DPDV: Dual-Pathway and Dual-View Representation Learning for Bridging Information Asymmetry in Text-Video Retrieval

Zequn Xie[1*] **Xin Liu**[2*] **Boyun Zhang**[1] **Fangming Feng**[1] **Tao Jin**[1†]
[1] Zhejiang University
[2] Southwestern University of Finance and Economics

Abstract

In recent years, CLIP-based text-video retrieval methods have developed rapidly, with research focusing on constructing diverse features and achieving effective interactions. However, the asymmetry of cross-modal information poses a challenge to accurately establishing retrieval relationships. To overcome this challenge, we propose a novel video retrieval framework, termed the Dual-Pathway and Dual-View model (DPDV), which consists of the Dual-Pathway Partitioning Module (DPPM) for constructing features at an appropriate granularity and the Dual-View Interaction Module (DVIM) for performing effective feature interactions. For DPPM, we simulate a human macro-level cognitive perspective by partitioning visual features into two categories based on their relevance to the text query and supplementing less relevant features with additional textual information. For DVIM, we simulate a human alignment strategy from macro to micro levels, focusing on local visual features while comprehensively modeling fine-grained interactions. We evaluate DPDV on five benchmark datasets, including MSRVTT, and achieve state-of-the-art performance on video retrieval.

1 Introduction

With the rapid development of the Internet, massive amounts of unlabeled video data are continuously uploaded and shared. The goal of text-to-video retrieval is to identify target videos from massive unlabeled collections based on a given textual query. Recently, large-scale text-image pre-trained model CLIP (Radford et al., 2021) have achieved remarkable success in various multimodal tasks (Lei et al., 2021; Piergiovanni et al., 2022; Li et al., 2023), providing new insights for video retrieval. Existing methods (Luo et al., 2022; Wu et al., 2023; Wang et al., 2024; Liu et al., 2025a) typically leverage CLIP to project both text and video into a shared latent space, thereby establishing feature-level similarity relationships. For example, Clip4clip (Luo et al., 2022) exemplifies a simple baseline by computing similarities between sentence and frame.

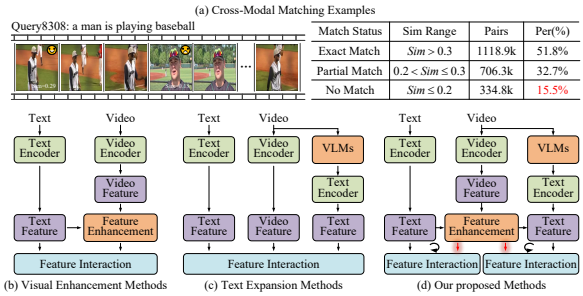

Figure 1: **Motivation.** (a) Specific and statistical examples of queries exhibiting information asymmetry in the MSRVTT-9k. (b, c) Existing methods for visual feature enhancement and text expansion. (d) Our proposed methods for addressing information asymmetry.

However, due to the unequal information capacity and the non-recoverable semantics that give rise to information asymmetry (Gong et al., 2024; Yuan et al., 2024; Gao et al., 2025) between modalities, accurate similarity computation becomes a challenging task. In Figure 1(a), we present specific matching example and statistics, which illustrate the potential impact of information asymmetry on feature interaction. Based on Query8308: “a man is playing baseball" and its corresponding video frames, we can visually observe that the clip showing “a man receiving an interview" is irrelevant to the query. Based on the Clip4clip (Luo et al., 2022) model, we evaluate query-frame matching on MSRVTT-9k (Xu et al., 2016) and find that 15.5% of the frames are unmatched with their corresponding query. This partial or unmatched phenomenon directly reflects modality information asymmetry and is particularly evident in sparse-text queries on LSMDC (Rohrbach et al., 2015) and long-video retrieval on DiDeMo (Anne Hendricks et al., 2017).

[*] Equal contribution.
[†] Corresponding author.

To mitigate the problem of information asymmetry in cross-modal interactions, existing methods can be categorized into two types: ❶ Visual enhancement (Gorti et al., 2022; Wang et al., 2023; Fang et al., 2023; Tian et al., 2024) (see Figure 1(b)): These methods focus on extracting more discriminative regions from video frames. X-Pool (Gorti et al., 2022) employs an attention mechanism to extract the most relevant video frames for a given text, thereby enhancing the representational capacity of visual. ❷ Text expansion (Wu et al., 2023; Xue et al., 2022; Wang et al., 2024; Shen et al., 2025) (see Figure 1(c)): These methods aim to enhance text modality representations by generating additional textual descriptions. Cap4Video (Wu et al., 2023) leverages vision-language models (VLMs) for zero-shot video captioning, thereby expanding the semantic scope of the textual modality. Although these methods achieve some success in mitigating information asymmetry, they still face limitations in precise cross-modal feature matching. The limitations of these methods are as follows: ❶ they assume all visual features in a video are relevant to the text, enforcing strong alignment while ignoring partial or unmatched content; ❷ they assume that the augmented text is superior to the original query, overlooking the potential introduction of redundant or misleading information.

In fact, bridging information asymmetry essentially involves selecting an appropriate feature granularity, which is realized by proportionally splitting default paired content to increase the "Exact Match" ratio while reducing the "No Match" ratio. Accordingly, global-level frame features can be partitioned based on the query to obtain the “Exact Match Path" and the “No Match Path". For simplicity, we refer to the first as Spot-Path (*i.e.*, “Spot the Key"), which focuses on the content deemed important, and the second as Recover-Path (*i.e.*, “Recover the Rest"), which leverages text augmentation methods (VLMs) to complement missing textual information for video segments. The division of these two paths significantly enhances the alignment of cross-modal features, providing support for subsequent fine-grained feature interaction.

Based on the above analysis, we propose a novel text-video retrieval framework, termed the **D**ual-**P**athway and **D**ual-**V**iew Model (**DPDV**), whose core architecture is illustrated in Figures 1(d) and 2. **First**, we propose a **D**ual-**P**athway **P**artitioning **M**odule (**DPPM**), which comprises a Spot-Path that selects high-similarity frames based on global sentence-frame similarity for fully informative interactions, and a Recover-Path that leverages VLMs, *e.g.*, VILA (Lin et al., 2024), to generate textual descriptions for low-similarity frames as complementary interaction targets. **Second**, we propose a **D**ual-**V**iew **I**nteraction **M**odule (**DVIM**) to further model fine-grained interactions between text and video from the perspective of pathway decomposition. Since DPDV performs pathway decomposition based on the global sentence-frame relationships, it naturally enables macro-level interactions between text and video features. For micro-level word-patch interactions, we merge patch features via clustering algorithm, enabling fine-grained alignment while effectively addressing patch redundancy without being constrained by trivial details. **Third**, DPDV and DVIM collaborate synergistically, jointly promoting cross-modal feature enhancement and interaction. We summarize our contributions as follows:

❶ We propose a novel video retrieval framework, DPDV, which integrates DPPM for feature enhancement and DVIM for feature interaction, providing a new paradigm for bridging information asymmetry.

❷ The proposed DPPM simulates the human macro-level perspective by partitioning visual features and supplementing textual ones, thereby enhancing cross-modal interaction accuracy.

❸ The proposed DVFI simulates the human alignment strategy from macro- to micro-level, effectively focusing on local visual features and comprehensively considering fine-grained interactions.

❹ We conduct extensive experiments on five benchmark datasets, including MSRVTT, DiDeMo, LSMDC, ActivityNet, and Charades, and achieve state-of-the-art retrieval performance.

2 Related Work

Text-Video Retrieval aims to retrieve the most semantically relevant video from a large collection given a textual query. Early works (Liu et al., 2019; Gabeur et al., 2020; Patrick et al., 2020) primarily focus on text and video feature representations based on machine learning and deep learning, which have facilitated the development of a series of benchmarks (Xu et al., 2016; Anne Hendricks et al., 2017). With the remarkable success of the large-scale text-image pre-trained model CLIP (Radford et al., 2021) in cross-modal representation, it has spurred improvements in retrieval tasks (Lei et al., 2021; Gorti et al., 2022;

Wang et al., 2023). Clip4clip (Luo et al., 2022) transfers the knowledge of the CLIP model to video-language retrieval in an end-to-end manner. Recently, transformer-based video retrieval methods (Gorti et al., 2022; Jin et al., 2023b,c) use cross-attention to abstract multi-modal cues, achieving significant performance gains. For example, TS2Net (Liu et al., 2022) employs a "token shift and selection transformer" to preserve token integrity and capture subtle actions, improving retrieval performance. These foundational studies have laid the groundwork for subsequent, more powerful feature enhancement and interaction.

