Enhanced Motion-Text Alignment for Image-to-Video Transfer Learning

Wei Zhang[1,2*], Chaoqun Wan[2], Tongliang Liu[3], Xinmei Tian[1,4†], Xu Shen[2†], Jieping Ye[2]
[1]University of Science and Technology of China, [2]Alibaba Cloud, [3]The University of Sydney
[4]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
zw0819@mail.ustc.edu.cn, tongliang.liu@sydney.edu.au, xinmei@ustc.edu.cn
{qionglong.wcq, shenxu.sx, jieping.yjp}@alibaba-inc.com

Abstract

*Extending large image-text pre-trained models (e.g., CLIP) for video understanding has made significant advancements. To enable the capability of CLIP to perceive dynamic information in videos, existing works are dedicated to equipping the visual encoder with various temporal modules. However, these methods exhibit "asymmetry" between the visual and textual sides, with neither temporal descriptions in input texts nor temporal modules in text encoder. This limitation hinders the potential of language supervision emphasized in CLIP, and restricts the learning of temporal features, as the text encoder has demonstrated limited proficiency in motion understanding. To address this issue, we propose leveraging "**MoT**ion-**E**nhanced **D**escriptions" (**MoTED**) to facilitate the extraction of distinctive temporal features in videos. Specifically, we first generate discriminative motion-related descriptions via querying GPT-4 to compare easy-confusing action categories. Then, we incorporate both the visual and textual encoders with additional perception modules to process the video frames and generated descriptions, respectively. Finally, we adopt a contrastive loss to align the visual and textual motion features. Extensive experiments on five benchmarks show that MoTED surpasses state-of-the-art methods with convincing gaps, laying a solid foundation for empowering CLIP with strong temporal modeling.*

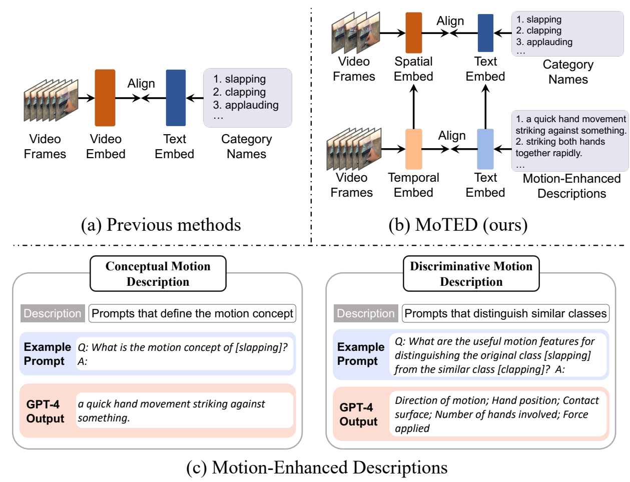

Figure 1. (a) Previous methods [27, 49, 75, 78] align the visual embeddings with the textual embeddings derived from category names. (b) Our work symmetrically aligns the spatial/temporal visual embeddings with class descriptions and motion descriptions correspondingly. (c) The descriptions generated by large language modes (*e.g.*, GPT-4 [50]) enhance the conceptual definitions and discriminative details of motions.

1. Introduction

Recent years have witnessed remarkable achievements in contrastive language-image pre-training models [9, 25, 36, 60, 87, 88], with CLIP [60] emerging as the front-runner. Through language supervision with a vast collection of 400 million image-text pairs, CLIP has achieved exceptional image comprehension and unprecedented zero-shot generalization. This breakthrough has opened up new possibilities for leveraging the power of large-scale pre-trained models to comprehend videos. It has also introduced a new paradigm [26, 40, 49, 52, 58, 75, 80] that endows image-based CLIP to effectively perceive dynamic information for video recognition.

