Vote-in-Context: VLMs as Explainable Zero-Shot Rank Fusers (Supplementary Material)

1. Prompt Design

The prompting strategy within the Vote-in-Context (ViC) framework enables a vision-language model to perform list-wise reranking and generate natural-language reasoning grounded in the underlying evidence. The prompt combines three information sources: the serialized query, the content representations of the candidates (images or text), and retriever metadata such as rank ordering and cross-list multiplicity. By conditioning jointly on these signals, the model determines a final ordering of candidates and produces concise explanatory statements that reflect both content relevance and metadata alignment. The complete prompting formulation for text-to-video (T2V) and video-to-text (V2T) retrieval, with and without subtitles, is summarized in Table 1.

2. Evaluation Metrics and Protocols

Recall@K. For a query set Q, each query $q \in Q$ has a set of relevant items G_q (e.g., all ground-truth captions for a video). Let $R_q^{(K)}$ be the top-K retrieved items for q. We report Recall@K as the fraction of queries for which at least one relevant item appears in the top-K:

$$\mathrm{R@K} \;=\; \frac{1}{|Q|}\sum_{q\in Q} \mathbf{1}\big\{\, G_q \cap R_q^{(K)} \neq \varnothing \,\big\}.$$

V2T with multiple references. In video-to-text (V2T) retrieval, a single video typically has many reference captions. We treat *any* of a query video's reference captions as relevant. A V2T hit at rank $k \le K$ is counted if at least one ground-truth caption for that video appears in $R_q^{(K)}$. This follows common practice on multi-caption datasets, such as VATEX, which provides 10 English and 10 Chinese captions per video. Unless specified otherwise, we evaluate on the English caption split for multilingual sets, following prior work.

Evaluation Setup. The VLM runs in deterministic mode `do_sample=False`, with `bfloat16` precision, `image_size=448`, `max_new_tokens=256`, and `FlashAttention` enabled. For Figure 1, T2V evaluation is done on MSR-VTT, DiDeMo, and ActivityNet. Latency is measured on a single NVIDIA A100 80GB GPU and averaged over 50 queries with a 1k video pool. Marker size reflects the VLM parameter count.

3. Full Retrieval Results

The main paper reports R@1. Here we include R@5 and R@10. Across single-list reranking and fusion, ViC's biggest gains concentrate at the very top of the list (R@1), while R@5 and R@10 often plateau or improve only modestly. This is expected for list-wise rerankers that explicitly optimize the head of the ranking and for retrieval suites where Top-K pools already contain the correct item (R@30 $\approx$ 100%). In a few cases, traditional fusion (e.g., CombSUM/CombMNZ or RRF) remains competitive at larger cutoffs because these methods vote over multiple lists and tend to preserve more of the original "headroom" in $K > 1$. Taken together, the extended results show ViC is most valuable where users care about the very first hit (R@1), while being broadly competitive at deeper cutoffs.

Moreover, Figure 1 contextualizes these performance gains against their inference cost. It clearly shows that **ViC**'s reranking and fusion methods establish a new Pareto frontier. While the original retrievers, such as InternVideo2, are fast, their performance is limited, clustering at the bottom-left. In contrast, **ViC** provides a massive leap in average R@1, pushing the SOTA from 57% to 90%. This gain comes at the expected latency cost of a second-stage reranker. However, the frontier itself shows promising scaling: the 8B and 14B models already achieve strong results, suggesting that the barrier to high performance is low and that future work on lightweight, fine-tuned rerankers could offer an even better performance-cost balance.

4. List Fusion Strategies

Given R off-the-shelf retrievers that produce ranked lists for a query, two standard list-fusion baselines are examined and compared against the proposed ViC.

(a) Soft Voting (score fusion). When calibrated similarity matrices are available, normalize each score distribution per query using min-max scaling and aggregate the results with nonnegative weights:

$$\begin{aligned}\tilde{S}(q,\cdot) \;&=\; \sum_{r=1}^{R} w_r \,\mathrm{norm}\big(S^{(r)}(\cdot)\big)\big),\\ \mathcal{C} \;&=\; \mathrm{TopK}\big(\tilde{S}(q,\cdot)\big).\end{aligned}$$

This family includes classical CombSUM/CombMNZ-style score fusion and serves as a strong yet simple baseline when scores are comparable across retrieval systems.

(b) Reciprocal Rank Fusion (RRF). When only heterogeneous *ranked lists* are available, RRF assigns each item x a fused score as

$$\mathrm{RRF}(x) \;=\; \sum_{r=1}^{R} \frac{1}{k + \mathrm{rank}_r(x)},$$

with a small smoothing constant k (commonly k=60), then returns the Top-K unique items.

