Building a Multi-modal Spatiotemporal Expert for Zero-shot Action Recognition with CLIP

Yating Yu*, Congqi Cao*†, Yueran Zhang, Qinyi Lv, Lingtong Min, Yanning Zhang

Northwestern Polytechnical University, Xi'an Shaanxi, 710129, China
yatingyu@mail.nwpu.edu.cn, congqi.cao@nwpu.edu.cn, zhangyueran3@mail.nwpu.edu.cn, {lvqinyi, minlingtong, ynzhang}@nwpu.edu.cn

Abstract

Zero-shot action recognition (ZSAR) requires *collaborative multi-modal spatiotemporal understanding*. However, fine-tuning CLIP directly for ZSAR yields suboptimal performance, given its inherent constraints in capturing essential temporal dynamics from both vision and text perspectives, especially when encountering novel actions with fine-grained spatiotemporal discrepancies. In this work, we propose **Spatiotemporal Dynamic Duo** (STDD), a novel CLIP-based framework to comprehend multi-modal spatiotemporal dynamics synergistically. For the vision side, we propose an efficient Space-time Cross Attention, which captures spatiotemporal dynamics flexibly with simple yet effective operations applied before and after spatial attention, without adding additional parameters or increasing computational complexity. For the semantic side, we conduct spatiotemporal text augmentation by comprehensively constructing an Action Semantic Knowledge Graph (ASKG) to derive nuanced text prompts. The ASKG elaborates on static and dynamic concepts and their interrelations, based on the idea of decomposing actions into spatial appearances and temporal motions. During the training phase, the frame-level video representations are meticulously aligned with prompt-level nuanced text representations, which are concurrently regulated by the video representations from the frozen CLIP to enhance generalizability. Extensive experiments validate the effectiveness of our approach, which consistently surpasses state-of-the-art approaches on popular video benchmarks (*i.e.*, Kinetics-600, UCF101, and HMDB51) under challenging ZSAR settings.

Code — https://github.com/Mia-YatingYu/STDD
Extended version — https://arxiv.org/abs/2412.09895

1 Introduction

Zero-shot action recognition (ZSAR) aims to classify video actions from novel categories that are not present in the training of models. A strong ZSAR learner should be endowed with *collaborative multi-modal spatiotemporal understanding*, where the statics and dynamics of videos and semantics should be aligned meticulously. Otherwise, it would inevitably lead to an ambiguous comprehension of actions. Let's first delve into a simple example. In Figure 1(a), a model lacking the ability to align visual contexts with static concepts *e.g.*, the *barbell*, may confuse whether the actor is performing *Dead Lifting* or just doing *Body Weight Squats*. Conversely, as shown in Figure 1(b), a model might struggle to generalize to novel actions of *Clean and Jerk* and *Snatch Weight Lifting*, if it falters in aligning nuanced multi-modal spatiotemporal dynamics of *Weight Lifting*, as they exhibit significant static visual affinities.

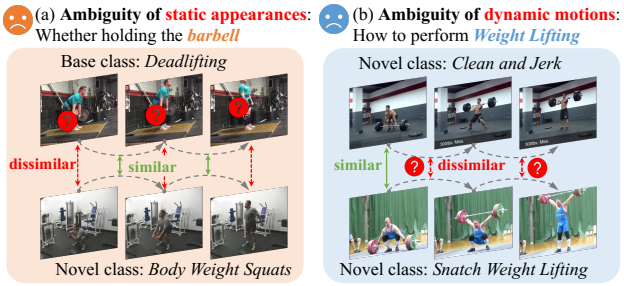

Figure 1: Illustration of the challenges without *collaborative multi-modal spatiotemporal understanding*. (a) A model lacking static context alignment may misidentify the novel class due to the ambiguity associated with the *barbell*. (b) It might also struggle generalizing to other novel weightlifting actions, due to the subtle dynamic differences and strong visual similarities.

Contrastive Language-Image Pretraining (CLIP) (Radford et al. 2021) has shown exceptional zero-shot inference in image-based tasks, benefiting from its strong generalization capability of the visual and linguistic alignment. Inspired by its success, CLIP can serve as a spatial expert for aligning static visual-semantic context of actions by processing the video frame-by-frame. However, due to its limitations in capturing temporal dynamics effectively, recent attempts (Wu, Sun, and Ouyang 2023; Wang, Xing, and Liu 2021; Yang et al. 2023) have been made to adapt CLIP for general action recognition. Despite notable advancements obtained by additional temporal modeling (Lin et al. 2022; Tu et al. 2023; Pan et al. 2022), they compromise with less-informative class-level prompts such as "*a*

*These authors contributed equally.
†Corresponding author.

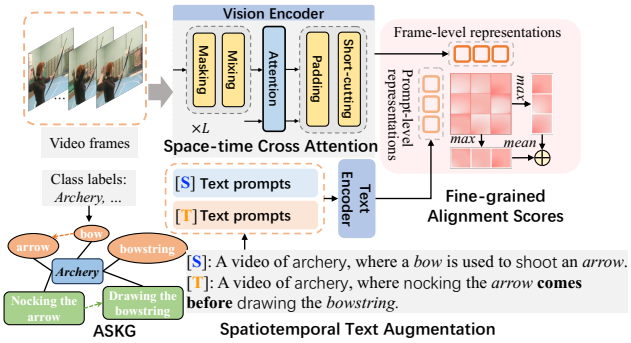

Figure 2: **Overview of our framework.** With a four-step operation applied within each block, we transform the spatial attention into novel Space-time Cross Attention. Spatiotemporal text augmentation is conducted to derive spatial and temporal text prompts, where multi-modal dynamics are meticulously aligned in a fine-grained manner.

video of {archery}", thus faltering when encountering novel actions with fine-grained spatiotemporal discrepancies. Other works (Chen et al. 2024; Jia et al. 2024; Wu et al. 2023) leverage Large Language Models (LLMs) (Brown et al. 2020; Achiam et al. 2023; Floridi and Chiriatti 2020) to extend CLIP with specialized knowledge, facilitating zero-shot generality. However, they often lay particular emphasis on semantic augmentation yet lacking efficient spatiotemporal understanding in visual world.

