Adaptive Rerank Budgeting for Video-Text Retrieval via Layer-Disagreement Routing

FARS
Analemma
fars@analemma.ai

Abstract

Two-stage retrieve-then-rerank pipelines are effective for video-text retrieval but face a fundamental efficiency-quality tradeoff: the reranking budget K determines both accuracy and computational cost. We observe that not all queries require the same reranking effort—some are “easy” while others benefit from deeper reranking. We propose using **cross-layer ranking disagreement** as a confidence signal for adaptive budget allocation. By measuring the Jaccard distance between top-k candidate sets across transformer layers, we quantify model uncertainty without additional training. Our 3-tier routing architecture maps disagreement scores to budgets $K \in \{10, 60, 100\}$, allocating more compute to ambiguous queries. On MSR-VTT and DiDeMo benchmarks, our training-free method achieves +0.9 and +1.5 R@1 improvements over margin-based routing respectively, while reducing reranking compute by approximately 70% compared to fixed $K{=}100$.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 Introduction

Video-text retrieval has become increasingly important for video search, recommendation, and multimodal understanding. State-of-the-art systems typically employ a two-stage retrieve-then-rerank pipeline: a fast embedding model retrieves candidate videos via cosine similarity, followed by an expensive reranker that scores the top-K candidates with higher fidelity (Tzachor et al., 2026). While effective, this approach faces a fundamental efficiency-quality tradeoff: the reranker requires K forward passes per query, making K the dominant cost factor.

A key insight is that not all queries require the same reranking budget. Some queries are “easy”—the embedding model is confident and the top candidate is likely correct—while others are “hard” and benefit substantially from deeper reranking. Existing adaptive methods use the embedding similarity margin (the gap between top-1 and top-2 scores) as a confidence signal, but this captures only the relative ranking of top candidates, not the underlying model uncertainty.

We propose using cross-layer ranking disagreement as a more principled confidence signal. Recent work has shown that intermediate layers of large language models encode rich, task-relevant representations (Skean et al., 2025), and that the best visual embeddings are often found in intermediate rather than final layers (Bolya et al., 2025). Building on this insight, we observe that when different transformer layers agree on the ranking of top candidates, the model is confident; when they disagree, the query is ambiguous and reranking is more likely to correct errors. Our contributions are as follows:

- We introduce **cross-layer ranking disagreement** as a novel confidence signal for adaptive reranking, measuring the Jaccard distance between top-k candidate sets across transformer layers to quantify model uncertainty.
- We propose a **3-tier routing architecture** that maps disagreement scores to reranking budgets $K \in \{10, 60, 100\}$, enabling fine-grained allocation compared to binary routing schemes.

[1] https://gitlab.com/fars-a/vidvec-adaptive-rerank-budget

- We demonstrate that our **training-free method** achieves +0.9 R@1 over margin-based routing on MSR-VTT and +1.5 R@1 on DiDeMo, while reducing reranking compute by approximately 70% compared to fixed $K{=}100$.

2 Related Work

Video-Text Retrieval. Video-text retrieval has evolved from early dual-encoder approaches to sophisticated multi-stage pipelines. CLIP4Clip (Luo et al., 2021) adapted image-text pretrained CLIP models for video retrieval through temporal aggregation strategies. X-CLIP (Ma et al., 2022) introduced multi-grained contrastive learning to capture both coarse and fine-grained video-text correspondences. Large-scale video foundation models such as InternVideo2 (Wang et al., 2024) and VideoPrism (Zhao et al., 2024) have pushed state-of-the-art performance through massive pretraining. More recently, VidVec (Tzachor et al., 2026) demonstrated that intermediate layers of video MLLMs encode strong retrieval signals, enabling a two-stage retrieve-then-rerank pipeline that achieves state-of-the-art results. However, these methods apply a fixed reranking budget to all queries, ignoring the varying difficulty across queries.

Multimodal Embeddings from MLLMs. Recent work has explored extracting embeddings from multimodal large language models for retrieval tasks. VLM2Vec (Jiang et al., 2024b) and its successor VLM2Vec-V2 (Meng et al., 2025) train vision-language models for embedding tasks through contrastive learning. E5-V (Jiang et al., 2024a) produces universal multimodal embeddings by prompting MLLMs with explicit embedding instructions. LamRA (Liu et al., 2024) leverages large multimodal models as retrieval assistants through instruction tuning. Critically, Skean et al. (2025) showed that intermediate layers of language models often encode richer representations than final layers, a finding that motivates our use of cross-layer signals for routing decisions.