To equip CLIP with motion perception, existing works propose to incorporate various temporal modules into the visual encoder. These temporal modules are additional tunable parameters, including designed efficient units between pre-trained transformer blocks [49, 75], and parallel structures to disentangle the temporal modeling out of the spatial modeling [40, 58]. However, while considerable effort has been put into capturing temporal visual features from videos, these methods have paid little attention to the in-

*This work was done when the author was visiting Alibaba as a research intern.
†Corresponding author.

put text and text encoder, leading to an imbalance between the visual and textual alignment. When transferring to the classification task of video recognition (Fig. 1(a)), the only available text is the "category names" of the actions, *e.g.*, "clapping" and "slapping". These coarse-grained word-level descriptors lack clear descriptions and explanations, making it difficult to distinguish between them. For instance, "slapping" refers to a quick hand movement striking against something, while "clapping" refers to the act of striking both hands together rapidly. The scarcity of textual motion information contradicts the language supervision principle of CLIP. Furthermore, it's observed that the CLIP text encoder delivers a strong bias towards spatial concepts (*e.g.*, nouns) [5, 30, 47, 63], with a weak understanding of temporal cues (*e.g.*, verbs). These issues significantly limit the effectiveness of extending CLIP with motion modeling for video understanding.

To overcome the limitations of language supervision on temporal modeling, we propose a novel approach that symmetrically aligns the spatial/temporal visual embeddings with class descriptions ("slapping") and motion descriptions correspondingly ("a quick hand movement striking against something"). We introduce the **MoT**ion-**E**nhanced **D**escriptions (**MoTED**) as the language supervision for the visual temporal modules. In the implementation, we encounter two main challenges: i) *how to generate motion-related descriptions*? ii) *how to effectively utilize descriptions as supervision*? To tackle the first challenge, we aim for clear, detailed, and distinctive descriptions in the text. With the recent advancements in large language models (LLMs) [4, 11, 50], it has become possible to automatically generate conceptual descriptions for corresponding actions. As for the second challenge, we disentangle the spatial and temporal learning by constructing a temporal encoder, in parallel to the CLIP image encoder, to extract the temporal features aligned with motion-enhanced descriptions.

To be specific, we employ GPT-4 [50] to generate descriptions by posing a query: "Q: what is the motion concept of `<CLASS>`? A: " (as shown in Fig. 1(c)). However, our empirical findings indicate that only the concept descriptions offer marginal assistance in motion perception, particularly on datasets such as Something-Something-V2 (SSv2) [20]) that necessitate the differentiation of similar classes using intricate motion cues. To mitigate this issue, we propose to incorporate related top-k actions as context for the LLMs to generate more discriminative descriptions. These generated texts are then passed through the CLIP text encoder to extract textual motion representations, where an adapter is utilized to eliminate the biases in the original model. Correspondingly, the temporal module in the visual encoder processes the video to extract dynamic features and aligns them with the generated textual descriptions through contrastive learning. Finally, these two sets of features are independently fused with the original CLIP's text and image features using cross-attention, and subsequently aligned. Evaluated on two supervised video recognition benchmarks, *i.e.*, Kinetics-400 [28] and SSv2 [20], as well as three zero-shot benchmarks, *i.e.*, Kinetics-600 [12], HMDB51 [24], UCF101 [65], the proposed MoTED surpasses state-of-the-art methods with convincing gaps, indicating the effectiveness of aligning enhanced motion descriptions with the temporal embedding of input videos.

We summarize the contributions as follows:

- We present a new perspective that underscores the significance of textual side on par with visual side. By delving into in-depth distinguishing descriptions of actions, we make the first attempt to reveal the potential of language supervision as emphasized in CLIP.
- We propose MoTED that leverages LLMs to automatically generate action descriptions, and mining distinctive descriptions among similar actions. Then, we use a parallel path for both visual and textual motion modeling.
- We evaluate our approach on supervised as well as generalization tasks. Extensive experiments demonstrate the superiority and good generalization ability of the proposed method.

