

# Supplementary Materials: Multi-Modal Inductive Framework for Text-Video Retrieval

Anonymous Authors

## A TEMPORAL ATTENTION MECHANISM

The initialization of the temporal matrix  $H$  integrates the sequence of key-images with their semantic distances. Specifically,  $H$  reflects both the positional information of key images in the video sequence and the semantic similarities between the images. Each key image is assigned a positional encoding based on its sequence in the video. This step can be accomplished through variants of sine and cosine functions, akin to the positional encoding method in the Transformer model [20]:

$$\begin{aligned} PE_{(pos,2i)} &= \sin\left(pos/10000^{2i/d}\right), \\ PE_{(pos,2i+1)} &= \cos\left(pos/10000^{2i/d}\right), \end{aligned} \quad (\text{A.1})$$

where  $pos$  denotes the position index of the image within the video sequence,  $i$  is the dimension index, and  $d$  is the dimension of the encoding. This encoding reflects the relative position of images within the video sequence.

The semantic distance between key images is computed by comparing the similarity of image embeddings using cosine similarity:

$$SD_{ij} = \cos(\text{emb}_i, \text{emb}_j), \quad (\text{A.2})$$

where  $\text{emb}_i$  and  $\text{emb}_j$  are the embedding representations of the  $i$ -th and  $j$ -th key images, respectively.

To more accurately reflect the temporal discrepancies and semantic connections within video content and provide a meaningful starting point for the temporal attention mechanism, each element  $H_{ij}$  in  $H$  is initialized by combining positional encoding and semantic distance,

$$H_{ij} = \lambda \cdot PE_{ij} + (1 - \lambda) \cdot SD_{ij}, \quad (\text{A.3})$$

where a parameter  $\lambda$  balances their contributions. In this manner,  $H$  considers both the sequence of key images and their semantic relationships upon initialization. Connections between images that are both semantically similar and temporally adjacent are emphasized due to higher  $H_{ij}$  values. This, in turn, allows for the further refinement and optimization of these temporal relationships throughout the training process.

## B SPATIAL ATTENTION MECHANISM

The initialization of the spatial matrix  $G$  is pivotal for capturing the intricate spatial relationships within video frames, particularly focusing on regions or objects that are relevant to textual queries. To achieve this, we synergistically integrate geometric proximity and semantic similarity among objects within the frames<sup>1</sup>. This approach ensures that each embedding is imbued with rich spatial connections, thereby significantly enhancing the model’s ability to discern and prioritize spatial relations that are most informative for comprehending the video content in the context of textual queries.

Geometric proximity aims to encode the physical closeness between objects in a frame, which is essential for recognizing spatial

layouts and configurations. In initializing  $G$ , geometric proximity is quantified to encode the spatial layout and distances between different regions. This is crucial for recognizing spatial patterns and relationships that are inherently based on the arrangement of elements within a frame. For each region or object  $i$  and  $j$ , represented by their centroids  $c_i$  and  $c_j$ , the geometric proximity can be calculated using the Euclidean distance:

$$P_{ij} = \sqrt{(c_{ix} - c_{jx})^2 + (c_{iy} - c_{jy})^2}, \quad (\text{B.1})$$

where  $c_{ix}$  and  $c_{iy}$  are the x and y coordinates of centroid  $i$ , and similarly for  $j$ . The proximity matrix  $P \in \mathbb{R}^{K \times K}$  is then normalized to ensure that its values are scaled between 0 and 1, facilitating its combination with the semantic similarity matrix:

$$P'_{ij} = 1 - \frac{P_{ij} - \min(P)}{\max(P) - \min(P)}. \quad (\text{B.2})$$

Semantic similarity measures the closeness in meaning or function between regions or objects. Unlike geometric proximity, which is based on physical distances, semantic similarity is derived from the content and attributes of the regions. For initializing  $G$ , semantic similarity is computed using feature vectors extracted from each region. The similarity between regions  $i$  and  $j$  could be quantified using cosine similarity between their feature vectors, resulting in a similarity matrix  $S \in \mathbb{R}^{K \times K}$ . Let  $f_i$  and  $f_j$  be the feature vectors for regions  $i$  and  $j$ , respectively. The semantic similarity can be calculated using cosine similarity:

$$S_{ij} = \frac{f_i \cdot f_j}{\|f_i\| \|f_j\|}, \quad (\text{B.3})$$

The similarity matrix is normalized for effective integration:

$$S'_{ij} = \frac{S_{ij} - \min(S)}{\max(S) - \min(S)}. \quad (\text{B.4})$$

The final initialization of  $G$  synthesizes both geometric proximity and semantic similarity to encapsulate the comprehensive spatial relationships. This can be achieved by combining the proximity matrix  $P$  and the similarity matrix  $S$  using a weighted approach:

$$G = \alpha P' + (1 - \alpha) S', \quad (\text{B.5})$$

where  $\alpha \in [0, 1]$  is a tunable parameter that balances the influence of geometric proximity and semantic similarity. This weighted combination ensures that  $G$  not only captures the physical layout of regions within frames but also their content-based relationships. The resulting matrix  $G$  thus provides a foundational structure for the spatial attention mechanism to dynamically focus on regions that are most relevant to the textual queries, enhancing the model’s performance in video understanding and retrieval tasks. By initializing  $G$  with this comprehensive approach, the spatial attention mechanism is equipped to dynamically focus on the most relevant regions, enhancing the model’s ability to align video content with textual queries based on both spatial and content-based considerations.

<sup>1</sup>The objects of all key-images are obtained by Faster R-CNN [19].

## C TEXTUAL INSTRUCTION POOL

We construct a textual instruction pool. It contains 20 questions for extracting important entities or words from the given text. The all textual instruction pool are:

1. What key concepts are mentioned in this text?
2. What are the key words utilized in this text?
3. What types of entities are represented by these key words?
4. What actions are being undertaken by these entities?
5. What events are depicted within this text?
6. Which events are occurring, and which entities are involved?
7. Identify the principal characters within narrative.
8. How do these events contribute to the overarching storyline?
9. In what ways might the identified concepts develop further?
10. What motivates the actions of these entities, and how are their internal conflicts or resolutions depicted?
11. What thoughts or emotions might the text evoke?
12. What messages or perspectives is the text attempting to convey?
13. How do these elements reflect upon the events they describe?
14. Identify the primary characters in the narrative.
15. How do these events relate to the themes presented within the text?
16. What roles do these entities assume within their respective environments?
17. How are the interactions between entities portrayed?
18. What are the implicit or explicit conflicts present in the text?
19. What motivations underlie these events?
20. What are the direct or indirect consequences of an entity's actions on other entities?

## D VISUAL INSTRUCTION POOL

We also construct a visual instruction pool. It contains 20 questions for extracting important entities or events from the given video. The all visual instruction pool are:

1. What are the important entities in this video?
2. What are the relationships of these entities in this video?
3. What entities in the video do?
4. How do these actions interact with the identified objects and entities?
5. What is the current state of these entities?
6. What story or message is conveyed through these interactions?
7. Identify the main characters in the given narrative.
8. What are the types of entities corresponding to the key words in the video?
9. If the video features persons, please describe them.
10. How do these elements affect the development of the story or the conveyance of its themes?
11. What deeper meanings are conveyed through these events?

12. How to explain the dynamic interaction between entities in a video?
13. What role do these entities play in the video narrative?
14. How do the events in the video affect the entity relationship?
15. How do the relationships between the characters in the video evolve over time, and what triggers these changes?
16. What are the motivations behind the characters' actions?
17. How does the dynamic develop between the entities in the video?
18. What are the relationships between the entities or events in the video?
19. Provide a description of the person or object featured in the video.
20. Offer a description of the relationship between the individuals or objects within the video.

## E KNOWLEDGE CLUSTERING

**K-means clustering** begins by initializing  $k$  centroids randomly. Each question prompt is assigned to the nearest centroid based on a distance measure, typically the Euclidean distance. This step partitions the question prompts into  $k$  clusters based on the current centroids. For each cluster  $i$ , the new centroid  $c_i$  is calculated as the mean of all question prompts assigned to that cluster. The formula for recalculating the centroid of a cluster  $i$  is given by:

$$c_i = \frac{1}{|S_i|} \sum_{q_j^i \in S_i} q_j^i, \quad (\text{E.1})$$

where  $S_i$  is the set of question prompts assigned to cluster  $i$ , and  $|S_i|$  is the number of question prompts in  $S_i$ . This step moves the centroid  $c_i$  to the center of the question prompts in the cluster. Repeated until the centroids  $c_i$  no longer change significantly between iterations, indicating that the algorithm has converged. The final centroids  $c_i$  after convergence are used as the representative centers of each cluster. These centroids are considered the most typical representation of the question prompts within their respective clusters.