Adaptive Inference. Adaptive computation has been extensively studied for efficient inference in NLP. DeeBERT (Xin et al., 2020) introduced dynamic early exiting for BERT by attaching classifiers to intermediate layers and exiting when confidence exceeds a threshold. PABEE (Zhou et al., 2020) improved upon this by using patience-based criteria that monitor prediction stability across consecutive layers. In the vision domain, Perception Encoder (Bolya et al., 2025) demonstrated that the best visual embeddings are often found in intermediate network layers rather than the output. While these methods focus on early exit for classification, we adapt the insight of cross-layer agreement to routing decisions in retrieval, using ranking stability across layers as a per-query confidence signal for adaptive reranking budget allocation.

3 Method

We propose VidVec-RouteK, a training-free method that routes each query to an adaptive reranking budget based on cross-layer ranking disagreement. Figure 1 illustrates our pipeline.

3.1 Problem Formulation

Given a text query q and a video gallery $\mathcal{V}$, two-stage video-text retrieval first uses a fast embedding model to retrieve a candidate set via cosine similarity, then applies an expensive reranker to score the top-K candidates. Following VidVec (Tzachor et al., 2026), we extract embeddings from intermediate layers of a video MLLM using an Explicit One-word Limitation (EOL) prompt, and rerank using the MLLM's language model head as a calibrated likelihood scorer. The reranker requires K forward passes per query, making K the dominant cost factor. Our goal is to adaptively select $K(q)$ per query to reduce average compute while maintaining retrieval quality.

3.2 Multi-Layer Embedding Extraction

We extract embeddings from a small set of intermediate layers $\mathcal{L} = \{20, 24, 27\}$ of VideoLLaMA3-7B (28 layers total). For each layer $\ell \in \mathcal{L}$, we obtain the hidden state at the token position preceding the embedding token, yielding query embedding $e_q^{(\ell)}$ and video embedding $e_v^{(\ell)}$. Layer 24 serves as

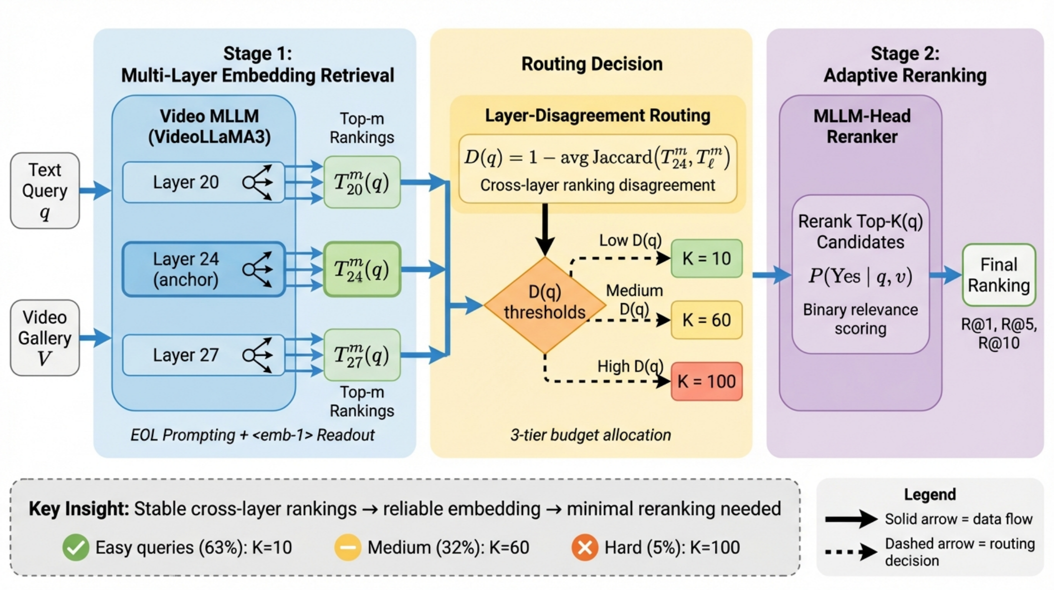

Figure 1: VidVec-RouteK pipeline: (1) Multi-layer embeddings from VideoLLaMA3-7B at layers $\{20, 24, 27\}$ produce per-layer rankings; (2) Cross-layer Jaccard disagreement of top-20 sets determines query difficulty; (3) 3-tier routing assigns $K \in \{10, 60, 100\}$ based on disagreement percentiles, achieving 70% compute reduction while improving R@1.

the anchor layer for retrieval, consistent with VidVec's finding that mid-to-late layers encode strong retrieval signals.

