KRONECKER MASK AND INTERPRETIVE PROMPTS ARE LANGUAGE-ACTION VIDEO LEARNERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive language-image pretraining (CLIP) has significantly advanced image-based vision learning. A pressing topic subsequently arises: how can we effectively adapt CLIP to the video domain? Recent studies have focused on adjusting either the textual or visual branch of CLIP for action recognition. However, we argue that adaptations of both branches are crucial. In this paper, we propose **CLAVER**: a **C**ontrastive **L**anguage-**A**ction **V**ideo Learn**er**, designed to shift CLIP's focus from the alignment of static visual objects and concrete nouns to the alignment of dynamic action behaviors and abstract verbs. Specifically, we introduce a novel Kronecker mask attention for temporal modeling. Our tailored Kronecker mask offers three benefits 1) it expands the temporal receptive field for each token, 2) it serves as an effective spatiotemporal heterogeneity inductive bias, mitigating the issue of spatiotemporal homogenization, and 3) it can be seamlessly plugged into transformer-based models. Regarding the textual branch, we leverage large language models to generate diverse, sentence-level and semantically rich interpretive prompts of actions, which shift the model's focus towards the verb comprehension. Extensive experiments on various benchmarks and learning scenarios demonstrate the superiority and generality of our approach. The code will be available soon.

1 INTRODUCTION

Video action recognition has long been a representative topic in video understanding. Over the past decade, there has been a continuous pursuit of learning spatiotemporal representations, giving rise to diverse architectures, such as traditional two-stream networks Simonyan & Zisserman (2014); Wang et al. (2016); Zhou et al. (2018); Karpathy et al. (2014), 3D convolutional neural networks Carreira & Zisserman (2017); Feichtenhofer (2020); Feichtenhofer et al. (2019); Hara et al. (2017); Qiu et al. (2017); Tran et al. (2015; 2018); Wang et al. (2018); Xie et al. (2018), and Video Vision Transformers Arnab et al. (2021); Bertasius et al. (2021); Fan et al. (2021); Liu et al. (2022); Patrick et al. (2021); Zhao et al. (2022); Li et al. (2022a); Yan et al. (2022). Recently, there has been increasing interest in leveraging visual-language models (VLMs) like CLIP Radford et al. (2021), Florence Yuan et al. (2021), and ALIGN Jia et al. (2021) for various video tasks, owing to the superior generalization abilities of these models. Several studies Wang et al. (2021); Lin et al. (2022); Ni et al. (2022); Ju et al. (2022); Rasheed et al. (2023); Tu et al. (2023); Chen et al. (2023) have devoted to adapt the CLIP for video action recognition, but they often focus on adjusting a single branch. According to predecessor studies, transferring CLIP from the image domain to the video domain involves two key considerations: 1) how to perform effective temporal modeling. 2) how to design suitable text descriptions for verb understanding that align with rich text semantics in the VLM's pre-training dataset. We argue that addressing both issues simultaneously is crucial.

In addressing the issue 1), several studies Wang et al. (2021); Ju et al. (2022); Chen et al. (2023); Rasheed et al. (2023) implement straightforward and simple strategies such as mean pooling or 1D-temporal convolution across the temporal dimension, or employing temporal attention among class tokens. X-CLIP Ni et al. (2022) and CLIP-ViP Xue et al. (2022) introduce extra tokens for cross-frame communication. Alternatively, some studies Lin et al. (2022); Tu et al. (2023) engineer tailored modules. In our work, we aim to elucidate the distinctions and intrinsic correlations between space and time, as well as design more general spatiotemporal modeling approaches.

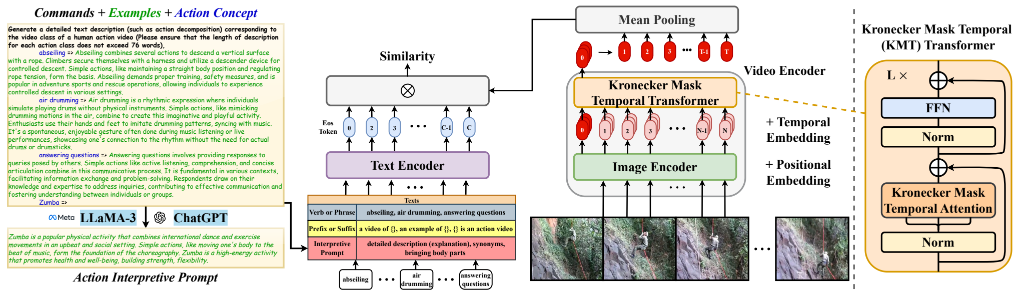

Figure 1: An overview of **CLAVER**. (**Right**) Image encoder and KMT transformer are assembled as a video encoder. (**Left**) How to get the interpretive prompts for actions.

Regarding the issue 2), some studies Hendricks & Nematzadeh (2021); Thrush et al. (2022) indicate that VLMs tend to focus on the correspondence between visual objects and nouns rather than action behaviors and verbs. We consider the essential gap is that visual objects are static and presented in lower dimensions, and nouns are concrete and easily understandable. However, action behaviors are dynamic that presented in higher dimensions, and verbs are intricate and abstract. Several existing methods Ni et al. (2022); Lin et al. (2022); Rasheed et al. (2023); Tu et al. (2023) use verbs or phrases as direct text descriptions. ActionCLIP Wang et al. (2021) integrates prompt templates to expand verbs or phrases into sentences. Ju et al. Ju et al. (2022) propose trainable continuous prompts to construct virtual prompt templates. However, these methods do not address aforementioned issues in essence. Alternatively VFC Momeni et al. (2023) and MAXI Lin et al. (2023) consider leveraging large language models (LLMs) to provide positive and negative text samples for contrastive learning or multiple instance learning, while ASU Chen et al. (2023) presents the concept of semantic units to supplement the semantic information of action labels.

To address these issues, we propose a **C**ontrastive **L**anguage-**A**ction **V**ideo Learn**er** (**CLAVER**, Fig. 1) to efficiently adapt the CLIP for video action recognition. Specifically, for the issue 1), we first obtain the frame-level visual representation from the image encoder, then apply tailored Kronecker mask for temporal modeling with a wider temporal receptive field to establish long-range and wide-range dependencies among frames, while mitigating spatiotemporal homogenization. Additionally, we reveal the intrinsic correlations between space and time from the perspective of Kronecker mask attention. Regarding the issue 2), we leverage LLMs to effectively generate diverse, sentence-level, and semantically rich interpretations of actions, augmenting text descriptions during training and testing. This approach allows the text descriptions to be presented in a more flexible, sentence-level form during inference. In summary, our main contributions are four-fold:

- We propose the **C**ontrastive **L**anguage-**A**ction **V**ideo Learn**er** (**CLAVER**) to adapt both the visual and textual branches, efficiently shifting the alignment in CLIP from visual objects and nouns to action behaviors and verbs.
- We propose the Kronecker mask temporal attention and Kronecker mask causal temporal attention for temporal modeling, aiming to capture the long-range and wide-range dependencies among frames with spatiotemporal heterogeneity.
- We introduce interpretive prompts of actions to facilitate the alignment of action behaviors and verbs, thereby improving zero-shot and few-shot generalization capabilities.
- Extensive qualitative and quantitative experiments demonstrate the effectiveness of CLAVER. Our method achieves superior or competitive performance on Kinetics-400 and Kinetics-600 under fully-supervised scenario, and on HMDB-51 and UCF-101 under zero-shot, few-shot scenarios.

2 Related Work

Video Recognition. Among early video recognition methods, 3D convolution is widely employed Qiu et al. (2017); Tran et al. (2015; 2018); Xie et al. (2018); Feichtenhofer et al. (2019); Feichtenhofer (2020). Some studies Qiu et al. (2017); Tran et al. (2018); Xie et al. (2018) propose to factorize convolutional operations across spatial and temporal dimensions, while others design the specific temporal modules to embed them into 2D CNNs Li et al. (2020b); Lin et al. (2019); Liu et al. (2021). Over the past few years, there has been an influx of transformer-based video works Arnab et al. (2021); Neimark et al. (2021); Bertasius et al. (2021); Fan et al. (2021); Liu et al. (2022); Yan et al. (2022); Li et al. (2022a), demonstrating promising performance. For example, some methods Arnab et al. (2021); Neimark et al. (2021); Girdhar & Grauman (2021) adopt a factorized encoder structure for spatial-temporal fusion. Alternatively, another family employ a factorized attention structure. Such as TimeSformer Bertasius et al. (2021), ViViT Arnab et al. (2021), and ATA Zhao et al. (2022) which proposes alignment-guided tepmoral attention. While Video Swin Transformer Liu et al. (2022) employs 3D window attention. The representative attention calculation forms in these works can be summarized into joint attention and factorized (divided) spatiotemporal attention. In this paper, to make minimal modifications to the original structure, we adopt a factorized encoder structure for video stream.

Visual-language Representation Learning. Visual-textual multi-modality is a hot topic in recent years. Several studies based on masked image modeling (MIM) have achieved commendable performance Li et al. (2020a); Lu et al. (2019); Su et al. (2019); Tan & Bansal (2019). There are also efforts focused on video-language representation learning Miech et al. (2019); Sun et al. (2019a;b); Zhu & Yang (2020); Xu et al. (2021). Concurrently, contrastive language-image pretraining Radford et al. (2021); Jia et al. (2021); Yuan et al. (2021) achieved remarkable progress, particularly in demonstrating impressive zero-shot generalization capacities. CLIP Radford et al. (2021) is one of the most representative works, with numerous follow-up studies have explored to adapt it for downstream tasks. For example, object detection, semantic segmentation, video retrieval and captioning, etc.Gu et al. (2021); Vinker et al. (2022); Li et al. (2022b); Luo et al. (2022); Xu et al. (2022). Additionally, there are also many applications in the video action recognition Wang et al. (2021); Ni et al. (2022); Lin et al. (2022); Ju et al. (2022); Pan et al. (2022); Rasheed et al. (2023); Chen et al. (2023); Tu et al. (2023); Lin et al. (2023); Momeni et al. (2023); Yang et al. (2023). For instance, ViFiCLIP Rasheed et al. (2023) aim to minimize modifications to original models and facilitate efficient transfer, while Chen Ju et al. Ju et al. (2022) suggest optimizing a few prompt vectors for adapting CLIP to various video understanding tasks. X-CLIP Ni et al. (2022) proposes an efficient cross-frame attention module. ILA Tu et al. (2023) designs implicit mask-based alignment to align features of two adjacent frames and EVL Lin et al. (2022) proposes a image encoder and video decoder structure. Regrading to the language branch, most previous works directly use verbs or phrases that lack rich semantics which overlook the importance of semantics. With the advancement of large language models like the GPT-3 Brown et al. (2020), PaLM Chowdhery et al. (2023), LLaMAs Touvron et al. (2023a;b); Abhimanyu Dubey et al. (2024) and ChatGPT OpenAI (2024). LLMs can replace manual labor Chen & Huang (2021); Qian et al. (2022) and automatically generate texts that meet human expectations to benefit visual-textual learning. For example, LaCLIP Fan et al. (2024) employs LLMs to rewrite text descriptions associated with each image for text augmentation. VFC Momeni et al. (2023) and MAXI Lin et al. (2023) leverage LLMs to generate positive and negative texts with diversity for language-video learning.

3 METHODOLOGY

In Sec. 3.1, we overview our proposed contrastive language-action video learner architecture. Then, we elaborate on the detail of the Kronecker mask attention in Sec. 3.2. Finally, we present the technique details of action interpretive prompt in Sec. 3.3.

3.1 OVERVIEW

Our contrastive language-action video learner architecture is illustrated in Fig. 1. We utilize a video encoder to obtain video representations, comprising two transformers-based components: an image encoder (ViT) from CLIP, a Kronecker mask temporal transformer. The text encoder aims to align text representations with the video representations. Concretely, given a video clip $V = [v_0, \cdots v_t, \cdots v_{T-1}] \in \mathbb{R}^{T\times H\times W\times 3}, v_t \in \mathbb{R}^{H\times W\times 3}$ and corresponding text descriptions $C = [c_0, \cdots c_m, \cdots c_{M-1}] \in \mathbb{R}^{M\times N}, c_m \in \mathbb{R}^N$, where T, H, W are the number of frames, height, width, respectively, M is the number of diverse text descriptions (share the same central concept) for an action category, N is the max sequence length. We feed texts C into the text encoder $f_{\theta_C}(\cdot)$ to obtain text representations $\mathbf{C} = [\mathbf{c}_0, \cdots \mathbf{c}_{M-1}]$. For the video stream, firstly, we input the video

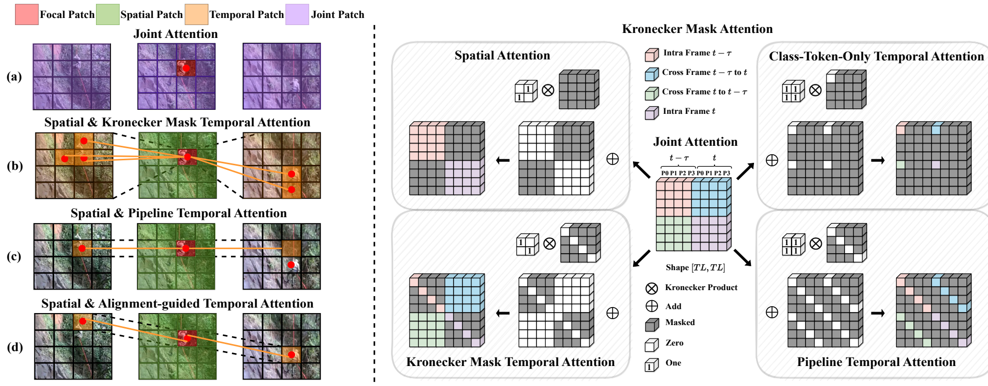

Figure 2: (**Left**) Red indicates the currently focal patch, green patches are visible in spatial attention, orange patches are visible in temporal attention, purple patches are visible in joint attention. (**Right**) Kronecker mask attention: Several attentions can be seen as employing tailored Kronecker masks for joint attention.

clip V to the image encoder $f_{\theta_I}(\cdot)$ to obtain frame-level representations $\mathbf{I}_t$.

$$\mathbf{I}_t = f_{\theta_I}(\mathrm{PE}(v_t) + \mathbf{e}^{pos}), \quad \mathbf{C} = f_{\theta_C}(C), \tag{1}$$

where $\mathrm{PE}(\cdot)$ is the patch embedding. Each frame is split into $L = \frac{H}{P} \times \frac{W}{P}$ patches, L is the number of patches, P is the patch size, $\mathbf{e}^{pos}$ is the absolute positional embedding, $\mathrm{PE}(v_t) + \mathbf{e}^{pos} = [v_{t,0} + \mathbf{e}_0^{pos}, \mathrm{PE}(v_{t,1}) + \mathbf{e}_1^{pos}, \cdots, \mathrm{PE}(v_{t,l}) + \mathbf{e}_l^{pos}, \cdots, \mathrm{PE}(v_{t,L}) + \mathbf{e}_L^{pos}] = [\mathbf{z}_{t,0}, \cdots, \mathbf{z}_{t,l}, \cdots, \mathbf{z}_{t,L}]$, $v_{t,0}$ is the class token. $\mathbf{I}_t = f_{\theta_I}([\mathbf{z}_{t,l}]_{t \in T, l \in L+1}) = [\mathbf{I}_{t,0}, \cdots, \mathbf{I}_{t,l}, \cdots, \mathbf{I}_{t,L}]$.