With these in mind, our key insight rests in extending the spatial expert *i.e.*, CLIP, into an effective spatiotemporal expert for ZSAR to comprehend multi-modal spatiotemporal dynamics synergistically. To this end, we propose a novel CLIP-based framework, termed **Spatiotemporal Dynamic Duo** (STDD), which meticulously aligns spatiotemporal visual contexts with refined static and dynamic text prompts, as shown in Figure 2. Our framework enables flexible capture of spatiotemporal dynamics for efficient multi-scale cross-frame interaction without requiring additional parameters or increasing computational complexity. Specifically, we realize this by implementing a four-step process *i.e.*, masking, mixing, padding and short-cutting, applied either before or after spatial attention. The masking strategy (Qing et al. 2023) discards a proportion of visual tokens, and the subsequent channel mixing (Bulat et al. 2021) is performed with visible tokens at the interleaved positions across frames. Then, the padding operation restores the original token quantity and spatial positions, while the ensuing short-cutting technique (He et al. 2016) is employed to seamlessly integrate dynamics into the primary vision encoding pathway. The tailored combination of these operations jointly transforms the spatial attention into a novel Space-Time Cross Attention (STCA). Besides, to acquire specialized action semantics systematically, we prompt LLMs to construct an Action Semantic Knowledge Graph (ASKG) that conceptually incorporates static appearances and dynamic motions of actions with a more structured and interpretable representation. Then, we perform spatiotemporal text augmentation to derive nuanced text prompts by parsing static and dynamic semantics in the ASKG. Finally, frame-level video representations are meticulously aligned with prompt-level text representations, which are concurrently regulated by video representations from the frozen CLIP to enhance generalizability.

Overall, our contributions can be summarized as follows:

- We introduce a novel CLIP-based framework, named **Spatiotemporal Dynamic Duo** (STDD), which synergistically comprehends dynamics for vision-text context refinement, facilitating *multi-modal spatiotemporal understanding*.
- For the vision side, we propose to transform spatial attention into Space-time Cross Attention through a four-step operation to capture cross-frame dynamics without additional parameters or increased computational complexity.
- For the semantic side, we propose to perform spatiotemporal text augmentation by comprehensively constructing an Action Semantic Knowledge Graph, which articulates static appearances and dynamic motions of actions.
- Extensive experiments on three popular benchmarks verify the effectiveness and superiority of our method, consistently achieving state-of-the-art performance.

2 Related Work

Adapting CLIP for Action Recognition. Recently, Vision Language Models (VLMs) (Sanghi et al. 2022; Bao et al. 2022; Yu et al. 2022) have demonstrated efficient multi-modal alignment, achieving impressive results in zero-shot inference. There is also a plethora of work (Mao et al. 2024; Qian, Xu, and Hu 2024; Li et al. 2024) using knowledge learned in VLMs to video understanding tasks in a zero-shot manner. Adapting CLIP to videos (Wang, Xing, and Liu 2021; Wang et al. 2023) is a common practice when designing a generalized video learner, where the key lies in utilizing additional temporal context for efficient video understanding. Recently, despite some works (Wu, Sun, and Ouyang 2023; Zhu et al. 2023; Wang, Xing, and Liu 2021) fully finetune the backbone with a video-header on top of CLIP, a collection of methods (Yang et al. 2023; Lin et al. 2022) work on parameter-efficient finetuning (PEFT) (Gao et al. 2022; Jie and Deng 2022), aiming to reduce trainable parameters, such as adapter-based (Yang et al. 2023; Lee, Lee, and Choi 2024; Cao et al. 2024b), prompt-based (Wasim et al. 2023; Ahmad, Chanda, and Rawat 2023) and decoder-based (Lin et al. 2022) methods. There are also other works (Huang et al. 2024; Rasheed et al. 2023) exclude temporal modeling, while aiming to adapt features from image to videos by distilling knowledge from pre-trained CLIP. In contrast, our proposed framework facilitates temporal dynamic modeling without additional parameters, where the refined spatiotemporal representations are regulated by video representations from the frozen CLIP to preserve generalizability.

Space-time Self-Attention. Recently, the self-attention mechanism inherent in the ViT architecture (Dosovitskiy et al. 2020) for spatial modeling has been extensively

adopted for video recognition. Due to the heavy complexity burden of full space-time attention, some prior works (Bertasius, Wang, and Torresani 2021; Arnab et al. 2021) focus on factorizing spatial and temporal attention to adapt 3D data. AIM (Yang et al. 2023) follows this idea by simply reusing pre-trained CLIP self-attention to perform temporal adaption, yet it nearly doubles the depth of the pre-trained encoder. Open-VCLIP (Weng et al. 2023) expands the temporal attention view *i.e.*, dimension, for aggregating the global temporal information, maintaining the weight dimensions of the CLIP. Other variants (Lin, Gan, and Han 2019; Wang, Cao, and Zhang 2022) adopt the "shift trick" (Wu et al. 2018) with zero-cost dimensionality reduction to achieve temporal modeling at a layer level. Most related to ours is X-ViT (Bulat et al. 2021), which constructs the key vectors by mixing information from tokens located at the same spatial location within a local temporal window. We share the similar "shift trick", but our approach applies masks to spatial tokens ahead of mixing, which enables the mixing of information from tokens at interleaved spatial locations within a local spatiotemporal window.

Semantic Knowledge for Video-text Alignment. Semantic knowledge provides a bridge among actions, allowing the model to generalize to novel categories based on their semantic connections. Early works usually design handcrafted attributes (Mandal et al. 2019; Mishra, Pandey, and Murthy 2020) or utilize object features (Jain et al. 2015) to represent action semantics. Later, word-embedding methods (Chen and Huang 2021; Wang, Cao, and Zhang 2023; Cao et al. 2024a) are adopted for semantic representations. Recently, due to the versatility of LLMs, some works (Chen et al. 2024; Jia et al. 2024) construct knowledge-rich text descriptions by harnessing the responses from LLMs. And others (Lin et al. 2023; Wu et al. 2024) introduce the captions generated by BLIP (Li et al. 2022, 2023) for multi-modal semantic knowledge. Diverging from existing semantic augmentation, we comprehensively construct a structured Action Semantic Knowledge Graph (ASKG) featuring refined static and dynamic semantics to derive spatial and temporal text prompts. Therefore, our spatiotemporal text augmentation allows for a more holistic representation of actions.

3 Method: Spatiotemporal Dynamic Duo

Pipeline Overview

Generally, as shown in Figure 2, our framework is capable of synergistic multi-modal spatiotemporal understanding which comprehends dynamics for vision-text context refinement. Our model is initialized from the CLIP, which consists of a vision encoder E_V and a text encoder E_T.