Feature Enhancement aims to produce more expressive feature representations and can be categorized into two types of existing methods. ❶ Video feature enhancement (Liu et al., 2022; Jin et al., 2023a; Tian et al., 2024) focuses on extracting more discriminative regions from videos. X-Pool (Gorti et al., 2022) employs an attention mechanism to select the most relevant video frames corresponding to a given text, thereby enhancing the representational capacity of visual features. ❷ Text feature enhancement (Ma et al., 2024; Wang et al., 2024; Shen et al., 2025) aims to improve text modality representations by generating additional textual descriptions. Cap4Video (Wu et al., 2023) leverages visual-language models (VLMs) for zero-shot video captioning, thereby expanding the semantic scope of the textual modality. Compared to these indiscriminate feature enhancement methods, DPDV adopts a more targeted pathway selection strategy to achieve joint enhancement of visual and textual features, providing a more convincing solution for mitigating information asymmetry.

Feature Interaction refers to the process of aligning text and video features at different granularities. Existing works mainly focus on three types of feature interaction methods, including coarse (Luo et al., 2022; Gorti et al., 2022), fine (Liu et al., 2022; Fang et al., 2023), and coarse-to-fine level (Yang et al., 2024; Tian et al., 2024). Coarse-to-fine methods have become the preferred strategy for feature interaction due to their comprehensive consideration of multiple granularities. For example, HBI (Jin et al., 2023a) generates hierarchical representations by clustering frame-level features and employs a game-theoretic fine-grained alignment, attracting considerable attention. UCoFiA (Wang et al., 2023) introduces a unified coarse-to-fine alignment model that effectively enhances text-video retrieval performance from a comprehensive perspective. Compared to the those methods, DPDV achieves strong performance by focusing on macro- and micro-level feature interactions along globally partitioned pathways to mitigate information asymmetry. This further suggests that appropriately adapted granular information can reduce the complexity of feature interactions.

3 Methodology

3.1 Feature Extraction and Interaction

Feature Extraction. Let $\mathcal{D} = (\mathcal{T}, \mathcal{V})$ denote a language and vision dataset, where $\mathcal{T}$ is a set of texts, and $\mathcal{V}$ is a set of videos. The goal of text-to-video retrieval is to locate the most relevant video v from a video set $\mathcal{V}$ given a text query $t \in \mathcal{T}$. Recent works (Luo et al., 2022; Jin et al., 2023a; Wang et al., 2024) have shown CLIP's (Radford et al., 2021) strong performance in feature representation, inspiring us to employ CLIP as our backbone to effectively construct retrieval matching relationships. Specifically, a video $v \in \mathcal{V}$ consists of N_f sequential frames $[f_1, f_2, ..., f_{N_f}] \in \mathbb{R}^{N_f \times H \times W \times C}$, where each frame is divided into N_p patches $[p_1, p_2, ..., p_{N_p}] \in \mathbb{R}^{N_p \times P \times P \times C}$ with $P \times P$ size. Following previous work (Luo et al., 2022; Gorti et al., 2022), we utilize the CLIP visual encoder to extract the patch features $V_p = [p_1, \ldots, p_{N_p}] \in \mathbb{R}^{N_p \times D}$ for each frame, and set p_0 as the [CLS] token of the current frame. We aggregate the [CLS] tokens of all video frames to obtain the frame features $V_f = [f_1, f_2, ..., f_{N_f}] \in \mathbb{R}^{N_f \times D}$. Similarly, given a query $t \in \mathcal{T}$, we leverage the CLIP text encoder to extract the word features $T_w = [w_1, w_2, \ldots, w_{N_w}] \in \mathbb{R}^{N_w \times D}$, where N_w denotes the number of words in the sentence. Following previous work (Luo et al., 2022; Gorti et al., 2022), we take the end of the sequence [EOS] token as the sentence feature $T_s = [s] \in \mathbb{R}^{1 \times D}$.

Feature Interaction. Feature interaction refers to the process of computing the similarity between cross-modal text and video features. Mean-Pooling is commonly used to compute feature similarities, e.g., between T_s and V_f, between T_s and V_f:

$$S_{T_s,V_f} = \frac{\boldsymbol{T_s} \cdot \boldsymbol{V_f}}{|\boldsymbol{T_s}| \cdot |\boldsymbol{V_f}|}, \; S_{T_w,V_p} = \frac{\boldsymbol{T_w} \cdot \boldsymbol{V_p}}{|\boldsymbol{T_w}| \cdot |\boldsymbol{V_p}|}, \quad (1)$$

where $\boldsymbol{T_s} = T_s$, $\boldsymbol{T_w} = \frac{1}{N_w}\sum_i^{N_w} T_{w,i}$, $\boldsymbol{V_f} = \frac{1}{N_f}\sum_i^{N_f} V_{f,i}$, and $\boldsymbol{V_p} = \frac{1}{N_f \cdot N_p}\sum_i^{N_f}\sum_j^{N_p} V_{f,i,j}$. Therefore, the word-patch (or sentence-frame)

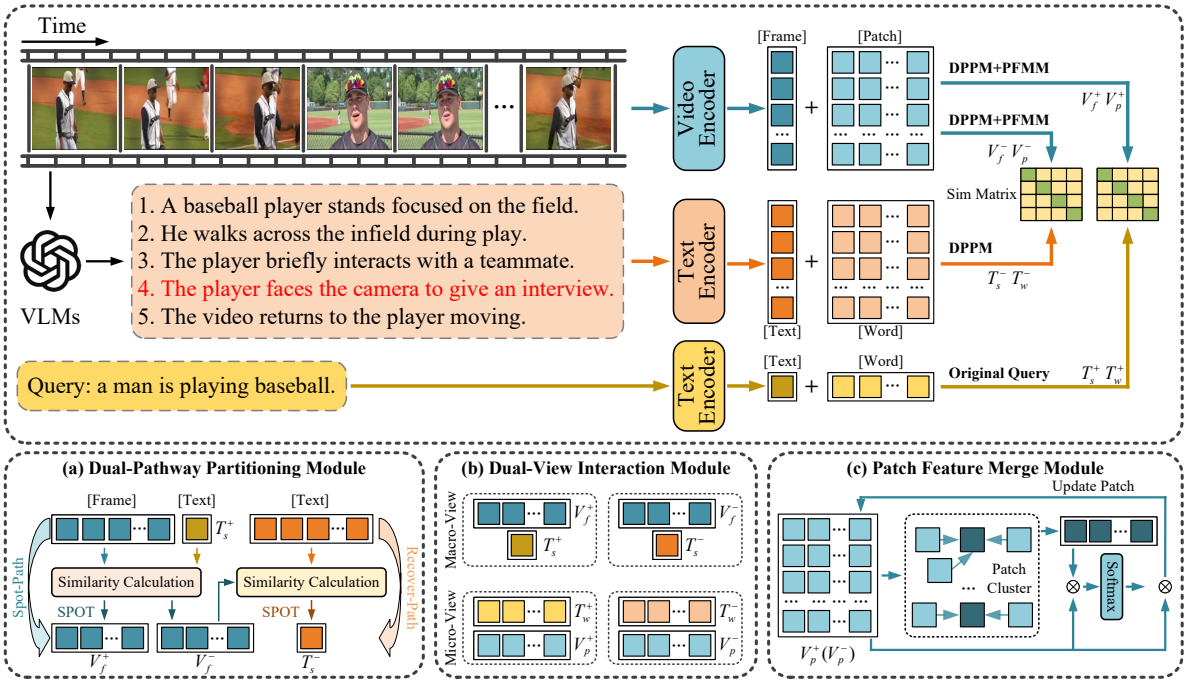

Figure 2: **Framework.** The top illustrates the overall DPDV pipeline. The bottom shows: (a) the Dual-Pathway Partitioning Module for query-aware frame selection and textual completion; (b) the Dual-View Interaction Module for macro- and micro-level feature alignment; (c) the Patch Feature Merge Module for aggregating patch feature.

cross-modal contrastive loss can be formulated as:

$$\mathcal{L}_{T_w,V_p} = -\frac{1}{2}\left(\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(S_{T_w^i,V_p^i}/\tau)}{\sum_{j=1}^{B}\exp(S_{T_w^i,V_p^j}/\tau)} + \frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(S_{T_w^i,V_p^i}/\tau)}{\sum_{j=1}^{B}\exp(S_{T_w^j,V_p^i}/\tau)}\right), \tag{2}$$

where B is the batch size, τ is the temperature hyper-parameter, and $S_{T_w^i,V_p^j}$ represents the similarity between the i^{th} text word features and the j^{th} video patch features in the entire batch B. This loss function maximizes the similarity of positive pairs and minimizes the similarity of negative pairs.

3.2 Dual-Pathway Partitioning Module

In Section 3.1, we sequentially use CLIP features and the Mean-Pooling interaction method to compute the key similarity scores for retrieval. However, the semantic and structural differences between the text and video modalities lead to information asymmetry, which in turn interferes with the computation of similarity. In other words, both modalities need to provide features at equivalent granularity to achieve more accurate similarity results. Therefore, in this section, we propose a **D**ual-**P**athway **P**artitioning **M**odule (**DPPM**), which partitions the original information-asymmetric interaction paths into a Spot-Path, providing fully informative interaction targets, and a Recover-Path, providing complementary interaction targets. Figure 2 (a) illustrates the complete framework.