**Answer terms**  $\nabla$  is constructed by extracting keywords from each text-video pair. Specifically, for textual content, we employ the Stanford CoreNLP tool<sup>2</sup> to extract entities and triples from sentences. By removing duplicates, we obtain a concise list of keywords representing the core content of the text. Regarding video content, we utilize the Faster R-CNN model [19] to identify entities within key frames of the video. To translate these visual elements into actionable data, we use Oscar (Object-Semantics Aligned Pre-training) [11] to transfer the image into a text description for each image, then apply the Stanford CoreNLP tool to these captions to recognize actions and entities. This dual-step process allows for a comprehensive extraction of relevant terms from video content, encompassing both the entities present and the actions performed.

<sup>2</sup><http://corenlp.run/>

## F DATASETS

We conducted experiments on five widely used Text-Video Retrieval datasets: MSR-VTT (Microsoft Research Video to Text) [21], MSVD (Multimedia Semantic Description of Visual Data) [2], LSMDC (Large Scale Movie Description Challenge), DiDeMo (Distinct Describable Moments), and ActivityNet.

**MSR-VTT (Microsoft Research Video to Text)** is a well-known dataset specifically designed for open-domain video captioning. It consists of a large-scale collection of 10,000 video clips spanning 20 different categories. Each video clip in the dataset has been meticulously annotated with 20 English sentences by Amazon Mechanical Turks. The primary objective of this dataset is to facilitate the retrieval of video segments that best correspond to a given textual description. The dataset is divided into three standard splits: 6,513 clips for training, 497 clips for validation, and 2,990 clips for testing.

**MSVD (Multimedia Semantic Description of Visual Data)** is another widely used dataset for Text-Video Retrieval tasks. It comprises 1,970 videos with varying durations, ranging from 1 to 62 seconds. Each video in the dataset is associated with 40 English captions. For the purpose of our experiments, we utilized 1,200 videos for training, 100 videos for validation, and 670 videos for testing.

**LSMDC (Large Scale Movie Description Challenge)** is a dataset created in a joint effort by Johns Hopkins University and FAIR (Facebook AI Research). It comprises a vast collection of 118,081 video segments that are accompanied by bilingual subtitles in both English and French. The primary task of this dataset is to retrieve video segments that align with a given textual description or bilingual subtitles.

**DiDeMo (Distinct Describable Moments)** consists of 10K Flickr videos annotated with 40K text captions. It was collected for localizing moments in video with natural language queries, and it is usually considered for “paragraph-to-video” retrieval, concatenating all descriptions, since different queries describe different localized moments in a clip.

**ActivityNet** contains 20K YouTube videos annotated with 100K sentences, with 10K videos in the training set. It was intended for dense captioning, which involves both detecting and describing (possibly multiple) events in a video; we used the common val1 set with 4.9K videos.

## G COMPARISON METHODS

We compare our method with eight text-video retrieval models, which focus on cross-modality semantic representation and alignment among the videos and the texts: (1) **CE** [12] compresses high-dimensional video information into a condensed representation. (2) **SSB** [18] employs a generative model to group related samples. (3) **FROZEN** [1] incorporates attention mechanisms in both spatial and temporal dimensions. (4) **CLIP4Clip** [16] applies transfer learning from the image-to-text pre-trained CLIP model. (5) **CLIP2Video** [5] leverages the interactions between images and text, as well as improving the temporal relationships between video frames and video-text. (6) **X-CLIP** [17] introduces a multi-grained contrastive model for retrieval. (7) **MIL-NCE** [13] presents an image animation strategy that facilitates the conversion of commonly. (8) **DiCoSA**