Specifically, given a video clip $V = \{\mathbf{x}_t\}_{t=1}^{T}, \mathbf{x}_t \in \mathbb{R}^{H\times W\times 3}$ of T frames, by dividing each frame into N non-overlapping patches $\{\mathbf{x}_{t,i}\}_{i=1}^{N}$ with the spatial size of $P\times P$, where $N = HW/P^2$. With prepending an additional [CLS] token $\mathbf{z}_{t,cls}^0$, to each frame, and adding positional embeddings to each patch, we can then obtain $\mathbf{z}_t^0 = [\mathbf{z}_{t,cls}^0, \mathbf{z}_{t,1}^0, \ldots, \mathbf{z}_{t,N}^0] \in \mathbb{R}^{(N+1)\times D}$ via the patch embedding layer.

The l-th ViT block with our tailored Space-time Cross Attention consists of a Multi-head Self-Attention (MHSA) followed by a MLP layer with short-cutting (+) and Layer Normalization (LN). It processes the tokens $\mathbf{z}_t^{l-1}$ from the previous block as follows:

$$\mathbf{z'}_t^l = \texttt{MHSA}(\texttt{LN}(\mathbf{z}_t^{l-1})) + \mathbf{z}_t^{l-1}, \tag{1}$$

$$\mathbf{z}_t^l = \texttt{MLP}(\texttt{LN}(\mathbf{z'}_t^l)) + \tilde{\mathbf{z}}_t^l, \tag{2}$$

where $\tilde{\mathbf{z}}_t^l$ is calculated by applying mixing-with-masking and padding before and after MHSA, respectively, which will be introduced in detail in the next subsection. Finally, the learned $\mathbf{z}_{t,cls}^L$ of the t-th frame from the last block is used as the frame-specific video representation.

Regarding the text flow, we perform *spatial* and *temporal* text augmentation for refined text prompts C^{st} consisting of C^s and C^t which articulate the static appearances and dynamic motions of actions, respectively. The j-th prompt-specific text representation $\mathbf{c}_{k,j}^{st}$ of the k-th class is obtained with the frozen E_T.

During training, we only optimize the parameters of E_V by calculating the fine-grained alignment scores between $\mathbf{z}_{t,cls}^L$ and $\mathbf{c}_{k,j}^{st}$. Meanwhile, the video representation learning are regulated by the video representations from the pre-trained CLIP to distill knowledge.

Space-time Cross Attention

As shown in Figure 3(a), our proposed Space-time Cross Attention is primarily achieved by implementing a four-step process within each ViT block. Before MHSA, (1) the Window Shift Masking (WSM) operation masks visual tokens and shifts along the temporal dimension to align tokens at interleaved spatial positions. (2) The subsequent Multi-scale Channel Mixing (MCM) processes window-shifted tokens within a local spatial window at multiple time scales to mix dynamic information actively. After MHSA, (3) the padding operation is employed ahead of (4) short-cutting, seamlessly assimilating dynamics into the encoding stream.

Window Shift Masking. Given the substantial temporal redundancy inherent in videos, we propose to perform a tailored masking strategy. Diverging from MAR (Qing et al. 2023) which masks the frame patches for reconstruction, our WSM is employed to obtain interleaved spatial tokens for information interaction within a local spatial window and multiple temporal scales. Formally, the spatial *window* (*e.g.*, $w_1 \times w_2 = 2\times 2$) is defined as the repeated unit used to partition all patches within each frame and generate masks. As shown in Figure 3(b), the masking positions shift along the temporal dimension, sequentially discarding a specified proportion (*e.g.*, $r = 0.5$) of tokens at different spatial locations. The masking flow can be formulated as:

$$M_t^l = \phi(M_{t-1}^l | M_{1:t-2}^l), l \in \{1, ..., L\}, \tag{3}$$

where M_t^l denotes the masking map for the t-th frame in the l-th ViT block and $\phi(\cdot|\cdot)$ is the periodic function to produce the masking map according to the former 1 to $(t-1)$ masking frames. Then, the masking map M_t^l is applied to visual

