# DIANE: Zero-Shot Video Retrieval via Index Time Alignment and Enrichment

**Anonymous ACL submission**

## Abstract

While recent progress in video retrieval has been advanced by the exploration of supervised representation learning, regarded as a strategy for training time alignment, in this paper, we focus on index time alignment, by transforming the video to text, bridging the representation gap between the video and query. However, naively generating captions from videos is sub-optimal – captions generated from the videos often miss crucial details and nuances. In this work, we take a step further by exploring the index time enrichment strategy – enhancing the text representation of video with diverse information. Specifically, we design a novel relevance-boosted caption generation method, bringing extra relevant details into video captions by using LLMs. To emphasize key information, we also extract key visual tokens from captions and videos. Moreover, to highlight the unique characteristics of each video, we propose a distinctiveness analysis method that infuses the key features into text representation. Benefiting from these methods, extensive experiments on several video retrieval benchmarks demonstrate the superiority of DIANE over existing fine-tuned and pretraining methods without any data. A comprehensive study with both human and automatic evaluations shows that the enriched captions capture the key details and barely bring noise to the captions. Codes and data will be released.

## 1 Introduction

Video Retrieval (Luo et al., 2022; Gao et al., 2021; Ma et al., 2022; Liu et al., 2022a; Zhao et al., 2022; Gorti et al., 2022; Fang et al., 2022) is to select the corresponding video from a pool of candidate videos given a text query. Recent years have witnessed the rapid development of VR with the support from powerful pretraining models (Luo et al., 2022; Gao et al., 2021; Ma et al., 2022; Liu et al., 2022a), improved retrieval methods (Bertasius et al., 2021; Dong et al., 2019; Jin et al., 2021), and video-language datasets construction (Xu et al., 2016). However, it remains challenging to precisely match video and language due to the raw data being in heterogeneous spaces and the use of modality-specific encoders.

One popular paradigm for video retrieval (Luo et al., 2022; Ma et al., 2022; Liu et al., 2022b) is on the **training time alignment**, which is to firstly learn a joint feature space across modalities and then compares representations in this space. However, with the discrepancy between different modalities and the design of modality-independent encoders, it is challenging to directly match representations of different modalities generated from different encoders (Liang et al., 2022). On the other hand, pioneering works (Wang et al., 2021, 2022e) explored **index time alignment**, converting images into captions for better presentation learning on image-language tasks, demonstrating that captioners can mitigate modality discrepancy.

Inspired by the trade-off between the training time scaling (Kaplan et al., 2020) and the test time scaling (Snell et al., 2024), we believe that leveraging more computation in indexing time can further boost performance. However, a naive strategy of translating video candidates to captions may not be optimal – the captioners often miss important information in the video, thus leading to poor retrieval performance. In this work, to take one step forward, building on top of indexing time alignment, we explore the index time enrichment, to further enhance the representation of video in the text modality.

To achieve index time alignment, we first generate video captions for videos, which can be directly used for retrieval. However, we notice that the captions might miss important information in the video, thus leading to unsatisfying retrieval performance (see Table 1). To this end, we propose three zero-shot strategies for **index time enrichment**, including caption enrichment, extracting visual tokens from captions and videos, and distinctiveness analysis. Specifically, caption enrichment augments video captions by encouraging large language models (LLMs) to add relevant details to captions. Moreover, to emphasize key entities, *e.g.*, objects, relationships, attributes, and phrases, we extract visual tokens from captions and videos and utilize them for detailed descriptions. We also propose a distinctiveness analysis method to identify key distinctive features among similar videos. Finally, DIANE utilizes off-the-shelf text retrieval methods, *e.g.*, BM25, for zero-shot text retrieval matching video captions enriched by the proposed methods.

In summary, our contributions are as follows:

- We propose a zero-shot video retrieval method focusing on test time alignment without requiring any training procedure or human-annotated data, only using the off-the-shelf captioning method and large language models.
- Our proposed DIANE achieves SOTA performance on several metrics across three video retrieval benchmarks, outperforms previous methods, including fine-tuning methods and few-shot methods.
- Detailed analysis reveals the effectiveness of different indexing time enrichment strategy. We will open-source the code and data to facilitate future research.

## 2 Related Work

**Video retrieval**, which involves cross-modal alignment and abstract understanding of temporal images (videos), has been a popular and fundamental task of language-grounding problems (Wang et al., 2020a,b, 2021; Yu et al., 2023). Most of the existing video retrieval frameworks (Yu et al., 2017; Dong et al., 2019; Zhu and Yang, 2020; Miech et al., 2020; Gabeur et al., 2020; Dzabraev et al., 2021; Croitoru et al., 2021) focus on learning powerful representations for video and text and extracting separated representations. For example, in Dong et al. (2019), videos and texts are encoded using convolutional neural networks and a bi-GRU (Schuster and Paliwal, 1997) while mean pooling is employed to obtain multi-level representations. MMT (Gabeur et al., 2020) uses a cross-modal encoder to aggregate features extracted by temporal images, audio, and speech for encoding videos. Following that, MDMMT (Dzabraev et al., 2021) further utilizes knowledge learned from multi-domain datasets to improve performance empirically. Further, MIL-NCE (Miech et al., 2020) adopts Multiple Instance Learning and Noise Contrastive Estimation, addressing the problem of visually misaligned narrations from uncurated videos.

Recently, with the success of self-supervised pretraining methods (Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020), vision-language pretraining (Li et al., 2020b; Gan et al., 2020; Singh et al., 2022) on large-scale unlabeled cross-modal data has shown promising performance in various tasks, *e.g.*, image retrieval (Radford et al., 2021), image captioning (Chan et al., 2023), and video retrieval (Luo et al., 2022; Wang and Shi, 2023a). Recent works (Lei et al., 2021; Cheng et al., 2021; Gao et al., 2021; Ma et al., 2022; Park et al., 2022a; Wang et al., 2022b,d; Zhao et al., 2022; Gorti et al., 2022) have attempted to pretrain or fine-tune video retrieval models in an end-to-end manner. CLIP-BERT (Lei et al., 2021; Bain et al., 2021), as a pioneer, proposes to sparsely sample video clips for end-to-end training to obtain clip-level predictions and then summarize them. Frozen in time (Bain et al., 2021) uses end-to-end training on both image-text and video-text pairs data by uniformly sampling video frames. CLIP4Clip (Luo et al., 2022) finetunes models and investigates three similarity calculation approaches for video-sentence contrastive learning on CLIP (Radford et al., 2021). Further, TS2-Net (Liu et al., 2022b) proposes a novel token shift and selection transformer architecture that adjusts the token sequence and selects informative tokens in both temporal and spatial dimensions from input video samples. While the mainstream of VR models (Xue et al., 2023; Wu et al., 2023) focuses on fine-tuning powerful image-text pre-trained models, on the other side, as a pioneer, (Tiong et al., 2022; Wang et al., 2022e) propose to use large language models (LLMs) for zero-shot video question answering.