[8] designs set-to-set alignment to simulate conceptualization and utilizes adaptive pooling mechanism to merge semantic concepts. (9) **LEAN** [10] employs a generative approach by modeling the joint probability. (10) **TS2-Net** [14] designs a transformative architecture that dynamically selects and adjusts video tokens to highlight informative content in videos, optimizing both temporal and spatial data analysis. (11) **EMCL** [7] introduces a contrastive learning approach using Expectation-Maximization to create compact, powerful video-and-language representations by decomposing features into a set of base elements, enhancing semantic representation capabilities. (12) **CenterCLIP** [23] utilizes a novel token clustering method to manage frame redundancy in videos by segmenting videos, clustering frames, and focusing on the most significant tokens, thereby reducing computational demands while preserving semantic integrity. (13) **X-Pool** [6] crafts a cross-modal attention model that allows text to identify and focus on semantically similar video frames, generating a video representation that is directly informed by textual data. (14) **ClipBERT** [9] pioneers a cost-effective learning framework for video-and-language tasks, leveraging sparse sampling to reduce the need for extensive video data during training. (15) **TT-CE** [3] introduces a distillation approach that combines cues from various text encoders to enhance the guidance provided to retrieval models, extending its efficiency to video modalities and reducing the necessity for multiple modalities during testing. (16) **DGL** [22] develops a cross-modal prompt tuning strategy that combines dynamic prompt generation with focused video attention, fostering richer interactions between text and video content. (17) **UATVR** [4] treats searches as distribution matching tasks, using specially designed tokens to dynamically gather nuanced semantic information for advanced reasoning.

## H IMPLEMENTATION DETAILS

For all baselines, we adopt the best hyper-parameters and copy results reported in the literature [8, 10, 13, 17]. We used PyTorch<sup>3</sup> as a deep learning framework to develop the TVR. All experiments were conducted on a server with four GPU (Tesla V100). The Llava version is llava-v1.5-13b in huggingface<sup>4</sup> for text and video initialization and the dimension was set to 768. The AdamW optimizer was chosen with a weight decay of 0.01. The batch size is set to 16. We performed a search for the learning rate in the range of  $1e-5$  to  $6e-5$ , and ultimately settled on  $3e-5$ . Training is performed using Adam optimizer with a learning rate of 0.001, learning rate decay of 0.2 every 50,000 steps, and a margin of 5.0 in the base loss function. We use a batch size of 512 and apply a dropout rate of 0.1 to prevent overfitting. The training was carried out for 30 epochs, with evaluation performed after the 10-th epoch. The model that performed the best on the validation set was selected and evaluated on the test set. The text and visual instruction pool have 10 questions respectively. The number of clusters  $k$  is the same as the number of questions in instruction pool. For hyper-parameters, the best coefficients  $\lambda_r$ ,  $\lambda_s$  are 0.6, and 0.3. To ensure fairness, all baseline models use the same data set partitioning. The learning rate is  $2e-4$ , batch size is 100, and dropout rate is 0.6. We use AdamW [15] to optimize the parameters. For the learning rate, we adopt the method of grid

<sup>3</sup><https://pytorch.org/>

<sup>4</sup><https://huggingface.co/liuhaotian/llava-v1.5-13b>

**Table I.1: Variant experiments on LSMDC dataset. “w/o” means removing corresponding module from the complete model. “repl.” means replacing corresponding module with the other module. “↑” denotes that higher is better. “↓” denotes that lower is better.**

Variants	Text-to-Video Retrieval					Video-to-Text Retrieval				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
<b>MMI-TVR (Ours)</b>	<b>26.6</b>	<b>44.6</b>	<b>54.1</b>	<b>7.0</b>	<b>51.8</b>	<b>24.3</b>	<b>43.2</b>	<b>53.9</b>	<b>10.0</b>	<b>45.3</b>
w/o Fine-tuning LVLM	24.2	42.6	52.9	9.0	53.3	22.0	41.7	51.5	12.4	47.1
w/o Inductive Reasoning Mechanism	25.5	42.8	53.6	8.0	52.9	23.2	42.8	52.9	11.0	46.2
w/o Temporal Attention Mechanism	25.9	43.0	53.4	8.0	52.2	23.1	42.6	52.8	11.0	46.5
w/o Spatial Attention Mechanism	25.4	43.7	54.2	7.0	52.0	23.4	43.4	53.1	11.0	46.1
repl. Self Mechanism	25.0	43.2	53.8	7.0	52.7	23.6	42.9	52.3	11.0	46.4
w/o Fine-Grained Knowledge Generation	24.1	42.6	52.6	9.0	53.1	22.9	41.2	51.2	12.0	47.3
w/o Textual Knowledge Generation	25.2	43.1	53.3	8.0	52.4	23.5	42.8	52.5	11.0	46.8
w/o Visual Knowledge Generation	25.8	43.9	53.5	8.0	52.5	23.7	42.3	52.6	11.0	46.6
w/o Knowledge Clustering	25.3	43.5	53.1	8.0	52.6	23.3	42.7	52.2	10.0	46.7

