Captioning for Text-Video Retrieval via Dual-Group Direct Preference Optimization

Ji Soo Lee[1] **Byungoh Ko**[1] **Jaewon Cho**[1] **Howoong Lee**[2] **Jaewoon Byun**[2] **Hyunwoo J. Kim**[3†]
[1]Korea University [2]Hanwha Vision [3]KAIST

{simplewhite9, ko990128, cho35750}@korea.ac.kr
{howoong.lee, jaewoon.byun}@hanwha.com hyunwoojkim@kaist.ac.kr

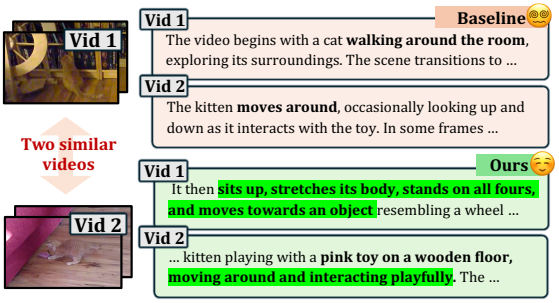
Abstract

In text-video retrieval, auxiliary captions are often used to enhance video understanding, bridging the gap between the modalities. While recent advances in multi-modal large language models (MLLMs) have enabled strong zero-shot caption generation, we observe that such captions tend to be generic and indistinguishable across visually similar videos, limiting their utility for fine-grained retrieval. Moreover, conventional captioning approaches are typically evaluated using language generation metrics, such as BLEU, which are not typically tailored for retrieval tasks that require making discriminative distinctions between candidates. To address this, we propose **CaRe-DPO**, a retrieval framework that directly optimizes caption generation using retrieval relevance scores. At its core is Dual-Group Direct Preference Optimization (DG-DPO), a novel learning strategy that supervises captioning by modeling preferences across groups of distinct video and caption pairs. In addition, we present an MLLM-based retrieval model that incorporates role-embeddings to better distinguish between textual inputs with different functional roles, such as an auxiliary caption and a text query. Through extensive experiments, we demonstrate that CaRe-DPO significantly enhances retrieval performance by effectively leveraging auxiliary knowledge to generate fine-grained captions for retrieval. Code is available at https://github.com/mlvlab/CaReDPO.

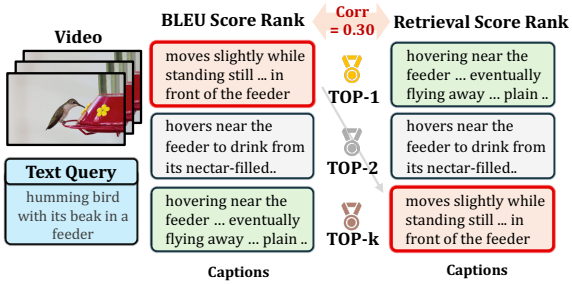
1 Introduction

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 captions, 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., 2024c; Zhang et al., 2024; Ko et al., 2023, 2025a; Park et al., 2025, 2024a) that encompass strong visual and text understandings, recently caught



(a) Generated captions for a pair of semantically similar videos



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

Figure 1: **Misalignment between conventional captioners and retrieval objectives.** (a) Captions from pretrained models are nearly identical for similar videos, while our method produces more distinct and retrieval-relevant descriptions. (b) BLEU-selected top captions often mismatch retrieval rankings; correlation between the two is as low as 30%.

† Corresponding authors.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 16022–16039
November 4-9, 2025

attention for handling multi-modal retrieval systems (Lin et al., 2025; Liu et al., 2025; Wei et al., 2024; Ko et al., 2025b). 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 captions as additional semantic context, providing a promising direction for advancing retrieval performance.

However, often captions produced by pretrained models fail to capture the detailed distinctions necessary for retrieval. As illustrated in Fig. 1a, two visually similar videos depicting cats in indoor environments are given captions that are generic and overlapping, describing actions like 'walking around the room' or 'moves around.' Nevertheless, we expect to generate captions that surface discriminative visual details, such as 'sits up, stretches its body' and 'interacting playfully', which provide discriminative cues critical for retrieval among similar videos. This issue is further amplified by the common practice of evaluating caption quality using metrics such as BLEU (Papineni et al., 2002). Specifically, as shown in Fig. 1b, among the three distinct captions generated for a video, the top-1 caption selected based on a conventional captioning metric often does not align with the top-1 caption when ranked by retrieval relevance score (placed at the bottom rank). To quantify this misalignment, we measured the Pearson correlation between the indices of the top-1 ranked captions under each metric, finding a correlation as low as 0.30. This highlights a substantial discrepancy between conventional captioning metric-based evaluations and retrieval-oriented objectives.

To this end, we propose **CaRe-DPO**, **Ca**ptioning for Text-Video **Re**trieval via **D**ual-Group Direct **P**reference **O**ptimization, a novel retrieval framework. At the core is Dual-Group Direct Preference Optimization (DG-DPO), directly supervising the captioning model with the retrieval scores to align with the retrieval objective. Unlike standard single-group DPO (local caption ranking within a video), DG-DPO incorporates dual-group preferences, learning global ranking over video-caption pairs. In addition, we introduce role-embeddings during retrieval model training to differentiate the roles of heterogeneous textual inputs, allowing the model to more effectively leverage the auxiliary captions. We empirically validate that CaRe-DPO encourages the MLLM-based retrieval model to further leverage the auxiliary captions during retrieval and enables enhancement of the caption quality, yielding a performance improvement across various text-video retrieval benchmarks.

The main contributions of this work are:

- To the best of our knowledge, we are the first to address the misalignment between conventional captioning metrics and retrieval objectives, and tackle the challenge of leveraging captions to improve retrieval performance.
- We propose **CaRe-DPO**, a retrieval framework that integrates role-embeddings with retrieval-aligned caption optimization to leverage auxiliary captions in MLLM-based text-video retrieval.
- Our DG-DPO supervises caption generation using retrieval relevance scores with both local (within-video) and global (cross-video-caption) ranking, enabling generation of captions that better reflect retrieval importance.
- Our CaRe-DPO improves caption quality and retrieval performance, achieving superior performance across multiple benchmarks.

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 them 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; Ko et al., 2025b) introduced MLLMs in multi-modal retrieval systems. MM-Embed (Lin et al., 2025) finetuned the MLLMs to universal retrievers, adopting thought prompt-and-reranking strategies. Concurrently with our work, BLiM (Ko et al., 2025b) investigates candidate prior bias induced by candidate likelihood estimation and improves retrieval performance through bidirectional likelihood estimation.

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 approaches introduce reward normalization (Meng et al., 2024), token-level probability down-sampling (Lu et al., 2024), and explicit length regularization (Park et al., 2024b). 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 multimodal, DPO has been adapted to align MLLMs for tasks such as visual question-answering (Li et al., 2024b), dense video captioning (Lee et al., 2025), and hallucination mitigations (Ouali et al., 2024; Wang et al., 2024a). In this work, we provide a retrieval-oriented preference modeling with MLLMs.