**Zero-shot cross-modal retrieval.** With the huge success of pretrained visual-language model (Radford et al., 2021; Luo et al., 2022), zero-shot cross-modal retrieval has attracted more and more research interest recently. Due to the powerful representation learning ability in image and text domains, CLIP (Radford et al., 2021) achieves satisfying zero-shot retrieval performance on several representative image-text retrieval bench-

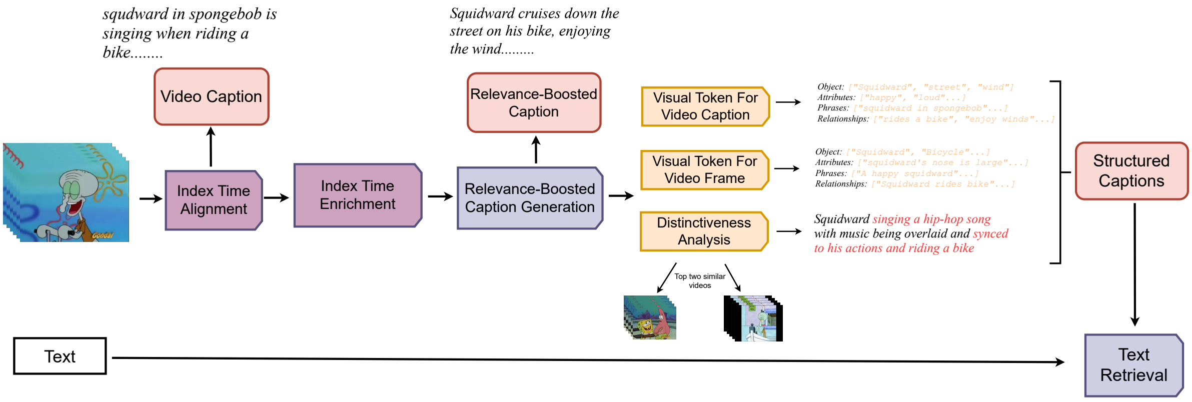


Figure 1: The illustration of our proposed DIANE. DIANE includes four steps. First, we implement index time alignment to generate video captions for video using off-the-shelf video captioning methods. Then, to enrich the captions and emphasize the important information in the captions, we propose an index time enrichment approach including relevance-boosted caption generation, extracting visual tokens from video captions and frames, and distinctiveness analysis. Finally, after obtaining structured video captions, we employ off-the-shelf text retrieval methods to perform zero-shot video retrieval.

marks (Huiskes and Lew, 2008; Lin et al., 2014). Inspired by this achievement, Liu et al. (2023a,b); Chen et al. (2023c); Liu et al. (2024); Guo et al. (2024) boost the performance of zero-shot image-text retrieval by better representation learning methods. On the other side, benefiting from large-scale video benchmarks (Xu et al., 2016; Chen and Dolan, 2011; Fabian Caba Heilbron and Niebles, 2015), video-language pre-trained models (Wang et al., 2022c; Chen et al., 2023a; Xu et al., 2023; Chen et al., 2023c; Li et al., 2023b; Liu et al., 2023c; Zhu et al., 2024) also achieve satisfying zero-shot video retrieval results.

## 3 Index Time Alignment

Instead of aligning the modality representation during the training time, we explore how to bridge the modality gap during the index time. One intuitive solution is to leverage the video captioning technique to translate the video into text.

Specifically, we employ Tewel et al. (2021, 2022) to generate video captions and then use GPT-2 (Radford et al., 2019) to enrich sentences using the prompts, *i.e.*, "Video presents".

## 4 Index Time Enrichment

Vanilla video captioning is deficient, since important details are often missed in captions. In this work, we further explore several strategies for representation enrichment via augmenting with relevance-boosted captions, visual tokens, and distinctiveness analysis.

### 4.1 Relevance-Boosted Caption Generation

As shown in Figure 3, we notice that the generated captions often miss some important information, leading to unsatisfying retrieval performance. A simple solution to this problem is to fine-tune the captioning models, which will improve their caption-generation abilities. However, this approach needs a huge amount of annotated video-caption data and expensive computation resources, and the fine-tuned models do not always generalize well (Tang et al., 2021). To this end, we propose the **relevance-boosted caption generation**, which is training-free and generates detailed captions that contain almost every detail of the video.

Specifically, we use large language models (LLMs) (Brown et al., 2020; Touvron et al., 2023) to conduct the relevance-boosted generation using the following prompt template.

```
The following is a caption from a
video: [" + <Video Caption> + "].
Based on this caption, generate two
paraphrased captions capturing the
key information and main themes,
each of which should be in one
sentence with up to twenty words.
Meanwhile, please be creative, you
can have some imagination and add
the necessary details. Generated
sentences should be in the number
list. Also please generate text
without any comment.
```

By scaling up the index time computation, we generate multiple relevance-boosted captions from LLM. However, some of these captions might introduce noise or lack strong relevance to the video's content. To mitigate potential negative impacts, we apply a filtering method to assess the semantic similarity between relevance-boosted captions and the original video caption by leveraging a pre-trained text encoder (Reimers and Gurevych, 2019). Specifically, each video in our dataset has two generated captions. For the retrieval process, we concatenate these captions for each video and then perform the ranking.

## 4.2 Visual Token for Caption

To understand what kind of information is essential to video retrieval, we analyze the contextual text of video captions by breaking down the video captions into four different visual tokens using model en_core_web_sm from the SPACY (Neumann et al., 2019), *i.e.*, phrase, object, relationship, and attribute. Finally, we structure the information into the following structure,

```
<Caption> <Phrases> <Attributes> <
Relationships> <Objects>
```

## 4.3 Visual Token for Video Frame

We propose a systematic approach to extract and structure information from video scenes using the Qwen2.5-7B-VLM model (Qwen et al., 2025). Video frames are uniformly sampled at 5 frames per second (fps), and we select a representative frame every five frames to balance action continuity and keyframe retention (Truong and Venkatesh, 2007). This ensures temporal coherence and preserves salient visual tokens.

**Visual Token Extraction**: The Qwen/Qwen2.5-VL-7B-Instruct model generates structured visual tokens for each frame using a predefined prompt. The visual tokens for the video frame include objects, attributes, relationships, and phrases which are serialized into a structured JSON format for downstream analysis:

```
Extract the information from this
image, Include:
Objects: List all visible objects
Attributes: Describe properties of
objects (color, size, texture, etc.)
Relationships: Describe spatial and
action relationships between objects
Phrases: Key descriptive phrases
about the scene;
Provide the output strictly as a
JSON list with the following format.
```json
{ "Objects": ["object1", "object2",
...], "Attributes": ["attribute1", "
attribute2", ...], "Relationships":
["relationship1", "relationship2",
...], "Phrases": ["phrase1", "
phrase2", ...], }
```

## 4.4 Distinctiveness Analysis

While videos may share common elements, identifying the unique and distinctive features of a specific video is valuable. To identify the unique characteristics of a video, we propose a distinctiveness analysis method. First, we leverage the video captions obtained in Section 4.1 and obtain captions embeddings using the Sentence Transformer (Reimers and Gurevych, 2019). For each video, we use cosine similarity to identify the most similar videos. We further leverage LLMs to contrast the video against others, especially the most similar ones, highlighting its distinctive features with text representation.

Specifically, we use the captions extracted in Section 4.1 and feed them into the Qwen/Qwen2.5-VL-7B-Instruct model to generate sentences revealing the uniqueness of each video. The prompt for this process is structured as follows.

```
Given the frame images from the
original video, as well as from
similar videos 1 and 2, and the
corresponding video descriptions:

 Current Video:
 {current_caption}

 Most Similar Videos:
 1. {most_similar_captions[0]}
 2. {most_similar_captions[1]}

 Generate one sentence (less than 50
  words) describing the unique
 characteristic of the Current Video
  without mentioning the Most
 Similar Videos:
```
```

# 5 Experiments

## 5.1 Video Retrieval

We compute the similarity score at the video level between text and video enriched representation using off-the-shelf retrieval methods, *i.e.*, BM25 (Robertson and Walker, 1994) and sentence transformers (Reimers and Gurevych, 2019). We provide the experimental results with BM25 for comparing with existing method. More results of sentence transformers can be found in Table 8.