**Table I.2: Variant experiments on LSMDC dataset. “w/o” means removing corresponding module from the complete model. “repl.” means replacing corresponding module with the other module. “↑” denotes that higher is better. “↓” denotes that lower is better.**

Variants	Text-to-video retrieval on ActivityNet					Text-to-video retrieval on DiDeMo				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
<b>MMI-TVR (Ours)</b>	<b>45.7</b>	<b>75.5</b>	<b>87.8</b>	<b>2.0</b>	<b>5.2</b>	<b>51.3</b>	<b>76.6</b>	<b>86.2</b>	<b>2.0</b>	<b>10.1</b>
w/o Fine-tuning LVLM	42.5	72.7	85.9	3.0	6.9	48.4	74.0	81.7	6.0	12.6
w/o Inductive Reasoning Mechanism	44.3	74.1	85.3	3.0	6.7	49.8	74.5	84.0	3.0	11.5
w/o Temporal Attention Mechanism	44.5	74.3	86.7	3.0	6.5	50.1	75.6	85.3	3.0	10.8
w/o Spatial Attention Mechanism	44.8	74.6	86.5	3.0	6.1	50.2	75.9	85.4	3.0	11.3
repl. Self Mechanism	44.2	74.8	86.4	3.0	6.8	50.3	75.2	85.6	3.0	11.7
w/o Fine-Grained Knowledge Generation	43.7	73.4	85.8	4.0	7.2	49.5	74.3	84.1	4.0	12.4
w/o Textual Knowledge Generation	44.5	74.9	86.2	3.0	6.5	51.0	75.7	85.9	3.0	11.2
w/o Visual Knowledge Generation	44.9	74.2	86.6	2.0	6.6	50.6	75.1	85.7	3.0	11.0
w/o Knowledge Clustering	44.1	74.5	87.0	2.0	5.7	50.7	75.4	85.2	3.0	10.9

search with a step size of 0.0001. All hyper-parameter settings are tuned on the validation data by the grid search with 5 trials.

## I DISCUSSION FOR MODEL VARIANTS

To investigate the effectiveness of each module in our proposed model on other text-video retrieval datasets, we conducted variant experiments and showcased the results in Table I.1 and I.2. From the table, we can observe that all these observations highlight the effectiveness of each component in our model.

## J DISCUSSION FOR INTERPRETABILITY

We delve into the interpretability of the multi-modal inductive framework for TVR by examining the outcomes generated from prompts in both text-to-text and video-to-video contexts. The interpretability aspect is crucial for understanding how the proposed framework processes and aligns different modalities—textual queries with video content—through the responses generated from prompts, as shown in Figure J.1.

From the figure, we can observe that: 1) For text-to-text interaction, we present prompts that are designed to elicit specific information relevant to the query and the corresponding answers that the model generates. These examples illustrate how the model comprehends and extracts pertinent information from textual content. These prompts and their responses demonstrate the model’s ability to recognize and name key entities within a narrative, aiding in the retrieval of videos that feature similar characters or roles. By prompting the model to summarize critical events, we can align text queries with video content that depicts similar sequences of events. 2) For video-to-video interaction, prompts are structured to extract significant entities and events from the video content, with the responses providing insights into the model’s understanding and interpretation of video data. These examples illustrate the model’s capability to identify primary entities and their interactions, crucial for matching videos with textual queries that describe similar scenes or activities. The response to this prompt provides a



Key-image of the Video

Text: a person is explaining something.

