Bridging Video-text Retrieval with *Multiple Choice Questions*
Appendix

A. Visualization

In our method, the pretext task MCQ is performed using a parametric module BridgeFormer, to answer multiple choice questions. We construct questions through erasing the content phrases (*i.e*. noun and verb phrases) of the text, and BridgeFormer is trained to select the correct answer from multiple choices by resorting to the local tokens of VideoFormer. Specifically, given question text tokens from TextFormer as the query, and video tokens from VideoFormer as the key and value, BridgeFormer performs cross-modality attention between them.

A.1. Answering Noun Questions

We first visualize the cross-modality attention between noun question tokens and video tokens in Fig. 1. In the second column, the noun phrase marked in blue (Q1) is erased as the question, and in the third column, the noun phrase marked in green (Q2) is erased as the question. In Fig. 1 (a), when "an old couple" is erased as the question (Q1), BridgeFormer focuses on video tokens that depict the appearance characteristics of the persons, and when "a plate of bread" is erased (Q2), it focuses on object video tokens on the table. In Fig. 1 (d), when "football" is erased (Q1), BridgeFormer focuses on the object video tokens that can be associated with "play", and when the location phrase "countryside lawn" is erased (Q2), it pays more attention to the video tokens in the background to infer the answer. BridgeFormer attends to video patches with specific object information to answer questions, which also shows that VideoFormer extracts accurate spatial content from videos.

A.2. Answering Verb Questions

We further visualize the cross-modality attention between verb question tokens and video tokens in Fig. 2. Three frames are sampled from a video and the verb phrase marked in blue is erased as the question. In Fig. 2 (a), when the verb "cutting" is erased, BridgeFormer focuses on the motion of the spoon on the pizza, and in Fig. 2 (b), when the verb "drinking" is erased, it follows the movement of the hand holding a cup of water. BridgeFormer focuses on object motions of video tokens to answer verb questions, which also shows that VideoFormer captures temporal dynamics of videos.

(a) "An old couple/[?] (Q1) are drinking coffee, and there is a plate of bread/[?] (Q2) on the table in front of them."

(b) "A girl is walking with a dog/[?] (Q1) near a lake/[?] (Q2), and there is a meadow on her left."

(c) "A woman wearing a pink dress/[?] (Q1) and carrying a black handbag/[?] (Q2) is walking in the park."

(d) "Parents and kids are playing football/[?] (Q1) on the countryside lawn/[?] (Q2)."

Figure 1. The visualization of the cross-modality attention between the text tokens of **noun questions** (as query) and video tokens (as key and value) from BridgeFormer. In the second column, the noun phrase marked in blue (Q1) is erased as the question, and in the third column, the noun phrase marked in green (Q2) is erased as the question. BridgeFormer attends to video patches with specific object information to answer noun questions.