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 pretrained captioning model $\mathcal{M}_{\text{cap}}$. During inference, $\mathbf{c}^{(i)}$ is paired with the $\mathbf{v}^{(i)}$, which T2V retrieval for instance is 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 denote 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:

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

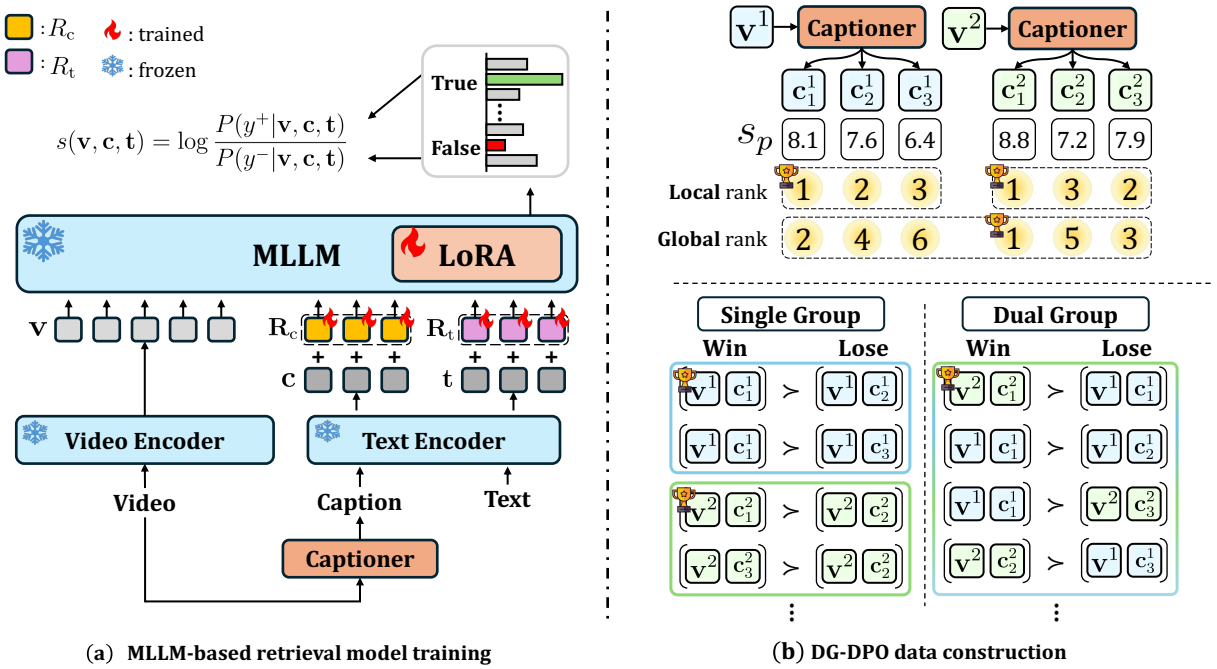
where $y^+ = \texttt{True}$. For T2V, the retrieved video is chosen as the candidate with the highest relevance score for a given text query $\mathbf{t}^{(i)}$, and vice versa for V2T. However, we observe that simply concatenating the caption $\mathbf{c}$ into the input hinders the model from differentiating between the heterogeneous textual inputs of the text query $\mathbf{t}$ and the auxiliary caption $\mathbf{c}$. We also empirically 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 Bradley and Terry (1952). Typically, the objective of DPO can be written as follows:

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

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,



(a) MLLM-based retrieval model training (b) DG-DPO data construction

Figure 2: **Illustration of our CaRe-DPO framework.** (a) depicts the MLLM-based retrieval model where we propose to adopt retrieval role-embeddings $\mathbf{R}_c$ and $\mathbf{R}_t$ 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'`). (b) visualizes our DG-DPO mechanism for optimizing the captioning model, where each caption given to the video is evaluated with the retrieval relevance score s_p. During training, unlike SG-DPO, which adopts the local rank preferences, DG-DPO adopts the global rank preference, exploring across video-caption pairs.

β is a hyperparameter for regularizing the disparity of π_θ and π_{ref}, and σ denotes the sigmoid function.

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. In Sec. 4.1, we describe our retrieval backbone built upon an MLLM, which jointly encodes video and text inputs to compute cross-modal similarity, where we also introduce retrieval role-embeddings to differentiate the heterogeneous textual inputs. Then in Sec. 4.2, we present DG-DPO, a preference optimization method that supervises the captioning model to further align with the retrieval objective. The overall framework is illustrated in Fig. 2.

4.1 MLLM-based Retrieval Model

Training. Following recent trends in retrieval, we adopt MLLMs for text-video retrieval. However, we observe an interesting phenomenon: incorporating auxiliary captions into MLLM-based retrieval models often leads to only marginal performance gains. In particular, even when descriptive captions are provided, replacing them with random captions yields nearly identical retrieval performance (see Sec. F and E in the Appendix). This suggests that the model does not effectively distinguish the caption's role as auxiliary context from the text query as the retrieval target, resulting in inefficient use of the additional information. Hence, we adopt a simple yet effective retrieval role-embedding. Specifically, given the input triplet $(\mathbf{v}, \mathbf{c}, \mathbf{t})$, we introduce a new role-embeddings $R_c \in \mathbb{R}^D$ and $R_t \in \mathbb{R}^D$, 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:

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

where $\mathbf{R}_c = \mathbf{1}_{N_c} R_c^\top$ and $\mathbf{R}_t = \mathbf{1}_{N_t} R_t^\top$. Such a simple approach avoids $\mathcal{M}_{\text{ret}}$ to explicitly distinguish according to its roles: the caption as contextual knowledge and the text query as retrieval targets. Note that unless stated, $\mathbf{c}$ and $\mathbf{t}$ corresponds to $\mathbf{c} + \mathbf{R}_c$ and $\mathbf{t} + \mathbf{R}_t$ for the following notations.

Inference. 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 enables the retrieval model to be more sensitive to the subtle differences between the input and its output decision, thereby enhancing retrieval performance (refer to Table 13 in the Appendix).

4.2 DG-DPO

Retrieval score-driven preference dataset. To further improve auxiliary captions for retrieval, which arises from the misalignment between captioning evaluation and retrieval objectives, 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 the retrieval model $\mathcal{M}_{\text{ret}}(\cdot)$ to evaluate the quality of the sampled captions for video-text retrieval. Specifically, we adopt the relevance score between $\mathbf{c}_k^{(i)}$ and the $\mathbf{t}_k^{(i)}$ while masking the video tokens in the attention mask ($\mathbf{v}^*$), which empirically showed to be effective in terms of precision than that of un-masked video tokens (see Tab. 14 in the Appendix). Formally, the relevance score for preference optimization, s_p, is defined as:

$$s_{\text{p}}(\mathbf{v}^{(i)}, \mathbf{c}_k^{(i)}, \mathbf{t}^{(i)}) = \log \frac{P(y^+|\mathbf{v}^*, \mathbf{c}_k^{(i)}, \mathbf{t}^{(i)})}{P(y^-|\mathbf{v}^*, \mathbf{c}_k^{(i)}, \mathbf{t}^{(i)})} \tag{7}$$

Single-Group Direct Preference Optimization. Similar to the conventional DPO approach, we define Single-Group Direct Preference Optimization (SG-DPO) as the variant where preference pairs are constructed by comparing outputs conditioned on a single input *i.e.*, *local retrieval rank preferences*, as illustrated under 'local rank' in Fig. 2 (b). Specifically, given a single video $\mathbf{v}^{(i)}$ with its associated two sampled captions of 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.

