Captioning for Text-Video Retrieval via DualGroup-Direct Preference Optimization

Anonymous ACL submission

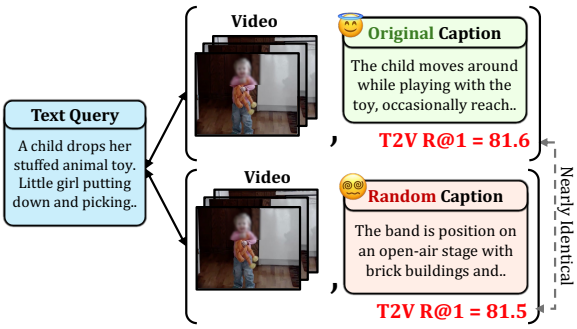
Abstract

In text-video retrieval, auxiliary captions are often used to enhance video understanding, bridging the gap between the modalities. Recently, with the remarkable capabilities in multi-modal understanding, retrieval with MLLMs has emerged as a promising direction. However, we identify two key limitations: (1) retrieval models often fail to effectively leverage the auxiliary captions, neglecting the semantic distinction between the caption (as contextual knowledge) and text queries (as retrieval targets); and (2) auxiliary captions are not typically tailored for retrieval, evaluated with language generation metrics such as BLEU that misalign with retrieval objectives, which require fine-grained discrimination. To address these challenges, we propose CaRe-DPO, a retrieval framework that integrates two key components. First, retrieval role-embeddings are introduced to explicitly differentiate between the roles of heterogeneous textual inputs, allowing the model to better utilize auxiliary captions during retrieval. Second, we present DualGroup-Direct Preference Optimization (DG-DPO), a novel caption optimization strategy that directly uses retrieval relevance scores to supervise caption quality. Moreover, unlike traditional DPO, DG-DPO incorporates group-level preferences, enabling the model to learn a global retrieval ranking over video-caption pairs. Through extensive experiments, we show that CaRe-DPO significantly improves retrieval performance by effectively utilizing the auxiliary knowledge while generating better captions for retrieval.

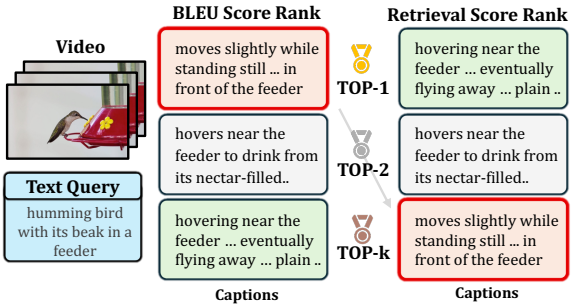
1 Introduction

Text-video retrieval is a fundamental task in multimodal learning, aiming to align natural language descriptions with video content. Traditional retrieval methods often adopt dual-encoder architectures, such as CLIP (Radford et al., 2021), which encode videos and text queries into a shared embedding space. However, these approaches often struggle with fine-grained semantic matching (Tian et al., 2024; Wang et al., 2023), particularly when videos contain complex temporal or contextual dynamics. To mitigate this, recent studies (Wu et al., 2023; Ma et al., 2024; Hur et al., 2025; Yang et al., 2025) have explored the use of video caption, natural language descriptions of video content, as auxiliary inputs to bridge the gap between the text queries and video content.

Multimodal Large Language Models (MLLMs) (Liu et al., 2024; Wang et al., 2024b; Li et al.,



(a) Failure case of Text-to-Video with caption retrieval



(b) Comparison on rank of captions (BLEU vs Retrieval)

Figure 1: (a) T2V retrieval with the *original* descriptive caption (video-to-caption retrieval R@1 of 90.7) compared to the *random* one. Nearly identical performance suggests that the model fails to effectively leverage the auxiliary knowledge. (b) illustrates that the top-1 caption selected based on captioning metric (BLEU) does not correspond to the top-1 caption when ranked by the retrieval score (placed at the bottom rank). Correlation between those two rankings remains as low as 30%.

2024c; Zhang et al., 2024) that encompass strong visual and text understandings, recently caught attention for handling multi-modal retrieval systems (Lin et al., 2025; Liu et al., 2025; Wei et al., 2024). Their capacity to jointly attend to both visual and textual inputs allows them to interpret diverse and complex text queries in relation to video content while also leveraging auxiliary captions as additional semantic context, providing a promising direction for advancing retrieval performance.

However, we observe that naively incorporating auxiliary captions into these retrieval models often leads to suboptimal gains. As shown in Fig. 1a, even when using descriptive captions (90.7 at R@1 for video-to-caption), replacing them with *random* captions results in nearly identical performance for text-to-video retrieval (81.5 vs. 81.6). This suggests that the model fails to effectively leverage the auxiliary knowledge, overlooking the semantic distinction between the heterogeneous textual inputs: the caption (as contextual knowledge) and text queries (as retrieval targets). The inefficiency in leveraging auxiliary captions is further highlighted when examining the alignment between the caption quality with the retrieval effectiveness. Specifically, as shown in Fig. 1b, we find that the top-1 caption selected based on conventional captioning metrics, *e.g.*, BLEU (Papineni et al., 2002), often does not correspond to the top-1 caption when ranked by the retrieval relevance score (placed at the bottom rank). We also further analyzed that the correlation between those two rankings is as low as 30% indicating a significant misalignment.

To this end, we propose **CaRe-DPO**, **Ca**ptioning for Text-Video **Re**trieval via **D**ualGroup-Direct **P**reference **O**ptimization, a retrieval framework that integrates two key components. First, the retrieval role-embeddings introduced in the retrieval model explicitly differentiate the roles of heterogeneous textual inputs, enabling the model to better utilize the auxiliary captions. Second, our DualGroup Direct Preference Optimization (DG-DPO), which not only directly supervises the captioning model with the retrieval scores to align with the retrieval objective, but also explores beyond standard single-group retrieval preference (local retrieval rank of captions given a single input video), to dual group preference (global retrieval rank over video-caption pairs across the dataset). We empirically validate that CaRe-DPO encourages the MLLM-based retrieval model to further leverage the auxiliary captions during retrieval and enables to enhance the quality of the caption, yielding a performance improvement across various text-video retrieval benchmarks.

The main contributions of ours are as follows:

- We propose CaRe-DPO, a novel retrieval framework that integrates retrieval role-embeddings and a retrieval-aligned caption optimization strategy to effectively leverage auxiliary captions in MLLM-based text-video retrieval.
- We introduce DualGroup-Direct Preference Optimization (DG-DPO), a new objective to caption for retrieval that supervises caption generation using retrieval relevance scores and incorporates both local (within-video) and global (cross-video-caption pair) ranks.
- Our extensive analyses show that CaRe-DPO significantly improves the retrieval performance by enhancing both the utility of auxiliary captions and fine-grained alignment between the captions with the retrieval objective.

2 Related Work

2.1 Text-Video Retrieval

To improve text-video retrieval, recent studies have explored the use of captions as auxiliary supervision. Cap4Video (Wu et al., 2023) treats captions as data augmentation to generate new training pairs, enhancing cross-modal interaction. NarVid (Hur et al., 2025) uses frame-level captions to enrich video understanding and applies a hard negative loss for better discrimination. ExCae (Yang et al., 2025) refines captions through self-learning to enhance expressiveness while minimizing manual intervention. Recently, with the advancement of Multimodal Large Language Models (MLLMs), several works (Lin et al., 2025; Liu et al., 2025) introduced MLLMs in multi-modal retrieval systems. MM-Embed (Lin et al., 2025) finetuned the MLLMs to universal retrievers, adopting the thought prompt-and-reranking strategies. LamRA (Liu et al., 2025) proposes reranking strategies of pointwise and listwise to further boost the retrieval performance. Yet, current approaches struggle to explore the adoption of captions into MLLM-based retrieval models while analyzing the effectiveness of those auxiliary captions.