Then, we add absolute temporal embedding $\mathbf{e}^{tem}$ to the $\mathbf{I}_t, t \in T$ and feed them into the Kronecker mask temporal transformer $f_{\theta_V}(\cdot)$. Finally, by selecting the class token from each frame and averaging them, we obtain a video representation $\mathbf{v}$ with the same dimension as $\mathbf{c}_m, m \in M$.

$$\mathbf{V} = f_{\theta_V}([\mathbf{I}_t]_{t \in T} + \mathbf{e}^{tem}), \quad \mathbf{v} = \mathrm{Avg}([\mathbf{V}_{t,0}]_{t \in T}), \tag{2}$$

where $\mathbf{V} = f_{\theta_V}([\mathbf{I}_t]_{t \in T} + \mathbf{e}^{tem}) = f_{\theta_V}([\mathbf{I}_t + \mathbf{e}_t^{tem}]_{t \in T}) = f_{\theta_V}([[\mathbf{I}_{0,0} + \mathbf{e}_0^{tem}, \cdots, \mathbf{I}_{0,L} + \mathbf{e}_0^{tem}], \cdots, [\mathbf{I}_{t,0} + \mathbf{e}_t^{tem}, \cdots, \mathbf{I}_{t,L} + \mathbf{e}_t^{tem}], \cdots, [\mathbf{I}_{T-1,0} + \mathbf{e}_{T-1}^{tem}, \cdots, \mathbf{I}_{T-1,L} + \mathbf{e}_{T-1}^{tem}]]) = [[\mathbf{V}_{0,0}, \cdots, \mathbf{V}_{0,L}], \cdots, [\mathbf{V}_{t,0}, \cdots, \mathbf{V}_{t,L}], \cdots, [\mathbf{V}_{T-1,0}, \cdots, \mathbf{V}_{T-1,L}]]$, $\mathrm{Avg}(\cdot)$ is the average pooling function. Our optimization goal is to maximize the cosine similarity between video $\mathbf{v}$ and its corresponding texts $\mathbf{c}_m \in \mathbf{C}$ representations:

$$\mathrm{sim}(\mathbf{v}, \mathbf{c}_m) = \frac{\langle \mathbf{v}, \mathbf{c}_m \rangle}{\|\mathbf{v}\| \cdot \|\mathbf{c}_m\|}. \tag{3}$$

3.2 Kronecker Mask Attention

For an image $\in \mathbb{R}^{H \times W \times 3}$, it is first split into patches and then flattened into a token sequence after patch embedding. The resulting feature shape is (L, D), $L = \frac{H}{P} \times \frac{W}{P}$, where D is the hidden dimension. This process is a standard operation in ViT Dosovitskiy et al. (2020), denoted as Spatial Attention (SA):

$$\mathbf{Z}_{(L,D)} = \mathrm{Softmax}(\mathbf{Q}_{(L,D)} \mathbf{K}^{\mathbf{T}}_{(L,D)} / \sqrt{D}) \mathbf{V}_{(L,D)}, \tag{4}$$

where $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ represent the query, key, value matrices, respectively.

For a video $\in \mathbb{R}^{T \times H \times W \times 3}$, T is the number of frames, previous studies Arnab et al. (2021); Bertasius et al. (2021); Neimark et al. (2021); Guo et al. (2021); Tong et al. (2022); Feichtenhofer et al. (2022) typically employ either joint attention Bertasius et al. (2021); Feichtenhofer et al. (2022) or factorized (divided) attention Bertasius et al. (2021); Arnab et al. (2021). Joint attention flattens a video into a longer token sequence, resulting in a feature shape of $(T \times L, D)$. This token interaction mode is illustrated in Fig. 2 **Left** (a). Joint attention encounters spatiotemporal homogenization issues, where random shuffling of tokens not affecting (or slightly) the final pooling result.

In contrast, factorized attention factorize joint attention into spatial attention and temporal attention to avoid spatiotemporal homogenization. They utilize the same spatial attention as Eqn. 4, while the temporal attention varies. Two common temporal attentions are pipeline temporal attention Arnab et al. (2021); Bertasius et al. (2021) and class-token-only temporal attention Arnab et al. (2021); Ni et al. (2022); Wang et al. (2021). Their feature shapes are (L, T, D) and (T, D), respectively. Pipeline temporal attention has a limited temporal receptive field, which is limited to a fixed time pipeline (tube), hence the term "pipeline temporal attention", as shown in Fig. 2 **Left** (c). It has limited scope of capture dynamic information since objects of interest do not always appear in the same 2D location across frames. Although ATA Zhao et al. (2022) utilizes alignment techniques to bend the time pipeline to capture dynamic objects, it still has a limited temporal receptive field, in Fig. 2 **Left** (d). Similarly, class-token-only temporal attention retains only the class token time pipeline and discard others, which both face limited receptive field and may discard lots of potentially valuable information. It is worth noting that in pipeline temporal attention, mean pooling is necessary across all tokens, otherwise, it is equivalent to class-token-only temporal attention.

To address aforementioned drawbacks, we propose Kronecker Mask Temporal Attention (KMTA). Specifically, we allow each patch (token) at timestamp t can interact with all other patches (tokens), excluding those sharing the same timestamp t, as illustrated in Fig. 2 **Left** (b). Compared to pipeline temporal attention, KMTA expands the temporal receptive field width of each token. KMTA can be achieved through joint attention incorporated a Kronecker mask, as shown in Fig. 2 **Right** (left down). Additionally, KMTA alleviates the impact of spatiotemporal homogenization due to the presence of the Kronecker mask. The trick for obtaining the Kronecker mask is Kronecker product $\otimes$:

$$\mathbf{A}_{m\times n} \otimes \mathbf{B}_{p\times q} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix}_{mp\times nq} \tag{5}$$

Eqn. 5 is the definition of $\otimes$, where $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{B} \in \mathbb{R}^{p\times q}$. Thus, it is referred to as the Kronecker mask. Kronecker Mask Temporal Attention (KMTA) can be formulated as:

$$\mathbf{M}_{(T\times L, T\times L)} = [\mathbf{I}_{(T,T)} \otimes (\mathbf{J}_{(L,L)} - \mathbf{I}_{(L,L)})]_{1==-\mathrm{inf}}, \tag{6}$$

$$\mathbf{Z}_{(T\times L,D)} = \mathrm{Softmax}(\mathbf{Q}_{(T\times L,D)}\mathbf{K}^{\mathbf{T}}_{(T\times L,D)}/\sqrt{D} + \mathbf{M}_{(T\times L,T\times L)})\mathbf{V}_{(T\times L,D)}, \tag{7}$$

where $\mathbf{I}_{(-,-)}$ is an identity matrix, $\mathbf{J}_{(-,-)}$ is an all-ones matrix, $[\quad]_{1==-\mathrm{inf}}$ means replacing 1 in the matrix with negative infinity ($-\mathrm{inf}$), and $\mathbf{M}_{(T\times L,T\times L)}$ is the Kronecker mask. A transformer equipped with KMTA is referred to as a Kronecker Mask Temporal (KMT) transformer.

In fact that both the spatial and temporal attention we mentioned above can be derived by combining a tailored Kronecker mask with joint attention. Therefore, we collectively refer to them as Kronecker Mask Attention, as depicted in Fig. 2 (**Right**). The Kronecker mask serves as a prior spatiotemporal heterogeneity inductive bias. Spatial attention allows intra-frame interactions but blocks inter-frame interactions, while Kronecker mask temporal attention allows inter-frame interactions but blocks intra-frame interactions, exhibiting a spatiotemporal structural complementarity.

Moreover, we design another type of Kronecker mask for temporal modeling, known as Kronecker mask causal temporal attention (KMCTA) that aims to alleviate the low-rank bottleneck, as shown in Fig. 3. The mask $\tilde{\mathbf{M}}_{(T\times L,T\times L)}$ of KMCTA can be formulated as:

$$\tilde{\mathbf{M}}_{(T\times L,T\times L)} = [\mathbf{I}_{(T,T)} \otimes (\mathbf{J}_{(L,L)} - \mathbf{I}_{(L,L)}) + (\mathbf{U}_{(T,T)} - \mathbf{I}_{(T,T)}) \otimes \mathbf{J}_{(L,L)}]_{1==-\mathrm{inf}}, \tag{8}$$

where $\mathbf{U}_{(-,-)}$ is an upper triangular matrix, and all elements of the upper triangle are 1. A transformer equipped with KMCTA is referred to as a KMCT transformer. KMCTA ensures the causality of time (in the time dimension, KMTA is bidirectional, while KMCTA is unidirectional). In addition, KMCTA always has a full-rank attention matrix, whereas KMTA and joint attention can not guarantee this property. The proof is detailed in *Appendix* A. Some studies Bhojanapalli et al. (2020); Han et al. (2023) indicate that it is important to avoid the low-rank bottleneck to improve the representation power of transformer architecture. The low-rank bottleneck describes a situation where, as the tokens

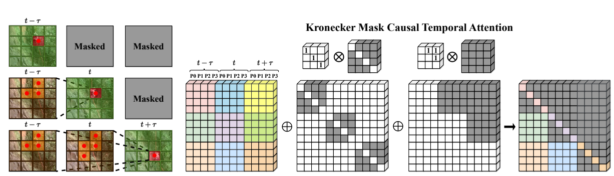

Figure 3: Kronecker mask causal temporal attention.

length significantly increases (i.e. the number of tokens n far exceeds the latent dimension d_{head} in multi-head attention), the performance improvement of transformer-based structures may encounter bottlenecks. This problem may become more knotty when transitioning images to video data.

3.3 INTERPRETIVE PROMPT

To address the issue 2) in Sec. 1, we prompt LLMs to generate interpretive texts that align the text semantics of video-text pairs with the rich text semantics in VLM's pre-training dataset. This design aids the model in understanding abstract verbs. This approach, termed interpretive prompt, designed from the following aspects: 1) Action decomposition: Providing detailed descriptions of actions by decomposing complex actions into simpler, more basic ones. This clarifies the include relationship between complex and basic actions, and help distinguish similar actions, as not all action concepts are of equal status. For example, the action 'playing basketball' may consist of 'running', 'jumping', and 'shooting'. 'Running', 'jumping', and 'hand movements' are considered as basic actions, while 'playing basketball' is a complex action. Similarly, 'dribbling' may also involves 'running', and 'hand movements', besides, it is also a subset of 'playing basketball'. Action decomposition enhances the separability of action concepts in semantic space, helping models understand the relationships between actions. 2) Synonym conversion: Generating synonyms for verbs and phrases that convey the same core concept but with more varied expressions. This improves zero-shot generalization as an action concept may have multiple expressions. When encountering similar action descriptions in the unseen domain, the model demonstrates stronger generalization robustness. 3) Involving body parts: Describing actions based on possible body parts involved, helping the model to localize the region where action occurs.

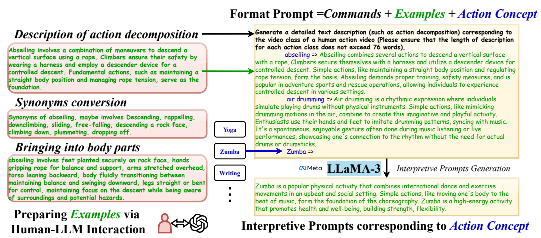

Figure 4: The Interpretive Prompt scheme.

We leverage ChatGPT OpenAI (2024) and LLaMA-3 Abhimanyu Dubey et al. (2024) to automatically generate action interpretations. Initially, we ask the knowledgeable ChatGPT OpenAI (2024) to provide several examples of text descriptions that align with our expect (based on the aforementioned aspects). Subsequently, we provide the format prompt to LLaMA-3 Abhimanyu Dubey et al. (2024) for text completion, as illustrated in Fig. 4. For example, "*Generate a detailed text description corresponding to the video class...*" –***Command***, accompanied by a few examples like, "*abseiling → Abseiling combines several actions...*" –***Examples***, and the action concept that LLM needs to interpret –***Action Concept***. We feed the format prompt to LLaMA-3 multiple times to obtain diverse interpretive prompts of actions. For a given action category, the texts used during training and inference include the original verbs or phrases, template filling prefixes or suffixes, and interpretive prompts. Our interpretive prompts provide texts at the sentence, phrase, and word-levels, enhancing the flexibility of text usage during the inference. More technique details refer to *Appendix* G.