**Textual Prompt:**

1. What are the key words mentioned in this text?
2. What types of entities are represented by these key terms?

**Textual Skeleton:**

1. The key words are "person" and "explaining."
2. The key entities are "person" (human) and "explaining" (action).

**Video Prompt:**

1. What are entities types of the key words in the video?
2. What the entities in the video do?

**Video Skeleton:**

1. Entity types: human (woman), vehicle (car), natural environment (trees).
2. The woman is driving, smiling, and possibly talking or recording.

(a) Example 1.



Key-image of the Video

Text: a boy and a girl play in ocean water on small surfboards.

**Textual Prompt:**

1. What key concepts are mentioned in this text?
2. What events are depicted within this text?

**Textual Skeleton:**

1. Key Concepts: Play, ocean water, surfboards, boy, girl."
2. The playful activity of surfing or surfboarding in the ocean by two children.

**Video Prompt:**

1. What the entities in the video do?
2. If the video contains person, Please description them.

**Video Skeleton:**

1. They are in the ocean, holding surfboards, and appear to be returning to the shore after a ride.
2. The video shows two children play in the ocean.

(b) Example 2.



Key-image of the Video

Text: an athlete in a red uniform walks by a man in a blue sports coat.

**Textual Prompt:**

1. What are the key words utilized in this text?
2. Identify the principal characters within narrative.

**Textual Skeleton:**

1. Key Words: Athlete, red uniform, walks, man, blue sports coat.
2. The athlete in the red uniform and the man in the blue sports coat are the main characters in this narrative.

**Video Prompt:**

1. What are the relationships of these entities in this video?
2. What is the current state of these entities?

**Video Skeleton:**

1. The athlete in red is interacting with the man in the suit, which suggests ...
2. The man in the suit is standing and seems to be in a state of attentiveness, possibly listening to the athlete.

(c) Example 3.

Figure J.1: The examples of text-videos pairs for the prompts and answers.

concise overview of the video's narrative, enabling more accurate alignment with text queries that seek specific actions or storylines.

The given examples of prompts and their respective answers shed light on how the proposed framework interprets textual and visual information, facilitating a more nuanced understanding of video-text alignment. By generating answers that highlight entities, events, and their interrelations, the model demonstrates its ability to parse and synthesize information across modalities effectively. This interpretability not only enhances the framework's utility in matching videos with text queries but also offers insights into its reasoning process, making it a valuable tool for a wide range of applications in multimedia retrieval and analysis.

## K DISCUSSION FOR HYPERPAMATERS

To better understand the effectiveness of parameters of the loss function, we construct experiments with different weights on the MSR-VTT, as shown in Figure K.1. We fix other parameters on the best value to evaluate each parameter respectively. From the figure, we can observe that: 1) The model achieves the best results as the  $\lambda_r$ , and  $\lambda_s$  are 0.6, and 0.3. It demonstrates that text-video retrieval loss  $\mathcal{L}_r$ , and self-identification enhancement loss  $\mathcal{L}_s$  are helpful for retrieval. 2) As the value of parameters increases in curves, the performance tends to peak first and then decline. And the weight of  $\mathcal{L}_r$  is significantly higher than others. It demonstrates that weight learning loss is most helpful for the TVR compared to other parameters. All the observation indicates the effectiveness of TVR loss and self-identification enhancement loss.