Spot-Path Representation. The Spot-Path corresponds to the video segments most relevant to the text. X-Pool (Gorti et al., 2022) uses the text query along with an attention mechanism to selectively pool frame features, thereby obtaining the implicitly interactive Spot-Path. However, this method lacks explicit selectivity, resulting in uncertainty in feature path selection. Therefore, we adopt a simpler and more transparent approach to construct interaction paths with higher confidence. Specifically, when aligning text and video, humans first read the overall sentence feature $T_s = [s] \in \mathbb{R}^{1\times D}$, and then compare it with each video frame feature $V_f = [f_1, f_2, \ldots, f_{N_f}] \in \mathbb{R}^{N_f\times D}$ to identify the effective frames $V_f^+ \in \mathbb{R}^{N_f^+\times D}$, where $N_f^+ < N_f$. This explicit selection can be directly based on the magnitude of the similarity $S_{T_s,V_f} \in \mathbb{R}^{1\times N_f}$ between T_s and V_f, which is formulated as:

$$V_f^+ = \text{SPOT}(V_f, S_{T_s,V_f}). \tag{3}$$

In addition, the patch features are also updated from $V_p \in \mathbb{R}^{(N_f\cdot N_p)\times D}$ to $V_p^+ \in \mathbb{R}^{(N_f^+\cdot N_p)\times D}$.

Recover-Path Representation. The construction of V_f^+ and V_p^+ in the Spot-Path facilitates a more balanced and accurate interaction between the text and visual content. However, the discarded visual features $V_f^-(V_f - V_f^+)$ and $V_p^-(V_p - V_p^+)$ inevitably lead to information loss. Cap4Video (Wu et al., 2023) first employs large vision-language models, e.g., ChatGPT (OpenAI, 2023), to generate textual descriptions of videos in a zero-shot manner, which can compensate for the limitations of text representations. Therefore, benefiting from this explicit textual augmentation approach, adding new textual descriptions to V^- is sufficient to construct the Recover-Path. Since V^- cannot be determined in advance, we generate N_s segmented descriptions for the entire video and encode them as $T_s = [s_1, s_2, ..., s_{N_s}] \in \mathbb{R}^{N_s \times D}$. Therefore, we follow the Spot-Path approach to obtain a complete textual description $T_s^- \in \mathbb{R}^{N_s^- \times D}$ adapted to V^-:

$$T_s^- = \text{SPOT}(T_s, S_{T_s, \boldsymbol{V_f^-}}), \tag{4}$$

where $\boldsymbol{V_f^-} = \frac{1}{N_f^-} \sum_{i=1}^{N_f^-} V_{f,i}^-$. Similarly, the word features are also updated from $T_w \in \mathbb{R}^{(N_s \cdot N_w) \times D}$ to $T_w^- \in \mathbb{R}^{(N_s^- \cdot N_w) \times D}$. The textual descriptions T_s^- and T_w^- differ from Cap4Video (Wu et al., 2023) in that: **1)** they selectively expand text queries rather than performing global expansion; **2)** they mitigate VLMs hallucination, resulting in more accurate textual matching.

The Dual-Pathway Partitioning Module provides an efficient solution to information asymmetry caused by mismatched feature granularity, effectively enhancing relevant feature pairs while accurately compensating for less relevant ones.

3.3 Dual-View Interaction Module

After feature partitioning, it is necessary to further consider interactions between features at different granularity levels. Previous works have proposed interaction strategies at coarse (Luo et al., 2022; Gorti et al., 2022), fine (Liu et al., 2022; Fang et al., 2023), and coarse-to-fine (Yang et al., 2024; Tian et al., 2024) feature granularities. For example, UCoFiA (Wang et al., 2023) proposes a unified coarse-to-fine alignment model that effectively improves retrieval performance from a comprehensive perspective. However, considering that the DPPM in Section 3.2 has already achieved a well-balanced division of cross-modal feature granularity and semantic roles, directly adopting existing interaction strategies may not only undermine the structural advantages between pathways but also introduce granularity mismatch and semantic interference. Inspired by the pathway partitioning, we prioritize interaction at the macro level before performing deeper micro-level interactions, jointly forming the **D**ual-**V**iew **I**nteraction **M**odule (**DVIM**).

Macro-View Interaction. Since pathway partitioning relies on the global correlation between sentence and frame features as defined in Equations 3 and 4, we initially compute the similarity score using Equation 1. However, the simplicity of Equation 1 introduces certain limitations, including feature information loss and difficulties in handling adaptive granularity (Wu et al., 2023; Wang et al., 2024). To overcome these issues, we propose a novel feature interaction method **W**eighted-**M**ax (**WM**), which retains the advantages of Equation 1 across different granularity levels while mitigating the complexity associated with the uniform granularity model used in UCoFiA (Wang et al., 2023).

Specifically, given sentence features $T_w \in \mathbb{R}^{N_s \times D}$ and frame features $V_f \in \mathbb{R}^{N_f \times D}$, the interaction matrix is defined as $A = [a_{i,j}] \in \mathbb{R}^{N_s \times N_f}$, where $a_{i,j} = \frac{s_i \cdot f_j}{|s_i| \cdot |f_j|}$ represents the alignment score between the i^{th} sentence feature and the j^{th} frame feature. For the i^{th} sentence feature, we compute its maximum interaction score $\max_j a_{ij}$, and aggregate all sentence features using weights $\theta_s = [\theta_s^1, \theta_s^2, ..., \theta_s^{N_s}] = \text{Softmax}(\text{MLP}_s(T_s))$ to obtain the sentence-to-frame similarity. Similarly, we can also obtain the frame-to-sentence similarity, and the overall sentence–frame similarity score S_{T_w, V_p} is defined as:

$$S_{T_s, V_f} = \frac{1}{2}\left(\sum_{i=1}^{N_s} \theta_s^i \max_j a_{i,j} + \sum_{j=1}^{N_f} \theta_f^j \max_i a_{i,j} \right). \tag{5}$$

If we set $N_s^+ = 1$ and $N_s^- = 1$, the Macro-View similarity scores $S_{T_s^+, V_f^+}$ and $S_{T_s^-, V_f^-}$ in the Spot-Path and Recover-Path can be simplified as:

$$S_{T_s^+, V_f^+} = \sum_{i=1}^{N_f^+} \theta_f^i a_i, \quad S_{T_s^-, V_f^-} = \sum_{i=1}^{N_f^-} \theta_f^i a_i. \tag{6}$$

where $a_i = \frac{s \cdot f_i}{|s| \cdot |f_i|}$ and $\theta_f = [\theta_f^1, \theta_f^2, ..., \theta_f^{N_f}] = \text{Softmax}(\text{MLP}_f(V_f))$. Compared with Equation 1, Equation 6 produces more reliable and accurate matching results thanks to feature partitioning and additional textual supplementation. Its effectiveness is confirmed by our ablation study.

Micro-View Interaction. The features involved in macro-view interaction exhibit a high degree of simplicity in both quantity and representation. However, high-volume and redundant patch features V_p cannot participate in interaction with text in the same way. For example, if we set $N_f = 12$, the number of patch features is $N_f \times \frac{H \times W \times C}{P \times P \times C} = 12 \times \frac{224 \times 224 \times 3}{32 \times 32 \times 3} = 588$ for *ViT-B/32* (2352 for *ViT-B/16*). Once a larger N_f is set, a greater number of patches need to be processed. UCoFiA (Wang et al., 2023) reduces the number of patch features by selecting a few key patches in each frame based on spatial attention weights based on TS2Net (Liu et al., 2022). TempMe (Shen et al., 2024) adopts the Temporal Token Merging method to merge redundant temporal tokens in adjacent video clips step by step, simplifying model complexity while reducing the number of patches to be processed. Although these methods attempt to minimize the number of patches required for interaction, using too few or too many patches may exacerbate information asymmetry, thereby negatively affecting interaction effectiveness.