3.4 TRAINING AND INFERENCE

In each training step, a batch of B videos is sampled, and M represents the total number of text descriptions of each verb or phrase. There are a total of K action categories. The training loss is formulated as follows:

$$L = -\frac{1}{N} \sum_{m}^{\mathrm{sub}\{M\}} \sum_{i}^{B} \log \frac{\exp(\mathrm{sim}(\mathbf{v}^i, \mathbf{c}_m^i)/\tau)}{\sum_k^K \exp(\mathrm{sim}(\mathbf{v}^i, \mathbf{c}_m^k)/\tau)}, \tag{9}$$

where, τ is the temperature parameter, $\mathrm{sub}\{M\}$ indicates that we sample a subset of all text descriptions for a single training step. $\mathbf{v}^i$ represents the video belonging to the i-th category. $\mathbf{c}_m^i$ denotes the descriptions corresponding to the i-th category. During inference, we take the sum of the similarities between each $\mathbf{v}^i$ and all $\mathbf{c}_m^i$ as the final similarity score S:

$$S = \sum_{m}^{M} \log \frac{\exp(\mathrm{sim}(\mathbf{v}^i, \mathbf{c}_m^i)/\tau)}{\sum_k^K \exp(\mathrm{sim}(\mathbf{v}^i, \mathbf{c}_m^k)/\tau)}. \tag{10}$$

Table 1: Comparison with state-of-the-art on Kinetics-400. * indicates pretraining with a video-text collection. The FLOPs per view of each method is reported. Red represents optimal performance. **Bold** represents optimal performance among CLIP-based methods at the same model scale.

Method	Pretrain	Frames	Top-1	Top-5	Views	FLOPs(G)
			Without Language			
Uniformer-B Li et al. (2022a)	IN-1k	32	83.0	95.4	4×3	259
TimeSformer-L Bertasius et al. (2021)	IN-21k	96	80.7	94.7	1×3	2380
Mformer-HR Patrick et al. (2021)	IN-21k	16	81.1	95.2	10×3	959
ATA Zhao et al. (2022)	IN-21k	32	81.9	95.5	4×3	793
Swin-L (384 ↑) Liu et al. (2022)	IN-21k	32	84.9	96.7	10×5	2107
MViTv2-L (312 ↑) Li et al.	IN-21k	40	86.1	97.0	5×5	2828
ViViT-H/14x2 Arnab et al. (2021)	JFT-300M	32	84.9	95.8	4×3	8316
TokenLearner-L/10 Ryoo et al. (2021)	JFT-300M	-	85.4	96.3	4×3	4076
CoVeR Zhang et al. (2021)	JFT-3B	-	87.2	-	1×3	-
			With Language			
MTV-H Yan et al. (2022)	WTS*	32	**89.1**	**98.2**	4×3	3705
X-CLIP-B/32 Ni et al. (2022)	CLIP-400M	8	80.1	94.8	4×3	39
X-CLIP-B/32 Ni et al. (2022)	CLIP-400M	16	81.0	95.1	4×3	75
ILA-B/32 Tu et al. (2023)	CLIP-400M	8	80.6	94.9	4×3	40
ILA-B/32 Tu et al. (2023)	CLIP-400M	16	81.8	95.4	4×3	75
CLAVER-B/32 (KMT)	CLIP-400M	8	**81.5**	**95.5**	4×3	50
CLAVER-B/32 (KMT)	CLIP-400M	16	82.4	**95.9**	4×3	98
CLAVER-B/32 (KMCT)	CLIP-400M	8	81.3	95.4	4×3	50
CLAVER-B/32 (KMCT)	CLIP-400M	16	**82.6**	**95.9**	4×3	98
Action-CLIP-B/16 Wang et al. (2021)	CLIP-400M	16	82.6	96.2	10×3	-
A6 Ju et al. (2022)	CLIP-400M	16	76.9	93.5	-	-
X-CLIP-B/16 Ni et al. (2022)	CLIP-400M	8	83.1	96.5	4×3	145
X-CLIP-B/16 Ni et al. (2022)	CLIP-400M	16	84.3	96.8	4×3	287
EVL-B/16 Lin et al. (2022)	CLIP-400M	8	82.9	-	-	444
EVL-B/16 Lin et al. (2022)	CLIP-400M	16	83.6	-	-	888
ViFiCLIP-B/16 Rasheed et al. (2023)	CLIP-400M	16	83.9	96.3	4×3	281
ASU-B/16 Chen et al. (2023)	CLIP-400M	8	84.1	96.3	4×3	146
ASU-B/16 Chen et al. (2023)	CLIP-400M	16	84.8	96.7	4×3	288
ILA-B/16 Tu et al. (2023)	CLIP-400M	8	83.4	96.3	4×3	149
ILA-B/16 Tu et al. (2023)	CLIP-400M	16	85.0	97.0	4×3	295
ALT-B/16 Chen et al. (2024b)	CLIP-400M	16	85.5	96.7	3×1	1308
CLAVER-B/16 (KMT)	CLIP-400M	8	**84.3**	96.3	4×3	203
CLAVER-B/16 (KMT)	CLIP-400M	16	85.9	**97.3**	4×3	434
CLAVER-B/16 (KMCT)	CLIP-400M	8	84.1	96.2	4×3	203
CLAVER-B/16 (KMCT)	CLIP-400M	16	**86.0**	97.2	4×3	434
X-CLIP-L/14 Ni et al. (2022)	CLIP-400M	8	87.0	97.7	4×3	658
X-CLIP-L/14 (336↑) Ni et al. (2022)	CLIP-400M	16	87.6	97.5	4×3	3086
EVL-L/14 Lin et al. (2022)	CLIP-400M	8	86.3	-	-	2022
EVL-L/14 (336↑) Lin et al. (2022)	CLIP-400M	32	87.7	-	-	18196
ASU-L/14 Chen et al. (2023)	CLIP-400M	8	87.8	97.8	4×3	660
ASU-L/14 (336↑) Chen et al. (2023)	CLIP-400M	16	88.3	98.0	4×3	3084
ILA-L/14 Tu et al. (2023)	CLIP-400M	8	87.6	97.8	4×3	673
ILA-L/14 (336↑) Tu et al. (2023)	CLIP-400M	16	88.1	97.8	4×3	3130
ALT-B/14 Chen et al. (2024b)	CLIP-400M	16	87.8	97.7	3×1	4947
CLAVER-L/14 (KMT)	CLIP-400M	8	**88.1**	97.7	4×3	931
CLAVER-L/14(336↑) (KMT)	CLIP-400M	16	88.8	**98.1**	4×3	5390
CLAVER-L/14 (KMCT)	CLIP-400M	8	87.9	97.7	4×3	931
CLAVER-L/14(336↑) (KMCT)	CLIP-400M	16	**88.9**	98.0	4×3	5390

4 Experiments

4.1 Implementation Details

Architectures and hyperparameters. We employ CLIP-B/32, CLIP-B/16, CLIP-L/14 as our backbones and derive corresponding variants: CLAVER-B/32, CLAVER-B/16, CLAVER-L/14, respectively. The frame length settings include 8 and 16. For all CLAVER variants, the number of layers in the KMT/KMCT transformer is equal to one-third of the number of layers in the image encoder. For example, if the image encoder has 12 or 24 layers, then the KMT/KMCT transformer employs 4 or 8 layers. The detailed hyperparameter settings are provided in *Appendix* D.

Datasets and metrics. We evaluate the performance of our method on four benchmarks: Kinetics-400 Kay et al. (2017), Kientics-600 Carreira et al. (2018), UCF-101 Soomro et al. (2012), HMDB-51 Kuehne et al. (2011). We report the Top-1 and Top-5 accuracy as evaluation metrics.

4.2 Comparison Results

Fully-supervised Experiments. We conducted fully supervised experiments on Kinetics-400 and Kinetics-600, respectively. In Tab. 1, we employ three variant models, CLAVER-B/32, CLAVER-B/16, and CLAVER-L/14, and sample 8 or 16 frames ($8f$,$16f$) with a sparse sampling for each model, employ KMT/KMCT, respectively. CLAVER-B/16$_{8f}$ (KMT/KMCT) and CLAVER-B/16$_{16f}$ (KMT/KMCT) surpass several methods Bertasius et al. (2021); Patrick et al. (2021); Zhao et al. (2022); Li et al. (2022a) pretrained on ImageNet-1k/21k Deng et al. (2009) with shorter frames.

Table 2: Comparison with state-of-the-art on Kinetics-600.

Method	Pretrain	Frames	Top-1	Top-5	Views
MViT-B-24 Fan et al. (2021)	-	32	83.8	96.3	5×1
Swin-L (384 ↑) Liu et al. (2022)	JFT-300M	32	85.8	96.5	4×3
ViViT-H/14x2 Arnab et al. (2021)	JFT-300M	32	85.8	96.5	4×3
TokenLearner-L/10 Ryoo et al. (2021)	JFT-300M	-	86.3	97.0	4×3
CoVeR Zhang et al. (2021)	JFT-3B	32	87.9	-	4×3
MTV-H Yan et al. (2022)	WTS*	32	89.6	98.3	4×3
Florence (384 ↑) Yuan et al. (2021)	FLD-900M	-	87.8	-	1×3
X-CLIP-B/16 Ni et al. (2022)	CLIP-400M	8	85.3	97.1	4×3
ASU-B/16 Chen et al. (2023)	CLIP-400M	8	85.7	-	4×3
CLAVER-B/16 (KMT)	CLIP-400M	8	85.9	97.3	4×3

Table 3: Few-shot performance on HMDB-51 and UCF-101.

Method	HMDB-51				UCF-101			
	K=2	K=4	K=8	K=16	K=2	K=4	K=8	K=16
TSM Lin et al. (2019)	17.5	20.9	18.4	31.0	25.3	47.0	64.4	61.0
TimeSformer Bertasius et al. (2021)	19.6	40.6	49.4	55.4	48.5	75.6	83.7	89.4
Swin-B Liu et al. (2022)	20.9	41.3	47.9	56.1	53.3	74.1	85.8	88.7
Action-CLIP Wang et al. (2021)	55.0	56.0	58.0	-	80.0	85.0	89.0	-
X-CLIP Ni et al. (2022)	53.0	57.3	62.8	64.0	76.4	83.4	88.3	91.4
X-Florence Ni et al. (2022)	51.6	57.8	64.1	64.2	84.0	88.5	92.5	94.8
MAXI Lin et al. (2023)	58.0	60.1	65.0	66.5	86.8	89.3	92.4	93.5
ASU Chen et al. (2023)	**60.1**	63.8	67.2	70.8	**91.4**	**94.6**	96.0	97.2
CLAVER (KMT)	58.6	**63.9**	**68.0**	**72.5**	89.7	92.9	**96.1**	**98.0**

Table 4: Zero-shot on HMDB-51 and UCF-101.

Method	HMDB-51	UCF-101
MTE Xu et al. (2016)	19.7 ± 1.6	15.8 ± 1.3
ASR Wang & Chen (2017)	21.8 ± 0.9	24.4 ± 1.0
ZSECOC Qin et al. (2017)	22.6 ± 1.2	15.1 ± 1.7
UR Zhu et al. (2018)	24.4 ± 1.6	17.5 ± 1.6
TS-GCN Gao et al. (2019)	23.2 ± 3.0	34.2 ± 3.1
E2E Brattoli et al. (2020)	32.7	48
ER-ZSAR Chen & Huang (2021)	35.3 ± 4.6	51.8 ± 2.9
Action-CLIP Wang et al. (2021)	40.8 ± 5.4	58.3 ± 3.4
X-CLIP Ni et al. (2022)	44.6 ± 5.2	72.0 ± 2.3
ASU Chen et al. (2023)	48.1 ± 2.8	75.0 ± 2.3
MAXI Lin et al. (2023)	51.2 ± 1.1	75.2 ± 0.9
OST Chen et al. (2024a)	52.9 ± 0.9	75.3 ± 2.1
CLAVER (KMT)	$\mathbf{54.1 \pm 2.4}$	$\mathbf{78.6 \pm 1.7}$

Table 5: Zero-shot on Kinetics-600.