## 5.2 Benchmarks, Baselines, and Evaluation Metrics

**Benchmarks.** Following previous work (Luo et al., 2022; Ma et al., 2022), we use three representative benchmarks for evaluating DIANE, *i.e.*, MSR-VTT (Xu et al., 2016), MSVD (Chen and Dolan, 2011), and ActivityNet (Fabian Caba Heilbron and Niebles, 2015). Details of the dataset split are presented in Appendix A.1.

**Baselines** To show the empirical efficiency of our DIANE, we compare it with fine-tuned models (LiteVL (Chen et al., 2022), NCL (Park et al., 2022b), TABLE (Chen et al., 2023b), VOP (Huang et al., 2023), X-CLIP (Ma et al., 2022), DiscreteCodebook (Liu et al., 2022a), TS2-Net (Liu et al., 2022b), VCM (Cao et al., 2022), HiSE (Wang et al., 2022b), CenterCLIP (Zhao et al., 2022), X-Pool (Gorti et al., 2022), S3MA (Wang and Shi, 2023b)), and MV-Apapter (Jin et al., 2024), pre-trained methods (VLM (Xu et al., 2021a), HERO (Li et al., 2020a), VideoCLIP (Xu et al., 2021b), EvO (Shvetsova et al., 2022), OA-Trans (Wang et al., 2022a), RaP (Wu et al., 2022), OmniVL (Wang et al., 2022c), mPLUG-2 (Xu et al., 2023), InternVL (Chen et al., 2023c), LangaugeBind (Zhu et al., 2024), UCOFIA (Wang et al., 2023), ProST (Li et al., 2023c), and UATVR (Fang et al., 2023), ), and a few-shot method, *i.e.*, VidIL (Wang et al., 2022e).

**Evaluation metric.** To evaluate the retrieval performance of our proposed model, we use recall at Rank K (R@K, higher is better), median rank (MdR, lower is better), and mean rank (MnR, lower is better) as retrieval metrics, which are widely used in previous retrieval works (Radford et al., 2021; Luo et al., 2022; Ma et al., 2022).

**Implementation details and related model details** are defferd to Appendix A.3.

| Methods | Venue | Text-to-Video Retrieval R@1↑ | R@5↑ | R@10↑ | MdR↓ | MnR↓ |
|---|---|---|---|---|---|---|
| *Training-based* | | | | | | |
| LiteVL-S | EMNLP'2022 | 46.7 | 71.8 | 81.7 | 2.0 | - |
| X-Pool | CVPR'2022 | 46.9 | 72.8 | 82.2 | 2.0 | 14.3 |
| CenterCLIP | SIGIR'2022 | 44.2 | 71.6 | 82.1 | 2.0 | 15.1 |
| TS2-Net | ECCV'2022 | 47.0 | 74.5 | 83.8 | 2.0 | 13.0 |
| X-CLIP | ACM MM'2022 | 46.1 | 74.3 | 83.1 | 2.0 | 13.2 |
| NCL | EMNLP'2022 | 43.9 | 71.2 | 81.5 | 2.0 | 15.5 |
| TABLE | AAAI'2023 | 47.1 | 74.3 | 82.9 | 2.0 | 13.4 |
| VOP | CVPR'2023 | 44.6 | 69.9 | 80.3 | 2.0 | 16.3 |
| DiscreteCodebook | ACL'2022 | 43.4 | 72.3 | 81.2 | - | 14.8 |
| VCM | AAAI'2022 | 43.8 | 71.0 | - | 2.0 | 14.3 |
| CenterCLIP | SIGIR'2022 | 48.4 | 73.8 | 82.0 | 2.0 | 13.8 |
| HiSE | ACM MM'2022 | 45.0 | 72.7 | 81.3 | 2.0 | - |
| TS2-Net | ECCV'2022 | 49.4 | 75.6 | 85.3 | 2.0 | 13.5 |
| S3MA | EMNLP'2023 | 53.1 | 78.2 | 86.2 | 1.0 | 10.5 |
| UCOFIA | ICCV'2023 | 49.4 | 72.1 | - | - | 12.9 |
| ProST | ICCV'2023 | 49.5 | 75.0 | 84.0 | 2.0 | 11.7 |
| UATVR | ICCV'2023 | 49.8 | 76.1 | 85.5 | 2.0 | 12.9 |
| MV-Adapter | CVPR'2024 | 46.2 | 73.2 | 82.7 | - | - |
| *Zero-Shot (Pretrained Models)* | | | | | | |
| VLM | ACL'2021 | 28.1 | 55.5 | 67.4 | 4.0 | - |
| HERO | EMNLP'2021 | 16.8 | 43.3 | 57.7 | - | - |
| VideoCLIP | EMNLP'2021 | 30.9 | 55.4 | 66.8 | - | - |
| EvO | CVPR'2022 | 23.7 | 52.1 | 63.7 | 4.0 | - |
| OA-Trans | CVPR'2022 | 35.8 | 63.4 | 76.5 | 3.0 | - |
| RaP | EMNLP'2022 | 40.9 | 67.2 | 76.9 | 2.0 | - |
| OmniVL | NeurIPS'2022 | 34.6 | 58.4 | 66.6 | - | - |
| mPLUG-2 | ICML'2023 | 48.3 | <u>75.0</u> | <u>83.2</u> | - | - |
| InternVL | arXiv'2023 | 42.4 | 65.9 | 75.4 | - | - |
| LanguageBind | ICLR'2024 | 42.6 | 65.4 | 75.5 | - | - |
| *Few-Shot* | | | | | | |
| VidIL | NeurIPS'2022 | 40.8 | 65.2 | - | - | - |
| *Zero-Shot* | | | | | | |
| DIANE w/o relevance-boosted caption genertion | | 20.3 | 40.9 | 51.7 | 9.0 | 60.3 |
| DIANE | | **58.7** | **76.6** | **84.4** | 1.0 | **17.9** |

Table 1: Text-to-Video retrieval results on MSR-VTT. The best results are marked in **bold**. The second best results are <u>underlined</u>.

| Methods | Venue | Text-to-Video Retrieval R@1↑ | R@5↑ | R@10↑ | MnR↓ |
|---|---|---|---|---|---|
| *MSVD* | | | | | |
| RaP | EMNLP'22 | 35.9 | 64.3 | 73.7 | - |
| LanguageBind | ICLR'24 | 52.2 | 79.4 | 87.3 | - |
| DIANE | | **57.2** | **80.0** | **88.2** | 15.6 |
| *ActivityNet* | | | | | |
| LanguageBind | ICLR'24 | 35.1 | 63.4 | 76.6 | - |
| DIANE | | **59.0** | **71.4** | **77.0** | 387.4 |

Table 2: Text-to-Video retrieval results on MSVD and ActivityNet. The best results are marked in **bold**.

## 5.3 Quantitative Results

In this part, we present the qualitative results of DIANE on three VR benchmarks.