Retrieval Task	Subtitle	Ranking prompt	Rationale prompt
Text-to-Video (T2V)			
	×	Rank the candidate videos, represented as grid images, from most to least relevant to the supplied natural language query. Return only the ordered list of indices.	For the top ten ranked items, produce one concise sentence per item explaining the reason for its assigned position. Each explanation must jointly reference the visual content and the multiplicity count indicating how many times that item appeared in the input list.
	✓	Rank the candidate videos represented as grid images and subtitle text according to their relevance to the query, considering both visual and textual information. Return only the ordered index list.	For the top ten ranked items, provide one concise explanatory sentence for each item that jointly references the video content, the accompanying subtitle text, and the multiplicity count describing repeated occurrences in the input list.
Video-to-Text (V2T)			
	×	Rank the candidate captions from most to least relevant to the given video represented as a grid image. Return only the ordered list of indices.	For the top ten ranked captions, produce a single sentence for each explaining the rationale for its position, referencing both the video content and the multiplicity count of the caption in the input list.
	✓	Rank the candidate captions according to their relevance to the video, considering both grid-based visual information and available subtitle context. Return only the ordered index list.	For the top ten ranked captions, provide one sentence per item explaining the ranking reasoning, jointly referencing the video content, the subtitle context, and the multiplicity count.

Table 1. Prompting structure used in the ViC framework for candidate ranking and natural-language rationale generation across T2V and V2T retrieval. The rationale instruction requires explicit justification referencing the visual content, subtitle information when available, and multiplicity derived from retriever outputs.

Backbone	MSRVTT			DiDeMo			ActivityNet			VATEX		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
BASELINES (NO RERANK)												
CLIP4Clip	34.4	55.1	65.0	27.1	52.0	63.7	21.6	46.5	60.3	—	—	—
VAST	49.9	66.8	72.7	51.0	70.3	76.6	50.2	72.2	79.3	77.0	91.9	94.4
GRAM	53.1	74.3	82.4	51.8	71.9	76.4	61.1	82.8	88.7	77.3	95.4	97.3
InternVideo2-6B	54.5	76.2	82.7	59.2	80.0	**86.4**	58.2	82.5	89.7	80.7	96.8	98.4
VIC SINGLE-LIST RERANKER (R=1)												
CLIP4Clip	62.8	69.0	69.7	60.4	65.0	66.3	64.6	66.7	66.8	—	—	—
CLIP4Clip*	64.2	68.9	69.9	—	—	—	—	—	—	—	—	—
VAST	67.3	74.7	75.0	70.2	74.5	75.9	79.7	82.3	82.4	91.9	94.7	94.8
VAST*	68.7	74.9	75.0	—	—	—	—	—	—	92.4	95.7	95.8
GRAM	75.4	**83.3**	**84.3**	70.9	75.9	77.5	82.4	85.4	85.5	—	—	—
GRAM*	**76.2**	83.2	83.9	—	—	—	—	—	—	—	—	—
InternVideo2-6B	74.0	**83.3**	83.8	**78.1**	**84.7**	86.0	**89.8**	**92.6**	**92.6**	95.5	**98.4**	**98.5**
InternVideo2-6B*	75.9	**83.3**	84.0	—	—	—	—	—	—	**95.8**	**98.4**	**98.5**

Table 2. **Zero-shot Text-to-Video (T2V) retrieval results.** Rows marked * use S-Grid; unmarked rows use Grid. Bold indicates the best per benchmark.

Backbone	MSRVTT			DiDeMo			ActivityNet			VATEX		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
BASELINES (NO RERANK)												
CLIP4Clip	29.9	53.3	64.8	20.3	45.9	56.3	20.3	40.7	54.1	—	—	—
VAST	46.2	67.7	73.5	47.8	71.3	77.4	48.7	72.8	80.0	77.6	98.8	99.4
GRAM	50.8	73.6	81.6	49.6	70.2	75.8	52.1	75.9	82.3	72.5	93.1	95.6
InternVideo2-6B	49.5	73.4	82.5	58.8	79.5	84.8	52.4	78.9	88.5	—	—	—
VIC SINGLE-LIST RERANKER (R=1)												
CLIP4Clip	61.3	73.4	74.1	53.8	62.5	63.1	62.8	66.8	66.9	—	—	—
CLIP4Clip*	62.5	73.7	74.2	—	—	—	—	—	—	—	—	—
VAST	62.2	75.7	75.9	63.4	76.4	76.9	75.2	83.8	83.9	99.4	**99.6**	**99.6**
VAST*	63.1	75.9	76.1	—	—	—	—	—	—	**99.6**	**99.6**	**99.6**
GRAM	72.3	86.6	87.2	63.9	77.6	77.8	77.2	86.1	86.2	—	—	—
GRAM*	73.6	86.5	87.0	—	—	—	—	—	—	—	—	—
InternVideo2-6B	74.1	88.7	89.0	**70.7**	**88.9**	**89.3**	**84.9**	**95.7**	**96.0**	—	—	—
InternVideo2-6B*	**76.6**	**89.0**	**89.3**	—	—	—	—	—	—	—	—	—

Table 3. **Zero-shot Video-to-Text (V2T) retrieval results.** Rows marked * use S-Grid; unmarked rows use Grid. Bold indicates the best per benchmark.