2.2 Direct Preference Optimization

Direct Preference Optimization (DPO) (Rafailov et al., 2023) has emerged as an efficient alternative to reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022; Stiennon et al., 2020) for aligning large language models with human preferences. Recent studies have explored several limitations of DPO. To mitigate length bias in preference data, prior approached introduce reward normalization (Meng et al., 2024), token-level probability down-sampling (Lu et al., 2024), and explicit length regularization (Park et al., 2024). Other studies attempt to eliminate the reliance on a reference model to reduce computational cost (Meng et al., 2024; Xu et al., 2024; Hong et al., 2024). In the multimodal setting, DPO has been adapted to align multimodal large language models (MLLMs) for tasks such as visual question-answering (Li et al., 2024b) and mitigating hallucinations (Ouali et al., 2024; Wang et al., 2024a). In this work, we provide a retrieval-oriented preference modeling for MLLMs, and propose a dual-group DPO formulation to capture both local and global preference.

3 Preliminary

3.1 Text-Video Retrieval

Text-Video Retrieval consists of two tasks, video-to-text retrieval (V2T) and text-to-video retrieval (T2V), which aim to find the most relevant text or video given the query among the candidates of video or text. Often to enhance the cross-modal retrieval, several works (Wu et al., 2023; Yang et al., 2025; Hur et al., 2025) propose to utilize the generated caption $\mathbf{c}^{(i)}$ of the given video $\mathbf{v}^{(i)}$ to bridge the modality gap with the textual query $\mathbf{t}^{(i)}$. Hence, the retrieval dataset can be defined as $\mathcal{D}_{\text{ret}} = \{\mathbf{v}^{(i)}, \mathbf{c}^{(i)}, \mathbf{t}^{(i)}\}_{i=1}^{N}$, where $\mathbf{c}^{(i)}$ is often sampled from a captioning model $\mathcal{M}_{\text{cap}}$. During inference of text-video retrieval with auxiliary caption, $\mathbf{c}^{(i)}$ is paired with the $\mathbf{v}^{(i)}$, which defined as:

$$i^*_{\text{T2V}} = \arg\max_i P(\mathbf{v}^{(i)}, \mathbf{c}^{(i)}|\mathbf{t}). \tag{1}$$

Recently, MLLMs have often been employed for multi-modal retrieval systems, where they are adopted to re-rank the top-k text-video candidate pairs based on joint text-video similarity. Typically, given the video $\mathbf{v} = [v_1, .., v_{N_v}] \in \mathbb{R}^{N_v \times D}$, caption $\mathbf{c} = [c_1, .., c_{N_c}] \in \mathbb{R}^{N_c \times D}$, and text $\mathbf{t} = [t_1, .., t_{N_t}] \in \mathbb{R}^{N_t \times D}$, where N_v, N_c, N_t, and D denotes the numbers of video, caption, text tokens, and the hidden dimension respectively, the objective for reranking with MLLM-based models for retrieval can be defined as follows:

$$\mathcal{L} = -\log P(y|\mathbf{v}, \mathbf{c}, \mathbf{t}, \mathbf{I}). \tag{2}$$

The output y is defined with $y \in \{\texttt{True}, \texttt{False}\}$ tokens, resembling a binary classification task, and note that the auxiliary caption $\mathbf{c}$ is simply concatenated to the video along with the text query $\mathbf{t}$. Also, $\mathbf{I}$ denotes the instruction prompt to answer '`True`' or '`False`' that is omitted for the following notations. Hence, for the matching triplets, *i.e.*, $(\mathbf{v}^{(i)}, \mathbf{c}^{(i)}, \mathbf{t}^{(j)})$ where $i = j$, the model is trained to output '`True`', while for the unmatching triples where $i \neq j$ the model is expected to output '`False`'. During inference, following Lin et al. (2025) and Liu et al. (2025) the typical approach of measuring the relevance score s is:

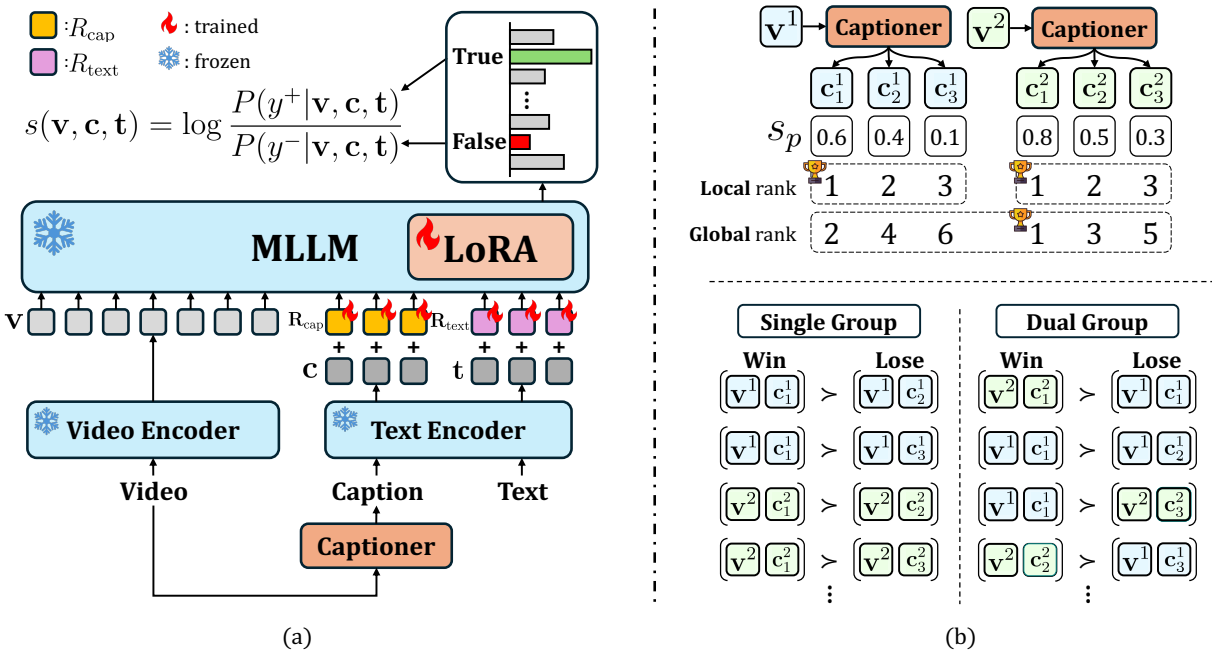
$$s(\mathbf{v}, \mathbf{c}, \mathbf{t}) = \log P(y^+|\mathbf{v}, \mathbf{c}, \mathbf{t}) \tag{3}$$

where $y^+ = \texttt{True}$. Thus, for T2V, we simply apply `softmax` over the relevance scores across all candidate videos with a fixed text query $\mathbf{t}^{(i)}$, and vice versa for V2T. However, we observe that simply concatenating the caption $\mathbf{c}$ into the input hinders the model from differentiating between the heterogeneous textual inputs of the text query $\mathbf{t}$ and the auxiliary caption $\mathbf{c}$. We further observe that the simple strategy of measuring the relevance score with the probability of predicting the '`True`' lacks fine-grained sensitivity required for retrieval.