**MSR-VTT.** We found that the contextual video text obtained directly through video captioning methods generally have mediocre performance (R@1: 20.3) compared to other baseline Text-Video Retrieval method. However, after using LLM to do relevance boosting from the video caption, the R@1 of our method nearly doubled (R@1 = 40.9) shown in Table 4 . Therefore, we further boosted each sentence and expanded it into two sentences. From the results presented in Table 1, it can be seen that this approach outperforms the second-best method by 9.9. This indicates the significant impact of relevance boosting and expanding captions on enhancing the performance of Text-Video Retrieval sys-

| Caption | VT4C | VT4V | DA | Text-to-Video Retrieval R@1↑ | R@5↑ | R@10↑ | MdR↓ | MnR↓ |
|---|---|---|---|---|---|---|---|---|
| ✓ | | | | 54.0 | 73.9 | 80.2 | 1.0 | 24.6 |
| ✓ | ✓ | | | 57.8 | 75.7 | 83.5 | 1.0 | 19.5 |
| ✓ | | ✓ | | 53.8 | 74.7 | 81.2 | 1.0 | 23.6 |
| ✓ | | | ✓ | 55.3 | 76.8 | 82.5 | 1.0 | 21.1 |
| ✓ | | ✓ | ✓ | 55.2 | 82.1 | 83.6 | 1.0 | 21.6 |
| ✓ | ✓ | ✓ | | 58.2 | 76.4 | 83.6 | 1.0 | 18.87 |
| ✓ | ✓ | | ✓ | 57.8 | 77.0 | 84.1 | 1.0 | 18.6 |
| ✓ | ✓ | ✓ | ✓ | **58.7** | **76.6** | **84.4** | 1.0 | **17.9** |

Table 3: Retrieval performance with different combinations of enrichment strategies (Visual tokens for captions and video frames, Distinctiveness Analysis) on MSR-VTT using DIANE. "VT4C", "VT4V", and "DA" represent visual tokens for captions, visual tokens for video frames, and distinctiveness analysis. Best in **Bold**.

| # of Relevance Boosted Captions | Text-to-Video Retrieval R@1↑ | R@5↑ | R@10↑ | MdR↓ | MnR↓ |
|---|---|---|---|---|---|
| 1 | 40.9 | 55.5 | 60.9 | 3.0 | 227.3 |
| 2 | **58.7** | **76.6** | **84.4** | 1.0 | **17.9** |
| 3 | 55.7 | 73.9 | 82.2 | 1.0 | 21.2 |

Table 4: Retrieval performance with different numbers of relevance-boosted captions on MSR-VTT using DIANE. Best in **Bold**.

| LLM | Text-to-Video Retrieval R@1↑ | R@5↑ | R@10↑ | MdR↓ | MnR↓ |
|---|---|---|---|---|---|
| LLaMA | 58.7 | 76.6 | 84.4 | 1.0 | 17.9 |
| GPT 3.5 | **61.2** | **80.4** | **86.8** | 1.0 | **15.0** |

Table 5: Retrieval performance with different LLM models on MSR-VTT using DIANE. Best in **Bold**.

| Template | Text-to-Video Retrieval R@1↑ | R@5↑ | R@10↑ | MdR↓ | MnR↓ |
|---|---|---|---|---|---|
| Basic Template | **58.7** | **76.6** | **84.4** | 1.0 | **17.9** |
| Structured Template | 55.7 | 74.6 | 81.2 | 1.0 | 21.1 |
| Template with Detailed Description | 55.9 | 74.6 | 81.7 | 1.0 | 21.2 |
| Narrative Format Template | 56.5 | 74.7 | 81.7 | 1.0 | 20.9 |

Table 6: Retrieval performance with different template formats on MSR-VTT using DIANE. Best in **Bold**.

tems. Compared to DiscreteCodebook (Liu et al., 2022a), which aligns modalities in an unsupervised manner, DIANE outperforms DiscreteCodebook on every metric. Meanwhile, DIANE also outperforms VidIL (Wang et al., 2022e), which uses few-shot prompting, demonstrating the usability of integrating zero-shot LLM on text-to-video retrieval. This suggests that leveraging zero-shot on LLMs is a promising approach to enhance text-to-video retrieval performance. Also, we notice that DIANE has bad results on mean rank. To understand why this happens, we visualize the distribution of rank in Figure 2. It is obvious that though most of the videos have very good rank, *e.g.*, lower than 10, there are still some captions ranked in the last.

**MSVD and ActivityNet.** The results on MSVD and ActicityNet are shown in Table 2. DIANE achieves the best R@1 on text-to-video retrieval on two datasets compared to the previous methods.

### 5.4 Ablation Studies

In this part, we present a series of ablation experiments on MSR-VTT to better understand the effectiveness of different components of DIANE, using LLaMA2-7b-chat-hf and BM25. Due to space limitations, we present the ablation study on retrieval methods and the exploration of different visual tokens in Appendix.

**Impact of combination of different components from Index Time Enrichment.** To determine the optimal combination of components for text-to-video retrieval, we conduct experiments with different configurations of visual tokens for captions and video frames, as well as distinctiveness analysis, as shown in Table 3. The results demonstrate that incorporating additional components generally improves retrieval performance. Notably, the best performance is achieved when all components are combined, yielding the highest R@1, R@5, and R@10 scores while minimizing MnR. This confirms the effectiveness of leveraging both caption and video visual tokens alongside distinctiveness analysis to enhance retrieval accuracy.

**Number of relevance-boosted captions.** In this part, we aim to explore how many relevance-boosted captions work the best. More captions have the potential to offer more detailed descriptions, which may enhance the viewer's comprehension of the visual content. Previous studies (Biten et al., 2019; Tang et al., 2023) have demonstrated that longer captions tend to be more descriptive and semantically rich, achieving improved comprehension and retrieval performance. However, more relevance-boosted captions might mean more noises are injected. So balancing the number of relevance-boosted captions would be highly important. From the results shown in Table 4, we notice that paraphrasing into two or three sentences significantly improved R@1, R@5, and R@10. Considering computational constraints and the similar effectiveness of paraphrasing into two and three sentences, we decide to boost it into two sentences. We also investigate relevance-boosted performance with different LLM models, including LLaMa and GPT-3.5 in Table 4.

**Complexity of prompt templates for extracting visual tokens.** The complexity of the prompt plays a pivotal role in shaping the output generated by the model, influencing the depth of analysis and

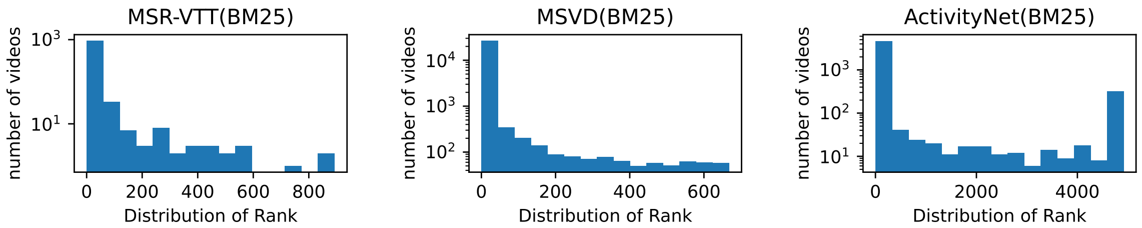

Figure 2: These figures illustrate the distribution of the rank of each (test) gallery video (captions) retrieved by (test) text queries.

| Relevance | Automatic Evaluation Metric HHEM | Human Evaluation Factual Accuracy | Relevance | Coherence | Specificity | Text-to-Video Retrieval R@1↑ | R@5↑ | R@10↑ | MdR↓ | MnR↓ |
|---|---|---|---|---|---|---|---|---|---|---|
| High-level | 16.1% | 0.33 | 0.42 | 0.24 | -0.95 | 56.9 | 75.1 | 82.6 | 1.0 | 21.4 |
| Medium-level | 14.7% | 0.52 | 0.78 | 1.21 | 0.07 | 57.3 | 75.2 | 82.4 | 1.0 | 18.1 |
| Low-level | 9.6% | 0.85 | 0.81 | **1.38** | **0.68** | 57.6 | 74.9 | 83.3 | 1.0 | 19.1 |
| DIANE | 10.9% | **0.87** | **0.86** | 1.28 | 0.52 | **58.7** | **76.6** | **84.4** | 1.0 | **17.9** |

Table 7: Retrieval performance with different levels of Relevance Boosting on MSR-VTT. Best in **Bold**.

the richness of information conveyed. An intricate prompt may provide the model with additional context and guidance, enabling it to produce more detailed responses. Specifically, we compare four templates (Basic, Structured, Detailed Description, and Narrative Format) offering different levels of complexity for organizing video content as shown in Appendix A.6. The results are shown in Table 6. The results show that with the simplest template (basic template), R@1, R@5, and R@10 on text-to-video retrieval has better performance. This might be because the simplest format template enables a more straightforward extraction of visual tokens, which can aid in the efficiency and accuracy of retrieval by presenting the information in a direct storytelling format. We observed that, while the narrative format performs worse than the basic template in text-to-video retrieval, it still outperforms other formats (such as the structured template and the detailed description template). This may be because the narrative format provides the model with more context and direction, but it can also cause the model to miss some key information that is important for accurate retrieval.