Method	MSRVTT			DiDeMo			ActivityNet			VATEX		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
BASELINES AND TRADITIONAL FUSION												
InternVideo2-6B†	54.5	76.2	82.7	59.2	80.0	86.4	58.2	82.5	89.7	80.7	96.8	98.4
RRF	78.3	91.4	94.1	72.8	87.5	91.0	**96.8**	**99.5**	**99.6**	94.7	95.5	95.8
CombSUM	84.4	**93.7**	**96.3**	80.4	**93.1**	**94.8**	95.8	98.7	99.4	96.1	99.4	99.7
CombMNZ	85.3	93.2	95.6	78.0	90.3	94.3	95.0	98.7	99.5	96.4	**99.5**	**99.8**
VIC RANK FUSER (R>1)												
ViC‡	84.2	91.4	92.6	85.5	91.0	92.2	94.8	97.5	97.6	96.1	99.3	99.3
ViC	**87.1**	92.2	93.2	**87.4**	91.5	92.7	96.0	97.7	97.8	**97.5**	99.4	99.5

Table 4. **Zero-shot Text-to-Video (T2V) rank fusion results.** InternVideo2-6B† denotes the previous strong backbone baseline. ViC‡ indicates ViC without duplicates in the candidate sequence. Bold indicates the best per benchmark.

Method	MSRVTT			DiDeMo			ActivityNet		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
BASELINES AND TRADITIONAL FUSION									
InternVideo2-6B†	49.5	73.4	82.5	58.8	79.5	84.8	52.4	78.9	88.5
RRF	80.2	91.7	94.5	73.2	88.5	91.6	**97.4**	**99.6**	**99.9**
CombSUM	83.0	94.9	96.9	83.1	**94.5**	**95.9**	95.2	96.7	97.3
CombMNZ	86.9	**95.5**	**97.1**	80.8	93.0	95.7	92.2	96.6	97.2
VIC RANK FUSER (R>1)									
ViC‡	80.7	94.3	95.0	76.1	88.0	89.1	91.9	97.6	97.8
ViC	**88.1**	95.3	96.1	**84.3**	92.1	93.2	96.2	98.5	98.6

Table 5. **Zero-shot Video-to-Text (V2T) rank fusion results.** InternVideo2-6B† denotes the previous strong backbone baseline. ViC‡ indicates ViC without duplicates in the candidate sequence. Bold indicates the best per benchmark.

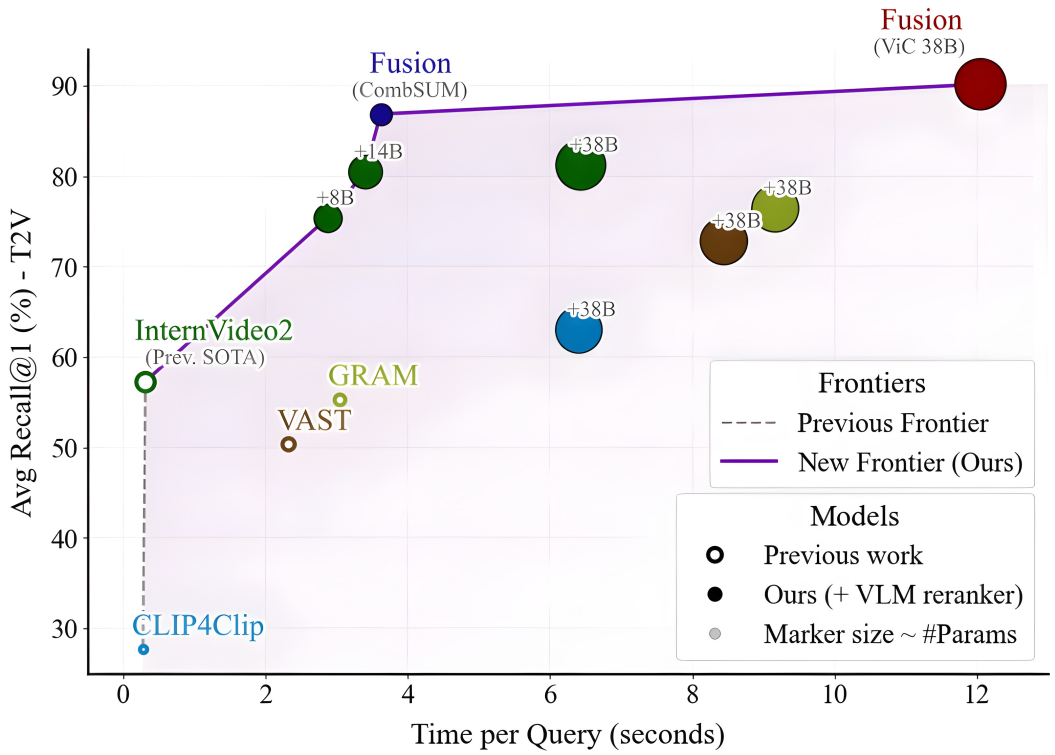

Figure 1. Efficiency vs. Performance Trade-off. Time per query vs. Avg Recall@1 for T2V retrieval over the benchmarks MSR-VTT, DiDeMo and ActivityNet in zero-shot settings. Marker size represents model parameters. The Pareto frontier highlights optimal trade-offs. ViC establishes strong Pareto points in the R@1-latency plane by trading a second-stage VLM pass for large top-1 gains.