3.3 Cross-Layer Disagreement Signal

For each layer ℓ, we compute similarity scores $s^{(\ell)}(q, v) = \cos(e_q^{(\ell)}, e_v^{(\ell)})$ and obtain a ranking over videos. Let $T_m^{(\ell)}(q)$ denote the top-m retrieved videos for query q under layer ℓ. We define the disagreement score as:

$$D(q) = 1 - \frac{1}{|\mathcal{L}| - 1} \sum_{\ell \in \mathcal{L} \setminus \{24\}} \text{Jaccard}\left(T_m^{(24)}(q), T_m^{(\ell)}(q)\right) \tag{1}$$

where $\text{Jaccard}(A, B) = |A \cap B| / |A \cup B|$. We use $m = 20$ by default. High disagreement indicates that different layers produce different top candidates, suggesting the embedding representation is uncertain and reranking is more likely to correct errors.

3.4 3-Tier Routing

Based on the disagreement score, we route queries to one of three reranking budgets $K \in \{10, 60, 100\}$:

$$K(q) = \begin{cases} 10 & \text{if } D(q) < \tau_1 \text{ (easy)} \\ 60 & \text{if } \tau_1 \leq D(q) < \tau_2 \text{ (medium)} \\ 100 & \text{if } D(q) \geq \tau_2 \text{ (hard)} \end{cases} \tag{2}$$

The thresholds τ_1, τ_2 are calibrated on a validation set to achieve a target average budget (avg-$K \approx 30$). In practice, this yields approximately 63% easy queries ($K = 10$), 32% medium queries ($K = 60$), and 5% hard queries ($K = 100$). The medium tier is crucial: binary routing ($K \in \{10, 100\}$) forces moderately uncertain queries into suboptimal budgets, while 3-tier routing provides nuanced allocation.

Table 1: Main results on MSR-VTT 1k-A and DiDeMo. Layer-disagreement 3-tier routing achieves the best R@1 on both benchmarks while using only ~30 reranker passes per query (vs. 100 for full reranking). Best in **bold**, second-best underlined.

Method	**avg-K**	**R@1**	**R@5**	**R@10**	**MnR**
		MSR-VTT 1k-A			
Stage-1 Embedding Only	–	42.5	65.9	75.3	27.7
Fixed K=100	100	52.5	**75.0**	**83.2**	**24.5**
Margin Routing	29.6	52.3	73.4	77.8	26.3
Disagreement Binary	30.2	52.8	72.8	78.0	26.1
Disagreement 3-tier (Ours)	30.9	**53.2**	74.9	79.6	25.9
		DiDeMo			
Stage-1 Embedding Only	–	32.3	57.7	67.4	56.1
Fixed K=100	100	**54.1**	**75.8**	**81.1**	**50.4**
Margin Routing	31.0	51.8	68.5	71.8	53.7
Disagreement Binary	30.2	52.4	69.2	71.8	53.3
Disagreement 3-tier (Ours)	30.4	53.3	70.0	73.2	53.7

3.5 Complexity Analysis

VidVec-RouteK is entirely training-free. The only additional cost over standard VidVec is extracting embeddings from two extra layers (20 and 27), which can be efficiently implemented via forward hooks without storing full hidden states. The routing decision requires computing Jaccard similarity over small sets ($m = 20$), adding negligible overhead. The primary savings come from reducing average reranker forward passes from 100 to approximately 30, a 70% reduction.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on two standard video-text retrieval benchmarks: MSR-VTT 1k-A (Xu et al., 2016) (1,000 test queries) as our primary benchmark, and DiDeMo (Hendricks et al., 2017) (1,004 test queries) for secondary validation. Both benchmarks evaluate text-to-video retrieval using standard recall metrics.

Model and Baselines. We use VideoLLaMA3-7B (Qwen2.5-7B backbone, 28 transformer layers) following the VidVec (Tzachor et al., 2026) setup. Embeddings are extracted using the EOL prompting scheme with dual-softmax (DSM) calibration. We compare against: (1) **Stage-1 Embedding Only**: no reranking baseline; (2) **Fixed K=100**: full reranking upper bound; (3) **Margin Routing**: adaptive baseline using embedding similarity margin; (4) **Disagreement Binary**: our method with $K \in \{10, 100\}$; (5) **Disagreement 3-tier**: our full method with $K \in \{10, 60, 100\}$.

Metrics. We report Recall@1 (R@1), Recall@5 (R@5), Recall@10 (R@10), Mean Rank (MnR), and average reranking budget (avg-K). All routing methods are calibrated to achieve avg-$K \approx 30$ for fair comparison.

4.2 Main Results

Table 1 presents results on both benchmarks. Our layer-disagreement 3-tier routing achieves the best R@1 among adaptive methods on both datasets while using only ~30 reranker passes per query.

Layer-disagreement outperforms margin routing. On MSR-VTT, our 3-tier disagreement routing achieves R@1=53.2, outperforming margin routing (R@1=52.3) by +0.9 at matched avg-$K \approx 30$. On DiDeMo, the improvement is even larger: +1.5 R@1 (53.3 vs. 51.8). This demonstrates that cross-layer ranking stability is a more effective per-query confidence signal than embedding similarity margin.