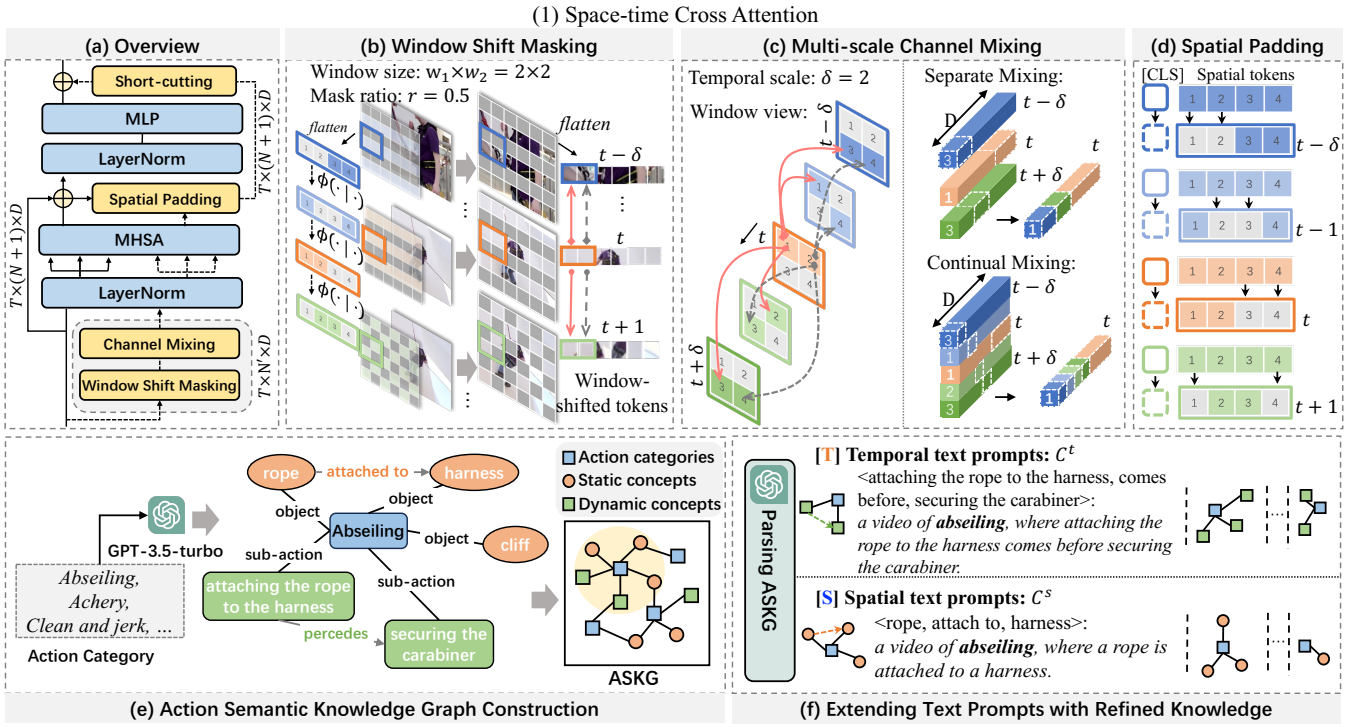

Figure 3: **Illustration of our method.** (1) We extend the spatial attention block to perform Space-time Cross Attention by applying Window Shift Masking to the input spatial tokens, and perform Multi-Scale Channel Mixing to capture temporal dynamics before MHSA. Then, we employ the spatial padding strategy to fill in the masked positions for seamless short-cutting, fusing additional dynamics effortlessly. (2) We conduct spatiotemporal text augmentation to obtain nuanced spatial and temporal text prompts by elaborating on static and dynamic concepts with their interrelations presented in ASKG.

tokens except for $\mathbf{z}_{t,cls}^{l-1}$ to yield window-shifted tokens:

$$\dot{\mathbf{z}}_t^{l-1} = \dot{\mathbf{z}}_{t,1:N'}^{l-1} \leftarrow M_t^l(\mathbf{z}_{t,1:N}^{l-1}), \tag{4}$$

where $\dot{\mathbf{z}}_t^{l-1} \in \mathbb{R}^{N' \times D}$ represent $N'(=r \times N)$ window-shifted spatial tokens of the t-th frame, and $N' = (\frac{L_1}{w_1} \times \frac{L_2}{w_2}) \times (r \times w_1 \times w_2), L_1 = \frac{H}{P}, L_2 = \frac{W}{P}$. The WSM maintains essential spatiotemporal correlations for effective mixing of channel information across windows and time scales with MCM.

Multi-scale Channel Mixing. After WSM, we follow the "shift trick" (Bulat et al. 2021; Wu et al. 2018) with zero-cost dimensionality reduction to perform MCM to enhance the fundamental inter-frame dynamic perception.

Formally, to expand the dynamic insight of MHSA, window-shifted tokens $\dot{\mathbf{z}}_t$ participate in fusing temporal information by indexing channels before and after the current time step. Here, we omit the denotation of layer index for simplicity. Figure 3(c) presents an example arranged in the window view to better illustrate the mechanism. When $\delta = 2$, the visual tokens $\dot{\mathbf{z}}_t$ marked as "1" within each window absorb channel information from tokens $\dot{\mathbf{z}}_{t-\delta}$ and $\dot{\mathbf{z}}_{t+\delta}$ at "3", which formulates a local spatial window for interaction. With the effect of temporal scales, the channel information flows actively among window-shifted tokens to capture spatiotemporal dynamics. Let $\dot{\mathbf{z}}_t(d_s:d_e) \in \mathbb{R}^{N' \times (d_e - d_s)}$ be the operator for indexing the channels from d_s to d_e with $\dot{\mathbf{z}}_t$. For separate mixing, $\dot{\mathbf{z}}_t$ only interacts with $\dot{\mathbf{z}}_{t-\delta}$ and $\dot{\mathbf{z}}_{t+\delta}$ as:

$$\dot{\mathbf{z}}_{t(\delta)} \triangleq [\dot{\mathbf{z}}_{t-\delta}(0:d_\delta); \dot{\mathbf{z}}_{t+\delta}(d_\delta:2d_\delta); \dot{\mathbf{z}}_t(2d_\delta:D)] \in \mathbb{R}^{N' \times D}, \tag{5}$$

where $d_\delta = \gamma \cdot D$ is a hyper-parameter for indexing channels. In contrast, for continual mixing (*e.g.*, $\delta = 2$), $\dot{\mathbf{z}}_t$ is mixed with all window-shifted tokens of 2δ range, *i.e.*, from $\dot{\mathbf{z}}_{t-2}$ to $\dot{\mathbf{z}}_{t+2}$:

$$\begin{aligned}\dot{\mathbf{z}}_{t(\delta)} \triangleq [&\dot{\mathbf{z}}_{t-2}(0:\frac{d_\delta}{\delta}); \dot{\mathbf{z}}_{t-1}(\frac{d_\delta}{\delta}:d_\delta); \dot{\mathbf{z}}_{t+1}(d_\delta:d_\delta+\frac{d_\delta}{\delta}); \\ &\dot{\mathbf{z}}_{t+2}(d_\delta+\frac{d_\delta}{\delta}:2d_\delta); \dot{\mathbf{z}}_t(2d_\delta:D)] \in \mathbb{R}^{N' \times D},\end{aligned} \tag{6}$$

where the channel dimension for mixing each token is divided by δ to achieve zero-cost dimensionality reduction. Then, we perform MHSA with the mixed tokens as follows:

$$\dot{\mathbf{z}}_{t(\delta)}^l = \mathrm{MHSA}(\mathrm{LN}(\dot{\mathbf{z}}_{t(\delta)}^{l-1})) + \dot{\mathbf{z}}_{t(\delta)}^{l-1} \in \mathbb{R}^{N' \times D}. \tag{7}$$

In a similar way, we further introduce multiple time scales $\{\delta_i\}_{i=1}^S$, with the purpose of unveiling the abundant dynamic insight and enriching the spatiotemporal information fusion. Then, we employ average pooling on all time scales to obtain $\bar{\mathbf{z}}_t^l \in \mathbb{R}^{N' \times D}$.

Spatial Padding and Short-cutting. Note that the quantity of visual tokens is reduced to N' due to the aforementioned WSM operation. Therefore, we employ a straightforward spatial padding strategy to obtain $\tilde{\mathbf{z}}_t^l \in \mathbb{R}^{(N+1)\times D}$, where tokens $\mathbf{z}_t'^l$ computed by Eq.(1) are selected carefully to fill in with $\bar{\mathbf{z}}_t^l$ according to M_t^l. As shown in Figure 3(d), it restores the original number and spatial positions of tokens for seamless short-cutting to fuse additional dynamics effortlessly.

Computational Complexity. Notably, the overall computational complexity of our Space-time Cross Attention is $O(TN^2 + ST(N')^2) = O(TN^2)$, which is equal to that of spatial-only attention. In contrast, the complexity of AIM and full space-time attention is $O(TN^2 + T^2N)$ and $O(T^2N^2)$, respectively.