To address the issue of excessive patch features and highlight the contribution of key visual entities, we propose a **P**atch **F**eatures **M**erging **M**odule (**PFMM**) for processing visual micro-level cues. First, given the Spot-Path patch features $V_p^+ = N_f^+ \times [p_1, p_2, ..., p_{N_p}] \in \mathbb{R}^{N_f^+ \times N_p \times D}$, we aim to capture key visual entities (e.g., humans, animals, etc.), which should not be limited to a single patch feature p_i with a restricted receptive field, but instead aggregate multiple adjacent patches $\{p_i, ..., p_j\}$ into a coherent unit. Building on this idea, we utilize a variant of the k-nearest neighbor-based density peaks clustering algorithm (DPC-KNN) (Du et al., 2016), commonly used in feature aggregation (Zeng et al., 2022; Jin et al., 2023a), to merge adjacent patches. Specifically, given patch features V_p^+, we compute the local density ρ_i of each patch p_i according to its k-nearest neighbors $\rho_i = \exp(-\frac{1}{k}\sum ||p_i - p_j||_2)$, where $p_j \in \text{KNN}(p_i)$ denotes the k-neatest neighbors of the patch p_i. Then, we compute the distance indicator δ_i of each patch:

$$\delta_i = \begin{cases} \min_i ||p_i - p_j||_2, & \rho_i < \rho_j, \\ \max_i ||p_i - p_j||_2, & \rho_i \geq \rho_j. \end{cases} \tag{7}$$

Intuitively, the patch p_i with a larger local density ρ_i and distance indicator δ_i is more likely to become a cluster center. We determine a cluster center by selecting the patches with the highest scores $\rho_i \times \delta_i$, and then merge the neighboring patches. The merged patch p_i^* is fed into a transformer block as query Q, the original patch p_i is used as key K and value V, and the attention is computed as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{D}}\right)V. \tag{8}$$

By introducing the feature clustering and attention mechanism, PFMM not only reduces the number of patch features but also focuses on key features and spatial relationships. Final, we iteratively apply the PFMM to merge patch features in order to reduce their number. Similar to the operations in Equation 5, we can compute the similarity between word features and aggregated patch features:

$$\begin{aligned} S_{T_w^+, V_p^+} = \frac{1}{2}\Bigg(& \sum_{i=1}^{N_w^+} \theta_w^i \max_j \max_k a_{i,j,k} \\ & + \sum_{j=1}^{N_f^+} \sum_{k=1}^{N_p^+} \theta_{f,p}^{j,k} \max_i a_{i,j,k} \Bigg). \end{aligned} \tag{9}$$

In addition, the micro-view interactions in the Recover-Path are processed in the same manner.

3.4 Training and Sampling

Training. Based on Equations 6 and 9, we can obtain the macro- and micro-view feature interaction losses in the Spot-Path:

$$\mathcal{L}_+ = \mathcal{L}_{T_s^+, V_f^+} + \mathcal{L}_{T_w^+, V_p^+}. \tag{10}$$

Similarly, the interaction loss $\mathcal{L}_-$ in the Recover-Path is obtained. However, in the Recover-Path, directly using the augmented textual descriptions as retrieval queries would violate the isolation between training and testing, thereby leading to potential data leakage. To further improve the sampling effectiveness, we additionally consider the interactions between the original text and video features:

$$\mathcal{L} = \mathcal{L}_{T_s, V_f} + \mathcal{L}_{T_w, V_p}, \mathcal{L}_{KL} = \sum KL\left(S_m || S_{m^\pm}\right), \tag{11}$$

where $m \in \{(T_s, V_f), (T_w, V_p)\}$, and KL denotes the Kullback-Leibler (Jin et al., 2023a) divergence loss. Therefore, the total loss can be formulated as:

$$\mathcal{L}_{total} = \mathcal{L} + \alpha(\mathcal{L}_+ + \mathcal{L}_-) + \beta\mathcal{L}_{KL}. \tag{12}$$

Sampling. After training, we compute the similarity matrices S_{T_s, V_f} and S_{T_w, V_p}, which are then aggregated into a final similarity matrix for calculating the corresponding retrieval metrics.

Methods	MSRVTT (Text-to-Video)					MSRVTT (Video-to-Text)				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
Clip4clip (Luo et al., 2022)	44.5	71.4	81.6	2.0	15.3	42.7	70.9	80.6	2.0	11.6
X-Pool (Gorti et al., 2022)	46.9	72.8	82.2	2.0	14.3	44.4	73.3	84.0	2.0	9.0
HBI (Jin et al., 2023a)	48.6	74.6	83.4	2.0	12.0	46.8	74.3	84.3	2.0	8.9
UATVR (Fang et al., 2023)	47.5	73.9	83.5	2.0	12.3	46.9	73.8	83.8	2.0	8.6
Cap4Video (Wu et al., 2023)	49.3	74.3	83.8	2.0	12.0	47.1	73.7	84.3	2.0	8.7
UCoFiA (Wang et al., 2023)	49.4	72.1	-	-	12.9	47.1	74.3	-	-	-
CLIP-ViP (Xue et al., 2022)	50.1	74.8	84.6	1.0	-	-	-	-	-	-
T-Mass (Wang et al., 2024)	50.2	75.3	85.1	1.0	11.9	47.7	**78.0**	86.3	2.0	8.0
DiscoVLA (Shen et al., 2025)	47.0	73.0	82.8	-	14.1	47.7	73.6	83.6	-	10.0
BiHSSP (Liu et al., 2025b)	48.1	74.0	84.1	2.0	12.1	48.0	74.1	83.5	2.0	9.0
DPDV	**50.5**	**76.1**	**86.2**	**1.0**	**11.3**	**48.3**	76.2	**86.7**	**2.0**	**7.8**

Table 1: Text-to-video and video-to-text retrieval performance on the MSRVTT.

Methods	DiDeMo (Text-to-Video)					LSMDC (Text-to-Video)				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
X-Pool (Gorti et al., 2022)	44.6	73.2	82.0	2.0	15.4	25.2	43.7	53.5	8.0	53.2
CLIP-ViP (Xue et al., 2022)	48.6	77.1	84.4	2.0	-	25.6	45.3	54.4	8.0	-
T-Mass (Wang et al., 2024)	50.9	77.2	85.3	1.0	12.1	28.9	48.2	57.6	6.0	43.3
DPDV	**51.0**	**77.3**	**85.7**	**1.0**	**11.6**	**29.3**	**49.6**	**58.4**	**6.0**	**40.3**
Methods	ActivityNet (Text-to-Video)					Charades (Text-to-Video)				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
ClipBERT (Lei et al., 2021)	21.3	49.0	63.5	6.0	-	6.7	17.3	25.2	32.0	149.7
Clip4clip (Luo et al., 2022)	40.5	72.4	83.6	2.0	7.5	9.9	27.1	36.8	21.0	85.4
T-Mass (Wang et al., 2024)	-	-	-	-	-	14.2	36.2	48.3	12.0	54.8
DPDV	**47.0**	**76.2**	**86.4**	**2.0**	**6.3**	**19.3**	**42.2**	**53.5**	**8.0**	**49.7**

Table 2: Text-to-video retrieval performance on the DiDeMo, LSMDC, ActivityNet and Charades.

4 Experiments

4.1 Experimental Settings

We adopt five benchmark datasets for the evaluation, including MSRVTT (Xu et al., 2016), DiDeMo (Anne Hendricks et al., 2017), LSMDC (Rohrbach et al., 2015), ActivityNet (Krishna et al., 2017) and Charades (Sigurdsson et al., 2016). We evaluate retrieval performance using Recall at rank K (R@K, K=1,5,10), Median Rank (MdR), and Mean Rank (MnR). Higher R@K values, together with lower MdR and MnR values, indicate better retrieval performance. The model backbone is initialized from pre-trained CLIP-ViT-B/32. More experimental settings are provided in the Appendix.

4.2 Comparison with State-of-the-art

We compare the performance of DPDV with recent state-of-the-art video retrieval methods in Tables 1 and 2. DPDV consistently achieves leading retrieval performance across all five datasets. Specifically, compared to the unified coarse-to-fine interaction model UCoFiA (Wang et al., 2023) and the explicit text augmentation method T-Mass (Wang et al., 2024), DPDV achieves improvements of **1.1** and **0.3** in R@1, respectively on the MSRVTT, demonstrating the effectiveness of our method for precise feature interactions. Similarly, DPDV maintains a leading performance on long-video DiDeMo and ActivityNet, as well as short-text LSMDC and Charades. Compared to the Transformer-based feature selection method X-Pool (Gorti et al., 2022), DPDV achieves a **6.4** improvement in R@1 on DiDeMo, highlighting the effectiveness of proactive pathway partitioning rather than relying on the model to automatically filter relevant frames.

4.3 Ablation Study

Ablation Study of DPPM. Table 3 presents the impact of the Dual-Pathway Partitioning Module on retrieval performance under feature interactions as in Equation 1. R1 indicates results without path-

way partitioning, while R2 and R3 correspond to macro- and micro-level interactions under the Spot-Path and Recover-Path, respectively. We observe that R4 achieves the highest performance of **47.8** when combining the SP and RP, confirming that pathway partitioning is an effective approach for feature enhancement and reduces the complexity of subsequent feature interactions.

Rx	DPPM_{SP}	DPPM_{RP}	R@1↑	R@10↑	MnR↓
R1			43.4	81.1	16.4
R2	✓		45.6	82.7	14.3
R3		✓	46.3	83.2	13.8
R4	✓	✓	**47.8**	**84.5**	**12.3**

Table 3: Ablation study of the Dual-Pathway Partitioning Module (DPPM). SP and RP denote the Spot-Path and Recover-Path, respectively, with feature interactions consistently computed according to Equation 1.