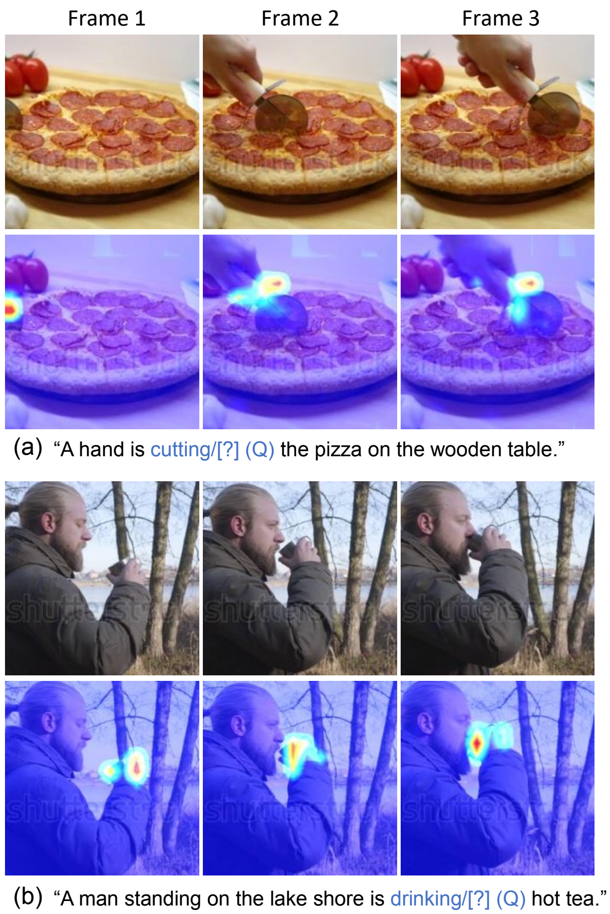

Figure 2. The visualization of the cross-modality attention between the text tokens of **verb questions** (as query) and video tokens (as key and value) from BridgeFormer. Three frames sampled from a video are shown and the verb phrase marked in blue (Q) is erased as the question. BridgeFormer focuses on object motions of video tokens to answer verb questions.

B. CLIP-based Pre-training

Because of the prominent success of the CLIP [9] (Contrastive Language-Image Pre-training) in learning image-text representations, which is pre-trained on 400 million image-text pairs, some recent work [6, 8] utilize the pre-trained CLIP for text-to-video retrieval. We also initialize our model from CLIP weights to pre-train a model following the setting of CLIP4Clip [6]. Specifically, we use the pre-trained CLIP (ViT-B/32) as the backbone of VideoFormer and TextFormer, and randomly initialize BridgeFormer. The comparisons between our method and other CLIP-initialized methods are shown in Table. 1. We can observe that our CLIP-based pre-trained model achieves higher performance for text-to-video retrieval on three datasets with under both the zero-shot and fine-tune evaluation. Our pretext task MCQ also benefits CLIP-based video-text pre-training for downstream text-to-video retrieval.

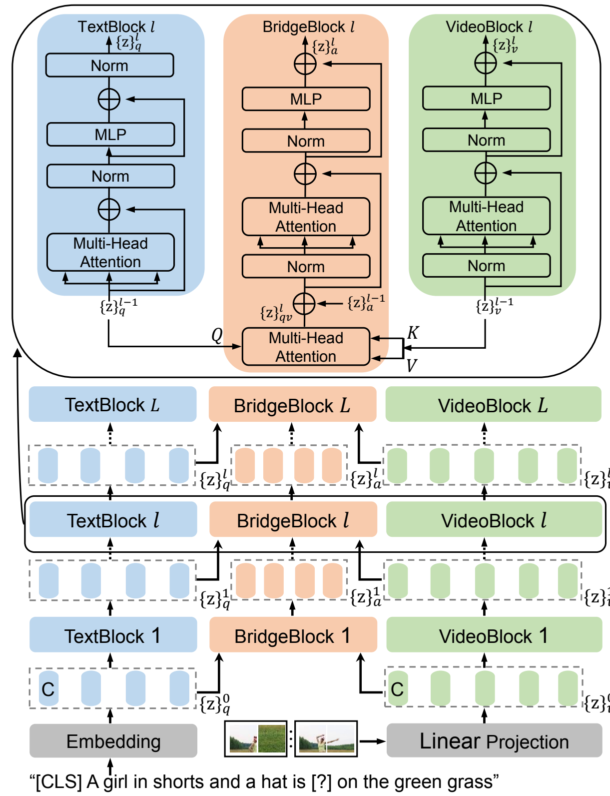

Figure 3. The architecture of TextFormer, VideoFormer and BridgeFormer, which contain a stack of TextBlocks, VideoBlocks and BridgeBlocks respectively. Tokens from all-level VideoBlock and TextBlock are fed into the corresponding BridgeBlock to perform cross-modal attention and then are added to the output tokens of the previous BridgeBlock (if any). Each block performs a series of operations such as multi-head attention [5], normalization (norm) and multi-layer perception [4] (MLP).

C. Detailed Model Architecture

Our method consists of three components, including a VideoFormer, a TextFormer and a BridgeFormer. Each component is made up of a stack of blocks as shown in Fig. 3. TextBlock and VideoBlock adopt the structure of BERT [4] and ViT [5] respectively, each performing a series of operations such as multi-head attention [5], normalization (norm) and multi-layer perception [4] (MLP). BridgeBlock takes question text tokens as the query and video tokens as the key and value to perform the cross-modality attention for the interacted tokens. The interacted tokens added with the output tokens from the previous BridgeBlock further go through a series of operations similar to those in the VideoBlock for temporal and spatial self-attention.