Spatiotemporal Text Augmentation

In addition to expand the dynamic perception via visual representation learning, we further propose spatiotemporal text augmentation to incorporate refined action knowledge for visual-semantic context alignment by prompting GPT-3.5 (Achiam et al. 2023). For clarity, we present a detailed text augmentation process of `abseiling` in Figure 3(2).

Action Semantic Knowledge Graph. Essentially, our primary objective is to construct an ASKG that conceptually disentangles action categories into static appearances and dynamic motions and their interrelations, as shown in Figure 3(e). Specifically, the ASKG abstracts action categories into a graph structure by representing the original actions, related static and dynamic concepts as graph nodes and their interrelations as edges. Therefore, it enables a more structured and interpretable representation of actions. We use the following prompts: "*Return the object entity list containing Top* $K(5 \leq K \leq 10)$ *most relevant objects / sub-actions involved in action:* {`abseiling`}" for semantic concepts, and "*Find the proper predicate names that concisely describe the relationship between each object / sub-action pair chosen from the entity list*" for semantic relations. [1]

Extending Text Prompts with Refined Knowledge. To incorporate the refined spatiotemporal semantic knowledge into our multi-modal pipeline, we propose to extend text prompts by parsing the ASKG to generate text prompts, as presented in Figure 3(f). Similar to the ASKG construction, we use the following prompts to extend the text prompts: "*Try to complete the whole sentence according to each relation triples: This is an example of* {`abseiling`} ,..."[1]. In this way, the extended text prompts are then obtained by concatenating the hard prompt templates with the output. The generated clauses reflect different spatiotemporal text hints, maintaining semantic consistency in describing the action. The spatial text prompts C^s describe the static appearances obtained by prompting *object relation triples*, while the temporal text prompts C^t capture the dynamic motions by prompting *action relation triples*.

[1]For a detailed demonstration of the prompts we used and response examples, please refer to Appendix.

Training Objectives

To overcome visual-semantic discrepancies at instance and frame level, the primary training objective is to meticulously align multi-modal spatiotemporal dynamics based on fine-grained alignment scores.

Specifically, let $\mathbf{z}_{n,t}$ be the final visual representation of the n-th video at the t-th frame, the alignment score for the k-th class is calculated across N^{st} spatiotemporal text representations $\{\mathbf{c}_{k,j}^{st}\}_{j=1}^{N^{st}}$ to achieve frame-to-prompt and symmetric prompt-to-frame fine-grained alignment:

$$S_{n,k}^{v2t} = \frac{1}{T}\sum_{t=1}^{T} \max_{1\leq j\leq N^{st}} \mathbf{z}_{n,t}^{\top}\mathbf{c}_{k,j}^{st}; \quad S_{n,k}^{t2v} = \frac{1}{N^{st}}\sum_{j=1}^{N^{st}} \max_{1\leq t\leq T} \mathbf{z}_{n,t}^{\top}\mathbf{c}_{k,j}^{st}. \tag{8}$$

The overall alignment score $S_{n,k}$ is calculated by averaging the scores above, which is also used for our zero-shot inference. The cross-entropy loss is implemented following (Wu, Sun, and Ouyang 2023; Jia et al. 2024) as:

$$L_{CE} = -\frac{1}{B}\sum_{n=1}^{B}\sum_{k=1}^{K} y_{n,k}\log\left(\frac{\exp(S_{n,k})}{\sum_{i=1}^{K}\exp(S_{n,i})}\right), \tag{9}$$

where B denotes the number of minibatch training videos of K seen classes. If the n-th video belongs to the k-th class, $y_{n,k}$ equals 1; otherwise, $y_{n,k}$ equals 0. Finally, our framework is optimized by L_{CE} together with feature distillation loss proposed by (Huang et al. 2024).

4 Experiment

Implementation Details

We use Kinetics-400 (K400) (Kay et al. 2017) dataset as the training set and evaluate our method under ZSAR settings on three popular benchmarks: UCF101 (UCF) (Soomro, Zamir, and Shah 2012), HMDB51 (HMBD) (Kuehne et al. 2011), and Kinetics-600 (K600) (Carreira et al. 2018), following the evaluation protocols: EP 1, EP 2 and EP 3 in (Brattoli et al. 2020; Ni et al. 2022).

We use the K400 pretrained models to directly perform cross-dataset ZSAR evaluation. Generally, we use two official CLIP backbones: ViT-B/16 and ViT-L/14. In our proposed Space-time Cross Attention, we define the spatial size of a repeated window as $w_1 \times w_2$ $(w_1 = w_2 = 2)$ with a mask ratio of 50%, and specify the temporal scales to $[\pm 1, \pm 2]$. Following (Weng et al. 2023), the models learned in different epochs are averaged to improve generalizability. Each video clip is uniformly sampled with 8 frames during training. For ZSAR evaluation, we use 3 temporal and 1 spatial views per video, and linearly aggregate the prediction results. More implementations are provided in Appendix.

Main Results

We compare our method with state-of-the-art ZSAR methods on three benchmarks following commonly-used evaluation protocols. In Table 1, we categorize the previous methods based on the visual backbones and semantic augmentation methods, presenting a comprehensive comparative analysis on UCF and HMDB under EP 1 and EP 2. Note that

Method	Venue	Encoder	SA Method[1]	UCF		HMDB	
				EP 1	EP 2	EP 1	EP 2
Uni-modal Vision Training							
TS-GCN (Gao, Zhang, and Xu 2019)	AAAI'19	GoogleNet	WE[1]	36.14 ± 4.8	-	23.2	-
CWEGAN (Mandal et al. 2019)	CVPR'19	I3D	WE	26.9 ± 2.8	-	30.2	-
ER-ZSAR (Chen and Huang 2021)	ICCV'21	TSM	ED[1]	51.8 ± 2.9	-	35.3 ± 4.6	-
DASZL (Kim et al. 2021)	AAAI'21	TSM	HA[1]	48.9 ± 5.8	-	-	-
Adapting Pretrained VLMs							
BIKE (Wu et al. 2023)	CVPR'23	ViT-L/14	AT[1]	86.6 ± 3.4	80.8	61.4 ± 3.6	52.8
Text4Vis* (Wu, Sun, and Ouyang 2023)	AAAI'23	ViT-L/14	CT	85.8 ± 3.3	79.6	58.1 ± 5.7	49.8
			ST Aug.	89.5 ± 2.9	84.2	63.7 ± 3.2	52.9
Open-VCLIP (Weng et al. 2023)	ICML'23	ViT-B/16	CT	89.9 ± 1.7	83.5	64.5 ± 4.5	53.2
		ViT-L/14		93.1 ± 1.9	87.9	68.5 ± 4.0	58.3
ViLT-CLIP (Wang et al. 2024)	AAAI'24	ViT-B/16	PE[1]	-	73.9	-	45.3
Ours		ViT-B/16	ST Aug.	90.3 ± 1.7	85.3	64.7 ± 3.8	54.7
		ViT-L/14		**93.4 ± 2.2**	**88.6**	**68.7 ± 4.5**	**58.7**

[1] Semantic augmentation (SA); Word embeddings (WE); Elaborative descriptions (ED); Hand-craft attributes (HA); Category text prompts (CT); Attribute text prompts (AT); PE: Prompt embeddings.