Dual-Group Direct Preference Optimization. On the other hand, our proposed Dual-Group Direct Preference Optimization (DG-DPO) builds upon the SG-DPO, which 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 any k-th caption and video for the i-th sample, and the latter denote any k-th caption and the video for the j-th sample, the preference pair among the video-caption pair *i.e.*, $\mathbf{c}_w^{(i)}|\mathbf{v}^{(i)} \succ \mathbf{c}_l^{(j)}|\mathbf{v}^{(j)}$, satisfy the following condition:

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

Notably, the preference pairs of DG-DPO include cases satisfying both $i = j$ and $i \neq j$, whereas SG-DPO only considers sample pairs with $i = j$. Overall, the model learns to prefer video-caption pairs, which results in higher retrieval relevance scores, while considering both the local rank preference of the caption and the global rank preference across distinct video-caption pairs, enhancing the retrieval performance. Hence, $\mathcal{L}_{\text{DG-DPO}}$ can be written as:

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

Note $\lambda_{i,j}$ serves as a weighting factor that balances the contribution between pairs with $i = j$ and $i \neq j$. Importantly, this does not require additional training samples; instead, we reuse the pre-computed log probability values from the computation from when $i = j$ to compute $\mathcal{L}_{\text{DG-DPO}}$ for $i \neq j$. 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 substantial 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 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-

	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
NarVid (Hur et al., 2025)	53.4	79.1	86.3	-	-	-	52.7	77.7	85.6	-	-	-	-	-	-	51.1	76.8	85.2
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	**79.2**	**93.6**	**96.5**	**64.1**	**83.8**	**88.8**	**82.5**	**95.2**	96.3	**74.4**	**92.4**	**96.3**	**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 †.

Video retrieval benchmarks: DiDeMo (Anne Hendricks et al., 2017), ActivityNet (Caba Heilbron et al., 2015), and MSRVTT (Xu et al., 2016). More explanation of the datasets is in Sec. A of the Appendix. For evaluation, we adopt the standard retrieval metrics: Recall@K$\in \{1, 5, 10\}$. Note that for auxiliary captions, we sample two per instance and average the performance over those to mitigate the caption variability while providing more robust results. Unless otherwise specified, all experiments are conducted on DiDeMo, which serves as our primary benchmark. For additional validation, we also report results on the multi-text text-to-video retrieval benchmark MSVD (Chen and Dolan, 2011) in Sec. B of the Appendix, which further demonstrates the effectiveness of our CaRe-DPO.

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 MLLM-based retrieval model, capable of leveraging auxiliary captions, is built upon VideoChat-Flash-7B (Li et al., 2024c). For captioning, we adopt pretrained LLaVA-OneVision-7B (Li et al., 2024a), where we utilize 16 frames per video for all datasets, and further align the model with our DG-DPO. For efficiency, we apply LoRA (Hu et al., 2022) for parameter-efficient fine-tuning for both training. See Sec. C of Appendix for more details, including the training hyperparameters.

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 percentage increase of 14.7%,7.7%, and 4.1% in R@1 for DiDeMo, ActivityNet, and MSRVTT, respectively. To further validate the effectiveness of our framework across MLLM-based retrieval models, we compare against MM-Embed and LamRA. Notably, our CaRe-DPO shows superior performance with an average percentage increase in R@1 of 3.9%, 3.1%, and 5.1% compared to MM-Embed, and 2.9%, 6.3%, and 6.2% compared to LamRA across datasets. Overall, CaRe-DPO consistently outperforms the baselines, highlighting its effectiveness in enhancing text-video retrieval with MLLM-based models.

	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	78.9	74.1	63.4	63.3	(+0.9)
+ $\mathcal{L}_{\text{DG-DPO}}$	**85.1**	**82.5**	**79.2**	**74.4**	**64.1**	**63.8**	(+1.3)

Table 2: **Ablation on training objectives for $\mathcal{M}_{\text{cap}}$.** R@1 retrieval performance from different training objectives for $\mathcal{M}_{\text{cap}}$. 'Avg. Δ' denotes the R@1 point increase compared to the baseline across datasets.

5.3 Quantitative Analysis

Ablation on training objectives for $\mathcal{M}_{\text{cap}}$. In Tab. 2, we analyze different objectives for training the captioning model on the performance of text-video retrieval. As shown, simply fine-tuning on

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
	+ $\mathcal{L}_{\text{DG-DPO}}$	51.2	43.4	53.0	43.6	49.1	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 3: **Analysis on caption quality for retrieval.** 'Baseline' denotes zero-shot captions adopted for retrieval. For $s(\mathbf{c}, \mathbf{t})$, we adopt the model trained with $(\mathbf{v}, \mathbf{c}, \mathbf{t})$ while masking the video tokens. For $s(\mathbf{v}, \mathbf{c})$, we utilize the model trained solely on $(\mathbf{v}, \mathbf{t})$. We report R@1 performance for both T2V and V2T.

Train Input $\mathcal{L}_{\text{ret}}(\cdot)$	$\mathbf{R_c}$, $\mathbf{R_t}$	Inf. cap. $\mathbf{c}$	Text-to-Video R@1	Text-to-Video R@5	Text-to-Video R@10	Video-to-Text R@1	Video-to-Text R@5	Video-to-Text R@10	Avg. Δ
$(\mathbf{v}, \mathbf{c}, \mathbf{t})$	✗	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)
$(\mathbf{v}, \mathbf{c}', \mathbf{t}')$	✔	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 4: **Ablation on the role-embeddings of $\mathcal{M}_{\text{ret}}$.** We adopt the zero-shot captions with the standard inference strategy. 'rand' and 'orig.' denote *random* and *original* captions, respectively, and 'Inf.' denotes the inference stage. $\mathbf{c}'$ and $\mathbf{t}'$ denote $\mathbf{c} + \mathbf{R}_{\text{c}}$ and $\mathbf{t} + \mathbf{R}_{\text{t}}$, respectively. 'Avg. Δ' denotes an average change in R@k performance, between 'rand' and 'orig' captions.

the given dataset, denoted as $\mathcal{L}_{\text{SFT}}$ (row 2), results in an average R@1 improvement of 0.2 points, with a per-dataset increase of 1.0 for DiDeMo and a decrease of 0.2 for both ActivityNet and MSRVTT. In contrast, adopting our $\mathcal{L}_{\text{SG-DPO}}$ or $\mathcal{L}_{\text{DG-DPO}}$, which optimizes the model with DPO using retrieval scores for preference determination, results in superior performance. Specifically, $\mathcal{L}_{\text{SG-DPO}}$ (row 3), which relies on the local preference of the retrieval score, shows point increases of 2.1, 0.3, and 0.2 for DiDeMo, ActivityNet, and MSRVTT, respectively. By further considering the global preference based on retrieval scores, $\mathcal{L}_{\text{DG-DPO}}$ (row 4) achieves even higher retrieval precision, with point increases of 2.5, 0.8, and 0.8 compared to the baseline across the datasets. These results highlight the effectiveness of leveraging retrieval scores with DPO to better align generated captions for retrieval, and demonstrate that DG-DPO, which accounts for global preferences beyond local video-caption pairs, further improves performance.

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. 3: 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 captions generated from $\mathcal{M}_{\text{cap}}$ trained with $\mathcal{L}_{\text{DG-DPO}}$ result in consistent improvements across both retrieval tasks. Specifically, in T2C, the caption generated after adopting our DG-DPO yields an average increase of 7.6 points, especially in MSRVTT, with an 8.6 points increase in T2V and a 7.8 points increase in V2T. Also notable in ActivityNet with a 9.8 points and a 6.6 points increase in T2V and V2T, respectively. 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 an average 0.9, 0.8, and 0.9 points increase in R@1 on DiDeMo, ActivityNet, and MSRVTT, respectively.