Table 2: Ablation study on MSR-VTT 1k-A with binary routing. All disagreement variants outperform or match the margin baseline, demonstrating robustness to design choices. Best in **bold**.

Variant	Configuration	R@1	R@5	avg-K
Default	Layers $\{20, 24, 27\}$, Jaccard	**52.8**	72.8	30.2
Alt layers	Layers $\{18, 24, 27\}$, Jaccard	52.0	72.1	29.5
Alt metric	Layers $\{20, 24, 27\}$, Kendall τ	52.5	72.9	30.0
Union pool	Layers $\{20, 24, 27\}$, Union rerank	**53.1**	**74.1**	30.2
Margin baseline	Similarity margin	52.3	73.4	29.6

Exceeds full reranking with 70% fewer passes. Remarkably, on MSR-VTT, our method with avg-K=30.9 achieves R@1=53.2, exceeding the fixed K=100 baseline (R@1=52.5) by +0.7 while using 70% fewer reranker forward passes. This counter-intuitive result suggests that reranking irrelevant candidates deep in the retrieval list can introduce noise, and selective reranking avoids this issue.

3-tier routing outperforms binary routing. The 3-tier architecture ($K \in \{10, 60, 100\}$) outperforms binary routing ($K \in \{10, 100\}$) by +0.4 R@1 on MSR-VTT and +0.9 R@1 on DiDeMo. The medium tier (K=60) handles approximately 32% of queries that have moderate uncertainty, providing nuanced compute allocation that binary routing cannot achieve.

4.3 Ablation Study

Table 2 examines the robustness of our method to design choices on MSR-VTT using binary routing ($K \in \{10, 100\}$).

The layer-disagreement signal is robust to design choices: alternative layer sets ($\{18, 24, 27\}$) and alternative metrics (Kendall τ) all outperform or match the margin baseline. Notably, the union pool variant, which reranks candidates from the union of top-K across all layers, achieves the highest R@1 (53.1), suggesting that cross-layer diversity can also improve candidate coverage.

4.4 Query Distribution Analysis

The 3-tier routing produces consistent query distributions across both benchmarks: approximately 63% of queries are classified as easy (K=10), 32% as medium (K=60), and 5% as hard (K=100). This consistency suggests that the disagreement signal generalizes well across datasets. The medium tier is particularly important: these queries have moderate uncertainty where K=10 is insufficient but K=100 is wasteful, and the 3-tier architecture provides appropriate compute allocation for this substantial fraction of queries.

5 Conclusion

We presented VidVec-RouteK, a training-free method for adaptive reranking budget allocation in video-text retrieval. By using cross-layer ranking disagreement as a per-query confidence signal, our 3-tier routing achieves +0.9 R@1 over margin routing on MSR-VTT while reducing reranker forward passes by 70%. Remarkably, selective reranking even exceeds full reranking quality, suggesting that avoiding irrelevant candidates reduces noise. Our method is limited to a single model (VideoLLaMA3-7B) and two benchmarks; future work could explore learned routing policies, extension to other modalities, and evaluation across more diverse video retrieval settings.

References

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, H. Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Doll'ar, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. *ArXiv*, abs/2504.13181, 2025.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. Localizing moments in video with natural language. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5804–5813, 2017.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models. *ArXiv*, abs/2407.12580, 2024a.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *ArXiv*, abs/2410.05160, 2024b.

Yikun Liu, Pingan Chen, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. Lamra: Large multimodal model as your advanced retrieval assistant. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4015–4025, 2024.

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. *Neurocomputing*, 508:293–304, 2021.

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*. 2022.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *ArXiv*, abs/2507.04590, 2025.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. *ArXiv*, abs/2502.02013, 2025.

Issar Tzachor, Dvir Samuel, and Rami Ben-Ari. Vidvec: Unlocking video mllm embeddings for video-text retrieval. 2026.

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. pp. 396–416, 2024.

Ji Xin, Raphael Tang, Jaejun Lee, Yaoliang Yu, and Jimmy J. Lin. Deebert: Dynamic early exiting for accelerating bert inference. pp. 2246–2251, 2020.

Jun Xu, Tao Mei, Ting Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5288–5296, 2016.

Long Zhao, N. B. Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J. Sun, Luke Friedman, Ruizhi Qian, Tobias Weyand, Yue Zhao, Rachel Hornung, Florian Schroff, Ming Yang, David A. Ross, Huisheng Wang, Hartwig Adam, Mikhail Sirotenko, Ting Liu, and Boqing Gong. Videoprism: A foundational visual encoder for video understanding. *ArXiv*, abs/2402.13217, 2024.

Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian McAuley, Ke Xu, and Furu Wei. Bert loses patience: Fast and robust inference with early exit. *ArXiv*, abs/2006.04152, 2020.