2. Related Work

Vision-Language Pre-training. In recent years, Vision-Language Pre-training (VLP) [9, 25, 36, 45, 46, 60, 76, 87, 88] has made remarkable progress. One of the most remarkable and influential works is CLIP [60], which adopts the contrastive language-image pretraining paradigm. Following that, this paradigm has shown impressive zero-shot generalization capabilities on various image-related tasks [31, 35, 37, 43, 82, 90]. However, pre-training a language-video model [34, 81] is prohibitively expensive, as it requires large-scale video-text data and extensive training resources (*e.g.*, thousands of GPU days). Meanwhile, transferring language-image pre-trained models to the video domain [10, 26, 40, 49, 52, 58, 75, 78] has captured significant attention due to its striking performance and training efficiency. For instance, X-CLIP [49] integrates the CLIP image encoder with a cross-frame attention module for temporal modeling. Vita-CLIP [78] utilizes multi-modal prompting techniques to learn video and text-specific context vectors. DiST [58] is the most related work that also disentangles spatial and temporal learning *in the visual side*. In contrast, our MoTED highlights the effectiveness of temporal disentangling both *in the visual and textual sides*, supervised by the detailed descriptions of motions.

Video Recognition. The conventional approaches in video recognition primarily focus on spatio-temporal learning under fully-supervised settings, where all categories are predefined. These approaches have achieved remarkable performance using various architectures, including convolution

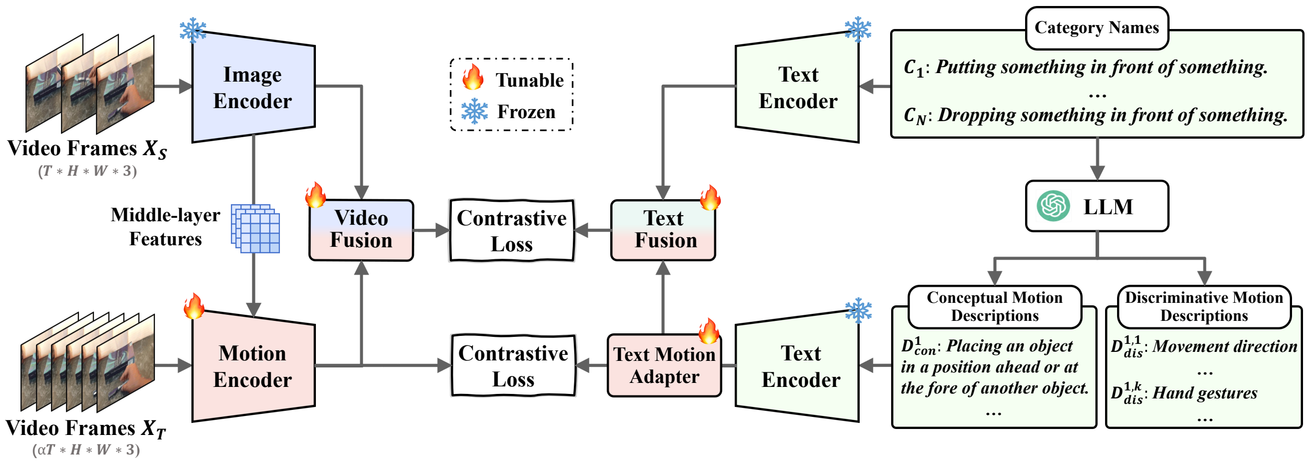

Figure 2. The overall framework of MoTED. Building upon the dual image and text encoders initialized by CLIP [60], we extend the capability to perceive motion information symmetrically in both sides. In the text side, the motion-related conceptual and discriminative descriptions are generated via querying LLMs (*e.g.*, GPT-4) using category names as input. In the vision side, a motion encoder is built to extract temporal dynamics given densely-sampled frames and integrated with the middle-layer features from image encoder. To align the motion embeddings and video embeddings from both sides, it employs two contrastive losses, respectively.

Related work	Task	Action annotation	Example
Dist. Sup. [39]	pretrain	event description	'insert window mounting bolts'
Weakly Sup. [14]	pretrain	event description	'take up the iron clamp'
LAVILA [89]	pretrain	event description	' a lady walks past a car'
VFC [47]	VL transfer learning	event description	'two brown horses eating grass'
MoTED (Ours)	VL transfer learning	motion concept description	'slapping: a quick hand movement striking against something'

Table 1. Comparison of our method with related works focused on action annotations. 'VL' is the abbreviation of 'vision-language'.