Ablation Study of DVIM. Table 4 presents the impact of the Dual-View Interaction Module on retrieval performance under DPPM. To highlight the superiority of the Weighted-Max interaction in Equation 5 over Mean-Pooling in Equation 1, R1 is set to R4 from Table 3. R2 and R3 correspond to macro- (Ma) and micro-level (Mi) feature interactions under the pathway partitioning, respectively. We observe that R4 achieves the highest performance of **50.5**, resulting from the comprehensive granularity of feature interactions.

Rx	DVIM_{Ma}	DVIM_{Mi}	R@1↑	R@10↑	MnR↓
R1			47.8	84.5	12.3
R2	✓		48.4	84.7	12.0
R3		✓	49.6	85.4	11.6
R4	✓	✓	**50.5**	**86.2**	**11.3**

Table 4: Ablation study of the Dual-View Interaction Module (DVIM). Ma and Mi denote macro- and micro-level feature interactions, respectively.

Ablation Study of PFMM. The Patch Feature Merge Module (PFMM) reduces the number of processed patches by determining cluster centers, demonstrating superior performance compared to average pooling (Luo et al., 2022) and the weighted selection strategy in UCoFiA (Wang et al., 2023). As shown in Table 5, repeatedly applying PFMM progressively reduces the number of patches, with the best performance achieved at a setting of 0.5.

Methods	N_p	R@1↑	MdR↓
Clip4clip	49→1	45.4	13.8
UCoFiA	49→4	49.4	12.9
PFMM(0.75)	49→37→28→21	50.2	11.5
PFMM(0.50)	49→25→13→07	**50.5**	**11.3**
PFMM(0.25)	49→04→01→01	49.6	12.1

Table 5: Ablation study of PFMM. In PFMM(x), x denotes the proportion of retained cluster centers.

Ablation Study of the Interaction Module. Table 6 presents the impact of various interaction methods on retrieval performance across different granularity levels. Although both models benefit from DPPM, the coarse-grained sentence-frame interaction DVIM_{Ma} and the fine-grained word-patch interaction DVIM_{Mi} perform better, further highlighting the advantage of the Patch Feature Merge Module in aggregating fine-grained features.

ITEM	Methods	R@1↑	R@10↑	MnR↓
$T_s \leftrightarrow V_f$	Mean-Pooling	45.3	81.8	15.1
	UCoFiA	47.1	82.6	14.1
	DVIM_{Ma}	**48.4**	**84.7**	**12.0**
$T_w \leftrightarrow V_p$	Mean-Pooling	47.1	82.4	14.2
	UCoFiA	48.2	83.3	13.2
	DVIM_{Mi}	**49.6**	**85.4**	**11.6**

Table 6: Ablation comparison with the unified coarse-to-fine interaction model UCoFiA (Wang et al., 2023) under different granularities of feature alignment.

Ablation Study of the Sampling Strategy. Table 7 reports the impact of the loss terms $\mathcal{L}$ and $\mathcal{L}_{KL}$ on retrieval performance. Combining both losses R4 achieves the best results across all metrics, highlighting their complementary contributions.

Rx	$\mathcal{L}$	$\mathcal{L}_{KL}$	R@1↑	R@10↑	MnR↓
R1			46.4	82.1	15.6
R2	✓		48.3	83.7	14.2
R3		✓	48.1	84.5	13.8
R4	✓	✓	**50.5**	**86.2**	**11.3**

Table 7: Ablation study of different sampling strategies.

4.4 Visualization

Visualization of Recover-Path Benefits. To illustrate the transition from information asymmetry to symmetry, we visualize the similarity changes on

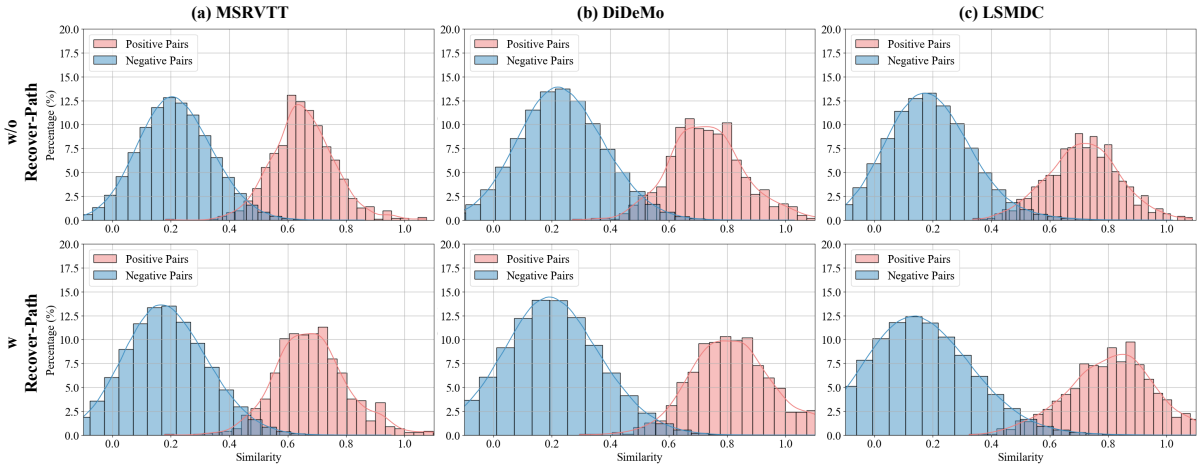

Figure 3: Visualization of similarity distributions under w/o Recover-Path and w/ Recover-Path on the MSRVTT, DiDeMo, and LSMDC. The x-axis represents similarity scores, and the y-axis indicates the proportion. From top to bottom, the similarity of positive pairs gradually increases, while that of negative pairs gradually decreases, indicating a reduction in information asymmetry and a more compact matching relationship.

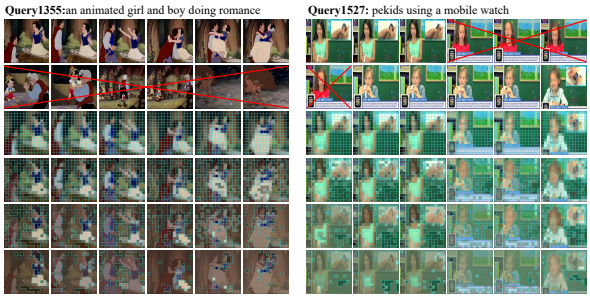

Figure 4: Visualization of frame selection and patch compression process within the Spot-Path. Red × indicates Recover-Path process similar to Spot-Path.

benchmark datasets in Figure 3. "Positive Pairs" and "Negative Pairs" denote the similarities of relevant and irrelevant pairs, respectively, where higher values are preferred for positive pairs and lower values for negative pairs. Specifically, the MSRVTT test set contains 1000 query texts and 1000 candidate videos, resulting in a 1000×1000 similarity matrix. The 1000 values along the main diagonal correspond to positive pair similarities, while the off-diagonal values correspond to negative pairs. The results show that the similarity of positive pairs shifts to the right, while that of negative pairs shifts to the left, indicating that the Recover-Path effectively enhances the discriminability of correct matches. Similarly, DiDeMo and LSMDC are processed in the same manner.

Visualization of Spot-Path Process. Figure 3 visualizes the frame partitioning and patch merging process of the Spot-Path during training. Red × marks indicate discarded frames, while the retained frames are used for patch feature merging. The last three rows show the visual results of three successive patch merging stages.

4.5 Further Details

The supplementary material provides additional details, including A. Experimental Settings, B. Task Performance, C. Ablation Study, D. Parameter Analysis and E. Result Visualization.

5 Conclusion

To address the challenge of asymmetric information interaction between sparse queries and complex visual cues, we propose a novel video retrieval framework, termed DPDV, which consists of the Dual-Pathway Partitioning Module for constructing features at an appropriate granularity and the Dual-View Interaction Module for effective feature interaction. For DPPM, we simulate a human-like macro-level cognitive perspective by partitioning visual features into two categories based on their relevance to the text query and supplementing less relevant features with additional textual information. For DVIM, we simulate a human alignment strategy from macro to micro levels, enabling the model to focus on local visual features while comprehensively modeling fine-grained interactions. Experiments on five benchmark datasets demonstrate that the DPDV model achieves state-of-the-art performance, validating the effectiveness of our approach. We hope that this work will provide inspiration to the video retrieval community.

Limitations

Despite its promising performance, DPDV presents several limitations:

❶ Granularity Sensitivity & PFMM Constraints: Model efficacy heavily relies on balancing feature interaction granularity (macro vs. micro). Furthermore, the spatial correlation-based Patch Feature Merge Module (PFMM) struggles with highly sparse or anomalous frames.

❷ Generalization of Recover-Path: The text supplementation strategy may generalize poorly to unseen domains or varying video lengths. For long videos, missing descriptions limit its benefits, often requiring dataset-specific tuning.

❸ Computational Overhead: Dual-path interactions and multi-view fusion introduce extra overhead. Using high-resolution backbones (e.g., CLIP-ViT-B/16) significantly increases memory usage and inference time, limiting real-time scalability.