3.2 Direct Preference Optimization

Direct Preference Optimization (DPO) (Rafailov et al., 2023), is a typical optimization strategy adopted to align LLMs output with human preferences, which is derived from the reinforcement learning objective in RLHF (Ziegler et al., 2019). $\mathcal{D}_{\text{DPO}}$ the preference dataset for DPO can be defined with $\{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}^{N}$, where x is an input, y_w, y_l are the preferred and dispreferred outputs, and the preference is estimated by the Bradley-Terry (BT) (Bradley and Terry, 1952). Typically, the objective of DPO, $\mathcal{L}_{\text{DPO}}$, can be written as follows:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \Big[\log \sigma(\hat{r}_\theta(x, y_w) - \hat{r}_\theta(x, y_l))\Big], \tag{4}$$



(a) (b)

Figure 2: **Illustration of our CaRe-DPO framework.** (a) depicts the MLLM-based retrieval model for text-video retrieval where we propose to adopt retrieval role-embeddings $\mathbf{R}_{\text{cap}}$ and $\mathbf{R}_{\text{text}}$ for the heterogeneous textual inputs applied to each token, accordingly: auxiliary caption (orange) and retrieval target text (purple). In addition, we illustrate the contrastive inference strategy (contrasting the probability of generation `'True'` to `'False'`) of which is more effective for retrieval. (b) visualizes our DualGroup-DPO mechanism where each caption given the video is evaluated with the retrieval relevance score s_p. Then, during training, SingleGroup-DPO adopts the local rank preference, while the DualGroup-DPO adopts the global rank preference, exploring across video-caption pairs.

where $\hat{r}_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$, given π_θ the policy model to optimize, and π_{ref} the reference model, β is a hyperparameter which determine distribution disparity of π_θ and π_{ref}, and σ denotes the sigmoid function. In standard DPO, the preference of the given outputs, y_w, y_l, is determined conditioned solely on a single input x, where the model learns to prefer one output over another, referred to as the local preference of x.

4 Method

In this section, we introduce our CaRe-DPO, **Ca**ptioning for **Re**trieval via **D**ualGroup-Direct **P**reference **O**ptimization, a novel retrieval framework that enhances text-video retrieval with auxiliary captions. First in Sec. 4.1, we introduce a simple retrieval role embedding mechanism that helps the model learn the distinct functional roles of the retrieval target and the auxiliary knowledge. Then in Sec. 4.2, we present DualGroup-DPO, a preference optimization method that supervises the captioning model to further align with the retrieval objective and propose to explore beyond standard single group preference to dual group preference learning. Overall framework is illustrated in Fig. 2.

4.1 Retrieval Role-embeddings

To enable the MLLM-based retrieval model, $\mathcal{M}_{\text{ret}}$, to differentiate the functional roles of the text query and the caption given as input for text-video retrieval, we adopt a simple yet effective retrieval role-embeddings. Specifically, given the input triplet $(\mathbf{v}, \mathbf{c}, \mathbf{t})$, we introduce a new role-embeddings $R_{\text{cap}} \in \mathbb{R}^D$ and $R_{\text{text}} \in \mathbb{R}^D$, which are combined to each corresponding tokens of $\mathbf{c}$ and $\mathbf{t}$, respectively. Hence, the training objective of Eq. 2 can be modified as follows:

$$\mathcal{L}_{\text{ret}} = -\log P(\mathbf{v}, \mathbf{c} + \mathbf{R}_{\text{cap}}, \mathbf{t} + \mathbf{R}_{\text{text}}) \quad (5)$$

where $\mathbf{R}_{\text{cap}} = \mathbf{1}_{N_c} R_{\text{cap}}^\top$ and $\mathbf{R}_{\text{text}} = \mathbf{1}_{N_t} R_{\text{text}}^\top$. Such a simple approach avoids $\mathcal{M}_{\text{ref}}$ from referencing the caption and the text as heterogeneous textual input, but enables it to explicitly distinguish according to its roles: the caption as contextual knowledge and the text query as retrieval targets. This role-specific encoding effectively guides the model to attend differently to the auxiliary caption and the query during training, leading to more precise cross-modal alignment and improved retrieval performance.

Inference Strategy. For the inference stage, we empirically observe that instead of simply adopting the probability of generating the y^+ token as the retrieval relevance score (Eq. 3), it is more effective to use the pairwise score margin between y^+ and y^- generation as follows:

$$s(\mathbf{v}, \mathbf{c}, \mathbf{t}) = \log \frac{P(y^+|\mathbf{v}, \mathbf{c}, \mathbf{t})}{P(y^-|\mathbf{v}, \mathbf{c}, \mathbf{t})} \tag{6}$$

Such a contrastive inference strategy allows the retrieval model to be more keen to the subtle differences of the input and its output decision, enhancing the retrieval performance.

4.2 DualGroup-DPO

Retrieval score driven Preference Dataset. To further handle the inefficiency in leveraging the auxiliary captions for retrieval, which stems from the misalignment between the training objective of the captioning and the retrieval models, we first construct the preference dataset that directly adopts the *retrieval scores* as supervision. First, we sample K number of captions $\{\mathbf{c}_k^{(i)}\}_{k=1}^K$ for each video $\mathbf{v}^{(i)}$, denoted as $\mathbf{c}_k^{(i)} \sim \mathcal{M}_{\text{cap}}(\mathbf{v}^{(i)})$ where $\mathcal{M}_{\text{cap}}(\cdot)$ refers to the pretrained captioning model. Then, we adopt $\mathcal{M}_{\text{ret}}(\cdot)$ to evaluate the quality of the sampled captions for video-text retrieval with the relevance score. We adopt the score between $\mathbf{c}_k^{(i)}$ and the text $\mathbf{t}_k^{(i)}$ while masking the video tokens in the attention mask ($\square$), which we empirically observe to be more effective in terms of precision than that of un-masked video tokens. Formally, the relevance score for preference optimization, s_p, is defined as:

$$s_{\text{p}}(\mathbf{v}^{(i)}, \mathbf{c}_k^{(i)}, \mathbf{t}^{(i)}) = \log \frac{P_{\mathcal{M}_{\text{ret}}}(y^+|\square, \mathbf{c}_k^{(i)}, \mathbf{t}^{(i)})}{P_{\mathcal{M}_{\text{ret}}}(y^-|\square, \mathbf{c}_k^{(i)}, \mathbf{t}^{(i)})} \tag{7}$$

DualGroup-Direct Preference Optimization. The conventional approach of DPO considers only *local retrieval rank preferences* that reference a single input, referred to as the SingleGroup-DPO. For instance, given a single video $\mathbf{v}^{(i)}$ with its associated two sampled captions, the preferred $\mathbf{c}_w^{(i)}$ and dispreferred $\mathbf{c}_l^{(i)}$, preference pair $\mathbf{c}_w^{(i)} \mid \mathbf{v}^{(i)} \succ \mathbf{c}_l^{(i)} \mid \mathbf{v}^{(i)}$ satisfy the following condition:

$$s_p\left(\mathbf{v}^{(i)}, \mathbf{c}_w^{(i)}, \mathbf{t}^{(i)}\right) > s_p\left(\mathbf{v}^{(i)}, \mathbf{c}_l^{(i)}, \mathbf{t}^{(i)}\right) + \gamma. \tag{8}$$

γ refers to the margin threshold, which enforces a minimum difference between retrieval scores. Building upon the SingleGroup-DPO, our DualGroup-DPO extends the framework to consider preferences across distinct video-caption pairs by leveraging their associated retrieval relevance scores across the dataset, *i.e.*, *global retrieval rank preferences*. For instance, given two video-caption pairs $(\mathbf{v}^{(i)}, \mathbf{c}_k^{(i)})$ and $(\mathbf{v}^{(j)}, \mathbf{c}_k^{(j)})$, where the former denote the k-th caption and video for the i-th sample, and the latter denote the k-th caption and the video for the j-th sample, the preference pair among the video-caption pair *i.e.*, $\mathbf{c}_w^{(i)}|\mathbf{v}_w^{(i)} \succ \mathbf{c}_l^{(j)}|\mathbf{v}_l^{(j)}$, can be defined as follows:

$$s_p\left(\mathbf{v}_w^{(i)}, \mathbf{c}_w^{(i)}, \mathbf{t}_w^{(i)}\right) > s_p\left(\mathbf{v}_l^{(j)}, \mathbf{c}_l^{(j)}, \mathbf{t}_l^{(j)}\right) + \gamma. \tag{9}$$