Related work	LLMs role	Example
Chatvideo [72]	manager& summarize	**Input**: 'Summarize the activity in video' **Output**: 'A person is cooking in the kitchen'
MM-REACT [86]	execution& summarize	**Input**: 'Please create a summary of the video' **Output**: 'The speaker is making a BLT sandwich ...'
MiniGPT4 [91]/ LLAVA [41]	KB& data clean	**Input**: 'Describe this image in detail.' **Output**: 'The image shows a group of musicians ...'
MoTED (Ours)	generate descriptions	**Input**: 'What is the motion concept of *slapping*' **Output**: 'A quick hand movement striking against ...'

Table 2. Comparison of our method with related works that take LLMs as a knowledge base (KB) and automatically annotated tool.

networks [7, 17, 21, 59, 64, 67–69, 71, 73], and vision transformers [1, 2, 6, 16, 32, 33, 38, 42, 48, 54, 55, 62]. In addition to the architecture design, self-supervised video representation learning [13, 18, 19, 23, 53, 57, 66, 74, 77, 79] has also gained popularity recently. However, these methods operate purely within the visual domain and meet bottlenecks of recognizing unseen or unfamiliar categories in real-world applications. Fortunately, the advance of large language models (LLMs) [4, 51] provides opportunities to mitigate this issue, due to their powerful capabilities of encoding world knowledge [29]. Recent studies [44, 56, 83, 85] have verified that factual sentences generated by LLMs can improve zero-shot image recognition accuracy. VFC [47] leverages PaLM [11] to create verb-focused hard negatives to enhance the understanding of verbs in video models. LSS [61] integrates language-based action concepts with self-supervised learning to adapt an image model to video domain. Our MoTED employs the descriptions that delineate the labeled classes and further accentuate the useful traits distinguishing similar classes.

3. Method

The generic pipeline of MoTED is presented in Fig. 2, which consists of three steps: (1) **Motion Description Generation** to obtain conceptual descriptions $\mathcal{D}_{con}$ and discriminative descriptions $\mathcal{D}_{dis}$, via querying language models given the category names of the target video dataset in Sec. 3.1; (2) **Textual Motion Adaptation** to obtain the text embeddings that extract the motion semantics given the generated descriptions $\mathcal{D}_{con}$ and $\mathcal{D}_{dis}$ in Sec. 3.2; (3) **Visual Motion Extraction** to obtain the visual embeddings that extract the motion semantics given input video frames in Sec. 3.3. Overall, the framework is trained in an end-to-end contrastive manner as illustrated in Sec. 3.4.

3.1. Motion Description Generation

The first step is to obtain a set of appropriate descriptions for each category. Given a video dataset with multiple different categories (*e.g.*, Kinetics400 [28] with 400 action classes), the descriptions are generated automatically utilizing GPT-4 [50]. Note that, the generation process is agnostic to this choice and other LLMs can be used instead. For each category name, we query GPT-4 to provide the motion concepts using the following prompt: " `Q: What is the motion concept in a video of <category name>? A:`". As shown in Fig. 1, the generated concept descriptions often cover moving objects, object interactions, moving directions and speeds, etc. But the output of LLMs can also be open-ended duplicates or anything in natural language (as illustrated in the Appendix). To control the generated descriptions to be concise and motion-related, we adopt a two-shot prompt in which we include two exemplars of question-answer for the same operation that is being queried.

Besides the conceptual descriptions $\mathcal{D}_{con}$, we further investigate into the confusing actions that are characterized by their distinct differences from similar categories. To this end, we first compute the cosine distance of text embeddings for every two categories. Then we select top-k similar classes (*e.g.*, k=5) and query GPT-4 to generate motion characteristics for distinguishing the similar classes using the following prompt: "`Q: What are the useful features for distinguishing the original class <category1 name> from the similar class <category2 name>? A:`". Different from the conceptual descriptions $\mathcal{D}_{con}$ that are mainly determined by the general knowledge of LLMs, the descriptions $\mathcal{D}_{dis}$ further contain task-specific information that the downstream task emphasizes.

In this way, given a dataset of N categories, we can obtain the motion-enhanced $N*(k+1)$ descriptions, consisting of 1 conceptual description and k discriminative descriptions for each category. The above generation process is completed before the training of whole framework.