Table 1. Text-to-video retrieval results of models initialized from CLIP [9] weights on different datasets under zero-shot and fine-tune evaluation, where **higher** R@k and **lower** MdR (Median Rank) and MnR (Mean Rank) indicate better performance.

	MSR-VTT					MSVD					LSMDC				
Method	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
CLIP-straight [8]	31.2	53.7	64.2	4.0	-	37.0	64.1	73.8	3.0	-	11.3	22.7	29.2	56.5	-
CLIP4Clip [6]	32.0	57.0	66.9	4.0	34.0	38.5	66.9	76.8	2.0	17.8	15.1	28.5	36.4	28.0	117.0
Ours	**33.2**	**58.0**	**68.6**	**4.0**	**25.7**	**48.4**	**76.4**	**85.8**	**2.0**	**7.4**	**15.5**	**30.7**	**38.7**	**22.0**	**97.9**
CLIP4Clip [6]	43.1	70.4	**80.8**	2.0	16.2	46.2	76.1	84.6	2.0	10.0	20.7	38.9	47.2	13.0	65.3
Ours	**44.9**	**71.9**	80.3	**2.0**	**15.3**	**54.4**	**82.8**	**89.4**	**1.0**	**6.1**	**21.8**	**41.1**	**50.6**	**10.0**	**60.5**

Table 2. Comparisons between the video encoder in our method and Frozen [1]. The evaluation is performed on zero-shot text-to-video retrieval on MSR-VTT, where **higher** R@k and **lower** MdR (Median Rank) indicate better performance. "# Params" denotes the number of parameters of the video encoder (M: million).

Method	R@1	R@5	R@10	MdR	# Params
Frozen [1]	18.7	39.5	51.6	10.0	114M
Ours	**22.3**	**43.8**	**52.0**	**9.0**	86M

Table 3. The effects of the prompt "[MASK]" for noun and verb representations, where "End", "Middle" and "Start" denote the location of the prompt. For zero-shot text-to-video retrieval on MSR-VTT, **higher** R@k is better. For zero-shot action recognition on HMDB51 and UCF101, **higher** top-1 accuracy is better.

	MSR-VTT			HMDB51	UCF101
Method	R@1	R@5	R@10	Top-1	Top-1
w/o Prompt	23.1	43.5	54.3	34.8	45.8
End	24.2	**45.7**	54.4	33.4	48.5
Middle	24.3	43.2	53.9	33.1	46.4
Start	**25.1**	45.4	**55.4**	**34.9**	**51.4**

D. VideoFormer

Video Input. VideoFormer takes a video $V \in R^{M\times 3\times H\times W}$ as input containing variable M frames of resolution $H \times W$. The input video is first divided into $M \times N$ patches of size $P \times P$, where $N = HW/P^2$. The video patches $v \in R^{M\times 3\times N\times P\times P}$ are fed into a linear projection head with a convolutional layer and are flattened into a sequence of tokens $z_v \in R^{M\times N\times D}$, where D is the number of embedding dimensions. Following BERT [4], a learnable [CLS] token is concatenated to the beginning of the token sequence, which is used to produce the final video representations. Learnable spatial positional embeddings $E_{pos} \in R^{(N+1)\times D}$ are added to each video token as the final input token sequence $z_v^0 \in R^{(1+M\times N)\times D}$ and all patches in the same spatial location in different frames are given the same spatial positional embedding.

Modification to ViT. VideoFormer is built upon a vision transformer ViT [5], and consists of a stack of VideoBlocks. We make a minor modification to the original ViT to allow for the input of video frames with variable length. Specifically, given $z_v^{l-1} \in R^{(1+M\times N)\times D}$ from previous VideoBlock, we perform multi-head attention (MSA) [5] for the [CLS] token through attending to all $(1 + M \times N)$ patches across time and space for temporal and spatial self-attention. For the rest $(M \times N)$ patch tokens, MSA is performed within each of M frames with $N + 1$ tokens (N patch tokens and 1 [CLS] token) for spatial self-attention. The video representations are obtained from the [CLS] token of the final VideoBlock.