Notably, the preference can be defined where $i = j$ and $i \neq j$, unlike the SingleGroup-DPO where the sample pairs always satisfy $i = j$. Overall, the model learns to prefer video-caption pairs, which results in higher retrieval relevance scores, while considering the local rank preference of the caption and the global rank preference across distinct video-caption pairs, enhancing the retrieval performance. Hence the $\mathcal{L}_{\text{DG-DPO}}$ can be written as:

$$\mathcal{L}_{\text{DG-DPO}} = -\mathbb{E}_{(\mathbf{v}_w^{(i)}, \mathbf{v}_l^{(j)}, \mathbf{c}_w^{(i)}, \mathbf{c}_l^{(j)}) \sim \mathcal{D}_{\text{DG-DPO}}} \left[\log \sigma\left(\hat{r}_\theta(\mathbf{c}_w^{(i)}, \mathbf{v}_w^{(i)}) - \hat{r}_\theta(\mathbf{c}_l^{(j)}, \mathbf{v}_l^{(j)})\right)\right]. \tag{10}$$

Note that in practice, we do not increase the number of training samples; instead, we reuse the precomputed log probability values from the computation from when $i = j$ to compute $\mathcal{L}_{\text{DG-DPO}}$ for samples where $i \neq j$. Hence, we effectively leverage the samples within the same batch-aggregated across multiple GPUs-to adopt the global rank of video-caption pairs. As a result, without any additional computational or memory overhead, the captioning model is encouraged to explore consistent ranking preferences across a wider range of sample combinations of video-caption pairs for video-text retrieval with auxiliary captions.

5 Experiments

5.1 Experiments Setup

Datasets and Metrics. To validate the effectiveness of CaRe-DPO, we evaluate on three Text-Video retrieval benchmarks: DiDeMo (Anne Hendricks et al., 2017), ActivityNet (Caba Heilbron et al., 2015), and MSRVTT (Xu et al., 2016). For evaluation, we adopt the standard retrieval metrics: Recall@K (R@1, R@5, R@10). Note that

	Text-to-Video									Video-to-Text								
	DiDeMo			ActivityNet			MSRVTT			DiDeMo			ActivityNet			MSRVTT		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Non-MLLM-based																		
CLIP4Clip (Luo et al., 2022)	42.8	68.5	79.2	40.5	72.4	83.4	44.5	71.4	81.6	42.5	70.6	80.2	42.6	73.4	85.6	43.1	70.5	81.2
ViCLIP (Wang et al., 2024c)	49.4	-	-	49.8	-	-	52.5	-	-	50.2	-	-	48.1	-	-	51.8	-	-
MV-Adapter (Jin et al., 2024)	44.3	72.1	80.5	42.9	74.5	85.7	46.2	73.2	82.7	42.7	73.0	81.9	43.6	75.0	86.5	47.2	74.8	83.9
InternVideo (Wang et al., 2022)	57.9	82.4	88.9	62.2	85.9	93.2	55.2	79.6	87.5	59.1	81.8	89.0	62.8	86.2	93.3	57.9	79.2	86.4
UMT (Li et al., 2023)	70.4	90.1	93.5	66.8	89.1	94.9	58.8	81.0	87.1	67.9	88.6	93.0	64.4	89.1	94.8	58.6	81.6	86.5
Cap4Video (Wu et al., 2023)	52.0	79.4	87.5	-	-	-	51.4	75.7	83.9	-	-	-	49.0	75.2	85.0	-	-	-
InternVideo2 1B* (Wang et al., 2024d)	75.3	92.5	95.8	68.8	89.7	94.7	59.4	80.9	86.6	73.1	92.1	94.9	65.3	88.0	94.2	56.9	76.9	84.6
InternVideo2 6B (Wang et al., 2024d)	74.2	-	-	74.1	-	-	62.8	-	-	71.9	-	-	68.7	-	-	60.2	-	-
MLLM-based†																		
MM-Embed (Lin et al., 2025)	81.6	94.9	**96.3**	78.5	-	-	61.2	82.7	**88.8**	79.7	94.9	96.2	70.7	-	-	60.5	82.3	87.1
LamRA (Liu et al., 2025)	83.5	94.8	96.2	76.0	92.8	96.3	59.7	81.4	87.2	79.4	94.8	**96.6**	68.7	90.1	95.3	60.7	82.3	**89.0**
CaRe-DPO (Ours)	**85.1**	**95.0**	96.2	**79.3**	**93.7**	**96.4**	**64.1**	**83.8**	**88.8**	**82.5**	**95.2**	96.3	**74.2**	**92.5**	**96.2**	**63.8**	**83.0**	87.3

Table 1: **Comparison with state-of-the-art Text-Video Retrieval models.** * denotes reproduced results. We also report the performance of MLLM-retrieval models, which we reproduced adequately for Text-Video Retrieval, adopting their approach while applying to the same baseline as ours, VideoChat-Flash, denoted with the †.

Train	Inf.	Text-to-Video			Video-to-Text			Avg.
$\mathcal{L}_{\text{ret}}(\cdot)$	cap. $\mathbf{c}$	R@1	R@5	R@10	R@1	R@5	R@10	Δ
$(\mathbf{v}, \mathbf{t})$	$\varnothing$	80.1	78.9	71.8	68.5	62.0	61.3	-
$(\mathbf{v}, \mathbf{c}, \mathbf{t})$	rand.	81.5	94.6	95.9	79.1	94.6	96.5	-
	orig.	81.6	94.3	95.9	79.2	94.7	96.7	(+0.1)
$(\mathbf{v}, \mathbf{c}+\mathbf{R}_{\text{c}}, \mathbf{t}+\mathbf{R}_{\text{t}})$	rand.	82.6	94.4	96.0	76.5	95.0	96.2	-
	orig.	83.1	94.4	96.2	79.6	94.6	96.6	(+1.8)

Table 2: **Ablation on the Role-embeddings of $\mathcal{M}_{\text{ret}}$.** We adopt the zero-shot captions with the standard inference strategy. 'Avg. Δ' denotes an average change in R@k performance. $\mathbf{R}_{\text{c}}$, $\mathbf{R}_{\text{t}}$ refers to $\mathbf{R}_{\text{cap}}$, $\mathbf{R}_{\text{text}}$, 'rand' and 'orig.' denote *random* and *original* captions, respectively, and 'Inf.' denotes the inference stage.

	DiDeMo		ActivityNet		MSRVTT		Avg.
	T2V	V2T	T2V	V2T	T2V	V2T	Δ%
Baseline	83.1	79.6	78.3	74.0	62.7	63.6	-
+ $\mathcal{L}_{\text{SFT}}$	82.6	82.0	78.0	73.9	62.9	63.0	(+0.2)
+ $\mathcal{L}_{\text{SG-DPO}}$	84.4	82.4	78.8	74.1	63.5	63.3	(+1.1)
+ $\mathcal{L}_{\text{DG-DPO}}$	85.1	82.5	79.3	74.2	64.1	63.8	(+1.7)

Table 3: **Anlaysis on training objectives for $\mathcal{M}_{\text{cap}}$.** R@1 retrieval performance from the different training objectives for $\mathcal{M}_{\text{cap}}$. 'Avg. Δ%' denotes the percentage increase compared to the baseline across the dataset.

for auxiliary captions, we sample two per instance and average the performance over those to mitigate the caption variability while providing more robust results. See the supplementary for more details.