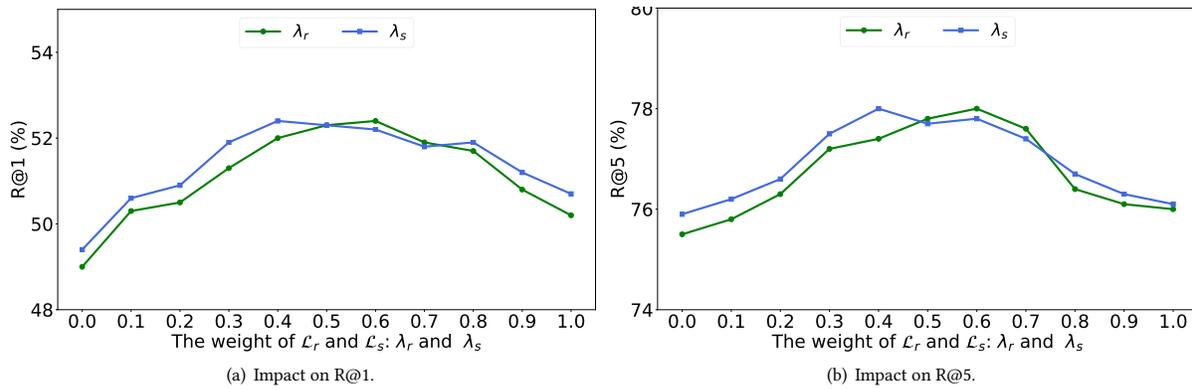


Figure K.1: The performance of different hyperparameters of  $\lambda_r$ , and  $\lambda_s$ . The larger the circle, the higher the performance.

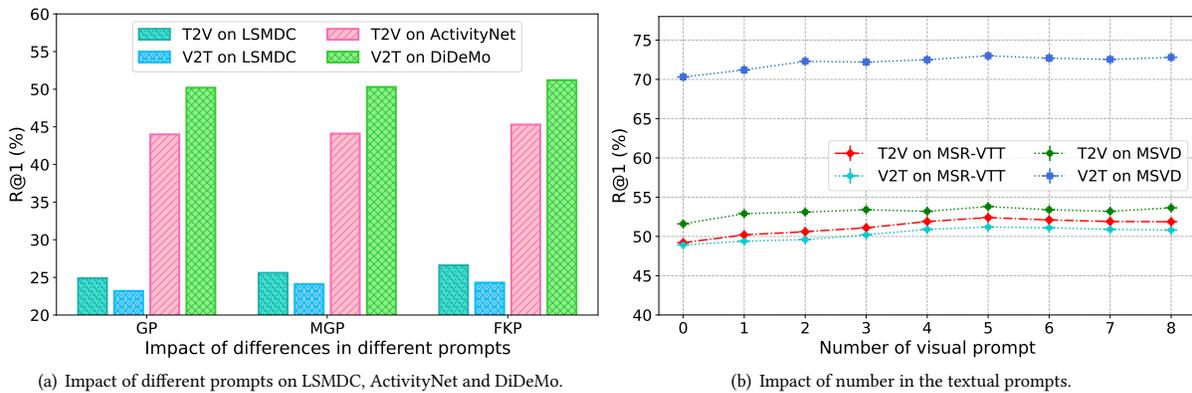


Figure L.1: Impact of different knowledge generation and number of prompts.

## L DISCUSSION FOR KNOWLEDGE GENERATION AND TEXTUAL PROMPTS

To further analyze the impact of different question prompts on the model's performance, we conducted a thorough comparison by selecting various prompts as input on more text-video retrieval datasets, as shown in Figure L.1 (a). The observation demonstrates that the information provided by the fine-grained knowledge prompt module is particularly valuable and the prompt generation serves as crucial information for enhancing retrieval.

We further investigated the performance of the model by examining its performance with different numbers of textual question prompts, as shown in Figure L.1 (b). From the figure, we can observe that using a moderate number of prompts, such as five, is sufficient to achieve optimal performance. The observation demonstrates that within a certain range of prompt numbers, the model is already capable of effectively capturing the correlated information between the text and video and gained valuable insights into the relationship between the number of prompts and the model's performance in text-video retrieval.

## REFERENCES

- [1] 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 *ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. 1708–1718.
- [2] David Chen and William B 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*. 190–200.
- [3] 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*. IEEE, 11563–11573. <https://doi.org/10.1109/ICCV48922.2021.01138>
- [4] Bo Fang, Wenhao Wu, Chang Liu, Yu Zhou, Yuxin Song, Weiping Wang, Xiangbo Shu, Xiangyang Ji, and Jingdong Wang. 2023. UATVR: Uncertainty-Adaptive Text-Video Retrieval. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 13677–13687. <https://doi.org/10.1109/ICCV51070.2023.01262>
- [5] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. 2021. CLIP2Video: Mastering Video-Text Retrieval via Image CLIP. *CoRR* abs/2106.11097 (2021). arXiv:2106.11097
- [6] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. 2022. X-Pool: Cross-Modal Language-Video Attention for Text-Video Retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 4996–5005. <https://doi.org/10.1109/CVPR52688.2022.00495>
- [7] Peng Jin, Jinfā Huang, Fenglin Liu, Xian Wu, Shen Ge, Guoli Song, David A. Clifton, and Jie Chen. 2022. Expectation-Maximization Contrastive Learning