Preference score	$\mathcal{M}_{\text{cap}}$	T2V	V2T	Avg. Confidence
BLEU	+ $\mathcal{L}_{\text{SG-DPO}}$	82.9	74.2	82.7
	+ $\mathcal{L}_{\text{DG-DPO}}$	83.7	75.4	83.1
Retrieval Score	+ $\mathcal{L}_{\text{SG-DPO}}$	83.7	75.4	86.9
	+ $\mathcal{L}_{\text{DG-DPO}}$	**84.5**	**77.4**	**89.6**

Table 5: **Effectiveness of DG-DPO in challenging retrieval cases.** Results of T2V and V2T R@1 in challenging retrieval cases with highly similar video candidates (pairwise average cosine similarity averaged over frames > 0.97).

Effectiveness of retrieval role-embeddings. Tab. 4 presents the impact of retrieval role-embeddings for the MLLM-based retrieval model. As shown, when replacing the caption with a random caption, the model shows a minimal performance drop of 0.1 in R@1 on average (compare rows 1 and 2). In contrast, using the same caption, training with role-embeddings results in a superior performance with 83.1 for T2V and 79.6 in V2T (compare rows 2 and 4) while showing higher sensitivity to the quality of the caption with a notable 1.8 improvement in average for R@1 compared to the one with a random caption (compare rows 3 and 4). These results highlight the effectiveness of role-embeddings in differentiating the two roles of auxiliary knowledge and retrieval target, leading to more accurate retrievals. Note that we use standard pretrained captions in this ablation to isolate the effect of retrieval role-embeddings; further experiments with DG-DPO-optimized captions are provided in Tab. 12 of the Appendix.

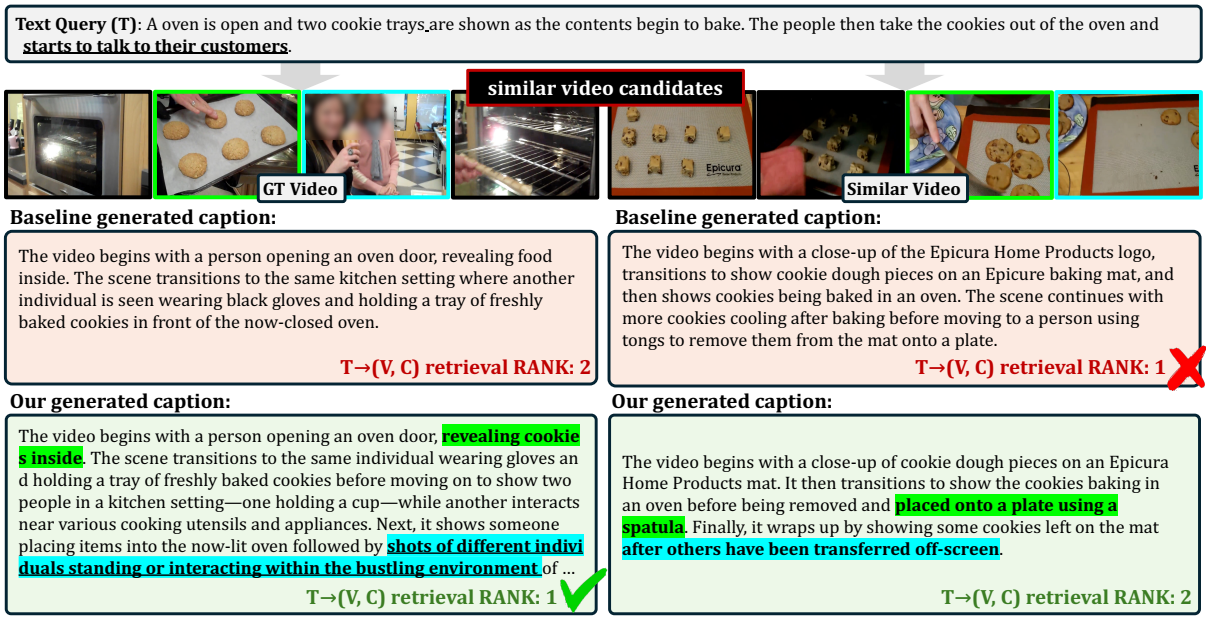


Figure 3: **Qualitative example on the effect of generated video caption.** Comparison of captions generated for visually similar videos on ActivityNet (left: ground-truth video, right: highly similar incorrect video) using the zero-shot captioning model (Baseline) and our DG-DPO-optimized model (Ours). Note the fine-grained details captured by our model but omitted in the baseline are highlighted in color. Additionally, the border color of each video frame corresponds to the caption depicted in our generated caption, and the underline denotes the fine-grained details that closely match the input text query (T). T $\rightarrow$ (V, C) denotes text-to-video retrieval, where C is the auxiliary caption for the given video. We also show the retrieval rank for each candidate with different generated captions. The qualitative results demonstrate that the DG-DPO-optimized caption improves retrieval, increasing the ground-truth video's rank from 2 to 1 in T2V.

Training Strategy	Self-BLEU (↓)	Distinct-1 (↑)	Distinct-2 (↑)
Zero-shot	0.50	0.08	0.41
+ $\mathcal{L}_{\text{SFT}}$	0.52	0.08	0.39
+ $\mathcal{L}_{\text{DG-DPO}}$	**0.47**	**0.10**	**0.43**

Table 6: **Caption diversity with different training strategies.** DG-DPO consistently improves diversity and reduces redundancy in caption generation.

Effectiveness of DG-DPO in challenging retrieval cases. Tab. 5 highlights the effectiveness of DG-DPO in retrieval scenarios involving highly similar video candidates, from which we left samples from the test set where any pairwise video cosine similarity (averaged over frames) is greater than 0.97. First, when comparing BLEU and retrieval score as the preference signal under SG-DPO, the retrieval score consistently yields higher average model confidence measured with Eq. 6 and scaled to a percentage (out of 100) for interpretability (86.9 vs. 82.7), indicating its stronger alignment with retrieval objectives. More importantly, DG-DPO further improves performance and model confidence over SG-DPO across both preference signals. For instance, using retrieval score supervision, DG-DPO boosts T2V and V2T R@1 from 83.7 to 84.5 and 75.4 to 77.4, respectively, and increases average confidence from 86.9 to 89.6. These results demonstrate that modeling group-level preferences across distinct video-caption pairs allows for providing more discriminative learning signals, leading to more accurate and confident retrieval in challenging cases.

Effect on caption diversity. Tab. 6 shows the impact of different training strategies on caption diversity measured with three different metrics of Self-BLEU (Zhu et al., 2018) that measure the redundancy across generated captions, and Distinct-1 and 2 (Li et al., 2015) that capture the ratio of unique unigrams and bigrams. While simple fine-tuning on the dataset (SFT) increases redundancy (0.52 vs. 0.50 in Self-BLEU compared to the zero-shot), likely due to overfitting to common patterns in the training data, our DG-DPO leads to significantly more diverse captions. It achieves the lowest Self-BLEU of 0.47 and the highest Distinct-1 of 0.10 and Distinct-2 of 0.43, indicating richer and less repetitive outputs. This suggests that DG-DPO not only improves discriminative supervision but also encourages the model to generate more detailed and distinctive descriptions that are essential for visually similar videos.