Comparison with Frozen. Frozen [1] also adopts ViT [5] as the video encoder, and adds temporal attention blocks based on the spatial attention blocks of ViT to encode videos with variable-length sequences. As shown in Table. 2, compared with Frozen, our VideoFormer decreases 28 million parameters. Furthermore, the model without the pretext task MCQ indeed takes the same pre-training approach as Frozen except for the video encoder, and achieves better results for zero-shot text-to-video retrieval on MSR-VTT [12], which proves the efficiency and effectiveness of our VideoFormer.

E. Prompt for Phrase Representation

In our method, BridgeFormer is trained to select the correct answer by contrasting noun answer representations with noun representations, and contrasting verb answer representations with verb representations. Accurate representations for noun and verb phrases are essential. Since TextFormer is trained with full sentences, it fails to encode accurate representations for phrases when it takes a single noun or verb phrase as the input due to the lack of context. Motivated by the success of prompt engineering [9], we add "[MASK]" before the noun and verb phrase (*e.g.* "[MASK] [MASK] [MASK] green grass") to extract noun or verb representations from TextFormer. We show ablation studies of the prompt "[MASK]" for noun and verb representations in Table. 3, where each model is pre-trained using 1 frame. The model without the prompt "[MASK]" takes a single noun or verb phrase as inputs, and achieves the worse results on both the zero-shot text-to-video retrieval and action recognition, showing that TextFormer cannot understand the semantics accurately with a single noun or verb phrase as inputs. The model with the prompt "[MASK]" at the beginning of the phrase achieves the best results in general, and we adopt this practice in our method.

F. More Discussions about Related Work

F.1. Video Question Answering (VQA)

Works on video question answering (VQA) [2, 7, 10, 14] aims to answer questions about videos through training a model with question and answer pairs, which cannot be directly applied for pre-training as they are deliberately optimized for increasing VQA accuracy. By contrast, our work aims to learn downstream-agnostic generic features for video-text retrieval, where a new pretext task, multiple choice questions, is proposed to enhance the semantic associations between video and text. Our paper *is the first to* use the form of VQA as a pre-training pretext task, with *two key innovations*: the MCQ loss and the BridgeFormer module. BridgeFormer smoothly bridges the final objective of learning well-aligned video and text features with the regularization of a VQA pretext task.

F.2. Video-text Retrieval with Nouns and Verbs

Works [3, 11, 13, 15] solved video-text retrieval by focusing on verbs and nouns of texts, which are specially designed for retrieval with verbs and nouns as the refined text representations to directly align with videos. By contrast, we exploit the rich semantics of nouns and verbs in the text to build questions for improving text and video encoders.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, pages 1728–1738, 2021. 3

[2] Aman Chadha, Gurneet Arora, and Navpreet Kaloty. iperceive: Applying common-sense reasoning to multi-modal dense video captioning and video question answering. *arXiv preprint arXiv:2011.07735*, 2020. 4

[3] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *CVPR*, pages 10638–10647, 2020. 4

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 2, 3

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 2, 3

[6] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 2, 3

[7] Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *CVPR*, pages 6884–6893, 2017. 4

[8] Jesús Andrés Portillo-Quintero, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. A straightforward framework for video retrieval using clip. In *Mexican Conference on Pattern Recognition*, pages 3–12. Springer, 2021. 2, 3

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 2, 3

[10] Arka Sadhu, Kan Chen, and Ram Nevatia. Video question answering with phrases via semantic roles. *arXiv preprint arXiv:2104.03762*, 2021. 4

[11] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *ICCV*, pages 450–459, 2019. 4

[12] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016. 3

[13] Ran Xu, Caiming Xiong, Wei Chen, and Jason Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *AAAI*, volume 29, 2015. 4

[14] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, pages 1686–1697, 2021. 4

[15] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *CVPR*, pages 3537–3545, 2019. 4