- 697 for Compact Video-and-Language Representations. 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*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). [http://papers.nips.cc/paper\\_files/paper/2022/hash/c355566ce402de341c3320cf69a10750-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2022/hash/c355566ce402de341c3320cf69a10750-Abstract-Conference.html)
- 698
- 699
- 700
- 701 [8] Peng Jin, Hao Li, Zesen Cheng, Jinfa Huang, Zhennan Wang, Li Yuan, Chang Liu, and Jie Chen. 2023. Text-Video Retrieval with Disentangled Conceptualization and Set-to-Set Alignment. *CoRR abs/2305.12218* (2023). arXiv:2305.12218
- 702
- 703 [9] 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*. Computer Vision Foundation / IEEE, 7331-7341. <https://doi.org/10.1109/CVPR46437.2021.00725>
- 704
- 705
- 706
- 707 [10] Qian Li, Lixin Su, Jiashu Zhao, Long Xia, Hengyi Cai, Suqi Cheng, Hengzhu Tang, Junfeng Wang, and Dawei Yin. 2024. Text-Video Retrieval via Multi-Modal Hypergraph Networks. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining, WSDM 2024, Merida, Mexico, March 4-8, 2024*, Luz Angelica Caudillo-Mata, Silvio Lattanzi, Andrés Muñoz Medina, Leman Akoglu, Aristides Gionis, and Sergei Vassilvitskii (Eds.). ACM, 369-377. <https://doi.org/10.1145/3616855.3635757>
- 708
- 709
- 710
- 711
- 712 [11] 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. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *ECCV (Lecture Notes in Computer Science, Vol. 12375)*. Springer, 121-137. [https://doi.org/10.1007/978-3-030-58577-8\\_8](https://doi.org/10.1007/978-3-030-58577-8_8)
- 713
- 714
- 715
- 716 [12] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. 2019. Use What You Have: Video retrieval using representations from collaborative experts. In *BMVC 2019, Cardiff, UK, September 9-12, 2019*. 279.
- 717
- 718 [13] Yu Liu, Huai Chen, Lianghua Huang, Di Chen, Bin Wang, Pan Pan, and Lisheng Wang. 2022. Animating Images to Transfer CLIP for Video-Text Retrieval. In *SIGIR 2022, Madrid, Spain, July 11 - 15, 2022*. 1906-1911.
- 719
- 720
- 721 [14] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. 2022. 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 (Lecture Notes in Computer Science, Vol. 13674)*, Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer, 319-335. [https://doi.org/10.1007/978-3-031-19781-9\\_19](https://doi.org/10.1007/978-3-031-19781-9_19)
- 722
- 723
- 724
- 725 [15] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *ICLR. OpenReview.net*. <https://openreview.net/forum?id=Bkg6RiCqY7>
- 726
- 727 [16] 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 (2022), 293-304.
- 728
- 729 [17] 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 *MM 2022, Lisboa, Portugal, October 10 - 14, 2022*. 638-647.
- 730
- 731 [18] Mandela Patrick, Po-Yao Huang, Yuki Markus Asano, Florian Metz, Alexander G. Hauptmann, João F. Henriques, and Andrea Vedaldi. 2021. Support-set bottlenecks for video-text representation learning. In *ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.
- 732
- 733 [19] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (Eds.). 91-99. <https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html>
- 734
- 735
- 736
- 737
- 738 [20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998-6008. <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
- 739
- 740
- 741 [21] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR 2016*. 5288-5296.
- 742
- 743 [22] Xiangpeng Yang, Linchao Zhu, Xiaohan Wang, and Yi Yang. 2024. DGL: Dynamic Global-Local Prompt Tuning for Text-Video Retrieval. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (Eds.). AAAI Press, 6540-6548. <https://doi.org/10.1609/AAAI.V38I7.28475>
- 744
- 745 [23] Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. 2022. CenterCLIP: Token Clustering for Efficient Text-Video Retrieval. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 970-981. <https://doi.org/10.1145/3477495.3531950>
- 746
- 747
- 748
- 749
- 750
- 751
- 752
- 753
- 754