Future work will explore dynamic granularity adjustment, improved PFMM robustness for irregular frames, and extending Recover-Path to diverse and mixed-length video tasks.

Acknowledgements

This work was supported by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang under (Grant No. 2025C02110), Public Welfare Research Program of Ningbo under (Grant No. 2024S062), and Yongjiang Talent Project of Ningbo under (Grant No. 2024A-161-G).

References

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint*. Available at https://arxiv.org/abs/2502.13923.

Mingjing Du, Shifei Ding, and Hongjie Jia. 2016. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowledge-Based Systems*, 99:135–145.

Bo Fang, Wenhao Wu, Chang Liu, Yu Zhou, Yuxin Song, Weiping Wang, Xiangbo Shu, Xiangyang Ji, and Jingdong Wang. 2023. Uatvr: Uncertainty-adaptive text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13723–13733.

Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multi-modal transformer for video retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 214–229. Springer.

Xiyuan Gao, Bing Cao, Pengfei Zhu, Nannan Wang, and Qinghua Hu. 2025. Asymmetric reinforcing against multi-modal representation bias. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(16):16754–16762.

Ziyu Gong, Chengcheng Mai, and Yihua Huang. 2024. Ascl: An asymmetry-sensitive contrastive learning method for image-text retrieval with cross-modal fusion. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. 2022. X-pool: Cross-modal language-video attention for text-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5006–5015.

Peng Jin, Jinfa Huang, Fenglin Liu, Xian Wu, Shen Ge, Guoli Song, David Clifton, and Jie Chen. 2022. Expectation-maximization contrastive learning for compact video-and-language representations. *Advances in neural information processing systems*, 35:30291–30306.

Peng Jin, Jinfa Huang, Pengfei Xiong, Shangxuan Tian, Chang Liu, Xiangyang Ji, Li Yuan, and Jie Chen. 2023a. Video-text as game players: Hierarchical banzhaf interaction for cross-modal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2472–2482.

Peng Jin, Hao Li, Zesen Cheng, Jinfa Huang, Zhennan Wang, Li Yuan, Chang Liu, and Jie Chen. 2023b. Text-video retrieval with disentangled conceptualization and set-to-set alignment. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 938–946. International Joint Conferences on Artificial Intelligence Organization.

Peng Jin, Hao Li, Zesen Cheng, Kehan Li, Xiangyang Ji, Chang Liu, Li Yuan, and Jie Chen. 2023c. Diffusionret: Generative text-video retrieval with diffusion model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2470–2481.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13700–13710.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7331–7341.

Hao Li, Peng Jin, Zesen Cheng, Songyang Zhang, Kai Chen, Zhennan Wang, Chang Liu, and Jie Chen. 2023. Tg-vqa: ternary game of video question answering. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 1044–1052.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26689–26699.

Yan-Bo Lin, Jie Lei, Mohit Bansal, and Gedas Bertasius. 2022. Eclipse: Efficient long-range video retrieval using sight and sound. In *European Conference on Computer Vision*, pages 413–430. Springer.

Xin Liu, Shibai Yin, Jun Wang, Jiaxin Zhu, Xingyang Wang, and Yee-Hong Yang. 2025a. Duq: Dual uncertainty quantification for text-video retrieval. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 5779–5787. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. 2019. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*.

Yang Liu, Shudong Huang, Deng Xiong, and Jiancheng Lv. 2025b. Learning dynamic similarity by bidirectional hierarchical sliding semantic probe for efficient text video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5667–5675.

Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. 2022. Ts2-net: Token shift and selection transformer for text-video retrieval. In *European conference on computer vision*, pages 319–335. Springer.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304.

Zongyang Ma, Ziqi Zhang, Yuxin Chen, Zhongang Qi, Chunfeng Yuan, Bing Li, Yingmin Luo, Xu Li, Xiaojuan Qi, Ying Shan, and 1 others. 2024. Ea-vtr: Event-aware video-text retrieval. In *European Conference on Computer Vision*, pages 76–94. Springer.

OpenAI. 2023. Gpt-4 technical report.

Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. 2020. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*.

AJ Piergiovanni, Kairo Morton, Weicheng Kuo, Michael S Ryoo, and Anelia Angelova. 2022. Video question answering with iterative video-text co-tokenization. In *European Conference on Computer Vision*, pages 76–94. Springer.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212.

Leqi Shen, Guoqiang Gong, Tianxiang Hao, Tao He, Yifeng Zhang, Pengzhang Liu, Sicheng Zhao, Jungong Han, and Guiguang Ding. 2025. Discovla: Discrepancy reduction in vision, language, and alignment for parameter-efficient video-text retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19702–19712.

Leqi Shen, Tianxiang Hao, Sicheng Zhao, Yifeng Zhang, Pengzhang Liu, Yongjun Bao, and Guiguang Ding. 2024. Tempme: Video temporal token merging for efficient text-video retrieval. *arXiv preprint arXiv:2409.01156*.

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer.

Kaibin Tian, Yanhua Cheng, Yi Liu, Xinglin Hou, Quan Chen, and Han Li. 2024. Towards efficient and effective text-to-video retrieval with coarse-to-fine visual representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5207–5214.

Atousa Torabi, Niket Tandon, and Leon Sigal. 2016. Learning language-visual embedding for movie understanding with natural-language. *arXiv:1609.08124*.

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

Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2023. Unified coarse-to-fine alignment for video-text retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2816–2827.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*.

Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. 2023. Cap4video: What can auxiliary captions do for text-video retrieval? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10704–10713.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. 2022. Clip-vip: Adapting pre-trained image-text model to video-language alignment. In *The Eleventh International Conference on Learning Representations*.

Xiangpeng Yang, Linchao Zhu, Xiaohan Wang, and Yi Yang. 2024. Dgl: Dynamic global-local prompt tuning for text-video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6540–6548.

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

Jialin Yuan, Ye Yu, Gaurav Mittal, Matthew Hall, Sandra Sajeev, and Mei Chen. 2024. Rethinking multimodal content moderation from an asymmetric angle with mixed-modality. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8532–8542.

Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. 2022. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11101–11111.

A Experimental Settings

Datasets. We adopt five benchmark datasets for the evaluation, including MSRVTT (Xu et al., 2016), DiDeMo (Anne Hendricks et al., 2017), LSMDC (Rohrbach et al., 2015), ActivityNet (Krishna et al., 2017) and Charades (Sigurdsson et al., 2016). **(1) MSRVTT** consists of 10,000 YouTube videos, each paired with 20 captions. We follow the training protocol in (Yu et al., 2018) and evaluate our model DPDV on both text-to-video and video-to-text retrieval tasks using the 1K-A test split. **(2) DiDeMo** contains 10,464 video clips and 40,543 captions. We concatenate the descriptions of individual video segments to construct a "video-paragraph" for retrieval. We use the training and testing protocols from (Gabeur et al., 2020). **(3) LSMDC** includes 118,081 video clips from 202 movies. The duration of videos in the LSMDC dataset is short. We use the split from (Torabi et al., 2016), with 1,000 videos reserved for testing. **(4) ActivityNet** contains densely annotated temporal segments for 20,000 YouTube videos. Following (Jin et al., 2023a), we report results on the "val1" split, using 10,009 videos for training and 4,917 for testing. **(5) Charades** consists of 9,848 video clips, where each corresponds to a text description. We adopt the split protocol from (Lin et al., 2022).

Metrics. We evaluate retrieval performance using Recall at rank K (R@K, K=1,5,10, higher is better), Median Rank (MdR, lower is better), and Mean Rank (MnR, lower is better). R@K measures the percentage of test samples whose correct results appear in the top-K retrieved items. MdR reports the median rank of the ground-truth results, and MnR reports their mean rank.

Implementation Details. Following previous methods (Luo et al., 2022; Gorti et al., 2022; Liu et al., 2025a), we use CLIP (Radford et al., 2021) as the backbone model for both text and video feature extraction. The DPDV model is trained for 5 epochs with a batch size of 32, and the feature dimension D is set to 512. For the MSRVTT (Xu et al., 2016), LSMDC (Rohrbach et al., 2015), and Charades (Sigurdsson et al., 2016) datasets, we set the frame length to $N_f = 12$ and the word length to $N_w = 32$. For the long video DiDeMo (Anne Hendricks et al., 2017) and ActivityNet (Krishna et al., 2017) datasets, we set the frame length to $N_f = 64$ and the word length to $N_w = 64$. We uniformly sample N_f frames from each video clip and resize them to 224×224

Methods	MSRVTT (Text-to-Video)					MSRVTT (Video-to-Text)				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
CLIP-ViT-B/16										
X-Pool (Gorti et al., 2022)	48.2	73.7	82.6	2.0	12.7	46.4	73.9	84.1	2.0	8.4
UATVR (Fang et al., 2023)	50.8	76.3	85.5	1.0	12.4	48.1	76.3	85.4	2.0	8.0
Cap4Video (Wu et al., 2023)	51.4	75.7	83.9	1.0	12.4	49.0	75.2	85.0	2.0	8.0
T-Mass (Wang et al., 2024)	52.7	77.1	85.6	1.0	10.5	50.9	**80.2**	**88.0**	1.0	7.4
DiscoVLA (Shen et al., 2025)	50.5	75.6	83.8	-	12.1	49.2	76.0	84.7	-	8.6
BiHSSP (Liu et al., 2025b)	50.8	75.9	84.4	1.0	11.0	50.3	75.5	84.5	1.5	7.8
DPDV	**53.2**	**78.9**	**87.2**	**1.0**	**10.3**	**51.2**	78.9	86.5	**1.0**	**7.4**

Table 8: Text-to-video and video-to-text retrieval performance on the MSRVTT (Xu et al., 2016).