Method	Top-1	Top-5
DEVISE Frome et al. (2013)	23.8 ± 0.3	51.0 ± 0.6
ALE Akata et al. (2015a)	23.4 ± 0.8	50.3 ± 1.4
SJE Akata et al. (2015b)	22.3 ± 0.6	48.2 ± 0.4
ESZSL Romera-Paredes & Torr (2015)	22.9 ± 1.2	48.3 ± 0.8
DEM Zhang et al. (2017)	23.6 ± 0.7	49.5 ± 0.4
GCN Ghosh et al. (2020)	22.3 ± 0.6	49.7 ± 0.6
ER-ZSAR Chen & Huang (2021)	42.1 ± 1.4	73.1 ± 0.3
X-CLIP Ni et al. (2022)	65.2 ± 0.4	86.1 ± 0.8
ASU Chen et al. (2023)	67.6 ± 0.2	87.2 ± 0.3
MAXI Lin et al. (2023)	70.9 ± 1.2	92.1 ± 0.5
OST Chen et al. (2024a)	70.5 ± 0.7	92.1 ± 0.3
CLAVER (KMT)	$\mathbf{73.8 \pm 0.6}$	$\mathbf{93.1 \pm 0.6}$

CLAVER-B/16_{16f} (KMT) outperforms Swin-L (384↑) Liu et al. (2022) by 1.0%, and is slightly lower than MViTv2-L (312↑) Li et al., as lower resolution, shorter frames and fewer views. CLAVER-B/14_{8f} (KMT/KMCT) outperforms some methods Arnab et al. (2021); Ryoo et al. (2021); Zhang et al. (2021) pretrained on JFT-300M/JFT-3B. CLAVER-B/14_{16f} (KMT) outperforms CoVeR Zhang et al. (2021) by 1.6%, however, inferior to MTV-H Yan et al. (2022), as it utilizes WTS*, which contains 70M video-text pairs with about 17B images, much larger than CLIP-400M. Compared to those approaches based on CLIP Radford et al. (2021), under configuration ViT-B/32_{8f} and ViT-B/32_{16f}, CLAVER (KMT/KMCT) surpasses X-CLIP Ni et al. (2022) and ILA Tu et al. (2023). Under configuration ViT-B/$16_{8f,16f}$ and ViT-L/$14_{8f,16f}$, CLAVER (KMT/KMCT) exceeds most methods under the same configuration. In Tab. 2, CLAVER-B/16_{8f} (KMT) achieves higher performance compared to ViViT-H/14x2 Arnab et al. (2021), MViT-B-24 Fan et al. (2021). Our method has lower performance than these methods Ryoo et al. (2021); Zhang et al. (2021); Yan et al. (2022); Yuan et al. (2021); Liu et al. (2022), as they use longer frames, or more data, or higher resolutions. In addition, CLAVER-B/16_{8f} (KMT/KMCT) outperforms ASU-B/16_{8f} Chen et al. (2023) and X-CLIP-B/16_{8f} Ni et al. (2022) at the same scale.

Few-shot Experiments. We pretrain CLAVER-B/16_{32f} (KMT) ($32f$ means with 32 frames) on Kinetics-400, and then perform few-shot transfer of 2, 4, 8, 16 samples on HMDB-51 and UCF-101. Tab. 3 depicts the results of few-shot experiments. CLAVER is comparable wtih ASU Chen et al. (2023), and consistently surpasses the X-CLIP Ni et al. (2022) and MAXI Lin et al. (2023) across all K ranges. Additionally, CLAVER significantly outperforms other previous methods like Lin et al. (2019); Bertasius et al. (2021); Liu et al. (2022). More details about the evaluation protocols are provided in the *Appendix* D.

Zero-shot Experiments. We also pretrain CLAVER-B/16_{32f} (KMT) on Kinetics400 for zero-shot. As shown in Tab. 4, on HMDB-51 Kuehne et al. (2011) and UCF-101 Soomro et al. (2012) benchmarks, CLAVER surpasses OST Chen et al. (2024a) , ASU Chen et al. (2023) and X-CLIP Ni et al. (2022) under the same configuration, and far outperforms other previous methods. Additionally, in Tab. 5, on Kinetics600 Carreira et al. (2018) benchmark, CLAVER outperforms OST Chen et al. (2024a), MAXI Lin et al. (2023) and all other methods. More details about the evaluation protocols are provided in the *Appendix* D.

4.3 Ablation Study

Components ablation studies. We performed ablation studies to evaluate the effects of each component under the CLAVER-B/32_{8f} (KMT) configuration. The results are shown in Tab. 6. Our baseline, denoted as CLIP-Mean, implements temporal mean pooling for CLIP. By equipping the CLIP with a KMT transformer at 1/3 scale, the Top-1 accuracy increases by 3.5%. We only introduce interpretive prompt, the performance increases by 1%. When we further incorporate both of them, CLAVER surpasses the CLIP-Mean by 4.1%. In addition, we test the effect of the number of KMT transformer layers. With only one layer, the performance improvement is minimal. Increasing

Table 6: Ablation of each component.

Components	Top-1 (%)
Baseline (CLIP-Mean)	77.4
Baseline + KMTA 1 layer	78.3
Baseline + KMTA 1/6 scale	79.6
Baseline + KMTA 1/3 scale	80.9
Baseline + Interpretive Prompt	78.4
Baseline + KMTA 1/3 scale + Interpretive Prompt	**81.5**

Table 7: Comparison of temporal attentions.

Temporal Modeling	Top-1 (%)	Top-5 (%)
Baseline (Mean Pooling)	78.4	94.3
Class-Token-Only	78.9	94.3
Pipeline Temporal Attention	79.4	94.4
Joint Attention	80.1	94.9
Kronecker Mask Causal Temporal Attention	81.4	**95.5**
Kronecker Mask Temporal Attention	**81.5**	**95.5**

Table 8: The impact of patch size and frame length on joint attention (JA), KMTA and KMCTA. Conducting on HMDB-51 and UCF-101.

HMDB-51, UCF-101	patch size = 32 (ViT-B/32)			patch size = 16 (ViT-B/16)		
	JA (%)	KMTA (%)	KMCTA (%)	JA (%)	KMTA (%)	KMCTA (%)
frame length = 8	67.9, 93.3	68.8, 94.1	68.7, 93.9	72.2, 96.2	73.2, 96.3	72.2, 96.1
frame length = 16	68.0, 93.6	69.1, 94.2	69.4, 94.8	72.1, 96.3	72.3, 96.6	72.8, 96.4

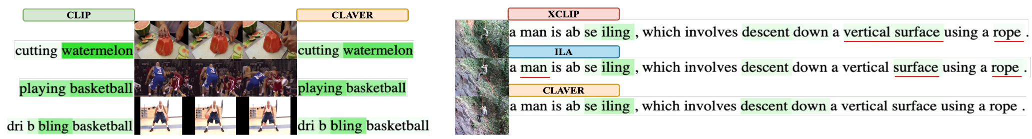

Figure 5: Word importance of CLIP, X-CLIP, ILA and CLAVER. Darker color, higher importance.

the number of layers to 1/6 scale results in further performance gains, and at 1/3 scale, we observe significant improvement.

Comparison of different temporal attentions and low-rank bottleneck issue. Tab. 7 compares the performance of different temporal modeling methods mentioned in Sec. 3.2. We find that class-token-only temporal attention and pipeline temporal attention have inferior performance. KMTA and KMCTA outperform joint attention. The top-1 (%) of KMTA slightly surpasses KMCTA by 0.1%. Then, we conduct further experiments on HMDB-51 and UCF101 that observing the impact of the low-rank bottleneck issue on them. We increasing the length of token sequences by reducing patch size or increasing frame length, and observe the effects on them, as shown in Tab. 8. When reducing the patch size, however, we do not observe the low-rank bottleneck. Joint attention, KMTA and KMCTA achieve better performance due to more fine-grained features as the smaller patch size results in each token representing smaller local region. In contrast, when we increase the frame length, KMCTA's performance is optimal in most configurations, and only the performance of KMCTA can steadily improve. The performance improvement of joint attention and KMTA is limited. Meanwhile, in Tab. 1, with the increase of frame length under the same backbone, the performance increase of KMCTA is also greater than that of KMTA. This indicates that KMCTA has a more significant advantage with longer frame length.

5 Visualization and Analysis

We employ Chefer et al. (2021) for bi-modal visualization and show the explainability of visual-textual attentions.

Word importance. We visualize word importance for CLIP, X-CLIP, ILA and CLAVER in Fig. 5. Observations indicate that CLIP tends to focus on nouns, whereas CLAVER prefers to verbs. Compared to previous works, CLAVER is more inclined towards verb concept, while X-CLIP Ni et al. (2022) and ILA Tu et al. (2023) show a slight inclination towards nouns. These findings indicate the effectiveness of interpretive prompts for nouns concept to verbs concept transition.

Spatiotemporal homogenization study. We define spatiotemporal homogenization as a phenomenon where random token shuffling has little impact on the semantics of the visual branch, which is illogical. It aims to illustrate the interpretability behind the performance of spatiotemporal modeling. In Fig. 6 (**Upper Left**), we illustrate the token shuffling. We denote token shuffling before adding time embedding as PreTE shuffling, token shuffling after adding time embedding as PostTE shuffling, and None represents no shuffle. Fig. 6 shows changes in word importance and similarity following both PreTE and PostTE shuffling. For joint attention, we observe that PreTE shuffling marginally affects similarity and word importance, while PostTE shuffling does not affect

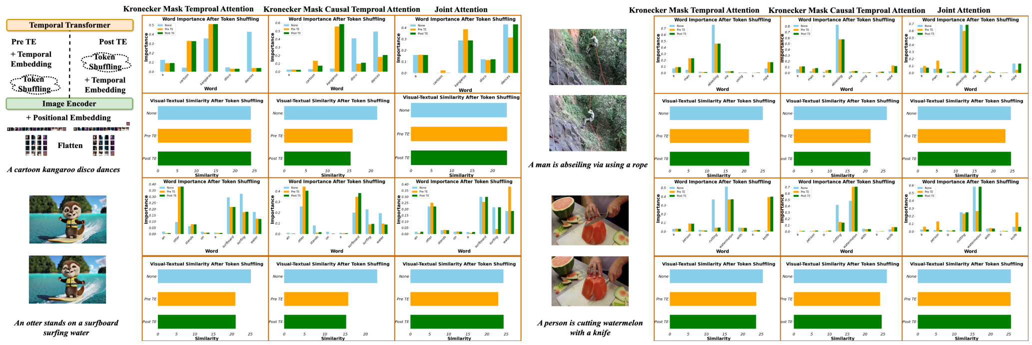

Figure 6: Spatiotemporal Homogenization. (**Upper Left**) Token shuffling. (**Vertical bar chart**) Word importance refers to the degree of correlation between each word in a sentence description and the semantics of the video content, while (**Horizontal bar chart**) Similarity refers to the similarity between visual and textual representations.

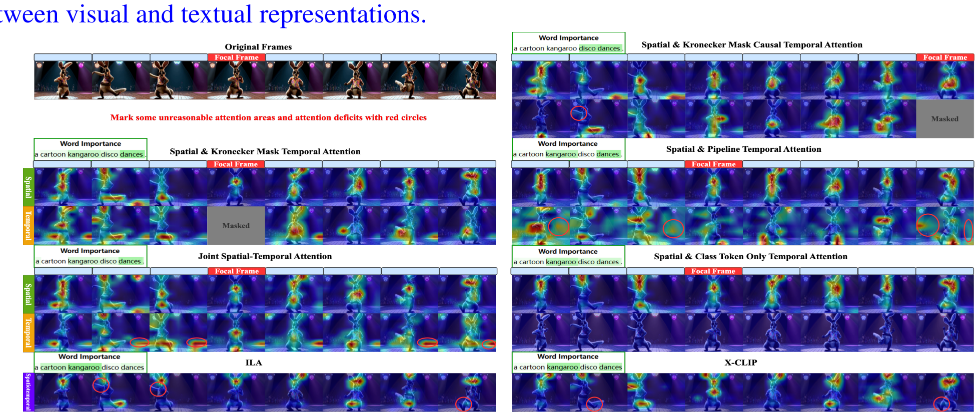

Figure 7: Visualizing spatiotemporal attention maps of different spatio-temporal modeling.

similarity score and word importance. Regarding KMTA, both PreTE and PostTE shuffling lead to changes in the similarity and word importance. Meanwhile, KMCTA is profoundly affected by both PreTE and PostTE shuffling, which result in lower similarity and disturbance of word importance. This phenomenon suggest that KMTA and KMCTA possess varying degrees ability in mitigating spatiotemporal homogenization and the Kronecker mask serves as a natural inductive bias for spatiotemporal structural heterogeneity. It also demonstrates that equipping learnable position/time encoding is not insufficient to alleviate spatiotemporal homogenization. Please refer to *Appendix* B for more and explanations and visualizations.

Visualization of spatiotemporal attention map. Fig. 7 presents various spatiotemporal attention maps. X-CLIP Ni et al. (2022) and ILA Tu et al. (2023) adopt factorized attention structures resulting in intertwined spatiotemporal attention maps. However, CLAVER employs a factorized encoder structure, allowing for the separation of spatial and temporal attention maps. Notably, class-token-only temporal attention discards other time tubes so that its temporal attention maps appear vacant. KMTA and KMCT have more reasonable spatial (focuses on the executor of the action) and temporal (focuses on areas where the action occurs) attention maps, while ILA and X-CLIP ignore some action dense areas. Moreover, joint attention and pipeline temporal attention will be attracted by some irrelevant backgrounds.

6 CONCLUSION

In this work, we present the **C**ontrastive **L**anguage-**A**ction **V**ideo Learn**er** (**CLAVER**) to shift from the alignment of visual objects and nouns in CLIP to the alignment of action behaviors and verbs. We propose Kronecker mask temporal attention and Kronecker mask causal temporal attention for temporal modeling. Interpretive prompts are employed to transition the focus on nouns to verbs. Extensive experiments under different evaluation settings demonstrate the effectiveness of our method. Limitation Discussion and Broader Impact are provided in the *Appendix* E, F.

REFERENCES

Abhinav Jauhri Abhimanyu Dubey, Abhinav Pandey, and et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7): 1425–1438, 2015a.

Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2927–2936, 2015b.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6836–6846, 2021.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, pp. 4, 2021.

Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Low-rank bottleneck in multi-head attention models. In *International conference on machine learning*, pp. 864–873. PMLR, 2020.

Biagio Brattoli, Joseph Tighe, Fedor Zhdanov, Pietro Perona, and Krzysztof Chalupka. 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*, pp. 4613–4623, 2020.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.

Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 397–406, 2021.

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

Tongjia Chen, Hongshan Yu, Zhengeng Yang, Zechuan Li, Wei Sun, and Chen Chen. 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*, pp. 18888–18898, 2024a.

Yifei Chen, Dapeng Chen, Ruijin Liu, Hao Li, and Wei Peng. Video action recognition with attentive semantic units. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10170–10180, 2023.

Yifei Chen, Dapeng Chen, Ruijin Liu, Sai Zhou, Wenyuan Xue, and Wei Peng. Align before adapt: Leveraging entity-to-region alignments for generalizable video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18688–18698, 2024b.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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. *arXiv preprint arXiv:2010.11929*, 2020.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6824–6835, 2021.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36, 2024.

Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 203–213, 2020.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6202–6211, 2019.

Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013.

Junyu Gao, Tianzhu Zhang, and Changsheng Xu. 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, pp. 8303–8311, 2019.

Pallabi Ghosh, Nirat Saini, Larry S Davis, and Abhinav Shrivastava. All about knowledge graphs for actions. *arXiv preprint arXiv:2008.12432*, 2020.

Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13505–13515, 2021.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

Xudong Guo, Xun Guo, and Yan Lu. Ssan: Separable self-attention network for video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12618–12627, 2021.

Dongchen Han, Xuran Pan, Yizeng Han, Shiji Song, and Gao Huang. Flatten transformer: Vision transformer using focused linear attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5961–5971, 2023.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 3154–3160, 2017.

Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*, 2021.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, pp. 105–124. Springer, 2022.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, 2014.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pp. 2556–2563. IEEE, 2011.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 11336–11344, 2020a.

Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatiotemporal representation learning. *arXiv preprint arXiv:2201.04676*, 2022a.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975, 2022b.

Y Li, CY Wu, H Fan, K Mangalam, B Xiong, J Malik, and C Feichtenhofer. Improved multiscale vision transformers for classification and detection. arxiv 2021. *arXiv preprint arXiv:2112.01526*.

Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. Tea: Temporal excitation and aggregation for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 909–918, 2020b.

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

Wei Lin, Leonid Karlinsky, Nina Shvetsova, Horst Possegger, Mateusz Kozinski, Rameswar Panda, Rogerio Feris, Hilde Kuehne, and Horst Bischof. 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*, pp. 2851–2862, 2023.

Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *European Conference on Computer Vision*, pp. 388–404. Springer, 2022.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3202–3211, 2022.

Zhaoyang Liu, Limin Wang, Wayne Wu, Chen Qian, and Tong Lu. Tam: Temporal adaptive module for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13708–13718, 2021.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

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 and captioning. *Neurocomputing*, 508: 293–304, 2022.

Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1049–1059, 2020.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2630–2640, 2019.

Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. Verbs in action: Improving verb understanding in video-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15579–15591, 2023.

Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3163–3172, 2021.

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2022.

OpenAI. Gpt-4 technical report https://arxiv.org/abs/2303.08774, 2024. URL `https://chat.openai.com`.

Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35:26462–26477, 2022.

Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. *Advances in neural information processing systems*, 34:12493–12506, 2021.

Yijun Qian, Lijun Yu, Wenhe Liu, and Alexander G Hauptmann. Rethinking zero-shot action recognition: Learning from latent atomic actions. In *European Conference on Computer Vision*, pp. 104–120. Springer, 2022.

Jie Qin, Li Liu, Ling Shao, Fumin Shen, Bingbing Ni, Jiaxin Chen, and Yunhong Wang. Zero-shot action recognition with error-correcting output codes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2833–2842, 2017.

Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pp. 5533–5541, 2017.

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. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6545–6554, 2023.

Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International conference on machine learning*, pp. 2152–2161. PMLR, 2015.

Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. *Advances in Neural Information Processing Systems*, 34:12786–12797, 2021.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019a.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7464–7473, 2019b.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28, 2022.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, 2015.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2018.

Shuyuan Tu, Qi Dai, Zuxuan Wu, Zhi-Qi Cheng, Han Hu, and Yu-Gang Jiang. Implicit temporal modeling with learnable alignment for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19936–19947, 2023.

Yael Vinker, Ehsan Pajouheshgar, Jessica Y Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. Clipasso: Semantically-aware object sketching. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022.

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pp. 20–36. Springer, 2016.

Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021.

Qian Wang and Ke Chen. Alternative semantic representations for zero-shot human action recognition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I 10*, pp. 87–102. Springer, 2017.

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity, 2020. URL `https://arxiv.org/abs/2006.04768`.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803, 2018.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 305–321, 2018.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18134–18144, 2022.

Xun Xu, Timothy M Hospedales, and Shaogang Gong. Multi-task zero-shot action recognition with prioritised data augmentation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 343–359. Springer, 2016.

Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language alignment. In *The Eleventh International Conference on Learning Representations*, 2022.

Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3333–3343, 2022.

Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. Aim: Adapting image models for efficient video action recognition. *arXiv preprint arXiv:2302.03024*, 2023.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

Bowen Zhang, Jiahui Yu, Christopher Fifty, Wei Han, Andrew M Dai, Ruoming Pang, and Fei Sha. Co-training transformer with videos and images improves action recognition. *arXiv preprint arXiv:2112.07175*, 2021.

Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2021–2030, 2017.

Yizhou Zhao, Zhenyang Li, Xun Guo, and Yan Lu. Alignment-guided temporal attention for video action recognition. *Advances in Neural Information Processing Systems*, 35:13627–13639, 2022.

Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 803–818, 2018.

Chen Zhu, Wei Ping, Chaowei Xiao, Mohammad Shoeybi, Tom Goldstein, Anima Anandkumar, and Bryan Catanzaro. Long-short transformer: Efficient transformers for language and vision. *Advances in neural information processing systems*, 34:17723–17736, 2021.

Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8746–8755, 2020.

Yi Zhu, Yang Long, Yu Guan, Shawn Newsam, and Ling Shao. Towards universal representation for unseen action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9436–9445, 2018.

7 Appendix

Roadmap

In the Appendix, we present proof in A, Spatiotemporal homogenization study details in B, some additional experiment results in C, experiment setting details in D, limitation discussions in E, broader impact in F, interpretive prompt technique details in G, analysis of synthetic video testing in H, dataset details in I, and more visualization of spatiotemporal attention map in J.

A Proof

We conclude that there is a low-rank bottleneck problem for joint attention and KMTA when facing longer token sequence lengths ($d \ll n$), while KMCTA always guarantees the property of full rank under our assumption. Proofs as follow:

A.1 Proof of joint attention cannot guarantee full rank

Based on the Representation Theorem of Bhojanapalli et al. Bhojanapalli et al. (2020):

Theorem 1 (Representation Theorem Bhojanapalli et al. (2020)). *If $d_q = d_k = d \geq n$, then given any full row rank matrix $\mathbf{X} \in \mathcal{R}^{n\times d}$ and an arbitrary $n \times n$ positive row stochastic matrix $\mathbf{P}$, subject to the sum of each row of $\mathbf{P}$ is equal to 1, i.e., $\mathbf{P1} = \mathbf{1}$, there always exists $d \times d$ projection matrices $\mathbf{W}_q$ and $\mathbf{W}_k$ such that*

$$\text{Softmax}[\frac{(\mathbf{X}\mathbf{W}_q)(\mathbf{X}\mathbf{W}_k)^T}{\sqrt{d_k}}] = \mathbf{P} \tag{11}$$

If $d_q = d_k = d < n$, there exist $\mathbf{X}$ and $\mathbf{P}$ such that Eqn. 11 does not hold for all $\mathbf{W}_q$ and $\mathbf{W}_k$.

The proof process of **Theorem 1** are referred from Bhojanapalli et al. Bhojanapalli et al. (2020). By the way, in Bhojanapalli et al. Bhojanapalli et al. (2020), they set $\mathbf{X} \in \mathcal{R}^{d\times n}$. However, for the convenience of our subsequent proof, we transpose the n and d dimensions, $\mathbf{X} \in \mathcal{R}^{n\times d}$.

For $d \geq n$ case. Since $\mathbf{X}$ exhibits full row rank, there exists a right pseudo inverse $\mathbf{X}^\dagger = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1} \in \mathcal{R}^{d\times n}$ such that $\mathbf{X}\mathbf{X}^\dagger = \mathbf{I}_n$. Let $\mathbf{W}_k = \mathbf{X}^\dagger\tilde{\mathbf{W}}_k$ and $\mathbf{W}_q = \mathbf{X}^\dagger\tilde{\mathbf{W}}_q$. Then

$$\begin{aligned}
(\mathbf{X}\mathbf{W}_q)(\mathbf{X}\mathbf{W}_k)^T &= \mathbf{X}\mathbf{W}_q\mathbf{W}_k^T\mathbf{X}^T && (12)\\
&= \mathbf{X}\mathbf{X}^\dagger\tilde{\mathbf{W}}_q\tilde{\mathbf{W}}_k^T(\mathbf{X}^\dagger)^T\mathbf{X}^T && (13)\\
&= \mathbf{X}\mathbf{X}^\dagger\tilde{\mathbf{W}}_q\tilde{\mathbf{W}}_k^T(\mathbf{X}\mathbf{X}^\dagger)^T && (14)\\
&= \mathbf{I}_n\tilde{\mathbf{W}}_q\tilde{\mathbf{W}}_k^T\mathbf{I}_n^T && (15)\\
&= \tilde{\mathbf{W}}_q\tilde{\mathbf{W}}_k^T = \tilde{\mathbf{W}}_{qk} && (16)
\end{aligned}$$

According Eqn. 11, we obtain that

$$\begin{aligned}
\text{Softmax}[\frac{(\mathbf{X}\mathbf{W}_q)(\mathbf{X}\mathbf{W}_k)^T}{\sqrt{d_k}}] &= \text{Softmax}[\frac{\tilde{\mathbf{W}}_{qk}}{\sqrt{d_k}}] && (17)\\
&= \mathbf{D}^{-1}_{\tilde{\mathbf{W}}_{qk}}\exp(\frac{\tilde{\mathbf{W}}_{qk}}{\sqrt{d_k}}) && (18)
\end{aligned}$$

where $\mathbf{D}_{\tilde{\mathbf{W}}_{qk}}$ is a $n \times n$ diagonal matrix such that

$$\begin{aligned}
(\mathbf{D}_{\tilde{\mathbf{W}}_{qk}})_{ii} &= \sum_{j=1}^{n}\exp(\frac{(\tilde{\mathbf{W}}_{qk})_{ji}}{\sqrt{d_k}}) && (19)\\
&= (\exp(\frac{\tilde{\mathbf{W}}_{qk}}{\sqrt{d_k}})\mathbf{1})_i && (20)
\end{aligned}$$

We can establish the desired result by showing that there always exists a $\tilde{\mathbf{W}}_{qk}$ that satisfies the following fixed point equation:

$$\mathbf{D}^{-1}_{\tilde{\mathbf{W}}_{qk}} \exp(\frac{\tilde{\mathbf{W}}_{qk}}{\sqrt{d_k}}) = \mathbf{P} \tag{21}$$

Given $\mathbf{P}$, to construct such a $\tilde{\mathbf{W}}_{qk}$, we can pick an random positive diagonal matrix $\mathbf{D}_0$:

$$\tilde{\mathbf{W}}_{qk} = \sqrt{d_k} \cdot \log(\mathbf{D}_0\mathbf{P}) \tag{22}$$

Since $\mathbf{P}$ is a positive matrix, and $\mathbf{D}_0$ is a positive diagonal matrix, such a $\tilde{\mathbf{W}}_{qk}$ always exists. According to Eqn. 19 we can conclude

$$\mathbf{D}_{\tilde{\mathbf{W}}_{qk}} = \mathrm{Diag}(\exp(\frac{\tilde{\mathbf{W}}_{qk}}{\sqrt{d_k}})\mathbf{1}) \tag{23}$$

$$= \mathrm{Diag}(\mathbf{D}_0\mathbf{P}\mathbf{1})) = \mathrm{Diag}(\mathbf{D}_0\mathbf{1})) = \mathbf{D}_0 \tag{24}$$

which indicates that $\exp(\frac{\tilde{\mathbf{W}}_{qk}}{\sqrt{d_k}}) = \mathbf{D}_0\mathbf{P} = \mathbf{D}_{\tilde{\mathbf{W}}_{qk}}\mathbf{P}$. This completes the proof of $d \geq n$ case of **Theorem 1**.

Regarding $d < n$ case. Consider the special case of $d = 1$ and $n = 2$. There has $\mathbf{X} \in \mathcal{R}^{2\times 1}$ and $\mathbf{W}_q, \mathbf{W}_k \in \mathcal{R}^{1\times 1}$. Suppose $\mathbf{X} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, so that:

$$\mathrm{Softmax}[\frac{(\mathbf{X}\mathbf{W}_q)(\mathbf{X}\mathbf{W}_k)^T}{\sqrt{d_k}}] = \mathrm{Softmax}[\frac{\begin{bmatrix} 1 \\ 0 \end{bmatrix} \mathbf{W}_q\mathbf{W}_k^T[1, 0]}{\sqrt{d_k}}] \tag{25}$$

$$= \mathrm{Softmax}[\begin{bmatrix} \frac{\mathbf{W}_q\mathbf{W}_k^T}{\sqrt{d_k}} & 0 \\ 0 & 0 \end{bmatrix}] \tag{26}$$

This matrix clearly cannot be used to generate $\mathbf{P}$. Then we extend the above special case to general values of n and d, ($d < n$). Let $\mathbf{X} = [\mathbf{1}_d, \cdots, \mathbf{1}_d, \mathbf{0}_d]^T = [\mathbf{1}_{\mathrm{mat}}, \mathbf{0}_d]^T \in \mathcal{R}^{n\times d}$, where $\mathbf{1}_d, \mathbf{0}_d \in \mathcal{R}^d$ denotes the all ones, zeros column vector, and $\mathbf{1}_{\mathrm{mat}}$ denotes the $d \times (n-1)$ all ones matrix. Then

$$\mathrm{Softmax}[\frac{(\mathbf{X}\mathbf{W}_q)(\mathbf{X}\mathbf{W}_k)^T}{\sqrt{d_k}}] = \mathrm{Softmax}[\frac{[\mathbf{1}_{\mathrm{mat}}, \mathbf{0}_d]^T\mathbf{W}_q\mathbf{W}_k^T[\mathbf{1}_{\mathrm{mat}}, \mathbf{0}_d]}{\sqrt{d_k}}] \tag{27}$$

$$= \mathrm{Softmax}[\begin{bmatrix} \mathbf{1}^T_{\mathrm{mat}}\frac{\mathbf{W}_q\mathbf{W}_k^T}{\sqrt{d_k}}\mathbf{1}_{\mathrm{mat}} & \mathbf{0}_{n-1} \\ \mathbf{0}_{n-1} & 0 \end{bmatrix}] \tag{28}$$

The basic idea of the proof remains consistency, and we can conclude the same conclusion. According to the **Theorem 1**, we can not ensure the attention matrix of joint attention to be full rank.