Implementation Details. For retrieval, we adopt InternVideo2-1B (Wang et al., 2024d) to initially compute the similarity between the video and the text query, and then we retrieve the top-16 candidates for re-ranking. Our baseline MLLM-based retrieval model, capable of adopting an auxiliary caption, is built upon VideoChat-Flash (Li et al., 2024c). For the captioning model, we adopt pretrained LLaVA-OneVision (Li et al., 2024a). More details are presented in the supplement.

5.2 Experimental Results

Main Results. Tab. 1 shows the performance of the State-of-the-Art text-video retrieval models, including non-MLLM-based and MLLM-based. The results show that our CaRe-DPO outperforms baseline models across various datasets, especially in R@1 for both T2V and V2T. Among non-MLLM-based models, ours effectively improves performance over the SOTA model of InternVideo2-6B, with an average of 14.7%, 7.5%, and 4.0% increase in R@1 for DiDeMo, ActivityNet, and MSRVTT, respectively. To further validate the effectiveness among the MLLM-based retrieval models, we compare against MM-Embed and LamRA. Notably, our CaRe-DPO shows superior performance with 3.9%, 3.0%, and 5.1% increase compared to MM-Embed, and 2.9%, 6.2%, and 6.2% increase on average for R@1 compared to LamRA across datasets. Overall, outperforming the baseline models over T2V and V2T, the results demonstrate its effectiveness in adopting CaRe-DPO for Text-Video retrieval, especially with MLLM-based models.

5.3 Quantitative Analysis

Effectiveness of Retrieval Role-embeddings. Tab. 2 presents an ablation study on the impact of retrieval role-embeddings for MLLM-based models. As shown, the model trained with the auxiliary caption, $+\mathcal{L}_{\text{ret}}(\mathbf{v}, \mathbf{c}, \mathbf{t})$, achieves solid performance compared to the baseline *i.e.*, R@1 of 81.6 for T2V and 79.2 for V2T. Nevertheless, the model fails to fully leverage the auxiliary caption of which is shown with a minimal performance drop, a 0.1 decrease in R@1 on average, when replacing the

Inference	$\mathcal{M}_{\text{cap}}$	DiDeMo T2V	DiDeMo V2T	ActivityNet T2V	ActivityNet V2T	MSRVTT T2V	MSRVTT V2T
$s(\mathbf{c}, \mathbf{t})$	Baseline	49.6	40.8	43.2	37.0	40.5	37.7
	+ $\mathcal{L}_{\text{DG-DPO}}$	51.3	43.4	52.2	43.6	49.0	45.5
$s(\mathbf{v}, \mathbf{c})$	Baseline	91.8	90.7	88.2	86.5	88.9	86.1
	+ $\mathcal{L}_{\text{DG-DPO}}$	92.1	92.2	88.7	87.5	89.7	87.1

Table 4: **Analysis on caption quality for retrieval.** Note that 'Baseline' denotes zero-shot caption adopted for retrieval. For $s(\mathbf{c}, \mathbf{t})$, we adopt the model trained with $(\mathbf{v}, \mathbf{c}, \mathbf{t})$, while we mask the video tokens. For $s(\mathbf{v}, \mathbf{c})$ we utilize the model trained solely on ($\mathbf{v}$ and $\mathbf{t}$). We report the R@1 performance for both T2V and V2T.

	Text-to-Video R@1	R@5	R@10	Video-to-Text R@1	R@5	R@10
Captioning Metric						
BLEU	84.1	95.0	96.3	82.3	94.7	96.4
METEOR	83.8	94.9	96.6	82.8	94.9	96.3
Retrieval Score (s_p)						
$\log \frac{P(y^+\|\mathbf{v}, \mathbf{c}, \mathbf{t})}{P(y^-\|\mathbf{v}, \mathbf{c}, \mathbf{t})}$	85.0	95.0	96.4	82.4	95.0	96.4
$\log \frac{P(y^+\|\square, \mathbf{c}, \mathbf{t})}{P(y^-\|\square, \mathbf{c}, \mathbf{t})}$	85.1	95.0	96.2	82.5	95.2	96.3

Table 5: **Comparison on adopting different preference scores s_p for constructing $\mathcal{D}_{\text{DG-DPO}}$.** Note that 'Mean R@1' signifies the average of the R@1 values. We report the retrieval performance on DiDeMo. Also, $\square$ denotes masked attention for video tokens.

$s(\mathbf{v}, \mathbf{c}, \mathbf{t})$	Text-to-Video R@1	R@5	R@10	Video-to-Text R@1	R@5	R@10
$\log P(y^+\|\mathbf{v}, \mathbf{c}, \mathbf{t})$	82.6	94.7	96.2	79.7	94.7	96.1
$\log P(y^-\|\mathbf{v}, \mathbf{c}, \mathbf{t})$	84.9	95.0	96.2	82.3	95.1	96.2
$\log \frac{P(y^+\|\mathbf{v}, \mathbf{c}, \mathbf{t})}{P(y^-\|\mathbf{v}, \mathbf{c}, \mathbf{t})}$	85.1	95.0	96.2	82.5	95.2	96.3

Table 6: **Comparison on the inference strategy.** Retrieval performance on DiDeMo, where $s(\mathbf{v}, \mathbf{c}, \mathbf{t})$ denotes the relevance score adopted for the inference. y^+ and y^- denote denote '`True`' and '`False`' respectively.

caption with a random one. In contrast, our model trained with the role-embeddings presents a superior performance with 83.1 for T2V and 79.6 in V2T, while showing higher sensitivity to the quality of the caption, with notable +1.8 improvement in average for R@1 compared to the random caption input. These results highlight the effectiveness of role-embeddings to encourage the model to differentiate the two roles of auxiliary knowledge and retrieval target, leading to more accurate retrievals.

Analysis on training objectives for $\mathcal{M}_{\text{cap}}$. In Tab. 3 we analyze different objectives for training the captioning model on the performance of text-video retrieval. As shown, simply fine-tuning the model on the given dataset denoted as $\mathcal{L}_{\text{SFT}}$ (row 2), results in an average of 0.2% improvement on average for R@1 while showing a performance degradation for ActivitiyNet, and MSRVTT of 0.3% for both. In contrast, adopting our $\mathcal{L}_{\text{SG-DPO}}$ or $\mathcal{L}_{\text{DG-DPO}}$, which optimizes the model with DPO while adopting the retrieval scores for preference determination, results in superior performance. Specifically, $\mathcal{L}_{\text{SG-DPO}}$ (row 2) that relies on local preference of the retrieval score, shows 2.5%, 0.4%, and 0.4% increase for DiDeMo, ActivityNet, and MSRVTT, respectively. Moreover, further considering the global preference based on the retrieval scores, $\mathcal{L}_{\text{DG-DPO}}$ (row 3), results in better precision for retrieval with 3.0%, 0.8%, and 1.3% performance improvement compared to the baseline across the datasets. The results highlight the effectiveness of adopting the retrieval scores with DPO to better align the generated captions for retrieval, and also demonstrate the effectiveness of DG-DPO, which considers the global preference beyond local preferences of video-caption pairs.

Analysis on the quality of caption for retrieval. To further investigate the effectiveness of captions in retrieval with CaRe-DPO, we design a series of experiments shown in Tab. 4: text-to-caption (T2C) (upper half) and video-to-caption retrieval (V2C) (lower half). T2C assesses how well the auxiliary caption semantically aligns with the query, V2C measures the degree to which the caption captures the distinctive content of the video itself. The results show that the caption generated from $\mathcal{M}_{\text{cap}}$ trained with $\mathcal{L}_{\text{DG-DPO}}$ results in consistent improvements across both retrievals. Specifically, in T2C, the caption generated after adopting our DG-DPO yields an average of 15.0% increase in performance, especially in MSRVTT with 21.0% increase in T2V and 20.7% increase in V2T. In V2C, the zero-shot caption itself shows strong explanability of the video, yet with our $\mathcal{L}_{\text{DG-DPO}}$, it further leads to performance enhancement with +1.0% in DiDeMo, +0.9% in ActivityNet, and +1.0% in MSRVTT.