# 6 Analysis on Quality of Relevance-Boosted Captions

As the details brought by relevance-boosted generation might bring irrelevant information, we analyze the quality of relevance-boosted captions.

## 6.1 Automatic Evaluation

Inspired by Li et al. (2023a), we generate video captions with varying levels of relevant details by using different prompts to control the level of relevance generation. Specifically, we generate captions at three levels: high, medium, and low (see Appendix B). We used the HHEM model (Honovich et al., 2022) to compute the hallucination rate by comparing the relevance-boosted captions and original video captions. As shown in Table 7, lower levels of generation do not significantly change retrieval results. We also observe that captions with a lower boosting rate perform worse than captions with higher levels.

## 6.2 Qualitative Results

To qualitatively validate the effectiveness of DIANE, we present an example in Figure 3. The retrieval results show that relevance-boosted captions have more information in the video than vanilla video captions. Besides, our proposed methods clearly emphasize the important visual tokens, *i.e.*, phrase, object, relationship, and attribute, further boosting the performance.

## 6.3 Human Evaluation

We also conduct a human evaluation to further evaluate the relevance-boosted captions.

**Participants:** Our human evaluation task involves reading relevance-boosted captions from different levels, video captions without relevance-boosting, and rating those relevance-boosted captions from them. We recruited 10 participants (7M, 3F). We conducted a rigorous qualification process, evaluating their English proficiency, to ensure high-quality annotations. We hired them by sending invited emails to graduate students. We allocated up to 30 minutes for each participant to complete the study, and for their valuable time and input, each participant received a compensation of $15.

**Task:** We randomly selected 50 pairs of relevance-boosted captions and original video captions from DIANE. Note that each pair has only one

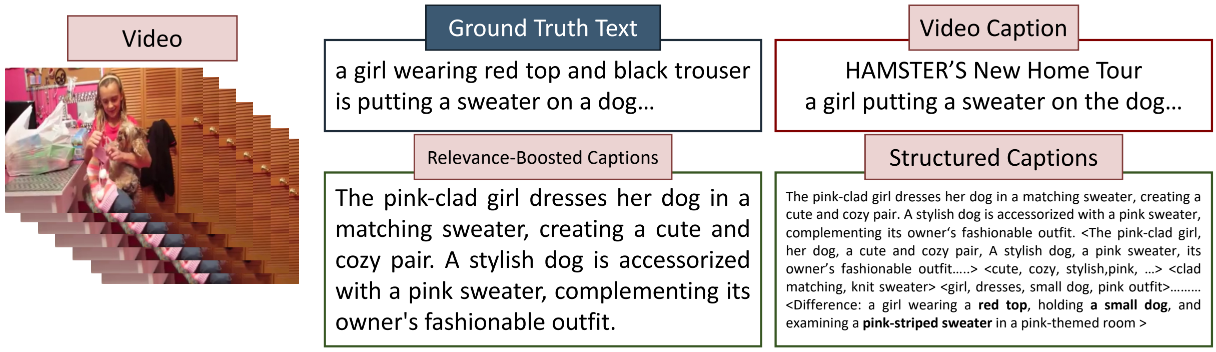


Figure 3: A retrieval example demonstrates that relevance-boosted captions contain more information compared to vanilla video captions in the video though some noises are also added.

relevance-boosted caption and one original video caption. Each participant is assigned 50 pairs. Each pair is evaluated by 10 individuals. In each trial, a participant reads 4 relevance-boosted captions for the original video caption: one by high-level boosting, one by medium-level boosting, one by low-level boosting, and one from DIANE. The order of these four is also randomized, so participants do not know which generated caption is from which method. The participant is asked to rate the 4 captions along four dimensions using a five-point Likert scale from -2 to 2.

- *Factual Accuracy*: The relevance-boosted caption is factually correct to convey the content from the video caption.
- *Relevance*: The relevance-boosted caption is relevant to the video caption.
- *Coherence*: The relevance-boosted caption is coherent to the video caption.
- *Specificity*: The relevance-boosted caption is specific and detailed to the video caption.

**Evaluation Results:** We conducted *Wilcoxon tests* (Woolson, 2007) with a significance level of 0.05 to compare the performance of high-level, medium-level, low-level boosting, and DIANE in the Factual Accuracy, Relevance, Coherence, and Specificity dimensions. The Wilcoxon test is a non-parametric statistical test used to compare two paired groups of data. The obtained p-values indicate the probability of observing the reported differences if there were no true differences between the models.

The results indicate significant differences in the Factual Accuracy dimension, where DIANE outperforms High-level boosting (V = 4836, p = 1.45e-30), Medium-level boosting (V = 4819, p = 7.22e-31). For the Coherence dimension, we notice that they are almost at the same level, likely because captions refined by the LLM are already sufficiently coherent for users. In the Relevance dimension, DIANE surpasses high-level boosting (V = 3247, p = 1.44e-21), medium-level boosting (V = 3693, p = 1.69e-20), low-level boosting (V = 3188, p = 1.53e-20). For the Specificity dimension which considers whether the relevance-boosted caption is detailed and specified, Low-level boosting outperforms all methods: High-level boosting (V = 4463, p = 1.25e-7), Medium-level boosting (V = 3830, p = 3.48e-14), DIANE (V = 2260, p = 2.63e-7). It is worth noting that while low-level boosting is more detailed than DIANE, it performs slightly worse in VR, possibly due to the higher importance of factual accuracy in evaluating the effectiveness of relevance-boosted captions. Future work can focus on designing an innovative framework for the relevance-boosted captioning method to integrate useful dimensions.

## 7 Conclusion

In this paper, we present an innovative zero-shot framework, DIANE, which revolutionizes video retrieval by capitalizing on existing captioning methods, large language models (LLMs), and text retrieval techniques. By sidestepping the need for model training or fine-tuning, our framework offers a streamlined approach to retrieval. To overcome the shortcomings of traditional captioning methods, we propose a groundbreaking index time enrichment to enhance retrieval performance by relevance-boosted caption generation technique, highlighting key visual tokens, and distinctiveness analysis. Through extensive experimentation across diverse benchmarks, we demonstrate the superior efficacy of DIANE compared to conventional fine-tuned and pretraining methods, even in the absence of training data.

## Limitations

In the future, it would be interesting to explore more detailed methods for zero-shot video retrieval, such as incorporating the audio modality and corresponding off-the-shelf foundation models. Moreover, as a pioneering work, our work mainly focuses on exploring index time alignment and enrichment. It would be great if we could explore more text retrieval methods, video captioning methods, and LLMs for relevance-boosted caption generation.

## References


Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1708–1718. IEEE.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 813–824. PMLR.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Ali Furkan Biten, Lluis Gomez, Marcal Rusinol, and Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Shuqiang Cao, Bairui Wang, Wei Zhang, and Lin Ma. 2022. Visual consensus modeling for video-text retrieval. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 167–175. AAAI Press.