5.4 Qualitative Results

Fig. 3 presents a comparison between captions generated by a zero-shot captioning model (baseline) and our DG-DPO-trained model (ours) for two visually similar videos depicting cookie baking. As illustrated, the baseline captions provide general scene descriptions, whereas our model generates more detailed and context-specific captions that highlight key visual cues such as 'revealing cookies inside,' 'placed onto a plate using a spatula,' and 'individuals standing or interacting within the bustling environment.' For the given query, the retrieval model initially ranked a visually similar but incorrect video higher (left video in Fig. 3), which lacked the scene where 'people talk to their customers'. However, after substituting the caption with one generated by ours, the retrieval model correctly retrieved the ground-truth video, guided by the discriminative details in the caption that closely match the text query (underlined in the figure). This demonstrates that the fine-grained captions generated by the DG-DPO optimized captioning model facilitate better differentiation between similar content, leading to improved retrieval performance. More qualitative results in Sec. H of the Appendix.

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 Dual-Group Direct Preference Optimization aligns caption generation with retrieval relevance scores while leveraging both local and global ranks. Through extensive experiments, we demonstrate that CaRe-DPO enhances overall retrieval accuracy across benchmarks.

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. Moreover, unlike simply training the retrieval model, ours requires training both the retrieval and captioning models and generating multiple captions for DPO training, which increases overall training time, yet results in improved retrieval performance.

Acknowledgments

This work was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2023R1A2C2005373), and Hanwha Vision.

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*.

David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *ACL*.

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*.

Dohwan Ko, Sihyeon Kim, Yumin Suh, Minseo Yoon, Manmohan Chandraker, Hyunwoo J Kim, and 1 others. 2025a. St-vlm: Kinematic instruction tuning for spatio-temporal reasoning in vision-language models. *arXiv preprint arXiv:2503.19355*.

Dohwan Ko, Ji Soo Lee, Minhyuk Choi, Zihang Meng, and Hyunwoo J Kim. 2025b. Bidirectional likelihood estimation with multi-modal large language models for text-video retrieval. In *ICCV*.

Dohwan Ko, Ji Soo Lee, Wooyoung Kang, Byungseok Roh, and Hyunwoo J Kim. 2023. Large language models are temporal and causal reasoners for video question answering. In *EMNLP*.

Ji Soo Lee, Jongha Kim, Jeehye Na, Jinyoung Park, and Hyunwoo J Kim. 2025. Vidchain: Chain-of-tasks with metric-based direct preference optimization for dense video captioning. In *AAAI*.

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*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

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*.

Jinyoung Park, Minseong Bae, Dohwan Ko, and Hyunwoo J Kim. 2024a. Llamo: Large language model-based molecular graph assistant. In *NeurIPS*.

Jinyoung Park, Jeehye Na, Jinyoung Kim, and Hyunwoo J Kim. 2025. Deepvideo-r1: Video reinforcement fine-tuning via difficulty-aware regressive grpo. In *NeurIPS*.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024b. 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*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *SIGIR*.

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.

MSVD. The MSVD (Chen and Dolan, 2011) dataset consists of 2k videos, of which each video is annotated with around 40 captions. Unlike previously mentioned, MSVD is a multi-text text-video retrieval benchmark where it treats each sentence as an independent sample. Specifically, it evaluates with one-to-many ground-truth text for video-to-text retrieval. 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 1k videos for training, and 670 videos for testing.

Method	T2V R@1	V2T R@1
Non-MLLM-based		
MV-Adapter (Jin et al., 2024)	49.4	71.8
NarVid (Hur et al., 2025)	53.1	-
InternVideo (Wang et al., 2022)	58.4	76.3
UMT (Li et al., 2023)	58.2	82.4
InternVideo2 1B* (Wang et al., 2024d)	59.0	**85.5**
InternVideo2 6B (Wang et al., 2024d)	**61.4**	85.2
MLLM-based†		
MM-Embed (Lin et al., 2025)	59.5	**85.5**
LamRA (Liu et al., 2025)	59.0	**85.5**
CaRe-DPO (Ours)	**59.8**	**85.9**

Table 7: **Comparison with SOTA retrieval models on MSVD.** We report R@1 for both T2V and V2T retrieval. Note that best values are **bold**, second-best are **bold and underlined**. † denotes reproduced on the same baseline as ours, VideoChat-Flash.

	T2V	V2T
Baseline	59.0	85.5
+ $\mathcal{L}_{\text{SFT}}$	59.7	85.5
+ $\mathcal{L}_{\text{SG-DPO}}$	59.7	85.6
+ $\mathcal{L}_{\text{DG-DPO}}$	**59.8**	**85.9**

Table 8: **Ablation on training objectives for $\mathcal{M}_{\text{cap}}$ on MSVD.** R@1 retrieval performance for T2V and V2T.

$s(\mathbf{c}, \mathbf{t})$	T2V	V2T
Baseline	49.5	76.6
+ $\mathcal{L}_{\text{DG-DPO}}$	**51.7**	**79.0**

Table 9: **Analysis on caption quality for retrieval on MSVD.** We report R@1 for T2V and V2T.

B Multi-Text Text-Video Retrieval

Comparison with SOTA models on MSVD. Tab. 7 presents a comparison of state-of-the-art text-to-video (T2V) and video-to-text (V2T) retrieval performance on the MSVD, a multi-text text-video retrieval benchmark, measured using R@1. When adopting our CaRe-DPO, the model achieves strong performance, especially in V2T, attaining 85.9, outperforming other MLLM-based approaches as well as non-MLLM-based models.

Ablation on training objectives on MSVD. Tab. 8 presents an ablation study on the impact of different training objectives for $\mathcal{M}_{\text{cap}}$ on MSVD. We observe that incorporating the supervised fine-tuning loss ($\mathcal{L}_{\text{SFT}}$) yields a slight improvement for T2V retrieval while V2T remains unchanged. Introducing our ($\mathcal{L}_{\text{SG-DPO}}$ and $\mathcal{L}_{\text{DG-DPO}}$) further enhances performance, with $\mathcal{L}_{\text{DG-DPO}}$ achieving the best results of 59.8 for T2V and 85.9 for V2T.

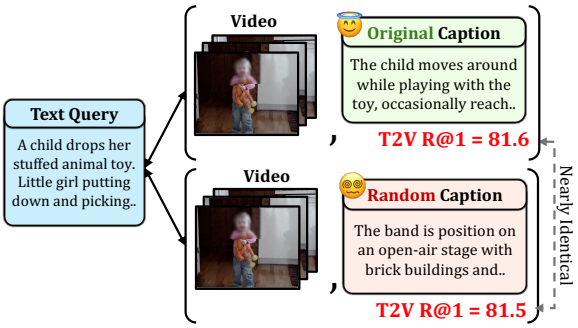


Figure 4: T2V retrieval with the *original* descriptive caption (video-to-caption retrieval R@1 of 90.7) compared to the *random* one. Nearly identical performance suggests that the model fails to effectively leverage the auxiliary knowledge.

Analysis on caption quality on MSVD. Tab. 9 analyzes the effect of caption quality on retrieval performance using the text-caption retrieval $s(\mathbf{c}, \mathbf{t})$, measuring how well the auxiliary caption semantically aligns with the query. Compared to the baseline, adding $\mathcal{L}_{\text{DG-DPO}}$ significantly improves R@1 metrics for both T2V and V2T, reaching 51.7 and 79.0, respectively.