Table 1: Comparisons of ZSAR accuracies (%) on UCF and HMDB with EP 1 and EP 2. "*" denotes our re-evaluation.

there is a significant performance gap between the previous uni-modal vision training methods and the methods adapting pretrained VLMs to ZSAR. Among these, our method exhibits the best performance on UCF and HMDB when using either the ViT-B/16 or the ViT-L/14 encoder. In comparison with other VLM-based methods, our method achieves better performance than the second-best competitor, *i.e.*, Open-VCLIP, on both UCF and HMDB, with a margin up to 1.8% and 1.5% using the ViT-B/16 encoder. Remarkably, our framework with ViT-B/16 backbone even outperforms BIKE (ViT-L/14) by 7.8% and 5.9% on UCF and HMDB, respectively. When compared with Text4Vis, the performance gap is up to 9.0% on UCF's EP 2. Furthermore, our semantic augmentation method has demonstrated strong adaptability. As a representative, by implementing our spatiotemporal text augmentation to Text4Vis, directly replacing its category prompts without any retraining, Text4Vis (w/ ST Aug.) has experienced promising improvements, with gains of +3.7%, +4.6%, +5.6% and +3.1%, respectively.

Table 2 shows the comparison results under EP 3 and K600 benchmark with different CLIP-based methods using the ViT-B/16 backbone. Despite most of the methods use category text prompts (CT) for semantic embeddings, recent methods based on text augmentation, including MAXI and Open-VCLIP++ with visual captions, AP-CLIP with action-conditional prompts, and FROSTER with action descriptions, yield consistent performance improvements. Compared to OST with STD, our ST Aug. not only describes spatial appearances and temporal evolutions of actions but also possess the capability to discover spatial and temporal interrelations with ASKG for nuanced text prompts, which surpasses the best previous knowledge-based OST by 7.3% on UCF and AP-CLIP by 1.7% on K600. We also compare our proposed text augmentation with CT and STD based on our vision backbone for fair comparison in Table 3. It can be

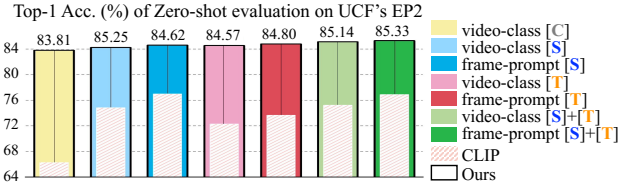

Figure 4: Effect of different combinations of text augmentation and alignment mechanisms for our method and CLIP.

observed that our ST Aug. consistently surpasses STD and CT on three datasets. Besides, integrating with our backbone rather than OST, STD leads to a notable increase in performance with gains of 7.0% and 1.1% in UCF and K600, further demonstrating the superiority of our method's effective spatiotemporal dynamics modeling.

Ablation Study

Effect of Space-time Cross Attention. Table 4 shows the effects of the Window Shift Masking (WSM) and Multi-scale Channel Mixing (MCM) with different adaptation methods. For a fair comparison, we start with the frozen CLIP as the baseline. In addition to our fully-tuned adaptation, we also evaluate prompt-based adaptation with the frozen backbone, where only a MLP is tuned to transform tokens after Space-time Cross Attention into prompts prepended with visual tokens within each block. However, fully-tuned adaptation consistently outperforms prompt-based adaptation, suggesting the importance of sufficient model capability for adapting to the video domain and the need for an effective space-time fusing strategy. Additionally, by simply implementing MCM with fully-tuned adaptation, the accuracy on UCF is improved by 10.3% and 10.4% with continual mixing by capturing dynamics

Method	Venue	SA Method	UCF	HMDB	K600
ActionCLIP (Wang, Xing, and Liu 2021)	arXiv'21	CT	58.3 ± 3.4	40.8 ± 5.4	66.7 ± 1.1
X-CLIP (Ni et al. 2022)	ECCV'22	CT	72.0 ± 2.3	44.6 ± 5.2	65.2 ± 0.4
MAXI (Lin et al. 2023)	ICCV'23	VC[1]	78.2 ± 0.7	52.3 ± 0.6	71.5 ± 0.8
VicTR (Kahatapitiya et al. 2024)	CVPR'24	CT + AT	72.4 ± 0.3	51.0 ± 1.3	-
OST (Chen et al. 2024)	CVPR'24	STD[1]	77.9 ± 1.3	54.9 ± 1.1	73.9 ± 0.8
AP-CLIP (Jia et al. 2024)	ACM MM'24	AP[1]	82.4 ± 0.5	55.4 ± 0.8	73.4 ± 1.0
Open-VCLIP++ (Wu et al. 2024)	TPAMI'24	VC	83.9 ± 0.6	55.6 ± 1.4	73.4 ± 0.8
FROSTER (Huang et al. 2024)	ICLR'24	AD[1]	84.8 ± 1.1	54.8 ± 1.3	74.8 ± 0.9
Ours		ST Aug.	**85.2 ± 1.2**	**55.9 ± 0.2**	**75.1 ± 0.7**

[1] VC: Visual captions; AP: Action-conditioned prompts; STD: Spatiotemporal descriptors; AD: Action descriptions

Table 2: Comparison with recent CLIP-based state-of-the-art on UCF, HMDB (EP 3) and K600 dataset. All methods are based on CLIP ViT-B/16. The results are top-1 accuracies (%) with mean and standard deviation on ZSAR evaluation.

SE Method	**UCF**	**HMDB**	**K600**
CT (Wang, Xing, and Liu 2021)	83.8	54.0	73.5
STD (Chen et al. 2024)	84.9	55.1	75.0
ST Aug. (Ours)	**85.2**	**55.9**	**75.1**

Table 3: Performance comparison (Top-1 Acc. (%)) with different semantic augmentation methods.