David M. Chan, Austin Myers, Sudheendra Vijayanarasimhan, David A. Ross, and John Canny. 2023. $IC^3$: Image Captioning by Committee Consensus. *arXiv preprint*. ArXiv:2302.01328 [cs].

David Chen and William Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Portland, Oregon, USA. Association for Computational Linguistics.

Dongsheng Chen, Chaofan Tao, Lu Hou, Lifeng Shang, Xin Jiang, and Qun Liu. 2022. LiteVL: Efficient video-language learning with enhanced spatial-temporal modeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7985–7997, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. 2023a. VAST: A vision-audio-subtitle-text omni-modality foundation model and dataset. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yizhen Chen, Jie Wang, Lijian Lin, Zhongang Qi, Jin Ma, and Ying Shan. 2023b. Tagging before Alignment: Integrating Multi-Modal Tags for Video-Text Retrieval. In *AAAI Conference on Artificial Intelligence*. arXiv. ArXiv:2301.12644 [cs].

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. 2021. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *CoRR*, abs/2109.04290.

Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. 2021. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 11563–11573. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. 2019. Dual encoding for zero-example video retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9346–9355. Computer Vision Foundation / IEEE.

Maksim Dzabraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko. 2021. MDMMT: multidomain multimodal transformer for video retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, pages 3354–3363. Computer Vision Foundation / IEEE.

Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970.

Bo Fang, Wenhao Wu, Chang Liu, Yu Zhou, Yuxin Song, Weiping Wang, Xiangbo Shu, Xiangyang Ji, and Jingdong Wang. 2023. Uatvr: Uncertainty-adaptive text-video retrieval. *Preprint*, arXiv:2301.06309.

Xiang Fang, Daizong Liu, Pan Zhou, and Yuchong Hu. 2022. Multi-modal cross-domain alignment network for video moment retrieval. *IEEE Transactions on Multimedia*.

Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multi-modal transformer for video retrieval. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*, volume 12349 of *Lecture Notes in Computer Science*, pages 214–229. Springer.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Zijian Gao, Jingyu Liu, Sheng Chen, Dedan Chang, Hao Zhang, and Jinwei Yuan. 2021. CLIP2TV: an empirical study on transformer-based methods for video-text retrieval. *CoRR*, abs/2111.05610.

Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. 2022. X-pool: Cross-modal language-video attention for text-video retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 4996–5005. IEEE.

Qingpei Guo, Furong Xu, Hanxiao Zhang, Wang Ren, Ziping Ma, Lin Ju, Jian Wang, Jingdong Chen, and Ming Yang. 2024. M2-encoder: Advancing bilingual image-text understanding by large-scale efficient pretraining. *Preprint*, arXiv:2401.15896.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. True: Re-evaluating factual consistency evaluation. *arXiv preprint arXiv:2204.04991*.

Siteng Huang, Biao Gong, Yulin Pan, Jianwen Jiang, Yiliang Lv, Yuyuan Li, and Donglin Wang. 2023. VoP: Text-Video Co-Operative Prompt Tuning for Cross-Modal Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6565–6574.

Mark J. Huiskes and Michael S. Lew. 2008. The MIR Flickr Retrieval Evaluation. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, MIR '08, pages 39–43, New York, NY, USA. Association for Computing Machinery. Event-place: Vancouver, British Columbia, Canada.

Weike Jin, Zhou Zhao, Pengcheng Zhang, Jieming Zhu, Xiuqiang He, and Yueting Zhuang. 2021. Hierarchical cross-modal graph consistency learning for video-text retrieval. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 1114–1124. ACM.

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. *Preprint*, arXiv:2301.07868.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 7331–7341. Computer Vision Foundation / IEEE.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464.

Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. 2023b. Unmasked teacher: Towards training-efficient video foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19948–19960.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020a. HERO: Hierarchical encoder for Video+Language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, Online. Association for Computational Linguistics.

Pandeng Li, Chen-Wei Xie, Liming Zhao, Hongtao Xie, Jiannan Ge, Yun Zheng, Deli Zhao, and Yongdong Zhang. 2023c. Progressive spatio-temporal prototype matching for text-video retrieval. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4077–4087.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 121–137. Springer.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in neural information processing systems*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Alexander Liu, SouYoung Jin, Cheng-I Lai, Andrew Rouditchenko, Aude Oliva, and James Glass. 2022a. Cross-modal discrete representation learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3013–3035, Dublin, Ireland. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

Ruyang Liu, Chen Li, Yixiao Ge, Ying Shan, Thomas H Li, and Ge Li. 2023c. One for all: Video conversation is feasible without video instruction tuning. *arXiv preprint arXiv:2309.15785*.

Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. 2022b. Ts2-net: Token shift and selection transformer for text-video retrieval. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XIV*, volume 13674 of *Lecture Notes in Computer Science*, pages 319–335. Springer.

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

Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. 2022. X-CLIP: End-to-end multi-grained contrastive learning for video-text retrieval. In *ACM international conference on multimedia*, MM '22, pages 638–647, New York, NY, USA. Association for Computing Machinery. Number of pages: 10 Place: Lisboa, Portugal.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9876–9886. Computer Vision Foundation / IEEE.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2630–2640. IEEE.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*. Association for Computational Linguistics.

Jae Sung Park, Sheng Shen, Ali Farhadi, Trevor Darrell, Yejin Choi, and Anna Rohrbach. 2022a. Exposing the limits of video-text models through contrast sets. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3574–3586, Seattle, United States. Association for Computational Linguistics.

Yookoon Park, Mahmoud Azab, Seungwhan Moon, Bo Xiong, Florian Metze, Gourab Kundu, and Kirmani Ahmed. 2022b. Normalized contrastive learning for text-video retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 248–260, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pages 232–241. Springer.

Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.*, 45(11):2673–2681.

Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio Feris, David Harwath, James Glass, and Hilde Kuehne. 2022. Everything at Once – Multi-modal Fusion Transformer for Video Retrieval. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19988–19997, New Orleans, LA, USA. IEEE.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A foundational language and vision alignment model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15617–15629. IEEE.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Benny J Tang, Angie Boggust, and Arvind Satyanarayan. 2023. Vistext: A benchmark for semantically rich chart captioning. *arXiv preprint arXiv:2307.05356*.

Mingkang Tang, Zhanyu Wang, Zhenhua Liu, Fengyun Rao, Dian Li, and Xiu Li. 2021. Clip4caption: Clip for video caption. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4858–4862.

Yoad Tewel, Yoav Shalev, Roy Nadler, Idan Schwartz, and Lior Wolf. 2022. Zero-shot video captioning with evolving pseudo-tokens. *arXiv preprint arXiv:2207.11100*.

Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2021. Zero-shot image-to-text generation for visual-semantic arithmetic. *arXiv preprint arXiv:2111.14447*, 1(3):6.

Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven C.H. Hoi. 2022. Plug-and-play VQA: Zero-shot VQA by conjoining large pretrained models with zero training. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 951–967, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ba Tu Truong and Svetha Venkatesh. 2007. Video abstraction: A systematic review and classification. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 3(1):3–es.

Alex Jinpeng Wang, Yixiao Ge, Guanyu Cai, Rui Yan, Xudong Lin, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2022a. Object-aware Video-language Pre-training for Retrieval. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3303–3312, New Orleans, LA, USA. IEEE.

Haoran Wang, Di Xu, Dongliang He, Fu Li, Zhong Ji, Jungong Han, and Errui Ding. 2022b. Boosting video-text retrieval with explicit high-level semantics. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 4887–4898. ACM.

Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. 2022c. Omnivl: One foundation model for image-language and video-language tasks. In *Advances in Neural Information Processing Systems*, volume 35, pages 5696–5710. Curran Associates, Inc.