A.2 PROOF OF KMCTA CAN GUARANTEE FULL RANK

Assumption 1 The elements of $\frac{(\mathbf{X}\mathbf{W}_q)(\mathbf{X}\mathbf{W}_k)^T}{\sqrt{d_k}}$ does not exist negative infinity in normal cases, which means that 0 will not appear at any other position in the attention matrix $\tilde{\mathbf{A}} = \mathrm{Softmax}[\frac{(\mathbf{X}\mathbf{W}_q)(\mathbf{X}\mathbf{W}_k)^T}{\sqrt{d_k}} + \tilde{\mathbf{M}}]$ except for the masked position.

The softmax operation can ensure that the elements at all other positions are greater than 0 except for the masked position. The attention matrix $\tilde{\mathbf{A}}$ of KMCTA is always a lower triangular matrix due to the presence of the $\tilde{\mathbf{M}}$ and the diagonal elements are always greater than 0, i.e., $\forall a_{ii} > 0$:

$$\tilde{\mathbf{A}} = \begin{bmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ a_{21} & a_{22} & 0 & \cdots & 0 \\ a_{31} & a_{32} & a_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{bmatrix} \tag{29}$$

The determinant of the $\tilde{\mathbf{A}}$, $|\tilde{\mathbf{A}}| = a_{11}a_{22}\cdots a_{nn} > 0 \neq 0$, because $\forall a_{ii} > 0$. Thus, the attention matrix $\tilde{\mathbf{A}}$ of KMCTA is always reversible, i.e. always be full rank.

A.3 PROOF OF KMTA CANNOT GUARANTEE FULL RANK

According to the **Assumption 1**. We consider a special case, when T=2 and L=2, the attention matrix $\mathbf{A} = \mathrm{Softmax}[\frac{(\mathbf{X}\mathbf{W}_q)(\mathbf{X}\mathbf{W}_k)^T}{\sqrt{d_k}} + \mathbf{M}]$ of KMTA can be formulated as:

$$\begin{bmatrix} a_{11} & 0 & a_{13} & a_{14} \\ 0 & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & 0 \\ a_{41} & a_{42} & 0 & a_{44} \end{bmatrix} \rightarrow \begin{bmatrix} a_{11} & 0 & a_{13} & a_{14} \\ 0 & a_{22} & a_{23} & a_{24} \\ 0 & 0 & a_{33} - \frac{a_{13}}{a_{11}}a_{31} - \frac{a_{23}}{a_{22}}a_{32} & -\frac{a_{14}}{a_{11}}a_{31} - \frac{a_{24}}{a_{22}}a_{32} \\ 0 & 0 & -\frac{a_{13}}{a_{11}}a_{41} - \frac{a_{23}}{a_{22}}a_{42} & a_{44} - \frac{a_{14}}{a_{11}}a_{41} - \frac{a_{24}}{a_{22}}a_{42} \end{bmatrix} \tag{30}$$

where $a_{11} + a_{13} + a_{14} = 1, a_{22} + a_{23} + a_{24} = 1, a_{31} + a_{32} + a_{33} = 1, a_{41} + a_{42} + a_{44} = 1$. When $a_{33} - \frac{a_{13}}{a_{11}}a_{31} - \frac{a_{23}}{a_{22}}a_{32} = -\frac{a_{13}}{a_{11}}a_{41} - \frac{a_{23}}{a_{22}}a_{42}$, $\mathbf{A}$ can not always be full rank. For example:

$$\mathbf{A} = \begin{bmatrix} 0.4 & 0 & 0.4 & 0.2 \\ 0 & 0.4 & 0.4 & 0.2 \\ 0.4 & 0.5 & 0.1 & 0 \\ 0.4 & 0.4 & 0 & 0.2 \end{bmatrix} \tag{31}$$

For a video, $n = T \times L$, If we fix the number of attention heads, the dimension of tokens, the patch size, and the resolution of the input image, the only factor that can affect n for a video is the frame length. When we pray for a longer frame length and an increase in the number of tokens to improve performance, we may face the problems pointed out in **Theorem 1**. Because increasing the frame length will further increase n, the expressive power of self attention may encounter bottlenecks. The low-rank bottleneck is vital, because it may lead to that many rows of the attention map are seriously homogenized. As the output of self-attention is the weighted sum of the same set of value vectors, the homogenization of attention weights inevitably leads to the resemblance among the aggregated features.

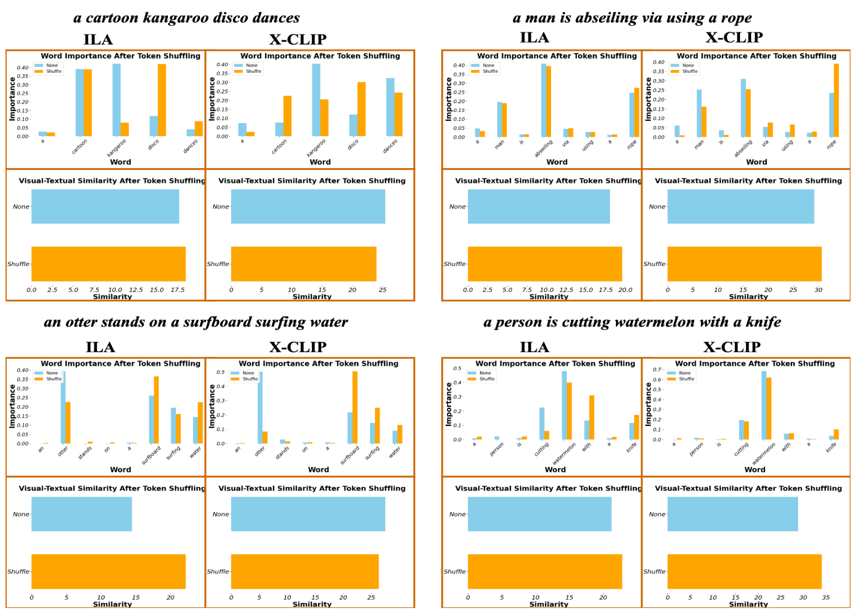

Figure 8: We visualize the impact of token shuffling on the ILA, XCLIP. In most cases, token shuffling has limited influence on their word importance. Surprisingly, in some figures, after token shuffling, the similarity actually increases, which is illogical. This indicates that the response of ILA and XCLIP to spatiotemporal position destruction is uncertain and has poor interpretability.

B SPATIOTEMPORAL HOMOGENIZATION STUDY

Fig. 6 shows that equipping learnable position/time encoding with joint attention is not insufficient to alleviate spatiotemporal homogenization. For example, (Upper Left) 'a cartoon kangaroo disco dances'. We can observe that token shuffling has a limited impact on word importance and similarity of joint attention, but it has a significant impact on the word importance of KMT and both similarity and word importance of KMCT. (Upper Right) 'a man is abseiling via using a rope'. (Lower Left) 'an otter stands on a surfboard surfing water'. (Lower Right) 'a person is cutting watermelon with a knife'. The key point is that classification accuracy does not necessarily indicate that these learnable position encoding or attention biases genuinely grasp the concept of spatiotemporal structure. Since the optimization goal of the network is usually to pursue performance, the learned positional encoding or attention bias serve to improve accuracy rather than understand the spatiotemporal concept. This is because their optimization is guided by gradients aimed at increasing accuracy, rather than by an objective function that tells them what spatiotemporal structure is. The Kronecker mask in KMTA and KMCTA pursues not only for good performance, but also for interpretability. The Kronecker mask act as a natural spatiotemporal heterogeneity inductive bias, KMTA/KMCTA exhibit better spatiotemporal complementarity with spatial attention. Additionally, in Fig. 8, we show that several tailored modules (ILA and XCLIP) also encounter spatiotemporal homogenization issues during our testing. Surprisingly, in some cases like 'a man is abseiling via using a rope,' 'an otter stands on a surfboard surfing water,' and 'a person is cutting watermelon with a knife,' the similarity actually increases after token shuffling, which is illogical.

Additionally, we evaluate the performance of joint attention, Kronecker mask temporal attention (KMTA) and Kronecker mask causal temporal attention (KMCTA) on K400 Kay et al. (2017) after introducing token shuffling. The results are shown in Tab. 9. The greater performance degradation observed in KMTA and KMCTA indicates that they are more sensitive to disturbances in the spatiotemporal structure, thus alleviating spatiotemporal homogenization.

Table 9: Performance of joint attention, KMTA and KMCTA while introducing token shuffling.

Token Shuffling (Yes-Y, No-N)	Top-1 (%)
Joint Attention, N	80.1
Kronecker Mask Temporal Attention, N	81.5
Kronecker Mask Causal Temporal Attention, N	81.4
Joint Attention, Y	58.9 (-21.2)
Kronecker Mask Temporal Attention, Y	43.5 (-38.0)
Kronecker Mask Causal Temporal Attention, Y	42.6 (-38.8)

C Additional Experiments

The effects of interpretive prompt. In Tab. 10, we can observe that, directly use verbs and phrases achieve the lowest performance. When employing prefix and suffix prompt templates, the performance slightly improved. When introducing action decomposition interpretive prompts, the performance significantly improved. When adding synonym conversion interpretive prompts, the performance is further increased. Last, we involve body parts interpretive prompts that the performance reaches the best.

Which branches should be finetuned? We separately freeze the parameters of the pretrained image and text encoder. From Tab. 11, we conclude the following observations: 1) Freezing both image and text encoders, only tuning KMT transformer achieves the worst performance. 2). Finetuning only one of them will improve the performance. 3). Finetuning them simultaneously achieves the best performance.

Table 10: Ablation on interpretive prompt.

Text Prompts	Top-1 (%)	Top-5 (%)
Noun and Phrase of Action	77.4	93.5
+ Prefix-Suffix	77.4	93.6
+ Action Decomposition Interpretive Prompt	78.0	94.0
+ Synonym Conversion Interpretive Prompt	78.2	94.2
+ Involving Body Parts Interpretive Prompt	**78.4**	**94.3**

Table 11: Which branches should be finetuned.

KMT transformer	Visual	Text	Top-1 (%)	Top-5 (%)
✓	×	×	78.7	94.4
✓	✓	×	81.0	95.1
✓	×	✓	80.8	95.2
✓	✓	✓	**81.5**	**95.5**

Performance on Something-Something-V2. We evaluate CLAVER on Something-Something-V2 (SSv2) benchmark. We employ CLAVER-B/16_{8f} (KMT) as the backbone and introduce the video-specific prompting module (2 layers) from X-CLIP Ni et al. (2022), since the SSv2 is a dataset with relatively small semantics. For SSv2, we uniformly sample the entire video at predefined temporal intervals without group division. We adopt the same hyperparameter settings as experiments on Kinetics-400 (Fully-supervised). As shown in Tab. 12, CLAVER-B/16_{8f} outperforms ILA-B/16_{8f} Tu et al. (2023), EVL-B/16_{16f} Lin et al. (2022), and X-CLIP/16_{8f} Ni et al. (2022). However, the performance of our method is lower than AIM Yang et al. (2023), but we require fewer FLOPs.

Generalization of Kronecker mask to different backbones. We experiment with a CLIP-based model (ActionCLIP-B/16_{8f}) as well as an ImageNet-based architecture (TimeSformer-ViT-B/16_{8f}) to demonstrate the generality of Kronecker mask temporal attention (KMTA). For ActionCLIP, we insert our Kronecker mask into the ActionCLIP while keep others unchanged. For TimeSformer, we replace the pipeline temporal attention with the our Kronecker mask temporal attention (KMTA). The results are depicted in Tab. 13. The utilization of KMTA results in a performance gain for ActionCLIP and TimeSformer, respectively, demonstrating KMTA can be plugged into various transformer-based backbones, seamlessly.

Table 12: Performance comparison with CLIP-based methods on Something-Something-V2.

Model	Frames	Top-1 (%)	Top-5 (%)	View	GFLOPs
X-CLIP-B/16 Ni et al. (2022)	8	57.8	84.5	4×3	145
AIM-ViT-B/16 Yang et al. (2023)	8	66.4	90.5	1×3	624
EVL-ViT-B/16 Lin et al. (2022)	16	61.7	-	1×3	345
ILA-B/16 Tu et al. (2023)	8	65.0	89.2	4×3	214
CLAVER-B/16	8	65.4	90.1	4×3	204

Table 13: Generalization ability of Kronecker mask temporal attention on various backbones for Kinetics-400.

Model	Pretraining	Top-1 (%)	Top-5 (%)
ActionCLIP Wang et al. (2021)	CLIP-400M	81.1	95.5
ActionCLIP+KMTA	CLIP-400M	**83.0**	**95.9**
TimeSformer Bertasius et al. (2021)	IN-21k	78.0	93.7
TimeSformer + KMTA	IN-21k	**80.2**	**94.9**

D EXPERIMENTAL SETTING DETAILS

Architectures. CLAVER-B/32 adopts ViT-B/32 (L_V=12, N_h=12, d=768, p=32) and is equipped with a KMT/KMCT transformer (L_K=2, N_h=12, d=768, p=32). CLAVER-B/16 employs ViT-B/16 (L_V=12, N_h=12, d=768, p=16), along with the KMT/KMCT transformer (L_K=2, N_h=12, d=768, p=32). CLAVER-L/14 is equipped with ViT-L/14 (L_V=24, N_h=16, d=1024, p=14) and the KMT/KMCT transformer (L_K=4, N_h=16, d=1024, p=14). Here L_V denotes the layers of ViT, L_K denotes the layers of Kronecker mask temporal transformer, N_h refers to the number of attention heads, d represents the embedding dimension and p is the patch size.