The detailed comparisons of action annotations and LLM-based generations are presented in Tab. 1 and Tab. 2, respectively. Tab. 1 shows that existing works employ annotations that describe the *action events* in a subject-verb-object manner, which are *object-centric*. In contrast, our generated annotations are *motion-related*, *object-agnostic* and to supervise the learning of visual motion features with the world knowledge of *motion concepts*. In Tab. 2, LLMs are mainly applied in generative tasks, to summarize query results and generate *general responses*, or to obtain *instruction data* to align with *human preference*. In contrast, we apply LLMs to obtain descriptions of *motion concept*, aligned with the temporal visual features.

3.2. Textual Motion Adaptation

After generating the motion-enhanced descriptions, the second step is to perceive the motion cues within the descriptions and aggregate them into a compact motion-enhanced embedding. Given a set of motion-enhanced descriptions $\mathcal{D}$ = $\left\{\mathcal{D}^i_{con}, \mathcal{D}^{i,j}_{dis}\right\}$ where $i \in [1, N]$ and $j \in [1, k]$, we extract normalized feature embedding e_i by using the text encoder f_{txt}: e_i = $f_{txt}(d_i)$, where d_i is a motion-enhanced description sampled from $\mathcal{D}$. In this way, we obtain the text embeddings $E_{txt} \in \mathbb{R}^{N\times(k+1)\times C}$, where C denotes the channel number of each embedding. Noticeably, the parameters of text encoder is frozen and initialized by CLIP [60] to inherit its capacity of encoding visual-aligned semantics.

However, as CLIP is pre-trained on image-text paired corpus, the text encoder has a strong bias towards spatial appearance of objects and backgrounds, instead of temporal motions [5, 30, 47, 63]. To adapt the text encoder to understand the motion-enhanced descriptions, we introduce a Text Motion Adapter, which consists of a multi-head self-attention (MHSA) [70]. The adapter is parameter-tunable and takes E_{txt} as input to learn the information dependencies between motion-enhanced descriptions. To aggregate the motion semantics for each category, we perform adaptive pooling to obtain the motion-enhanced embeddings $E^m_{txt} \in \mathbb{R}^{N\times 1\times C}$.

3.3. Visual Motion Extraction

In the visual side, as shown in Fig. 3, it comprises of two components: (i) The image encoder adopts the CLIP pre-trained Vision Transformer (ViT), which extracts frozen features for sparse frames with powerful spatial semantics. (ii) The motion encoder takes dense frames as input to capture the local temporal cues, integrated with the global temporal dynamics from middle-layer image features.

3.3.1 Image Encoder

Given a video clip $\mathbf{X}_S \in \mathbb{R}^{T\times H\times W\times 3}$ (T, H, and W represent the frame number, height, and width, respectively), the image features are extracted individually for several sparse frames. Following ViT [15], each frame is divided into $K = \frac{H}{P} \times \frac{W}{P}$ patches, and the size of each patch is denoted as $P \times P$. These patches are then projected using a fully connected layer, referred to as the 2D stem in Fig. 3. This projection generates a sequence of patch embeddings $[\mathbf{x}^{(0)}_{t,\text{cls}}, \mathbf{x}^{(0)}_{t,1}, \mathbf{x}^{(0)}_{t,2}, \cdots, \mathbf{x}^{(0)}_{t,K}] + \mathbf{e}^{\text{spatial}}$, where $t = \{1, \cdots, T\}$, $\mathbf{x}_{\text{cls}}$ is an additional learnable token (termed as "class embedding"), and $\mathbf{e}^{\text{spatial}}$ denotes spatial position embedding. Assuming that the spatial encoder has L Transformer blocks, the features of the l_{th} layer for the t_{th} frame can be extracted by:

$$\mathbf{X}^{(l)}_t = \text{Transformer}^{(l)}(\mathbf{X}^{(l-1)}_t) \in \mathbb{R}^{(K+1)\times C}, \quad (1)$$

where $l = \{1, \cdots, L\}$ denotes the layer index. The class embeddings of T frames in the l_{th} layer is termed as $\mathbf{X}^{(l)} = [\mathbf{X}^{(l)}_1, \cdots, \mathbf{X}^{(l)}_T] \in \mathbb{R}^{T\times 1\times C}$. Benefit from the CLIP pre-trained parameters, the class embeddings $\mathbf{X}^{(L)}$ aggregate powerful spatial semantics within each frame.