Xiaohan Wang, Linchao Zhu, Zhedong Zheng, Mingliang Xu, and Yi Yang. 2022d. Align and tell: Boosting text-video retrieval with local alignment and fine-grained supervision. *IEEE Transactions on Multimedia*, pages 1–11.

Yimu Wang, Shiyin Lu, and Lijun Zhang. 2020a. Searching privately by imperceptible lying: A novel private hashing method with differential privacy. In *Proceedings of the 28th ACM International Conference on Multimedia*, page 2700–2709.

Yimu Wang and Peng Shi. 2023a. Video-Text Retrieval by Supervised Multi-Space Multi-Grained Alignment. *arXiv preprint*. ArXiv:2302.09473 [cs].

Yimu Wang and Peng Shi. 2023b. Video-text retrieval by supervised sparse multi-grained learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 633–649, Singapore. Association for Computational Linguistics.

Yimu Wang, Xiu-Shen Wei, Bo Xue, and Lijun Zhang. 2020b. Piecewise hashing: A deep hashing method for large-scale fine-grained search. In *Pattern Recognition and Computer Vision - Third Chinese Conference, PRCV 2020, Nanjing, China, October 16-18, 2020, Proceedings, Part II*, pages 432–444.

Yimu Wang, Bo Xue, Quan Cheng, Yuhui Chen, and Lijun Zhang. 2021. Deep unified cross-modality hashing by pairwise data alignment. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 1129–1135.

Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, Shih-Fu Chang, Mohit Bansal, and Heng Ji. 2022e. Language models with image descriptors are strong few-shot video-language learners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2023. Unified coarse-to-fine alignment for video-text retrieval. *Preprint*, arXiv:2309.10091.

Robert F Woolson. 2007. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*, pages 1–3.

Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. 2023. Cap4Video: What Can Auxiliary Captions Do for Text-Video Retrieval? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10704–10713.

Xing Wu, Chaochen Gao, Zijia Lin, Zhongyuan Wang, Jizhong Han, and Songlin Hu. 2022. RaP: Redundancy-aware video-language pre-training for text-video retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3036–3047, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Fei Huang, and Jingren Zhou. 2023. mplug-2: A modularized multi-modal foundation model across text, image and video. *ArXiv*, abs/2302.00402.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. 2021a. VLM: Task-agnostic video-language model pre-training for video understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4227–4239, Online. Association for Computational Linguistics.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021b. Video-CLIP: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, Las Vegas, NV, USA. IEEE.

Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. 2023. CLIP-ViP: Adapting Pre-trained Image-Text Model to Video-Language Alignment. In *The Eleventh International Conference on Learning Representations*.

Qiying Yu, Yang Liu, Yimu Wang, Ke Xu, and Jingjing Liu. 2023. Multimodal federated learning via contrastive representation ensemble. In *The Eleventh International Conference on Learning Representations*.

Youngjae Yu, Jongseok Kim, and Gunhee Kim. 2018. A joint sequence fusion model for video question answering and retrieval. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany,* *September 8-14, 2018, Proceedings, Part VII*, volume 11211 of *Lecture Notes in Computer Science*, pages 487–503. Springer.

Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. 2017. End-to-end concept word detection for video captioning, retrieval, and question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3261–3269. IEEE Computer Society.

Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. 2022. Centerclip: Token clustering for efficient text-video retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 970–981, New York, NY, USA. Association for Computing Machinery.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Wan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. 2024. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations*.

Linchao Zhu and Yi Yang. 2020. Actbert: Learning global-local video-text representations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8743–8752. Computer Vision Foundation / IEEE.

## A Experiments

### A.1 Details of Benchmarks

- **MSR-VTT** (Xu et al., 2016) contains 10,000 videos with length varying from 10 to 32 seconds, each paired with about 20 human-labeled captions. Following the evaluation protocol from previous works (Yu et al., 2018; Miech et al., 2019), we use the training-9k / test 1k-A splits for training and testing respectively.

- **MSVD** (Chen and Dolan, 2011) contains 1,970 videos with a split of 1200, 100, and 670 as the train, validation, and test set, respectively. The duration of videos varies from 1 to 62 seconds. Each video is paired with 40 English captions.

- **ActivityNet** (Fabian Caba Heilbron and Niebles, 2015) is consisted of 20,000 Youtube videos with 100,000 densely annotated descriptions. For a fair comparison, following the previous setting (Luo et al., 2022; Gabeur et al., 2020), we concatenate all captions together as a paragraph to perform a video-paragraph retrieval task by concatenating all the descriptions of a video. Performances are reported on the "val1" split of the ActivityNet.

### A.2 Baselines

To show the empirical efficiency of our DIANE, we compare it with fine-tuned models (LiteVL (Chen et al., 2022), NCL (Park et al., 2022b), TABLE (Chen et al., 2023b), VOP (Huang et al., 2023), X-CLIP (Ma et al., 2022), Discrete-Codebook (Liu et al., 2022a), TS2-Net (Liu et al., 2022b), VCM (Cao et al., 2022), HiSE (Wang et al., 2022b), CenterCLIP (Zhao et al., 2022), X-Pool (Gorti et al., 2022), S3MA (Wang and Shi, 2023b)), and MV-Apapter (Jin et al., 2024), pre-trained methods (VLM (Xu et al., 2021a), HERO (Li et al., 2020a), VideoCLIP (Xu et al., 2021b), EvO (Shvetsova et al., 2022), OA-Trans (Wang et al., 2022a), RaP (Wu et al., 2022), OmniVL (Wang et al., 2022c), mPLUG-2 (Xu et al., 2023), InternVL (Chen et al., 2023c), Langauge-Bind (Zhu et al., 2024), UCOFIA (Wang et al., 2023), ProST (Li et al., 2023c), and UATVR (Fang et al., 2023), ), and a few-shot method, *i.e.*, VidIL (Wang et al., 2022e).

### A.3 Implementation Details

For video caption generation, we use Tewel et al. (2021, 2022) to generate video captions and GPT-2 (Radford et al., 2019) to enrich sentences. For relevance-boosted caption generation, we employ LLaMA2-7b-chat-hf (Touvron et al., 2023) and get two boosted captions. For extracting visual tokens, we use SPACY (Bird et al., 2009). For text retrieval, we use BM25 (Robertson and Walker, 1994).

We use **GPT2** (Radford et al., 2019) for sentence enrichment during video caption generation. GPT-2 (Radford et al., 2019), developed by OpenAI, is a large-scale transformer-based language model renowned for its ability to generate coherent and contextually relevant text. With 1.5 billion parameters, GPT-2 can be fine-tuned for a variety of natural language processing tasks, such as text generation, summarization, and captioning. In our task, we enrich image captions with GPT-2 with one NVIDIA A100 GPU using around 20 hours.

We use Llama (Touvron et al., 2023)(version: Llama-2-7b-chat-hf) to conduct the relevance-boosted caption generation task. **Llama** (Touvron et al., 2023) is an advanced language model with approximately 7 billion parameters. Its default backend is designed for efficiency and scalability. The computational budget for LlaMA in our task is approximately 23 hours with one NVIDIA A100 GPU. Its ability to understand context, generate coherent and contextually relevant responses, and perform a wide range of language-related tasks is significantly enhanced. LlaMA is a powerful and accessible tool, widely used in various applications. Therefore, it is included as an advanced baseline.

We use **Qwen2.5-VL-7B-Instruct** (Qwen Team, 2025) to conduct Index Time Enrichment (ITE) during video frame analysis. Qwen2.5-VL-7B-Instruct (Qwen et al., 2025), developed by the Qwen Team, is a large-scale visual-language model consisting of 7 billion parameters. This model is designed for efficient and context-aware visual token extraction from video frames. In our experiment, we use the Qwen2.5-VL-7B-Instruct model to generate key visual tokens from video captions and video frames, which include objects, attributes, relationships, and descriptive phrases from sampled video frames.