Hyperparameters. The experiments are conducted on 8 NVIDIA 80G A100 GPUs. We present the training hyperparameters in Tab. 14. Additionally, the learning rate for updating the KMT/KMCT transformer (randomly initialized) parameters is set $10\times$ higher than the learning rate for parameters of the text encoder or image encoder. Because the text/image encoder already possesses a ability to extract high-quality text/image representations, conversely, KMT/KMCT transformer is trained from scratch. In the experiment, we freeze the parameters of the CLIP's image encoder to reduce certain computational costs, as it already has strong image feature extraction capabilities. And then we also conducted experiments on this aspect in subsequent ablation study.

Table 14: The training hyperparameters settings of experiments.

Config	Fully-sup	Few-shot	Zero-shot
Optimizer	AdamW		
Base learning rate	12e-6	2e-6	12e-6
Minimal learning rate	12e-8	2e-8	12e-8
Weight decay	0.001		
Optimizer betas	β_1, β_2 =0.9, 0.98		
Batch size	128 (ViT-B) 32 (ViT-L)		
Learning rate schedule	Cosine decay		
Warmup epochs	5		
Training epochs	60 (ViT-B) 40 (ViT-L)	80 (20 on K400)	0 (20 on K400)
Augmentation	RandomFlip, MultiScaleCrop, ColorJitter GrayScale, Label smoothing, Mixup, Cutmix		

Fully-supervised experiments setting. We conduct the fully-supervised experiments on Kinetics-400&600. During training, a sparse sampling strategy is used. The number of frames is set to 8 or 16. We spatially scale the shorter side of each frame to 256 and take a 224 crop center crop. Following, we adopt the multi-view inference with 3 spatial crops and 4 temporal clips.

Few-shot experiments setting. We randomly sample 2, 4, 8 and 16 videos from each class on UCF-101 and HMDB-51 constructing the training set. For evaluation, we use the first split of the test set on UCF-101 and HMDB-51.

Zero-shot experiments setting. We train CLAVER-B/16 with 32 frames on Kinetics-400. The same as, we apply the following two evaluation protocols in zero-shot experiments. 1) Evaluation for HMDB-51 and UCF-101. Following, the prediction is conducted on the three splits of the test data, and we report the average top-1 accuracy and standard deviation. 2) Evaluation for Kinetics-600. Following, the 220 new categories outside Kinetics-400 in Kinetics-600 are used for evaluation. The evaluation is conducted three times. For each iteration, we randomly sampled 160 categories for evaluation from the 220 categories in Kinetics-600.

E LIMITATIONS

The drawbacks of KMTA and KMCTA is that they have the same higher computational complexity as joint attention in self-attention calculation, $\mathcal{O}(T^2L^2D)$, which is also a problem with Kronecker mask attention. Although several spatiotemporal attention forms can be derived by joint attention and tailored Kronecker mask, they do not use this form for practical calculations. Since their specific Kronecker masks allow them to be converted to forms with lower computational complexity, as shown in Fig. 9. For example, the feature shape of spatial attention is (T, L, D), the computation

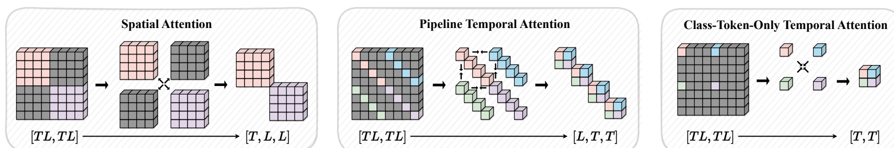

Figure 9: Convert Kronecker mask attentions to lower computational complexity forms.

complexity is $\mathcal{O}(TL^2D)$. The feature shape of pipeline temporal attention is (L, T, D), the computation complexity is $\mathcal{O}(LT^2D)$, and the feature shape of class-token-only temporal attention is (T, D), the computation complexity is $\mathcal{O}(T^2D)$. For solutions to reduce computational complexity, the survey *Efficient transformers : A survey* Tay et al. (2022) summarizes many related methods. Some methods reduce quadratic complexity to lower complexity, some use sparse attention, and some use window attention to reduce sequence length. As illustrate in Tab. 15 We attempt some strategies to reduce the computational complexity, indicating that using some existing efficient transformer algorithms can significantly reduce FLOPs while still maintaining superior performance.

Table 15: Lower computation complexity of CLAVER.

Method	CLAVER-B/16$_{8f}$				CLAVER-B/14$_{8f}$			
	KMTA		KMCTA		KMTA		KMCTA	
	Top-1 (%)	FLOPs (G)	Top-1 (%)	FLOPs (G)	Top-1 (%)	FLOPs (G)	Top-1 (%)	FLOPs (G)
Shifted Sliding Window	83.2	146	83.5	146	87.3	502	87.3	502
Linformer Wang et al. (2020)	83.4	139	83.3	139	87.5	494	87.5	494
Reformer Kitaev et al. (2020)	83.5	102	83.7	102	87.7	415	87.7	415
Transformer-LS Zhu et al. (2021)	84.1	97	84.3	97	87.6	330	87.7	330

Furthermore, considering that KMTA and KMCTA are derived from tailored Kronecker masks, a good strategy to reduce the computational complexity of KMTA and KMCTA is worth exploring in the future.

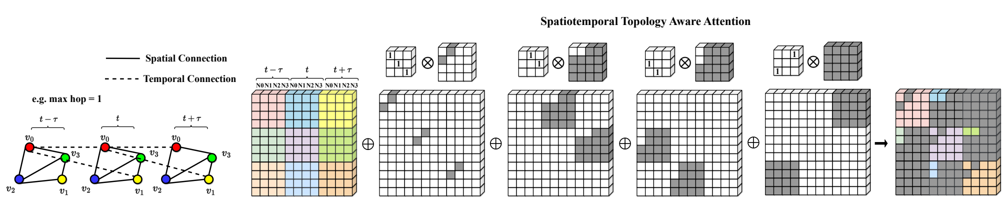

Figure 10: Spatiotemporal topology attention.

F BROADER IMPACT

KMCTA has potential to some extent which may be reflected in generation tasks (video generation). Due to the fact that causality is commonly required in the decoder-only language model for autoregressive modeling, and the spatiotemporal heterogeneity of KMCTA may be beneficial for the rationality of video generation. It can alleviate the low-rank bottleneck, simultaneously. Furthermore, Kronecker mask has certain generality, such as for spatiotemporal graph data, we can design topology structure by tailored Kronecker mask. Fig. 10 shows an example.

G Interpretive Prompt Details

First, we input command prompts into the ChatGPT OpenAI (2024) chatbot API to obtain a demonstration case of a small amount (in our experiment, we set it to 3 here) of detailed descriptions of actions.

Question to ChatGPT:

```
Below, I will provide you with some action nouns. Please provide a simple
and detailed description (explanation) about the action decomposition
of these action nouns. Please note that the sentence length of the
description should not exceed 76 words.
```

Answered by ChatGPT:

```
Certainly! Please go ahead and provide the action nouns you'd like
descriptions for, and I'll provide concise explanations within the
specified sentence length.
```

Question to ChatGPT:

```
abseiling.
```

Answered by ChatGPT:

```
Abseiling combines several actions to descend a vertical surface with a
rope. Climbers secure themselves with a harness and utilize a descender
device for controlled descent. Simple actions, like maintaining a
straight body position and regulating rope tension, form the basis.
Abseiling demands proper training, safety measures, and is popular in
adventure sports and rescue operations, allowing individuals to
experience controlled descent in various settings.
```

Question to ChatGPT:

```
air drumming.
```

Answered by ChatGPT:

```
Air drumming is a rhythmic expression where individuals simulate playing
drums without physical instruments. Simple actions, like mimicking
drumming motions in the air, combine to create this imaginative and
playful activity. Enthusiasts use their hands and feet to imitate
drumming patterns, syncing with music. It's a spontaneous, enjoyable
gesture often done during music listening or live performances,
showcasing one's connection to the rhythm without the need for actual
drums or drumsticks.
```

Question to ChatGPT:

```
answering questions.
```

Answered by ChatGPT:

```
Answering questions involves providing responses to queries posed by
others. Simple actions like active listening, comprehension, and concise
articulation combine in this communicative process. It is fundamental
in various contexts, facilitating information exchange and problem-
solving. Respondents draw on their knowledge and expertise to address
inquiries, contributing to effective communication and fostering
understanding between individuals or groups.
```

Question to ChatGPT:

```
Below, I will provide you with some action nouns or phrases. Please
provide many words and phrases that share the same central concept
as these action nouns or phrases but have more diverse expressions.
Please note that your reply should not exceed 76 words in length.
```

Answered by ChatGPT:

```
1. Running: Jogging, sprinting, dashing, hurrying, racing, pounding the
    pavement, speeding, trotting.
2. Eating: Dining, feasting, devouring, munching, nibbling, gobbling up,
    indulging, savoring.
3. Writing: Scribbling, jotting down, composing, drafting, penning,
    inscribing, scrawling, scripting.
4. Laughing: Chuckling, giggling, guffawing, snickering, roaring with
    laughter, belly laughing, cackling, chortling.
5. Reading: perusing, browsing, scanning, skimming, devouring books,
    absorbing, immersing oneself, delving into texts.
```

Question to ChatGPT:

```
Cutting in the kitchen.
```

Answered by ChatGPT:

```
Slicing, dicing, chopping, mincing, cleaving, carving, trimming,
preparing ingredients.
```

Question to ChatGPT:

```
driving car.
```

Answered by ChatGPT:

```
Operating a vehicle, maneuvering behind the wheel, navigating the road,
piloting an automobile, steering, cruising, commuting by car, motoring.
```

Question to ChatGPT:

```
Walking With Dog.
```

Answered by ChatGPT:

```
Strolling with a canine companion, ambling with a pet, promenading with a
pup, hiking with a furry friend, sauntering alongside a dog, wandering
with a four-legged buddy, leash-walking, trotting with a pooch.
```

Question to ChatGPT:

```
Below, I will provide you with some action nouns. Please describe
these actions based on their nouns and possible body parts involved.
Please note that the sentence length of your response should not exceed
76 words.
```

Answered by ChatGPT:

```
1. Running: The rhythmic motion of legs propelling the body forward, feet
     striking the ground with each stride, muscles contracting and
    relaxing to generate momentum.
2. Swimming: Arms slicing through water, legs kicking to maintain
    propulsion, torso rotating to streamline movement, lungs rhythmically
     inhaling and exhaling for oxygen.
```

```
3. Writing: Fingers gripping a pen or pencil, hand moving across paper or
     keyboard, muscles in the forearm and wrist coordinating precise
    movements to form letters and words.
4. Jumping: Legs bending at the knees, feet pushing off the ground, arms
    swinging for momentum, body soaring through the air momentarily
    before landing.
5. Dancing: Feet tapping, legs extending and bending, hips swaying, arms
    gesturing and flowing in sync with music, body moving rhythmically
    with expression and grace.
```

Question to ChatGPT:

```
Cutting in the kitchen.
```

Answered by ChatGPT:

```
Using a sharp knife, fingers gripping the handle, hand guiding the blade
through ingredients on a cutting board, wrist controlling the motion,
fingers curling slightly to hold the food steady, precision applied
to achieve desired shapes or sizes, ensuring safety and efficiency
during food preparation.
```

Question to ChatGPT:

```
driving car.
```

Answered by ChatGPT:

```
Gripping the steering wheel, hands adjusting position, fingers pressing
pedals for acceleration and braking, eyes scanning surroundings for
obstacles, feet coordinating between clutch, brake, and accelerator,
body positioned comfortably in the driver’s seat, mind focused on
    navigation
and traffic signals, reacting swiftly to changing road conditions.
```

Question to ChatGPT:

```
Walking With Dog.
```

Answered by ChatGPT:

```
Leash in hand, fingers securing grip, arm relaxed as it swings alongside
the body, legs moving in tandem with the dog’s pace, feet stepping
forward with purpose, eyes attentive to the dog’s behavior and
    surroundings,
occasional stops for sniffing or marking, a bond of companionship
evident in synchronized movement.
```

Subsequently, as illustrated in Fig 4, we input the examples generated by ChatGPT OpenAI (2024) into LLaMA-3 Abhimanyu Dubey et al. (2024) multiple times in the format of ***Command*** + ***Examples*** + ***Action Concept***, automatically generating diverse text descriptions about the ***Action Concept*** through the program. When dealing with datasets containing a large number of categories, this scheme can greatly save labor costs. In addition, it should be noted that LLaMA-3 offers controllable parameters to control the randomness (τ) and diversity (p) of output texts. We set τ and p to 0.90 and 0.95, respectively to ensure that there are significant variations in output content for the same input, while maintaining consistency in the central concept.

H Analysis of Synthetic Video Testing

We use the original prompts of synthetic videos (several action-related examples generated by Imagen Saharia et al. (2022) and Sora Brooks et al. (2024)) as their corresponding text descriptions, and show Top-5 most relevant texts in Fig. 11. The results shows the robustness and generalization of CLAVER.