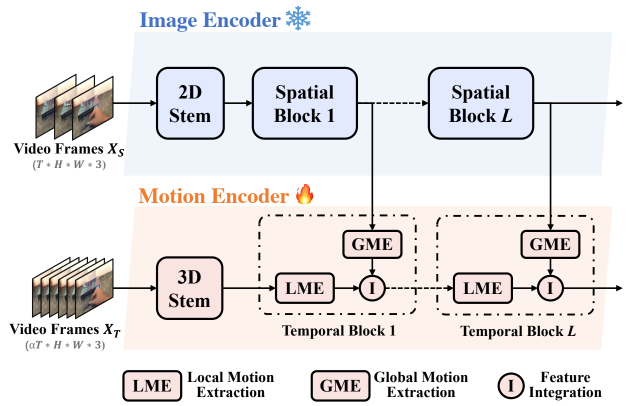

Figure 3. The structural details of MoTED in the vision side.

3.3.2 Motion Encoder

To fully extract the motion information in videos, the temporal input $\mathbf{X}_T \in \mathbb{R}^{\alpha T \times H \times W \times 3}$ for the motion encoder is sampled around the spatial input $\mathbf{X}_S$ by α times. In this study, we set α = 2 by default, as empirically studied in [17, 58]. Then, $\mathbf{X}_T$ is projected by a 3D convolution, *i.e.*, the 3D stem in Fig. 3, for patch embedding. The kernel size and stride of the 3D stem are both P in spatial dimension and α in temporal dimension. Thus the number of temporal patch tokens are the same with that in the image encoder, which makes it convenient to integrate the spatial and temporal features. Thus the projected temporal features can be formulated as: $\mathbf{Z}^{(0)} = \text{Conv3d}(\mathbf{X}_T) \in \mathbb{R}^{T \times K \times C}$, where $K = \frac{H}{P} \times \frac{W}{P}$. Then, a series of Temporal Blocks are designed to extract motion patterns, which can be written as:

$$\mathbf{Z}^{(l)} = \text{Temp-Block}^{(l)}(\mathbf{Z}^{(l-1)}, \mathbf{X}^{(l-1)}) \in \mathbb{R}^{T \times K \times C}, \quad (2)$$

where the function Temp-Block$(\cdot)$ performs temporal modeling integrated with the spatial features $X^{(l-1)}$.

Following the effective and efficient design philosophy in [32, 33], the temporal block consists of the following three modules: (i) **Local Motion Extraction (LME).** Since 3D convolution can capture detailed and local spatiotemporal features, by processing each pixel with context from a small 3D neighborhood (*e.g.*, $2 \times 16 \times 16$), the temporal patch tokens $\mathbf{Z}^{(0)}$ remain the nature of local motion extraction. Then these tokens are flattened to a sequence $\mathbb{R}^{TK \times C}$ and perform spatiotemporal learning via joint self-attention modules [66] and token merging for efficiency [3]. (ii) **Global Motion Extraction (GME).** Given the class embeddings $\mathbf{X}^{(l)}$ that aggregate powerful spatial semantics for each individual frame, we apply a cross-frame self-attention module [49] to learn the global inter-frame interactions. (iii) **Feature Integration.** After obtaining the local and global motion features, we adopt a cross-attention module, which takes local features as query and global features as key/value, to complement the features for better temporal modeling. In this way, the output of the final temporal block $\mathbf{Z}^{(L)}$ aggregates powerful temporal dynamics within the dense frames.