Ablation on preference score for DG-DPO. In Tab 5, we compare T2V results while adopting different types of preference scores s_p for constructing $\mathcal{D}_{\text{DG-DPO}}$. We observe that directly using retrieval-based scores (rows 3 and 4) consistently outperforms traditional captioning metrics such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005). Specifically, in the T2V setting, using BLEU and METEOR leads to a performance drop of 1.0% and 1.3% in R@1, respectively, compared to using retrieval-

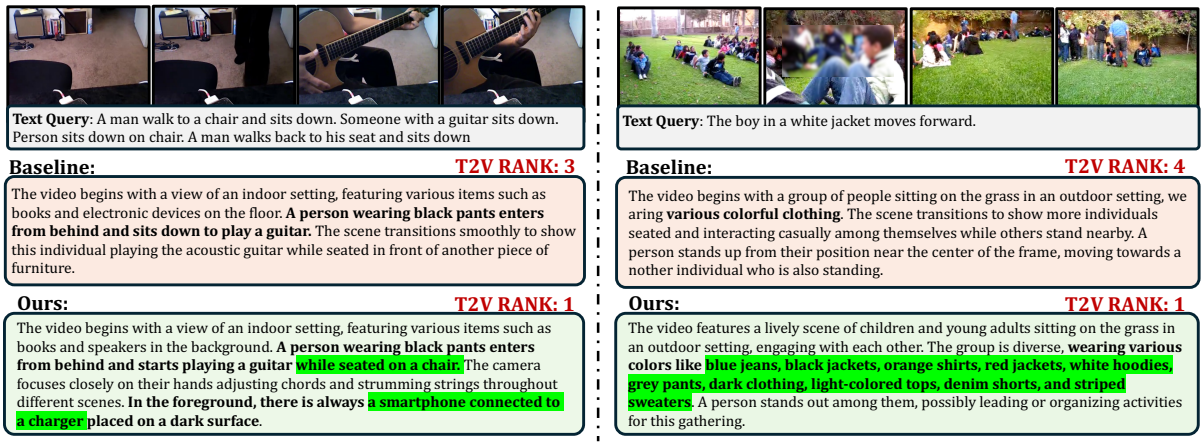


Figure 3: **Qualitative example of video captioning.** Comparison of the predictions of the caption generated by the zero-shot captioning model with our model trained with DG-DPO on DiDeMo. We also report the Text-to-Video retrieval rank for both cases, for which our model results in a higher rank. The highlighted green depicts the fine-grained detail generated by our model, which is not provided in the caption generated by the baseline.

based preference scores. Note we adopt the masked version for training due to its better precision.

Analysis on the inference strategy. Tab. 6 explores the different inference strategies in MLLM retrieval, and we determine that our contrastive inference strategy yields the best result. The standard approach (row 1) results in significant performance degradation compared to those that adopt the probability of generating '`False`' (row 2 and 3). Specifically, simply adopting $\log P(y^-|\mathbf{v},\mathbf{c},\mathbf{t})$ (row 2), results in +2.5% increase in R@1 on average, and adopting $\log P(y^+|\mathbf{v},\mathbf{c},\mathbf{t}) - \log P(y^-|\mathbf{v},\mathbf{c},\mathbf{t})$ (row 3), results in +2.7% increase.

5.4 Qualitative Results

Qualitative results of DG-DPO. Fig. 3 illustrates captions generated by the base model and ours. For the same video, our DG-DPO trained model with direct supervision of retrieval scores consistently provides more informative results that align more with retrieval. In Fig. 3 (left), while the base model only describes the act of playing the guitar resulting in T2V retrieval rank of 3, our model captures finer details such as the person being "*seated on a chair*" and the presence of "*a smartphone connected to a charger*", resulting in rank 1. In Fig. 3 (right), our caption depicts rich elements like "*blue jeans, black jackets, .., white hoodies*", whereas the baseline model simply writes as "*various colorful clothing*", improving the retrieval rank from 4 to 1. These richer descriptions align better with text-video retrieval that requires fine-grained discrimination among candidates, validating the effectiveness of DG-DPO.

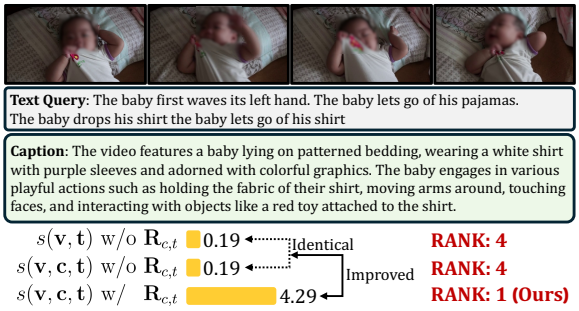


Figure 4: **Qualitative example on effectiveness of Role-embeddings.** We report the relevance scores for different baseline models on V2T, while adopting the caption generated with DG-DPO on DiDeMo.

Qualitative results of Role-embeddings. Fig. 4 illustrates the effect of role-embeddings for video-to-text retrieval, which illustrates that ours grounds the auxiliary caption better (retrieval rank improves from 4 to rank 1). Despite the descriptive caption of "*baby engages in various.. actions*", "*interacting with objects like .. attached to the shirt*" that closely relates with the text query, the baseline models (without the role-embeddings) fails to utilize the caption, showing retrieval relevance score of only 0.19, whereas ours shows 4.1 point increase to 4.29.

6 Conclusion

We present CaRe-DPO, a novel retrieval framework that enhances text-video retrieval with auxiliary captions. Our role-embeddings enable retrieval models to explicitly distinguish the roles of heterogeneous textual inputs. Furthermore, our DualGroup-Direct Preference Optimization aligns caption generation with retrieval relevance scores while leveraging both local and global ranks.

Limitations

In this work, we propose CaRe-DPO that relies on the MLLM-based models for text-video retrieval. CaRe-DPO builds upon MLLM-based retrieval models, which inherently rely on the pre-trained multimodal knowledge encoded in the MLLM, which also includes the captioning model adopted. As a result, the performance of our approach may be constrained by the underlying capabilities and biases of the base MLLM, especially in domain-specific or low-resource settings. Furthermore, the group-level preference modeling in DG-DPO can benefit from larger sample groups to learn robust global ranking signals; however, this increases the computational cost, posing challenges when scaling to large-scale video-caption datasets.

References

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *ICCV*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL*.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*.

Feng Cheng, Xizi Wang, Jie Lei, David Crandall, Mohit Bansal, and Gedas Bertasius. 2023. Vindlu: A recipe for effective video-and-language pretraining. In *CVPR*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *NeurIPS*.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*.

Chan Hur, Jeong-hun Hong, Dong-hun Lee, Dabin Kang, Semin Myeong, Sang-hyo Park, and Hyeyoung Park. 2025. Narrating the video: Boosting text-video retrieval via comprehensive utilization of frame-level captions. In *CVPR*.

Xiaojie Jin, Bowen Zhang, Weibo Gong, Kai Xu, Xueqing Deng, Peng Wang, Zhao Zhang, Xiaohui Shen, and Jiashi Feng. 2024. Mv-adapter: Multimodal video transfer learning for video text retrieval. In *CVPR*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. 2023. Unmasked teacher: Towards training-efficient video foundation models. In *ICCV*.

Shengzhi Li, Rongyu Lin, and Shichao Pei. 2024b. Multi-modal preference alignment remedies degradation of visual instruction tuning on language models. In *ACL*.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, and 1 others. 2024c. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2025. Mm-embed: Universal multimodal retrieval with multimodal llms. In *ICLR*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Yikun Liu, Pingan Chen, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. 2025. Lamra: Large multimodal model as your advanced retrieval assistant. In *CVPR*.