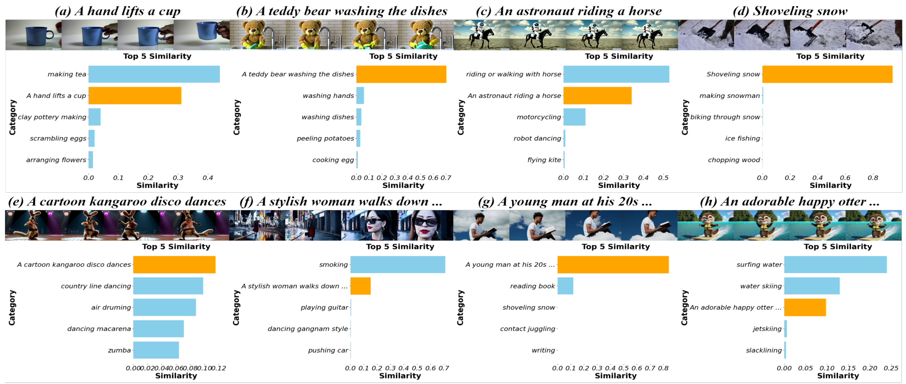

Figure 11: Testing on synthetic videos from Imagen (**Upper**) and Sora (**Lower**). The bar chart represents the Top-5 similarity texts, orange bar denotes the original prompts of synthetic videos.

For cases generated by Imagen, in Fig. 11 **Upper**, we denote the Kinetics-400 action text set as $\mathcal{A}$. The action category "lift a cup" in Fig. 11 (a) does not appear in $\mathcal{A}$, while, (b) "washing dishes", (c) "riding horse", (d) "shoveling snow" do appear in $\mathcal{A}$. All videos do not come from the real-world, which are synthetic. We observe that all the original prompts appear in the Top-5 similarity (more precisely, Top-2). Specifically, for (a), "making tea" has the highest similarity, possibly due to the presence of cups and human hands in the scene. In addition, the original text has the second-highest similarity. In (b), apart from the original prompt in Top-1, other texts such as "washing dishes" convey the same meaning as the original prompts but lack a description of the subject ("teddy") of the action. "Washing hands" and "peeling potatoes" may involve actions and scenes like hand movements and sinks, thus exhibiting high similarity. For (c), although the one with the highest similarity is not "An astronaut riding a horse" , "riding or walking horses" is the exactly describes the action in the video, with the only difference being the lack of the subject of the action. For (d), "Shoveling snow" belongs to $\mathcal{A}$, and since the original prompt of the video is "Shoveling snow", its similarity is very high, exceeding 90%.

For the cases generated by Sora, as depicted in Fig. 11 **Lower**, (e) "disco dances" concept is not in $\mathcal{A}$ but has the highest similarity. Besides, other action concepts in the Top-5 ("country line dancing","air druming", "dancing macarena","zumba") are all dance-related. (f) is a relatively challenging sample here. Although the ground truth description is located in the second, the "smoking" in Top-1 is entirely unrelated to the content of the video, because the original prompt of the video primarily describes the scene and the action-related text is relatively short. In addition, the shot is a process from far to near, which may lead the model to arrive at an unreasonable Top-1. In (g), the Top-2 descriptions are original prompt and "reading book", both strongly related to the action in video. For (h), the original prompt only rank third, however, Top-1 "surface water" and the second "water skimming" are descriptions strongly related to the action in video.

I Dataset Details

Kinetics-400&600. The Kinetics dataset consists of 10-second video clips collected from YouTube. Specifically, Kinetics-400 Kay et al. (2017) consists of approximately 240k training videos and 20k validation videos with 400 categories, while Kinetics-600 Carreira et al. (2018) is an extension of Kinetics-400, consisting of approximately 410k training videos and 29k validation videos with 600 categories.

UCF-101. UCF-101 Soomro et al. (2012) consists of 101 action categories, over 13k clips and 27 hours of video data. The database comprises realistic user uploaded videos containing camera motion and cluttered backgrounds. The training and test data are divided into three splits.

HMDB-51. HMDB-51 Kuehne et al. (2011) is a collection of realistic videos from various sources, including movies and web videos. It is composed of 6,766 video clips from 51 action categories, with each category containing at least 101 clips. The dataset is divided into three splits for training and test data. In each split, each action class has 70 clips for training and 30 clips for testing.

Something-Something-V2. Something-Something-V2 Materzynska et al. (2020) is a collection of 220,847 labeled video clips of humans performing pre-defined, basic actions with everyday objects. The dataset consists of 174 action categories. For each video in the training and validation sets there is an object annotation in addition to the video label, if applicable. For example, for a label like ”Putting [something] onto [something],” there is also an annotated version, such as ”Putting a cup onto a table.” In total, there are 318,572 annotations involving 30,408 unique objects.

Synthetic Videos. We select some action-related videos from the demo cases generated by Imagen Saharia et al. (2022) and Sora Brooks et al. (2024), and employ the original prompts and their corresponding synthetic videos as a pair of test samples.

J More Attention Heat Map Visualization

We visualize several spatiotemporal attention map of samples from Kinetics400, Imagen, Sora.

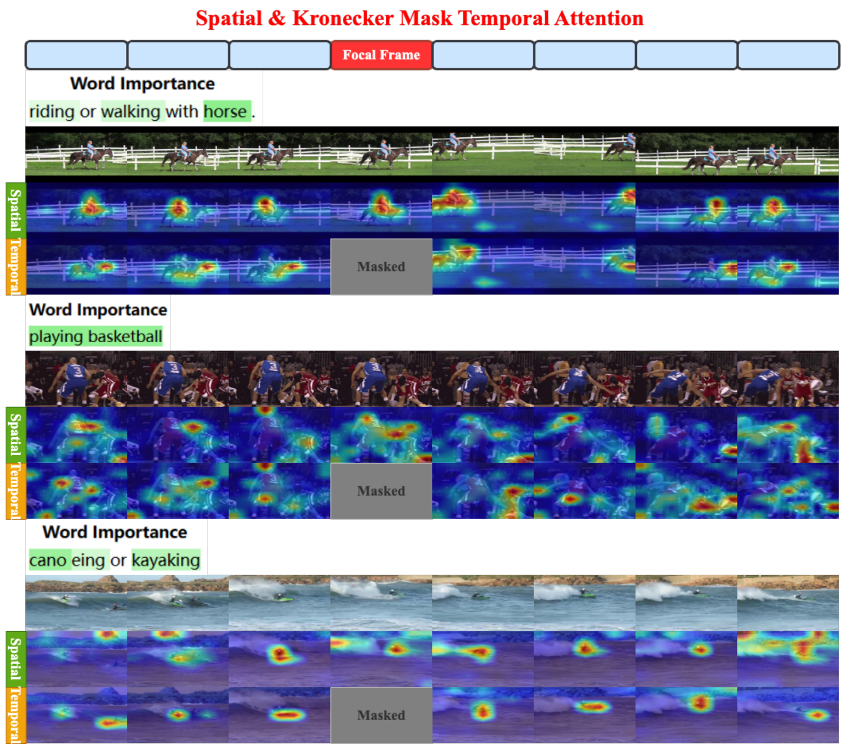

Figure 12: Attention map of videos from Kinetics400.

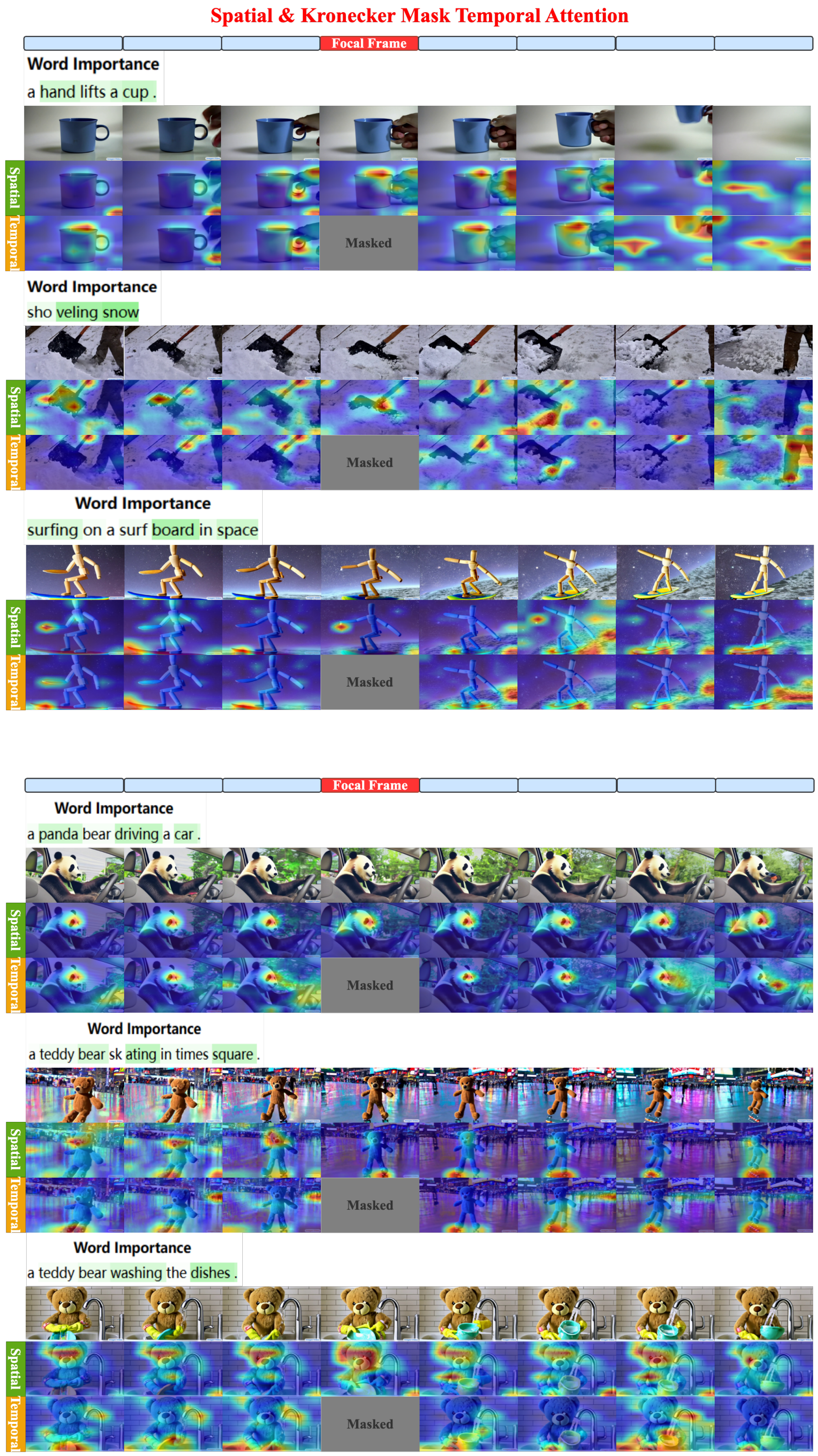

Figure 13: Attention map of synthetic videos from Imagen.

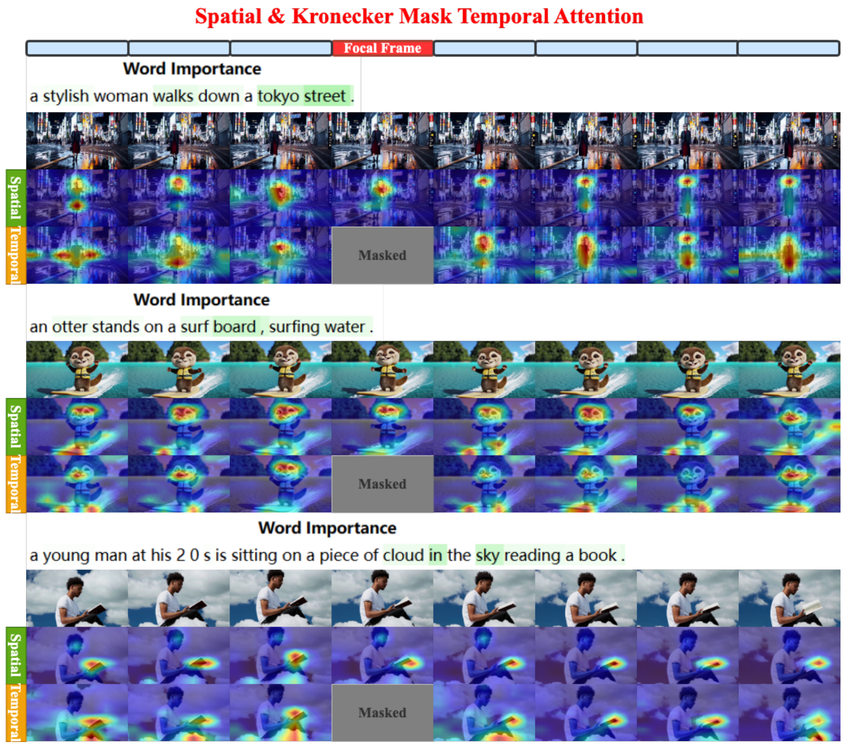

Figure 14: Attention map of synthetic videos from Sora.

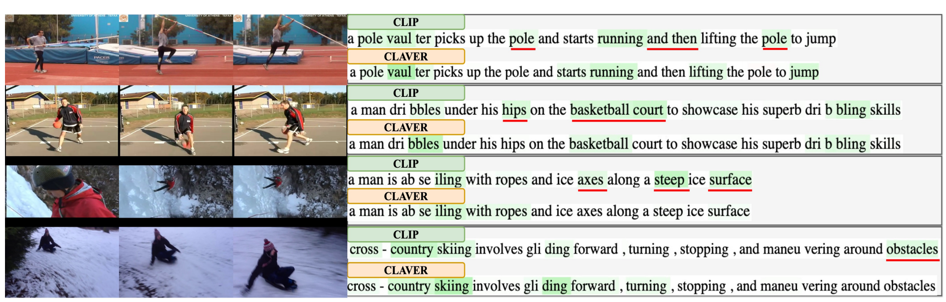

Figure 15: Visualization examples of transition of CLIP's attention on nouns to CLAVER's preference on verbs.