong-Duy Nguyen, See-Kiong Ng, and Anh Tuan Luu. Video-language understanding: A survey from model architecture, model training, and data perspectives. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3636–3657, Bangkok, Thailand, 2024. Association for Computational Linguistics. 1

[18] Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*, 2019. 1

[19] Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, et al. Large language models are effective text rankers with pairwise ranking prompting. *arXiv preprint arXiv:2306.17563*, 2023. 2, 3

[20] Fawaz Sammani, Tanmoy Mukherjee, and Nikos Deligiannis. Nlx-gpt: A model for natural language explanations in vision and vision-language tasks. In *proceedings of the IEEE/CVF conference on computer vi-*

sion and pattern recognition, pages 8322–8332, 2022. 3

[21] Kaibin Tian, Yanhua Cheng, Yi Liu, Xinglin Hou, Quan Chen, and Han Li. Towards efficient and effective text-to-video retrieval with coarse-to-fine visual representation learning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5207–5214, 2024. 2

[22] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 3, 5

[23] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Lin. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2019. 5

[24] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024. 2, 5

[25] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 2

[26] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[27] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European conference on computer vision (ECCV)*, pages 471–487, 2018. 2

[28] Crystina Zhang, Sebastian Hofstätter, Patrick Lewis, Raphael Tang, and Jimmy Lin. Rank-without-gpt: Building gpt-independent listwise rerankers on open-source large language models. In *European Conference on Information Retrieval*, pages 233–247. Springer, 2025. 3

[29] Chunhui Zhu, Qi Jia, Wei Chen, et al. Deep learning for video-text retrieval: a review. *International Journal of Multimedia Information Retrieval*, 12(3):1–26, 2023. 1

[30] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755, 2020. 2