C 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:* `[text query]`. *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:* `[text query]`. *Based on the video and its description, is the video relevant to the caption? Answer True or False*." Again, 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. Note that $\lambda_{i,j}$ in Eq. 10 denotes the weighting factor that balances the contribution between local pairs $(i = j)$ and cross-group pairs $(i \neq j)$. Specifically, we denote the former as $\lambda_{i=j}$ and the latter as $\lambda_{i\neq j}$. The value of $\lambda_{i\neq j}$ adopted for our experiments are in Tab 10, and $\lambda_{i=j}$ is $(1 - \lambda_{i\neq j})$.

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

	DiDeMo		ActivityNet		MSRVTT		MSVD	
	$\mathcal{M}_{\text{ret}}$	$\mathcal{M}_{\text{cap}}$	$\mathcal{M}_{\text{ret}}$	$\mathcal{M}_{\text{cap}}$	$\mathcal{M}_{\text{ret}}$	$\mathcal{M}_{\text{cap}}$	$\mathcal{M}_{\text{ret}}$	$\mathcal{M}_{\text{cap}}$
Learning rate	8e-5	8e-6	2e-5	8e-6	1e-4	8e-6	2e-4	8e-6
Warmup Epochs	1	0.1	1	0.1	1	0.1	1	0.1
Epoch	5	1	5	1	3	1	5	1
Batch Size	32	8	32	8	512	8	32	8
LoRA r	8	64	8	64	8	64	8	64
LoRA α	32	128	32	128	32	128	32	128
$\lambda_{i\neq j}$	-	0.3	-	0.1	-	0.3	-	0.1
β	-	0.1	-	0.1	-	0.1	-	0.1
γ	-	0.7	-	2.0	-	0.5	-	0.01

Table 10: Training hyperparameters for $\mathcal{M}_{\text{ret}}$ with $\mathcal{L}_{\text{ret}}$ and $\mathcal{M}_{\text{cap}}$ with $\mathcal{L}_{\text{DG-DPO}}$.

retrieval scores across both in order to account for the variability in caption generation and provide a more robust performance estimate.

Training strategy 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.

D Hyperparameters

In Tab. 10, 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.

E Further Ablation on Role-Embeddings.

Ablation on each component. We conduct further analysis on the role-embeddings for text-video retrieval on DiDeMo, evaluating with and without each function role embedding in Tab. 11. The results suggest that the model trained with R_{t} results in higher V2T retrieval at R@1 (79.2 to 79.8), whereas the model trained with R_{c} results in higher

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 11: **Ablation on each component of the role-embedding.** Note that we adopt the zero-shot captions with the standard inference strategy.

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.

Ablation with DG-DPO optimized caption. In Tab. 12, we analyze the effectiveness of retrieval role-embeddings when adopting the caption that is generated by our DG-DPO optimized captioning model, similar to Tab. 4 in the main. As shown, when replacing the caption with a random one, the model shows a minimal performance drop of 0.7 in R@1 on average (compare rows 1 and 2), whereas the model trained with role-embeddings shows higher sensitivity to the quality of the caption with a notable 1.6 point change in performance of R@1 (compare rows 3 and 4). Moreover, while adopting the same caption, the model trained with the role-embeddings yields superior performance, for instance, 85.1 compared to the baseline of 84.9 (compare rows 2 and 4).

Analysis on the inference strategy. Tab. 13 explores the different inference strategies in MLLM retrieval, and we determine that our contrastive in-

Train Input $\mathcal{L}_{\text{ret}}(\cdot)$	$\mathbf{R}_c$, $\mathbf{R}_t$	Inf. cap. $\mathbf{c}$	Text-to-Video R@1	R@5	R@10	Video-to-Text R@1	R@5	R@10	Avg. Δ
$(\mathbf{v}, \mathbf{c}, \mathbf{t})$	✗	rand.	81.6	94.9	96.4	83.4	94.1	95.9	-
	✗	ours	84.9	95.4	96.2	82.1	95.3	96.4	(+0.7)
$(\mathbf{v}', \mathbf{c}', \mathbf{t}')$	✔	rand.	80.9	94.6	96.3	80.9	95.1	96.1	-
	✔	ours	85.1	95.0	96.2	82.5	95.2	96.3	(+1.6)

Table 12: **Ablation on the Role-embeddings of $\mathcal{M}_{\text{ret}}$ with DG-DPO optimized caption.** We adopt our DG-DPO optimized captions, denoted with 'our', with the contrastive inference strategy. 'rand.' denotes *random* caption and 'Inf.' denotes the inference stage. $\mathbf{c}'$ and $\mathbf{t}'$ denote $\mathbf{c}+\mathbf{R}_c$ and $\mathbf{t}+\mathbf{R}_t$, respectively. 'Avg. Δ' denotes an average change in R@k performance, between 'rand' and 'ours' captions.

$s(\mathbf{v}, \mathbf{c}, \mathbf{t})$	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 13: **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.

	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
$\log \frac{P(y^+\|\mathbf{v}^*, \mathbf{c}, \mathbf{t})}{P(y^-\|\mathbf{v}^*, \mathbf{c}, \mathbf{t})}$	85.1	95.0	96.2	82.5	95.2	96.3

Table 14: **Comparison on adopting different preference scores s_p for constructing $\mathcal{D}_{\text{DG-DPO}}$.** We report the retrieval performance on DiDeMo. Also, $\mathbf{v}^*$ denotes masked attention for video tokens.

ference 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.

F Further Ablation on DG-DPO

Ablation on preference scores. In Tab 14, we compare retrieval performance 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, adopting BLEU and METEOR leads to a performance drop of 1.0 and 1.3 points in T2V R@1, respectively. For V2T, the impact of caption quality appears marginal, as captions primarily augment the video rather than the text query, contributing more significantly to improvements in T2V.

Method	T2V R@1	V2T R@1
Baseline	83.1	79.6
$+\mathcal{L}_{\text{SFT}}$	82.6	82.0
$+\mathcal{L}_{\text{DG-DPO}}$ (Batch=4)	84.8	82.6
$+\mathcal{L}_{\text{DG-DPO}}$ (Batch=8)	84.2	**82.8**
$+\mathcal{L}_{\text{DG-DPO}}$ (Batch=16)	84.2	82.5
$+\mathcal{L}_{\text{DG-DPO}}$ (Batch=32)	**85.1**	82.5

Table 15: **Effect of DG-DPO with varying batch sizes.** We report R@1 for text-to-video (T2V) and video-to-text (V2T) retrieval.

Training Dataset	Test Dataset	Preference score	T2V R@1	R@5	R@10	V2T R@1	R@5	R@10
DiDeMo	MSRVTT	BLEU	63.9	82.9	87.2	63.7	83.0	87.2
DiDeMo	MSRVTT	retrieval	64.2	83.8	88.5	63.9	83.1	87.3
MSRVTT	MSRVTT	retrieval	64.1	83.8	88.8	63.8	83.0	87.3

Table 16: **Cross-dataset generalization of DG-DPO captions.** To evaluate out-of-domain generalization of $\mathcal{M}_{\text{cap}}$ with DG-DPO on the DiDeMo dataset and test retrieval performance on MSRVTT, which differs in video length and query granularity.