3.4. Training Loss

Following CLIP [60], we first apply adaptive pooling to obtain a video-level motion embedding $\mathbf{Z}_{avg}^{(L)} \in \mathbb{R}^{1 \times C}$. Then we introduce a contrastive loss to align the visual and textual motion embeddings:

$$\mathcal{L}_{motion} = -\log \frac{\exp(\text{sim}(\mathbf{Z}_{avg}^{(L)}, \mathbf{e}_i)/\tau)}{\sum_{c=1}^{N} \exp(\text{sim}(\mathbf{Z}_{avg}^{(L)}, \mathbf{e}_c)/\tau)}, \quad (3)$$

where $\text{sim}(\cdot,\cdot)$ is the normalized cosine similarity, τ refers to the temperature parameter. $\mathbf{e}_i$ is a motion-enhanced text embedding for the i_{th} category.

Moreover, we adopt a cross-attention module to fuse the visual motion features (as key/value) and image features (as query) to obtain the compact video embedding $\mathbf{v}_i$. Similarly, the fused text embedding $\mathbf{e}'_i$ is acquired via a cross-attention module. We introduce a contrastive loss to align the visual and textual video-level embeddings:

$$\mathcal{L}_{video} = -\log \frac{\exp(\text{sim}(\mathbf{v}_i, \mathbf{e}'_i)/\tau)}{\sum_{c=1}^{N} \exp(\text{sim}(\mathbf{v}_i, \mathbf{e}'_c)/\tau)}. \quad (4)$$

Compatible with CLIP training paradigm, the overall framework is trained based on these two contrastive losses:

$$\mathcal{L} = \mathcal{L}_{motion} + \mathcal{L}_{video}. \quad (5)$$

4. Experiments

4.1. Dataset and Implementation.

Datasets. In the supervised setting, we train on the train set of Kinetics-400 (K400) [28] and Something-Something-V2 (SSv2) [20] and report supervised performance against existing methods on the validation sets of K400 and SSv2. In the zero-shot setting, we train on the Kinetics-400 training set and evaluate on three datasets: Kinetics-600 (K600) [12], HMDB51 [24] and UCF101 [65]. For zero-shot evaluation on K600, following [12], we use the 220 new categories outside of K400 for evaluation, and conduct evaluation three times, each time randomly sampling 160 categories for evaluation from the 220 categories. For zero-shot evaluation on HMDB51 and UCF101, we follow [49] and report average top-1 accuracy and standard deviation on three splits of the test set.

Implementation Details. Following previous work [40, 58], we use the CLIP [60] pre-trained ViT-B/16, ViT-L/14

Class Names	Concept.	Discr.	SSv2	K400
✓	✗	✗	65.7	82.9
✓	✓	✗	66.0	84.5
✓	✗	✓	69.9	84.0
✓	✓	✓	**70.1**	**85.1**

(a) "Concept." is the abbreviation of "conceptual descriptions". "Discr." is the abbreviation of "discriminative descriptions".

M.Enc.	Adapter.	SSv2	K400
✗	✗	64.3	81.5
✓	✗	68.6	83.4
✗	✓	66.2	83.8
✓	✓	**70.1**	**85.1**

(b) "M.Enc." is the separated motion encoder where motion modules are parallel with CLIP visual encoder. "Adapter" is the text motion adapter.

Text.Fuse.	Vision.Fuse.	SSv2	K400
✗	✗	68.3	83.6
✓	✗	68.9	84.2
✗	✓	69.5	84.5
✓	✓	**70.1**	**85.1**

(c) "Text.Fuse." indicates feature fusion in the textual side. "Vision.Fuse." indicates feature fusion in the visual side.

Table 3. Ablations on **Something-Something-V2** and **Kinetics-400**. The spatial encoder is a 8-frame vanilla ViT-B/16 pre-trained by CLIP [60]. The inference protocol of all models and datasets are 3 clips × 1 center crop.