For the ITE process, the model is run on one NVIDIA H100 GPU for approximately 4 hours. The generated visual tokens are structured into a JSON format for further analysis and integration

| Retrieval Methods | Text-to-Video Retrieval | | | | |
|---|---|---|---|---|---|
| | R@1↑ | R@5↑ | R@10↑ | MdR↓ | MnR↓ |
| BM25 | **58.7** | **76.6** | **84.4** | 1.0 | **17.9** |
| Sentence Transformer | 41.2 | 62.1 | 70.5 | 2.0 | 33.5 |

Table 8: Retrieval performance with different retrieval models on MSR-VTT using DIANE. Best in **Bold**.

| Caption | Phrase | Object | Relationship | Attribute | Text-to-Video Retrieval | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | R@1↑ | R@5↑ | R@10↑ | MdR↓ | MnR↓ |
| ✓ | | | | | 54.0 | 73.9 | 80.2 | 1.0 | 24.5 |
| ✓ | ✓ | | | | 57.4 | 76.2 | 83.0 | 1.0 | 19.3 |
| ✓ | | ✓ | | | 56.9 | **77.5** | 83.8 | 1.0 | 18.6 |
| ✓ | | | ✓ | | 54.2 | 73.2 | 79.6 | 1.0 | 24.9 |
| ✓ | | | | ✓ | 55.0 | 74.2 | 80.2 | 1.0 | 24.1 |
| ✓ | ✓ | ✓ | | | 57.4 | 76.2 | 83.5 | 1.0 | 18.7 |
| ✓ | ✓ | | ✓ | | 57.3 | 76.3 | 82.6 | 1.0 | 19.8 |
| ✓ | ✓ | | | ✓ | 57.6 | 76.3 | 83.5 | 1.0 | 19.1 |
| ✓ | | ✓ | ✓ | | 56.9 | 76.6 | 83.2 | 1.0 | 19.3 |
| ✓ | | ✓ | | ✓ | 57.6 | 77.4 | 83.8 | 1.0 | 18.2 |
| ✓ | | | ✓ | ✓ | 54.0 | 73.3 | 79.6 | 1.0 | 24.9 |
| ✓ | ✓ | ✓ | ✓ | | 58.0 | 75.9 | 83.7 | 1.0 | 19.3 |
| ✓ | ✓ | ✓ | | ✓ | 57.8 | 76.3 | 84.1 | 1.0 | 18.3 |
| ✓ | ✓ | | ✓ | ✓ | 57.8 | 76.0 | 82.5 | 1.0 | 19.5 |
| ✓ | | ✓ | ✓ | ✓ | 57.3 | 76.7 | 83.2 | 1.0 | 18.9 |
| ✓ | ✓ | ✓ | ✓ | ✓ | **58.7** | **76.6** | **84.4** | 1.0 | **17.9** |

Table 9: Retrieval performance with different combinations of four visual tokens from video captions (Phrase, Object, Relationship, Attribute) on MSR-VTT using DIANE. Best in **Bold**.

into the video retrieval pipeline.

### A.4 Impact of Combination of Visual Tokens

To choose the best combination method for the extracted visual tokens (phrases, attributes, objects, and relationships), we conduct experiments using different arrangements of these visual tokens, as shown in Table 9. By reducing the inclusion of visual tokens, the retrieval performance of DIANE decreases, thereby proving the usefulness of integrating these four visual tokens together.

### A.5 Choice of Retrieval Methods

In this part, we investigate the impact of different retrieval methods, *i.e.*, BM25 (Robertson and Walker, 1994) and sentence transformers (Reimers and Gurevych, 2019). The results are shown in Section 7. It shows that BM25 outperforms the sentence transformer.

### A.6 Prompts for Visual Token Extraction

1. Basic Template: the simplest, providing a straightforward list of video elements, the one shown in Section 4.2.

2. Structured Template: It adds categorized elements, making the information easier to navigate for the retrieval method.

```
Video Caption : <Caption>. Key
Phrases: <{Phrases}>. Main
Objects: <Objects>. Notable
Features: <{Attributes}>. Key
Relationships: <Relationship>
```

3. Template with Detailed Description: This further elaborates on each element, offering in-depth insights.

```
Detailed Video Description:
Caption: <{Caption}> Objects and
 Attributes Overview: Each
object, <{Objects}>, is detailed
 with attributes such as <{
Attributes}> to provide a
clearer image. Relationship
Analysis: The video's narrative
is driven by relationships like
<{Relationships}>, which are
elaborated for better
understanding. Phrases Insight:
Phrases like <{Phrases}> are
explained for their significance
 to the content.
```

4. Narrative Format Template: it weaves the elements into a cohesive story, enhancing engagement and providing a thematic understanding.

```
Caption: <Caption> In this video
, we observe <{Objects}> with <{
Attributes}>, a vivid
representation of <{
Relationships}>. Phrases such as
 <{Phrases}> punctuate the
narrative, offering insights
into the unfolding story.
```

### A.7 Are Relevance-Boosted Caption Generation and Visual Token(for caption and video) Extraction Necessary?

We also conduct another ablation study to investigate the effect of the video caption repeating itself several times to form text that is the same length

| Retrieval Methods | Text-to-Video Retrieval | | | | |
|---|---|---|---|---|---|
| | R@1↑ | R@5↑ | R@10↑ | MdR↓ | MnR↓ |
| DIANE | **58.2** | **75.8** | **83.5** | 1.0 | **18.9** |
| DIANE (video caption only repeats to the same length as structured caption) | 54.0 | 73.9 | 80.2 | 1.0 | 24.6 |
| DIANE (visual tokens for captions and videos only repeat to the same length as video caption) | 18.6 | 25.1 | 27.1 | 15.0 | 444.6 |

Table 10: Comparative Analysis of Caption Repetition and Extracted Visual Token Repetition on Retrieval Performance

as the structured caption stage. According to Table 10, we find that our DIANE method outperforms the others, indicating that a blend of relevance boosting (imagined or generated content) and visual tokens significantly improves retrieval results. Specifically, in text-to-video retrieval, DIANE achieves much higher recall rates and lower median and mean ranks than the other methods, which rely solely on caption repetition or visual tokens. Also, we find that caption repetition outperforms visual tokens extraction repetition. This suggests that incorporating relevance boosting is crucial for enhancing retrieval effectiveness.

# B Prompt to Generate Captions in Different Levels of Relevance Boosting

## B.1 Low-level Relevance

```
The following is a caption from a
video: [" + text + "]. Based on this
 caption, generate two paraphrased
captions capturing the key
information and main themes, each of
 which should be in one sentence
with up to twenty words (Do not
include any details not mentioned in
 the text. Focus on the main points
and key details.). Also Please
generate text without any comment.
```

## B.2 Medium-level Relevance

```
The following is a caption from a
video: [" + text + "]. Based on this
 caption, generate two paraphrased
captions capturing the key
information and main themes, each of
 which should be in one sentence
with up to twenty words (Feel free
to elaborate on points that seem
important, even if not explicitly
mentioned.). Also Please generate
text without any comment.
```

## B.3 High-level Relevance

```
The following is a caption from a
video: [" + text + "]. Based on this
 caption, generate two paraphrased
captions capturing the key
information and main themes, each of
 which should be in one sentence
with up to twenty words (Feel free
to add any details or
interpretations that you think
enhance the summary, even if they
are not directly mentioned in the
text.). Also Please generate text
without any comment.
```