Ablation on DG-DPO batch size. To analyze the effect of batch size in preference optimization, we conduct an ablation study by applying DG-DPO with varying batch sizes during training. As shown in Table 15, all DG-DPO variants outperform the baseline and SFT-only models, demonstrating the effectiveness of direct preference optimization. Notably, performance improves consistently as the batch size increases, with the best results observed at a batch size of 32, achieving 85.1 R@1 in text-to-video and 82.5 R@1 in video-to-text retrieval. This highlights the benefit of utilizing diverse preference pairs per update and suggests that larger batch sizes help the model better capture fine-grained semantic preferences aligned with retrieval objectives.

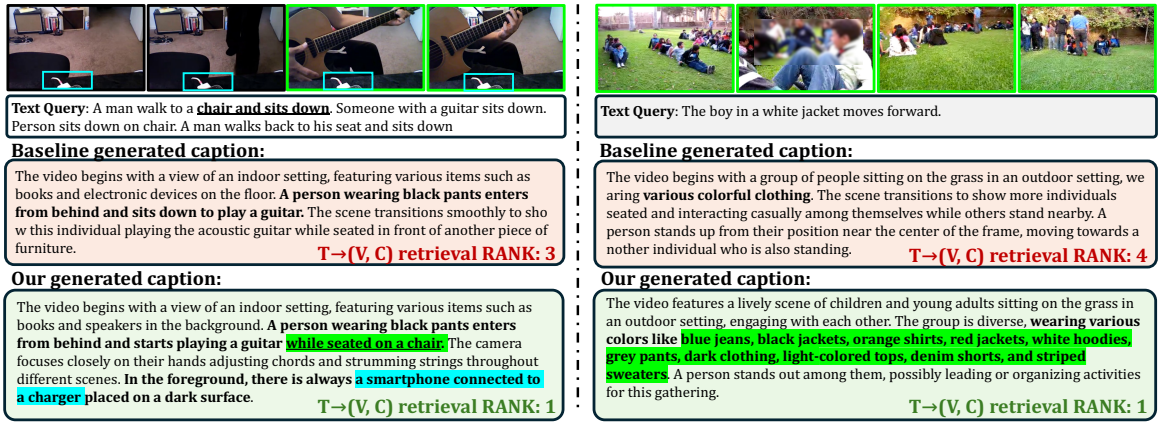
G Out-of-Domain Generalization

To assess the out-of-domain generalization ability of our method, we conducted a cross-dataset evaluation by training the captioner with DG-DPO on DiDeMo and evaluating retrieval performance

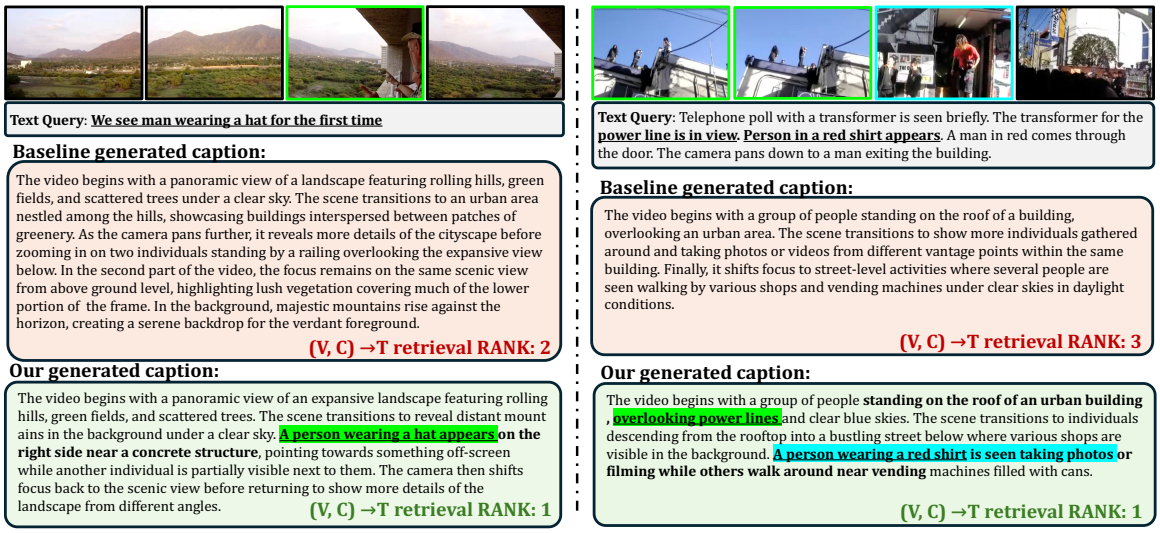
on the MSRVTT benchmark, of which the former contains longer videos with more moment-specific queries, whereas the latter contains shorter videos with queries that refer to the entire video. As shown in Tab 16, DG-DPO-trained captions generalize well to unseen datasets, even outperforming the performance when trained with the specific dataset (see row 2 and row 3). In addition, retrieval score-based optimized captions outperform BLEU-optimized captions even when trained out-of-domain. This confirms that DG-DPO captures meaningful retrieval-aligned semantics rather than overfitting to domain-specific biases.

H Further Qualitative Results

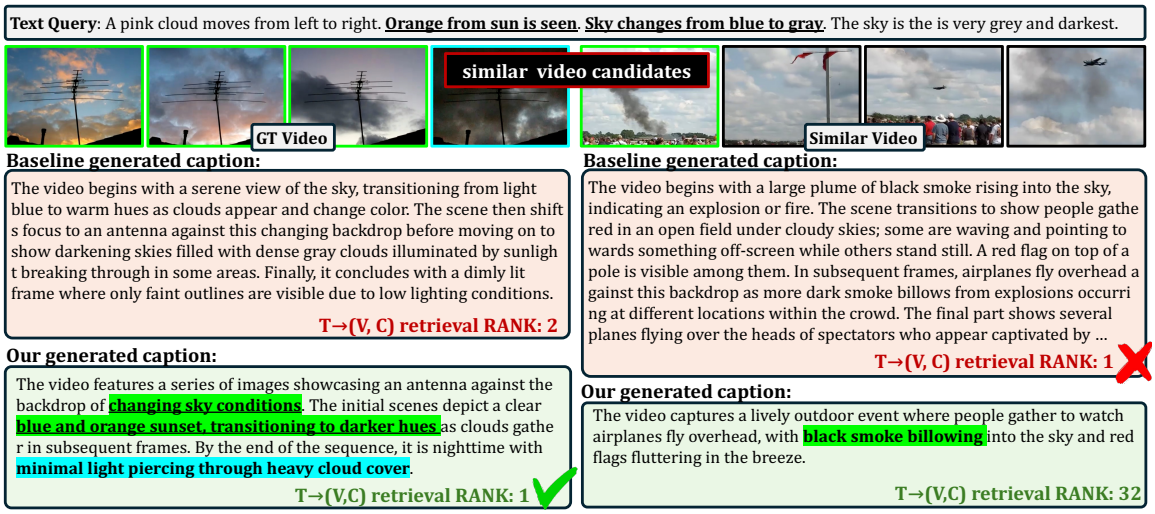
In Fig 5, Fig 6, we further show the qualitative results between captions from the baseline and our model on DiDeMo and ActivityNet, focusing on those benchmarks that require fine-grained detail for retrieval. All figures (a) depict examples for text-to-video retrieval, whereas (b) depict examples of video-to-text retrieval. (c) illustrates cases where, when adopting a baseline (zero-shot) caption, the model confuses between the two similar videos for text-to-video retrieval, whereas with DG-DPO optimized captions, the fine-grained details successfully lead to discriminating between the videos. Overall, our model trained with the DG-DPO of our CaRe-DPO enables the retrieval model to better align with the text-video retrieval task.