Methods	DiDeMo (Video-to-text)					LSMDC (Video-to-text)				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
Clip4clip (Luo et al., 2022)	41.4	68.2	79.1	2.0	12.4	20.8	39.0	48.6	12.0	54.2
EMCL-Net (Jin et al., 2022)	45.7	74.3	82.7	2.0	10.9	22.2	40.6	49.2	12.0	-
DiffusionRet (Jin et al., 2023c)	46.2	74.3	82.2	2.0	10.7	23.0	43.5	51.5	9.0	40.2
T-Mass (Wang et al., 2024)	49.1	76.4	85.9	2.0	**8.0**	**26.0**	48.4	**57.5**	6.0	**37.8**
DPDV	**50.4**	**77.8**	**86.5**	**1.0**	8.6	25.9	**48.5**	57.4	**6.0**	38.1
Methods	ActivityNet (Video-to-text)					Charades (Video-to-text)				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
Clip4clip (Luo et al., 2022)	41.4	73.7	85.3	2.0	6.7	-	-	-	-	-
DiffusionRet (Jin et al., 2023c)	43.8	75.3	86.7	2.0	6.3	-	-	-	-	-
T-Mass (Wang et al., 2024)	-	-	-	-	-	13.2	37.3	48.5	11.0	56.1
DPDV	**44.5**	**77.3**	**87.4**	**1.0**	**5.4**	**17.5**	**40.6**	**49.9**	**10.0**	**45.7**

Table 9: Video-to-text retrieval performance on the DiDeMo (Anne Hendricks et al., 2017), LSMDC (Rohrbach et al., 2015), ActivityNet (Krishna et al., 2017) and Charades (Sigurdsson et al., 2016).

pixels. Accordingly, the number of patches per frame is set to $N_p = \frac{H \times W \times C}{P \times P \times C} = \frac{224 \times 224 \times 3}{32 \times 32 \times 3} = 49$. We adopt vision-language models (VLMs), including Qwen2.5-VL (Bai et al., 2025), Next-GPT (Wu et al., 2024), Chat-univi (Jin et al., 2024), and VILA (Lin et al., 2024), to generate $N_s = 6$ text descriptions for each video. The initial learning rate is set to $1e{-}5$ for both the text and video encoders. The hyperparameters α and β are set to 0.5 and 0.1, respectively, while the frame feature partitioning and patch feature aggregation ratios are both set to 0.5. We train our model on four NVIDIA RTX 4090 24GB GPUs, and the training process takes approximately 10 hours on the MSRVTT.

B Task Performance

Retrieval Performance under CLIP-ViT-B/16. Table 8 in the main text reports the retrieval performance of DPDV using CLIP-ViT-B/32. To further highlight the advantages of DPDV on MSRVTT (Xu et al., 2016), we present its retrieval results using CLIP-ViT-B/16 in Table 1. Since CLIP-ViT-B/16 processes smaller patches ($P \times P = 16 \times 16$), the model must handle a larger number of visual tokens with more dispersed semantics, which exacerbates information redundancy and increases the difficulty of cross-modal alignment during retrieval. Notably, compared with single text augmentation methods such as Cap4Video (Wu et al., 2023) and T-Mass (Wang et al., 2024), DVDP achieves superior performance (53.2 of R@1) due to its targeted textual enhancement via the Recover-Path.

Video-to-text Retrieval Performance. In Table 9, we report the video-to-text retrieval performance on the DiDeMo (Anne Hendricks et al., 2017), LSMDC (Rohrbach et al., 2015), ActivityNet (Krishna et al., 2017) and Charades (Sigurdsson et al., 2016) datasets. The video-to-text retrieval task involves finding the matching text given visual features, which is the opposite of the text-to-video retrieval task. Table 9 shows that DPDV consistently improves retrieval performance across these four datasets. For example, on the long-video DiDeMo dataset, DPDV achieves a **1.3** improvement in R@1

Methods	MSRVTT>MSRVTT			MSRVTT>DiDeMo			MSRVTT>LSMDC		
	R@1↑	R@Sum↑	MdR↓	R@1↑	R@Sum↑	MdR↓	R@1↑	R@Sum↑	MdR↓
Clip4clip (Luo et al., 2022)	43.8	195.8	2.0	31.8	154.9	4.0	15.3	87.1	21.0
X-Pool (Gorti et al., 2022)	46.9	201.9	2.0	35.3	168.5	3.0	16.4	93.5	18.0
DiffusionRet (Jin et al., 2023c)	49.0	206.9	2.0	33.2	160.9	3.0	17.1	90.5	21.0
T-Mass (Wang et al., 2024)	50.2	210.6	1.0	39.5	178.2	2.0	19.7	102.5	14.0
DPDV	**50.5**	**212.8**	**1.0**	**41.2**	**183.6**	**2.0**	**20.2**	**105.0**	**12.0**

Table 10: Text-to-video cross-domain generalization performance. $X > Y$ indicates that X is used as the training data and Y as the testing data. R@Sum = R@1 + R@5 + R@10, with higher values indicating better performance.

Methods	GFLOPs ↓	Inference Time(s) ↓	Training Time(h) ↓	R@1 ↑
Clip4clip (Luo et al., 2022)	**53.0**	**52.8**	**12.4**	44.5
X-Pool (Gorti et al., 2022)	68.7	64.7	13.3	46.9
UCoFiA (Wang et al., 2023)	88.5	78.6	13.6	49.4
DPPM_{RP}	83.2	73.2	14.0	48.5
DVIM_{Mi}	76.8	75.7	14.1	49.6
DPDV	86.8	76.5	14.2	**50.5**

Table 11: Comparison of computational efficiency between the DPDV module and existing methods (Clip4clip (Luo et al., 2022), X-Pool (Gorti et al., 2022) and UCoFiA (Wang et al., 2023)) on MSRVTT (Xu et al., 2016).

compared to T-Mass (Wang et al., 2024), resulting from the fine-grained feature interactions.

Cross-domain Generalization Performance. Cross-domain generalization performance measures the ability of a model to perform on data from unseen domains. In Table 10, we use MSRVTT (Xu et al., 2016) as the source domain for training and DiDeMo (Anne Hendricks et al., 2017) and LSMDC (Rohrbach et al., 2015) as the target domains for testing to evaluate the generalization performance of DVDP. Compared with recent state-of-the-art methods, DVDP demonstrates consistent performance advantages.

C Ablation Study

Ablation Study on Module Efficiency. In the Table 11, we compare the core modules of DPDV, DPPM_{RP} and DVIM_{Mi}, with several existing baselines in terms of computational cost and performance. Since the text encoder needs to process more than $N_s > 1$ sentences, the GFLOPs of DPPM_{RP} increase significantly. Meanwhile, DVIM_{Mi} introduces additional computational overhead due to token merging. In terms of the trade-off between efficiency and performance, DPDV significantly outperforms existing methods, demonstrating state-of-the-art performance.

Ablation Study of Different VLMs. Table 12 compares the text supplementation effects of the Recover-Path under different vision-language models, validating the generality and robustness of the proposed strategy across diverse model settings.

VLMs	R@1↑	MdR↓
Next-GPT (Wu et al., 2024)	50.1	11.4
Chat-univi (Jin et al., 2024)	49.6	11.8
VILA (Lin et al., 2024)	50.5	**11.3**
Qwen2.5 (Bai et al., 2025)	**50.8**	11.6

Table 12: Ablation study of different text augmentation VLMs on the MSRVTT (Xu et al., 2016).

Ablation of the Feature Extractor. In the Table 13, we compare two CLIP feature extractors, CLIP-ViT-B/32 and CLIP-ViT-B/16, in terms of retrieval efficiency and performance. The results show that, regardless of the retrieval model, CLIP-ViT-B/16 incurs substantially higher computational cost than CLIP-ViT-B/32. Therefore, in real-time or large-scale applications, special attention should be paid to the computational overhead of dense, high-resolution feature extractors.