Adaptation	**WSM**	**MCM**	**UCF**
Frozen	✗	✗	74.2
Prompt-based	✗	Separate	79.8 (+5.6)
	✗	Continual	80.1 (+5.9)
	✓	Separate	82.3 (+8.1)
	✓	Continual	82.7 (+8.5)
Fully-tuned	✗	Separate	84.5 (+10.3)
	✗	Continual	84.6 (+10.4)
	✓	Separate	85.0 (+10.8)
	✓	Continual	85.2 (+11.0)

Table 4: Effect of different operations in Space-time Cross Attention and adaptation methods.

densely, achieving a slightly better performance than separate mixing. The most significant improvement is observed when WSM and MCM are used in conjunction, culminating in 85.0%(+10.8%) and 85.2%(+11.0%) on UCF.

Effect of spatiotemporal text augmentation and alignment mechanisms. Figure 4 illustrates the effect of different implementations for text augmentation and alignment mechanisms. Notably, the proposed spatiotemporal text augmentation [**S**]+[**T**] and frame-prompt alignment significantly enhance performance of the vanilla CLIP on UCF without additional finetuning on the videos, highlighting its superior inference-time adaptability. However, for CLIP, the improvement attributable to spatiotemporal text prompts [**S**]+[**T**] is negligible compared with spatial text prompts [**S**], indicating potential object bias of the CLIP. In contrast, our proposed method achieves better results with spatiotemporal text prompts [**S**]+[**T**], showing a +1.08% improvement over spatial text prompts [**S**], which validates its significant multi-modal spatiotemporal understanding capability.

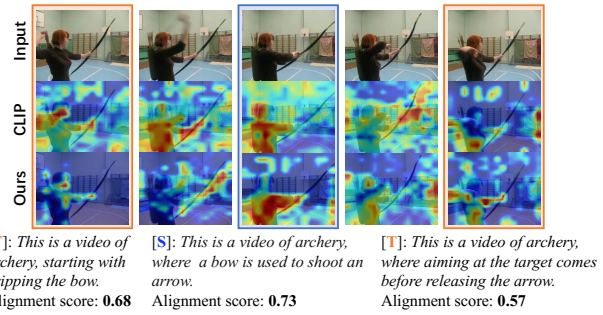

Figure 5: Visualizations of the attention maps and frame-prompt alignment scores of *Archery*. Our framework consistently prioritizes local body parts and objects participated in the dynamic movements.

Visualization

Figure 5 presents the attention map visualizations of video frames and text prompts with the highest alignment scores obtained by our method. The attention maps of CLIP primarily attend to the actor's body and surroundings unrelated to the actions being performed. Conversely, our framework consistently prioritizes the key body parts (*e.g.*, arms and hands) and objects (*e.g.*, bow and target) involved in the dynamic movements required for the action *archery*, which is crucial for multi-modal spatiotemporal understanding.

5 Conclusion

In this work, we present a novel multi-modal spatiotemporal expert for ZSAR. The Space-time Cross Attention integrates a four-step operation, capturing cross-frame dynamics without introducing additional parameters or increasing computational complexity. The spatiotemporal text augmentation elaborates on static and dynamic concepts and their interrelations for action categories. Extensive evaluations consistently validate the superior effectiveness of our framework.

Acknowledgments

This work is supported by National Natural Science Foundation of China (No. 62376217, 62301434), Young Elite Scientists Sponsorship Program by CAST (2023QNRC001), Key R&D Project of Shaanxi Province (No. 2023-YBGY-240), and Young Talent Fund of Association for Science and Technology in Shaanxi, China (No. 20220117).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ahmad, S.; Chanda, S.; and Rawat, Y. S. 2023. EZ-CLIP: Efficient Zeroshot Video Action Recognition. *arXiv preprint arXiv:2312.08010*.

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6836–6846.

Bao, H.; Wang, W.; Dong, L.; Liu, Q.; Mohammed, O. K.; Aggarwal, K.; Som, S.; Piao, S.; and Wei, F. 2022. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35: 32897–32912.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is space-time attention all you need for video understanding? In *ICML*, volume 2, 4.

Brattoli, B.; Tighe, J.; Zhdanov, F.; Perona, P.; and Chalupka, K. 2020. Rethinking zero-shot video classification: End-to-end training for realistic applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4613–4623.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Bulat, A.; Perez Rua, J. M.; Sudhakaran, S.; Martinez, B.; and Tzimiropoulos, G. 2021. Space-time mixing attention for video transformer. *Advances in neural information processing systems*, 34: 19594–19607.

Cao, C.; Zhang, H.; Lu, Y.; Wang, P.; and Zhang, Y. 2024a. Scene-dependent prediction in latent space for video anomaly detection and anticipation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Cao, C.; Zhang, Y.; Yu, Y.; Lv, Q.; Min, L.; and Zhang, Y. 2024b. Task-Adapter: Task-specific Adaptation of Image Models for Few-shot Action Recognition. In *ACM Multimedia 2024*.

Carreira, J.; Noland, E.; Banki-Horvath, A.; Hillier, C.; and Zisserman, A. 2018. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*.

Chen, S.; and Huang, D. 2021. Elaborative rehearsal for zero-shot action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13638–13647.

Chen, T.; Yu, H.; Yang, Z.; Li, Z.; Sun, W.; and Chen, C. 2024. OST: Refining Text Knowledge with Optimal Spatio-Temporal Descriptor for General Video Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18888–18898.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Floridi, L.; and Chiriatti, M. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30: 681–694.

Gao, J.; Zhang, T.; and Xu, C. 2019. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 8303–8311.

Gao, Y.; Shi, X.; Zhu, Y.; Wang, H.; Tang, Z.; Zhou, X.; Li, M.; and Metaxas, D. N. 2022. Visual prompt tuning for test-time domain adaptation. *arXiv preprint arXiv:2210.04831*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, X.; Zhou, H.; Yao, K.; and Han, K. 2024. FROSTER: Frozen CLIP is a Strong Teacher for Open-Vocabulary Action Recognition. In *International Conference on Learning Representations*.

Jain, M.; Van Gemert, J. C.; Mensink, T.; and Snoek, C. G. 2015. Objects2action: Classifying and localizing actions without any video example. In *Proceedings of the IEEE international conference on computer vision*, 4588–4596.

Jia, C.; Luo, M.; Chang, X.; Dang, Z.; Han, M.; Wang, M.; Dai, G.; Dang, S.; and Wang, J. 2024. Generating action-conditioned prompts for open-vocabulary video action recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 4640–4649.