Junru Lu, Jiazheng Li, Siyu An, Meng Zhao, Yulan He, Di Yin, and Xing Sun. 2024. Eliminating biased length reliance of direct preference optimization via down-sampled kl divergence. In *EMNLP*.

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. *arXiv preprint arXiv:2104.08860*.

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 *ECCV*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. In *NeurIPS*.

Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2021. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *CVPR*.

Yassine Ouali, Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. 2024. Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms. In *ECCV*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. Disentangling length from quality in direct preference optimization. In *ACL Findings*.

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 *ICML*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *NeurIPS*.

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 *AAAI*.

Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024a. mdpo: Conditional preference optimization for multimodal large language models. In *EMNLP*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. 2024c. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *ICLR*.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, and 1 others. 2024d. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, and 1 others. 2022. Internvideo: General video foundation models via generative and discriminative learning.

Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2023. Unified coarse-to-fine alignment for video-text retrieval. In *CVPR*.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2024. Uniir: Training and benchmarking universal multimodal information retrievers. In *ECCV*.

Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. 2023. Cap4video: What can auxiliary captions do for text-video retrieval. In *CVPR*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. In *ICML*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Baoyao Yang, Junxiang Chen, Wanyun Li, Wenbin Yao, and Yang Zhou. 2025. Expertized caption auto-enhancement for video-text retrieval.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *ICCV*.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Dataset Details

DiDeMo. DiDeMo (Anne Hendricks et al., 2017) is a text-video retrieval benchmark, namely the Distinct Describable Moments, which comprises 10K videos, which are segmented into 5-second clips for annotation, totaling 26K annotated moments. Each moment is richly described with references to camera movement, temporal transitions, and actions. We treat the retrieval task as a paragraph-to-video retrieval where we concatenate all the captions within the video, following prior works (Luo et al., 2022; Wu et al., 2023; Li et al., 2023; Wang et al., 2024d; Cheng et al., 2023; Hur et al., 2025). Note that the dataset provides 8,394 training and 1,003 test samples.

ActivityNet. Activitynet (Caba Heilbron et al., 2015) is a text-video retrieval benchmark that is based on 19K YouTube videos, categorized into 200 activity classes. For each class, there exists an average of 137 videos, and each video contains about 1.41 temporal activities. Similar to DiDeMo, we aggregate all the captions per video and implement the task as a paragraph-to-video retrieval, while we evaluate on the val1 split following Luo et al. (2022); Li et al. (2023); Wang et al. (2024d); Cheng et al. (2023); Hur et al. (2025).

MSRVTT. The MSRVTT (Xu et al., 2016) dataset, namely Microsoft Research Video to Text, contains 10k video clips that span across 20 categories, of which each clip is annotated by 20 sentences. Following previous protocols (Luo et al., 2022; Wang et al., 2024d; Li et al., 2023; Cheng et al., 2023; Hur et al., 2025), we use the 9k sample set for training (which is about 180k caption-video pairs), and adopt the 1,000 clips for testing.

	DiDeMo		ActivityNet		MSRVTT	
	$\mathcal{M}_{\text{ret}}$	$\mathcal{M}_{\text{cap}}$	$\mathcal{M}_{\text{ret}}$	$\mathcal{M}_{\text{cap}}$	$\mathcal{M}_{\text{ret}}$	$\mathcal{M}_{\text{cap}}$
Learning rate	8e-5	8e-6	2e-5	8e-6	1e-4	8e-6
Warmup Epochs	1	0.1	1	0.1	1	0.1
Epoch	5	1	5	1	3	1
Batch Size	32	8	32	8	512	8
LoRA r	8	64	8	64	8	64
LoRA α	32	128	32	128	32	128
β	-	0.1	-	0.1	-	0.1
γ	-	0.7	-	2.0	-	0.5

Table 7: Training hyperparameters for $\mathcal{M}_{\text{ret}}$ with $\mathcal{L}_{\text{ret}}$ and $\mathcal{M}_{\text{cap}}$ with $\mathcal{L}_{\text{DG-DPO}}$.

B Implementation Details

Training Details for retrieval. For training an MLLM-based model for retrieval, we adopt the recent MLLM of VideoChat-Flash-7B (Li et al., 2024c). The baseline model is equipped with a visual encoder of UMT-L (Li et al., 2023) and an LLM of Qwen2 (Yang et al., 2024). For each benchmark, we only train the linear projection layer while adopting LoRA (Hu et al., 2022) for fine-tuning the model for efficiency. We adopt 16 frames per video for all datasets. All the experiments were done using 8 NVIDIA H100 80GB GPUS.

Prompts for text-video retrieval. We built several different models capable of implementing text-video retrieval. For the model trained with the loss of $\mathcal{L} = -\log P(y|\mathbf{v}, \mathbf{t})$, which is the baseline text-video retrieval model that does not accept auxiliary caption as input, we adopted the prompt of "*Caption:* `[caption]`. *Does the above video match the caption? True or False*". Note that we utilized the word 'Caption' for referencing the text query that is different from the auxiliary caption that we dealt with in this paper. For training the model with the loss of $\mathcal{L} = -\log P(y|\mathbf{v}, \mathbf{c}, \mathbf{t})$, which is capable of adopting the auxiliary caption for training, we use the prompt of "*Video description:* `[caption]`. *Caption* `[caption]`. *Based on the video and its description, is the video relevant to the caption? Answer True or False*." To clarify, the 'video description' corresponds to the auxiliary caption dealt with in the paper, whereas the 'caption' refers to the text query (retrieval target).

Training Details for captioning. We adopt LLaVA-OneVision-7B (Li et al., 2024a) for training the captioning model with Direct Preference Optimization. Similar to MLLM-based retrieval model finetuning, we adopt LoRA (Hu et al., 2022) for parameter-efficient finetuning. LLaVA-Onevision consists of Qwen2 (Yang et al., 2024) as the LLM, and SigLIP vision encoder (Zhai et al., 2023). We adopt 16 frames per video for all datasets. All the experiments were done using 8 NVIDIA H100 80GB GPUS.

Prompt for captioning. We empirically explored several ways of generating the caption for the dataset for zero-shot. Simply prompting the captioning model to generate a *detailed* description about the video will cause the model to generate a very long paragraph for the given video. Hence, we utilized the prompt of "*Describe this video in detail with three sentences.*".

Inference Details for retrieval. MLLM-based retrieval models are adopted as a re-ranker (Lin et al., 2025; Liu et al., 2025; Miech et al., 2021), benefiting from the ability to jointly attend to both visual and textual data. Hence, based on the InternVideo-1B (Wang et al., 2024d) similarity computed between the video and the text query, we retrieve the top-16 candidates for re-ranking. Finally, we weight the output scores of the two models following the protocol of Miech et al. (2021).

Inference Details for captioning. To construct the dataset $\mathcal{D}_{\text{DG-DPO}}$, we sample $k = 3$ captions per video using a generation temperature of 0.2, following the settings provided by LLaVA-OneVision (Li et al., 2024a). For the caption generation of evaluating the retrieval model, we sample $k = 2$ captions per video. For our experiments, we average the retrieval scores across both in order to account for the variability in caption generation and provide a more robust performance estimate.

Training details for $\mathcal{L}_{\text{DG-DPO}}$. To construct preference pairs from ranked retrieval results, we explore two strategies. First, we compute a global rank for each sample in the dataset and refer to these ranks within each batch to determine preference between video-caption pairs. The first strategy treats the top half of the ranked samples (i.e., higher-ranked pairs) as *chosen* and the bottom half as *rejected*. In contrast, the second strategy forms preference pairs by grouping the ranked indices into adjacent pairs, where the higher-ranked sample in each pair is treated as the *chosen* one and the lower-ranked as the *rejected*. Empirically, we observe that the latter strategy yields greater performance improvements. We hypothesize that this is because it produces training pairs with relatively smaller marginal differences compared to the former approach, allowing the model to learn more nuanced preference signals.