(a) Qualitative example on the effect of generated video caption in **text-to-video** retrieval on DiDeMo.

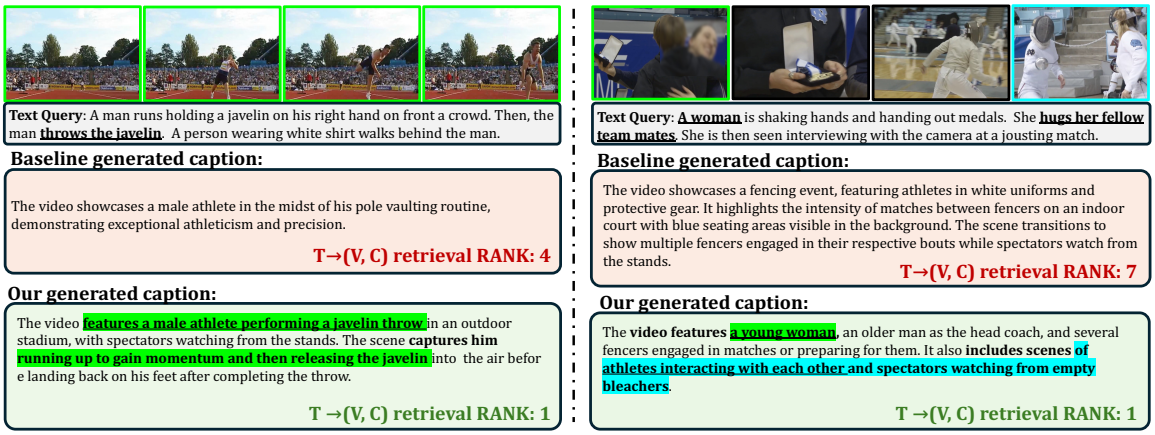


(b) Qualitative example on the effect of generated video caption in **video-to-text** retrieval on DiDeMo.

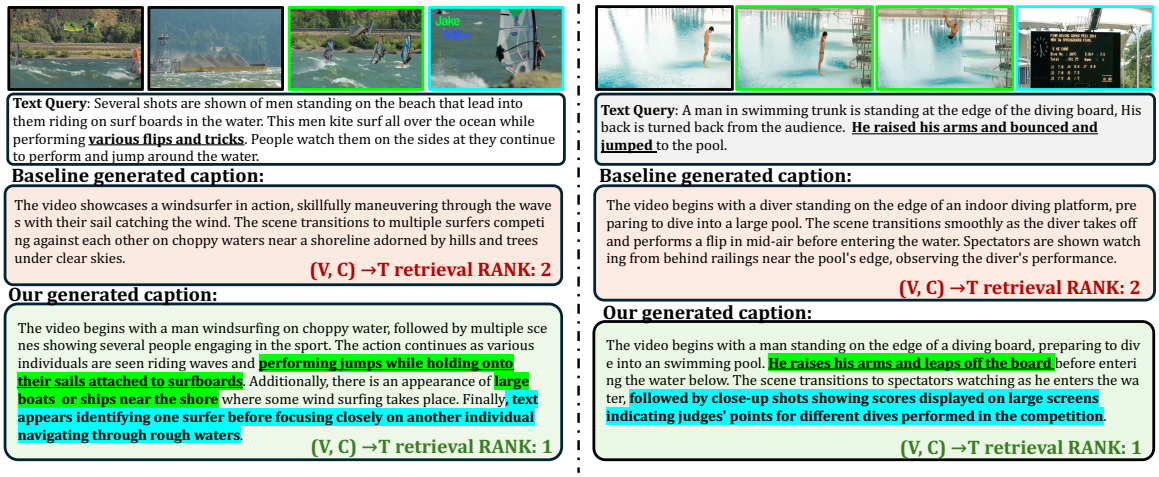


(c) Qualitative example on the effect of generated video caption in **text-to-video** retrieval on DiDeMo for Challenging Retrieval Cases.

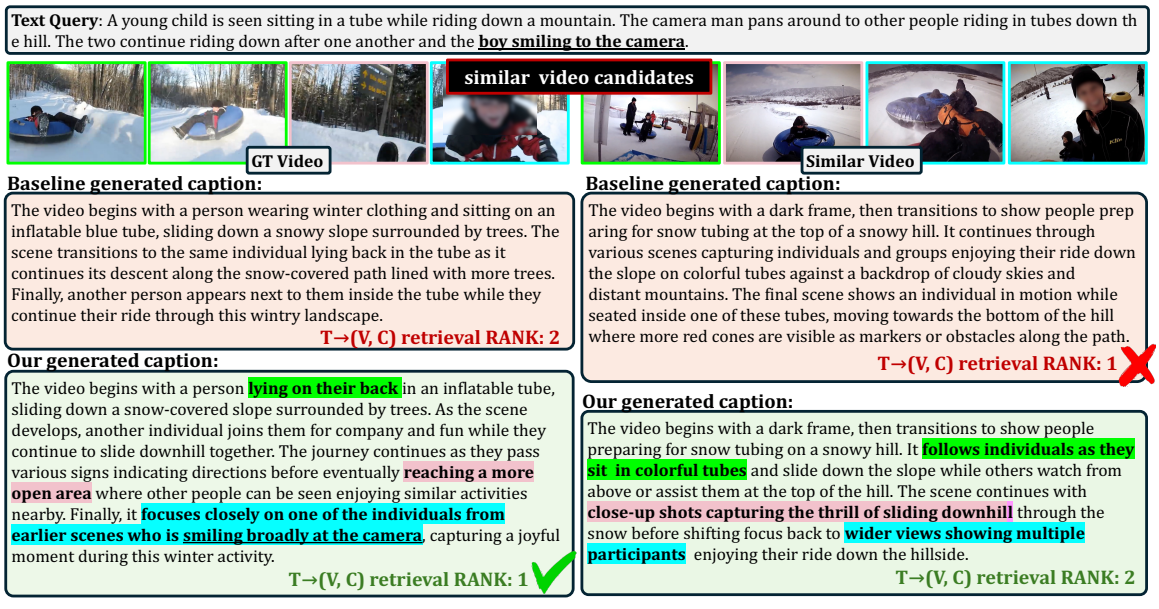
Figure 5: **Further qualitative example of video captioning.** Comparison of the predictions of the caption generated by the zero-shot captioning model (Baseline) with our model trained with DG-DPO (Ours) on DiDeMo. The highlighted parts depict the fine-grained detail generated by our model, which is not provided in the caption generated by the baseline. The border color of each video frame corresponds to the caption highlighted, and the underline denotes the details that closely match the given query.



(a) Qualitative example on the effect of generated video caption in **text-to-video** retrieval on ActivityNet.



(b) Qualitative example on the effect of generated video caption in **video-to-text** retrieval on ActivityNet.



(c) Qualitative example on the effect of generated video caption in **text-to-video** retrieval on ActivityNet for Challenging Retrieval Cases.

Figure 6: **Further qualitative example of video captioning.** Comparison of the predictions of the caption generated by the zero-shot captioning model (Baseline) with our model trained with DG-DPO (Ours) on ActivityNet. The highlighted parts depict the fine-grained detail generated by our model, which is not provided in the caption generated by the baseline. The border color of each video frame corresponds to the caption highlighted, and the underline denotes the details that closely match the given query.