Jie, S.; and Deng, Z.-H. 2022. Convolutional bypasses are better vision transformer adapters. *arXiv preprint arXiv:2207.07039*.

Kahatapitiya, K.; Arnab, A.; Nagrani, A.; and Ryoo, M. S. 2024. Victr: Video-conditioned text representations for activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18547–18558.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Kim, T. S.; Jones, J.; Peven, M.; Xiao, Z.; Bai, J.; Zhang, Y.; Qiu, W.; Yuille, A.; and Hager, G. D. 2021. Daszl: Dynamic action signatures for zero-shot learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 1817–1826.

Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; and Serre, T. 2011. HMDB: a large video database for human motion recognition. In *2011 International conference on computer vision*, 2556–2563. IEEE.

Lee, D.; Lee, J.; and Choi, J. 2024. CAST: Cross-Attention in Space and Time for Video Action Recognition. *Advances in Neural Information Processing Systems*, 36.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, L.; Xiao, J.; Chen, G.; Shao, J.; Zhuang, Y.; and Chen, L. 2024. Zero-shot visual relation detection via composite visual cues from large language models. *Advances in Neural Information Processing Systems*, 36.

Lin, J.; Gan, C.; and Han, S. 2019. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7083–7093.

Lin, W.; Karlinsky, L.; Shvetsova, N.; Possegger, H.; Kozinski, M.; Panda, R.; Feris, R.; Kuehne, H.; and Bischof, H. 2023. Match, expand and improve: Unsupervised finetuning for zero-shot action recognition with language knowledge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2851–2862.

Lin, Z.; Geng, S.; Zhang, R.; Gao, P.; De Melo, G.; Wang, X.; Dai, J.; Qiao, Y.; and Li, H. 2022. Frozen clip models are efficient video learners. In *European Conference on Computer Vision*, 388–404. Springer.

Mandal, D.; Narayan, S.; Dwivedi, S. K.; Gupta, V.; Ahmed, S.; Khan, F. S.; and Shao, L. 2019. Out-of-distribution detection for generalized zero-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9985–9993.

Mao, Y.; Deng, J.; Zhou, W.; Li, L.; Fang, Y.; and Li, H. 2024. CLIP4HOI: Towards Adapting CLIP for Practical Zero-Shot HOI Detection. *Advances in Neural Information Processing Systems*, 36.

Mishra, A.; Pandey, A.; and Murthy, H. A. 2020. Zero-shot learning for action recognition using synthesized features. *Neurocomputing*, 390: 117–130.

Ni, B.; Peng, H.; Chen, M.; Zhang, S.; Meng, G.; Fu, J.; Xiang, S.; and Ling, H. 2022. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, 1–18. Springer.

Pan, J.; Lin, Z.; Zhu, X.; Shao, J.; and Li, H. 2022. St-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35: 26462–26477.

Qian, Q.; Xu, Y.; and Hu, J. 2024. Intra-Modal Proxy Learning for Zero-Shot Visual Categorization with CLIP. *Advances in Neural Information Processing Systems*, 36.

Qing, Z.; Zhang, S.; Huang, Z.; Wang, X.; Wang, Y.; Lv, Y.; Gao, C.; and Sang, N. 2023. Mar: Masked autoencoders for efficient action recognition. *IEEE Transactions on Multimedia*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rasheed, H.; Khattak, M. U.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6545–6554.

Sanghi, A.; Chu, H.; Lambourne, J. G.; Wang, Y.; Cheng, C.-Y.; Fumero, M.; and Malekshan, K. R. 2022. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18603–18613.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Tu, S.; Dai, Q.; Wu, Z.; Cheng, Z.-Q.; Hu, H.; and Jiang, Y.-G. 2023. Implicit temporal modeling with learnable alignment for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19936–19947.

Wang, H.; Liu, F.; Jiao, L.; Wang, J.; Hao, Z.; Li, S.; Li, L.; Chen, P.; and Liu, X. 2024. ViLT-CLIP: Video and Language Tuning CLIP with Multimodal Prompt Learning and Scenario-Guided Optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5390–5400.

Wang, M.; Xing, J.; and Liu, Y. 2021. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*.

Wang, Q.; Du, J.; Yan, K.; and Ding, S. 2023. Seeing in Flowing: Adapting CLIP for Action Recognition with Motion Prompts Learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5339–5347.

Wang, Y.; Cao, C.; and Zhang, Y. 2022. Beyond vision: a semantic reasoning enhanced model for gesture recognition with improved spatiotemporal capacity. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 420–434. Springer.

Wang, Y.; Cao, C.; and Zhang, Y. 2023. Visual-semantic network: a visual and semantic enhanced model for gesture recognition. *Visual Intelligence*, 1(1): 25.

Wasim, S. T.; Naseer, M.; Khan, S.; Khan, F. S.; and Shah, M. 2023. Vita-CLIP: Video and text adaptive CLIP via Multimodal Prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23034–23044.

Weng, Z.; Yang, X.; Li, A.; Wu, Z.; and Jiang, Y.-G. 2023. Open-vclip: Transforming clip to an open-vocabulary video model via interpolated weight optimization. In *International Conference on Machine Learning*, 36978–36989. PMLR.

Wu, B.; Wan, A.; Yue, X.; Jin, P.; Zhao, S.; Golmant, N.; Gholaminejad, A.; Gonzalez, J.; and Keutzer, K. 2018. Shift: A zero flop, zero parameter alternative to spatial convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9127–9135.

Wu, W.; Sun, Z.; and Ouyang, W. 2023. Revisiting classifier: Transferring vision-language models for video recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 2847–2855.

Wu, W.; Wang, X.; Luo, H.; Wang, J.; Yang, Y.; and Ouyang, W. 2023. Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6620–6630.

Wu, Z.; Weng, Z.; Peng, W.; Yang, X.; Li, A.; Davis, L. S.; and Jiang, Y.-G. 2024. Building an open-vocabulary video CLIP model with better architectures, optimization and data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yang, T.; Zhu, Y.; Xie, Y.; Zhang, A.; Chen, C.; and Li, M. 2023. Aim: Adapting image models for efficient video action recognition. *arXiv preprint arXiv:2302.03024*.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Zhu, Y.; Zhuo, J.; Ma, B.; Geng, J.; Wei, X.; Wei, X.; and Wang, S. 2023. Orthogonal Temporal Interpolation for Zero-Shot Video Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7491–7501.