C Hyperparameters

In Tab. 7, we report the hyperparameters adopted for training the retrieval model $\mathcal{M}_{\text{ret}}$, and the captioning model $\mathcal{M}_{\text{cap}}$, across the text-video retrieval dataset.

D Further ablation on Role-embeddings.

We conduct further analysis on the role-embeddings for text-video retrieval on DiDeMo, evaluating with and without each function role embedding. The results suggest that the model trained with R_{text}, results in higher V2T retrieval at R@1 (79.2 to 79.8) whereas the model trained with R_{cap}, results in higher T2V retrieval at R@1 (81.6 to 82.6). Notably, combining both role embeddings results in the best overall performance, achieving 83.1 R@1 in T2V and 79.6 R@1 in V2T. These findings highlight the importance of role-embeddings integrated for both heterogeneous textual inputs, enhancing the model's ability to distinguish the auxiliary caption and the retrieval target, enabling the model to utilize more of the auxiliary caption for retrieval.

Train			Text-to-Video			Video-to-Text		
$\mathcal{L}_{\text{ret}}(\cdot)$	$\mathbf{R}_{\text{c}}$	$\mathbf{R}_{\text{t}}$	R@1	R@5	R@10	R@1	R@5	R@10
$(\mathbf{v}, \mathbf{c}, \mathbf{t})$	✗	✗	81.6	94.3	95.9	79.2	94.7	96.7
$(\mathbf{v}, \mathbf{c}, \mathbf{t} + \mathbf{R}_{\text{t}})$	✗	✔	81.2	94.5	95.6	79.8	94.3	96.5
$(\mathbf{v}, \mathbf{c} + \mathbf{R}_{\text{c}}, \mathbf{t})$	✔	✗	82.6	94.4	95.9	79.6	94.6	96.6
$(\mathbf{v}, \mathbf{c} + \mathbf{R}_{\text{c}}, \mathbf{t} + \mathbf{R}_{\text{t}})$	✔	✔	83.1	94.4	96.2	79.6	94.6	96.6

Table 8: **Ablation on the Role-embeddings of $\mathcal{M}_{\text{ret}}$.** Note that we adopt the zero-shot captions with the standard inference strategy. $\mathbf{R}_{\text{c}}$, and $\mathbf{R}_{\text{t}}$ denotes $\mathbf{R}_{\text{cap}}$ and $\mathbf{R}_{\text{text}}$.

E Further qualitative examples

In Fig 5, Fig 6, we further show the qualitative comparisons between captions from the baseline and our model on DiDeMo, and ActivityNet. In addition, we provide further qualitative examples on the effectiveness of our role-embeddings presented in Fig. 7. The qualitative examples reveal that the fine-grained descriptions of the video generated from our model trained with DG-DPO of our CaRe-DPO enable the retrieval model to better align with the text-video retrieval task. Furthermore, the role-embeddings improve the retrieval models in distinguishing the roles of heterogeneous textual inputs that include the auxiliary caption and the text query as retrieval targets.

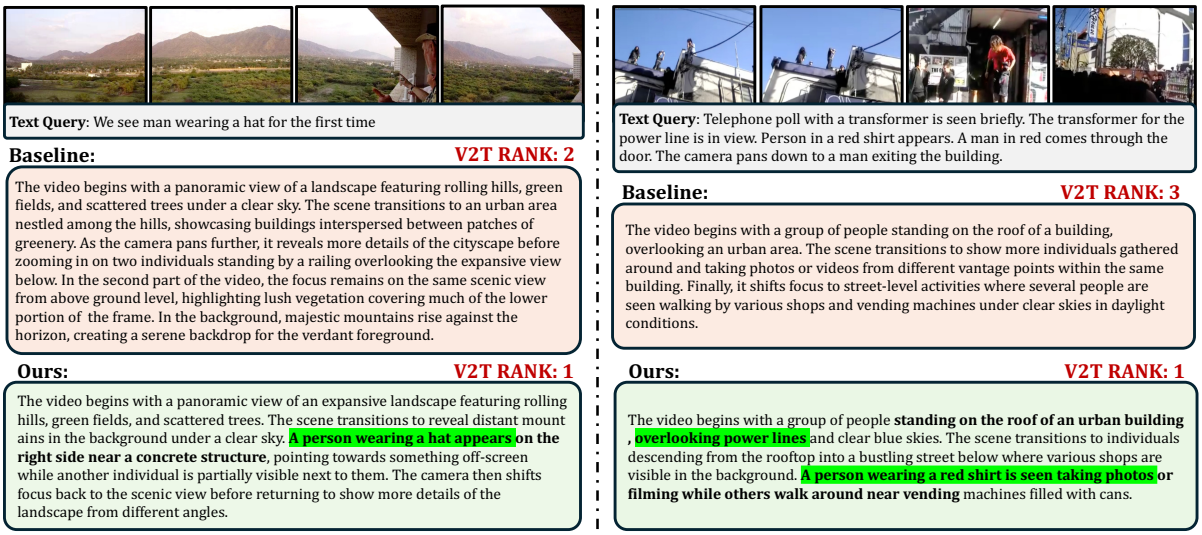


Figure 5: **Further qualitative example of video captioning.** Comparison of the predictions of the caption generated by the zero-shot captioning model with our model trained with DG-DPO on DiDeMo. The highlighted green depicts the fine-grained detail generated by our model, which is not provided in the caption generated by the baseline.

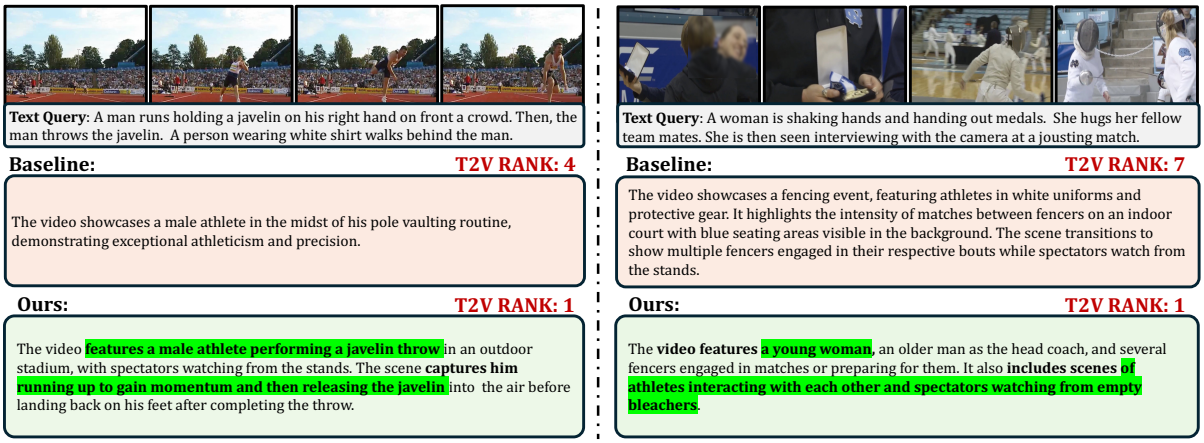


Figure 6: **Further qualitative example of video captioning.** Comparison of the predictions of the caption generated by the zero-shot captioning model with our model trained with DG-DPO on ActivityNet. The highlighted green depicts the fine-grained detail generated by our model, which is not provided in the caption generated by the baseline.

Figure 7: **Further qualitative example on effectiveness of role-embeddings.** We report the relevance scores for different baseline models of V2T while adopting the caption generated with DG-DPO on DiDeMo. The results reveal that our models effectively utilize the caption for retrieval.