Ablation Study of KL Divergence. In Equation 11, we use knowledge distillation to jointly optimize the original text–video similarity, aiming to prevent retrieval data leakage and simplify sampling. Data leakage occurs in two common scenarios: ❶ the model has prior access to the text paired with a video and uses it to enhance the video repre-

Methods	Extractor	GFLOPs ↓	Inference Time(s) ↓	R@1 ↑
Clip4clip (Luo et al., 2022)	CLIP-ViT-B/32	**53.0**	**52.8**	44.5
Clip4clip (Luo et al., 2022)	CLIP-ViT-B/16	171.5	142.8	48.2
UCoFiA (Wang et al., 2023)	CLIP-ViT-B/32	88.5	78.6	49.4
UCoFiA (Wang et al., 2023)	CLIP-ViT-B/16	231.7	203.4	51.4
DPDV (Ours)	CLIP-ViT-B/32	86.8	76.5	50.5
DPDV (Ours)	CLIP-ViT-B/16	225.6	182.5	**53.2**

Table 13: Comparison of computational cost and retrieval performance on the MSRVTT (Xu et al., 2016).

sentation; ❷ employing VLMs to generate textual descriptions of videos, which are then directly used as retrieval queries. In Table 14, we compare the performance with and without KL divergence. The setting with KL (*w* KL) achieves superior performance, improving by **4.3%** compared to the setting without KL (*w/o* KL).

Methods	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
✗ KL	48.4	75.5	84.2	2.0	13.2
✓ KL	**50.5**	**76.1**	**86.2**	**1.0**	**11.3**

Table 14: Ablation study of KL Divergence.

D Parameter Analysis

Analysis of the Number of Sampled Frames. In Table 15, we present a comparison of text-to-video retrieval performance under different numbers of sampled frames on the MSRVTT (Xu et al., 2016). Since most MSRVTT videos are around 12 seconds long, the performance improvement is more pronounced when N_f varies from 4 to 8 to 12, compared to the variation from 12 to 16 to 20. For a fair comparison with existing methods (Gorti et al., 2022; Luo et al., 2022), we uniformly set $N_f = 12$.

N_f	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
4	43.2	65.7	74.6	5.0	17.9
8	47.8	74.2	83.1	3.0	15.6
12	50.5	76.1	86.2	**1.0**	11.3
16	51.0	76.7	**86.7**	**1.0**	11.0
20	**51.5**	**76.8**	86.4	**1.0**	**10.7**

Table 15: Ablation study of different sampling frame numbers on the MSRVTT (Xu et al., 2016).

Analysis of the Frame Partitioning. In DPPM, video frames are partitioned based on the similarity between macro-level textual sentences and video frames, where the partition ratio plays a critical role in subsequent interaction alignment. Table 16 reports the results under different partition settings, with a fixed total number of frames $N_f = N_f^+ + N_f^- = 12$. When $(N_f^+, N_f^-) = (6, 6)$, DPDV achieves the best performance, which is consistent with the statistical observations in Figure 1.

(N_f^+, N_f^-)	R@1↑	R@5↑	R@10↑	MnR↓
(4, 8)	48.2	75.7	84.6	12.4
(6, 6)	**50.5**	**76.1**	**86.2**	**11.3**
(8, 4)	49.5	75.4	85.9	11.5

Table 16: Ablation study of different frame partition ratios on the MSRVTT (Xu et al., 2016).

Analysis of the Number of Supplementary Texts. In the Recover-Path, we sample a single description from the augmented text set $T_s = [s_1, s_2, \ldots, s_{N_s^-}] \in \mathbb{R}^{N_s^- \times D}$ as $T_s^- \in \mathbb{R}^{1 \times D}$. When multiple supplementary texts are selected to describe V^-, the interactions between $T_s^- \in \mathbb{R}^{N_s^- \times D}$ and $V_f^- \in \mathbb{R}^{N_f^- \times D}$ can be formulated as:

$$S_{T_s^-, V_f^-} = \frac{1}{2}\left(\sum_{i=1}^{N_s^-} \theta_s^i \max_j a_{i,j} + \sum_{j=1}^{N_f^-} \theta_f^j \max_i a_{i,j}\right), \tag{13}$$

and the interactions between $T_w^- \in \mathbb{R}^{N_s^- \times N_w \times D}$ and $V_p^- \in \mathbb{R}^{N_f^- \times N_p \times D}$ can be formulated as:

$$S_{T_w^-, V_p^-} = \frac{1}{2}\left(\sum_{i=1}^{N_s^-}\sum_{j=1}^{N_w} \theta_{s,w}^{i,j} \max_k \max_l a_{i,j,k,l} + \sum_{k=1}^{N_f^-}\sum_{l=1}^{N_p} \theta_{f,p}^{j,k} \max_i \max_j a_{i,j,k,l}\right). \tag{14}$$

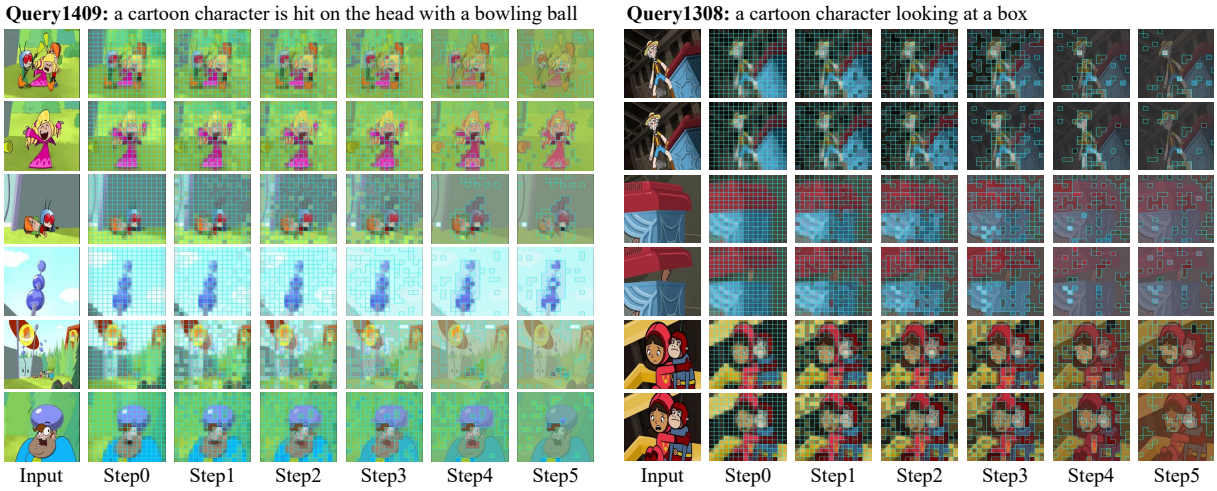

Figure 5: Visualization of the patch feature merge process. Six frames are selected from Video1409 and Video1308 on the MSRVTT, where the number of patches is reduced by half at each step. Note that a larger number of steps does not necessarily lead to better results; intuitively, Step4 is sufficient to achieve the desired outcome.

Since videos in DiDeMo (Anne Hendricks et al., 2017) are relatively long and contain more frames lacking textual descriptions, Table 17 compares the retrieval performance of DPPM with $N_s^- \in \{1, 2, 3\}$ supplementary texts, illustrating the impact of the number of supplementary texts. Cost denotes the training time of the model.

N_s^-	R@1↑	R@10↑	MnR↓	Cost↓
1	51.0	84.7	11.6	**4.51h**
2	52.4	85.3	10.4	4.62h
3	**53.6**	**85.9**	**10.1**	4.72h

Table 17: Ablation study on the number of supplementary texts for DiDeMo (Anne Hendricks et al., 2017).

E Result Visualization

Visualization of Patch Merge. Figure 5 illustrates the five-stage merging process of visual feature patches. Six frames from Video1409 and Video1308 in MSRVTT (Xu et al., 2016) are selected, with a merge ratio of 0.5 at each step. As the stages progress, the number of patch features decreases while the representation increasingly emphasizes key entity features, consistent with human perception of fine-grained alignment. Unlike UCoFiA (Wang et al., 2023), which selects individual patches based on weights at Step0 in Figure 5, PFMM strengthens overall entity features by capturing spatial correlations among patches.

Visualization of Successful Retrieval Results. In Figure 6 (left), we present a successful video retrieval example, where the given query text is "a man extinguishes a fire outside." The retrieved results are ranked from top to bottom in descending order of similarity. Since the target video is ranked first (Rank 1), R@1 reaches 100%, indicating a successful retrieval.

Figure 6: Visualization of successful case Query7060 and failed case Query9243. The retrieval results are displayed in descending order of similarity, where videos with green bounding boxes indicate the correct matches.

Visualization of Failed Retrieval Results. In Figure 6 (right), we present a failed video retrieval example, where the given query text is "they are singing a song and playing a guitar." The retrieved results are ranked according to similarity scores. The target video is ranked third (Rank 3); therefore, R@1 is 0 while R@5 reaches 100%. Notably, most failures are attributed to the simplicity of the textual descriptions and the high similarity among video scenes, rather than limitations of the model itself.