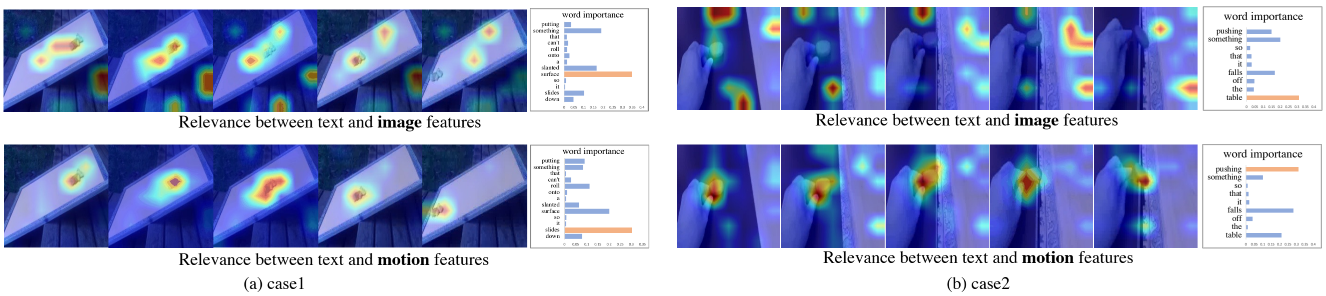

Figure 4. Two cases to visualize the relevance [8] between text and image/motion features to highlight the information relevant to the prediction. The different "regions of interest" and "words of importance" indicate that the motion and image features could be disentangled.

and ViT-L/14-336p as our image encoder. Unless otherwise specified, we mark the default settings in the temporal encoder in gray in Sec. 4.2. For simplicity, the ablation studies are conducted based on ViT-B/16 with a low-rate sampled 8 frames. We conduct the experiments with the NVIDIA 32G V100 GPUs. More implementation details (*e.g.*, training and testing hyper-parameters) are described in the Appendix.

4.2. Ablation and Analysis

Language Supervision. In this section, we aim to demonstrate the significance of language supervision by comparing different types of text descriptions. Tab. 3a presents a comparison of three types of texts: category names, motion concept descriptions, motion discriminative descriptions. It reveals that the performance of action category names, which contain the least information, is notably lower than other results, particularly on the SSv2 dataset with abundant temporal information. Although descriptions with only action concepts show a slight improvement of 0.3%/1.6%, they lack the vital information regarding the distinguishing characteristics of the actions. As a result, the model exhibits poor classification accuracy for similar categories. On the other hand, when the discriminative descriptions are adopted, the accuracy on SSv2 is improved noticeably (+4.2%), revealing the significance of discriminative characteristics for fine-grained datasets. Additionally, when the full descriptions are utilized, the model's performance obtain the further gains of 4.4%/2.2%.

Disentangled Motion Modeling. Our method has an advantage in allowing the learning of dynamic information without interference, while preserving CLIP's original spatially transferable representation capabilities. To validate this, we compared two different structures in Tab. 3b. For the serial structure, additional dynamic modules are inserted between the Transformer Blocks of each layer of the visual encoder, resulting in a unified video representation. All the texts are combined through the text encoder to obtain text representations for alignment. It can be observed that the parallel structure outperforms the serial structure, with improvements of 4.3% and 1.9% respectively. This demonstrates the rationality of the parallel approach. Tab. 3b also presents a comparison of introducing an adapter in the text encoder. The model's classification performance improved by 1.9% and 2.3% after adding the adapter, indicating its effectiveness in reducing bias in the text encoder.

Feature Fusion Direction. As shown in Tab. 3c, both directions of information integration can improve performances. The combination of the two can boost accuracy more significantly +1.8%/+1.5% on SSv2/K400, respectively. This verifies the importance of spatio-temporal blending for the parallel architectures. In our opinion, independent learning of motion can effectively avoid excessive reliance of the model on the previously learned spatial information. By using cross attention for fusion, it is possible to effectively integrate features from two different dimensions, ultimately forming the features of the target video and achieving the best results.

Method	Pre-train	Architecture	Input Size	FLOPs×Cr.×Cl. (T)	Param (M)	Frozen	Top-1	Top-5
SlowFast [17]	ImageNet-21K	R101+NL	16×224^2	$0.1 \times 3 \times 1$	60	✗	63.1	87.6
ViViT FE [1]	IN21K+K400	ViT-L	16×224^2	$1.0 \times 3 \times 4$	612	✗	65.4	89.8
MTV-B(320p) [84]	IN21K+K400	-	32×224^2	$0.9 \times 3 \times 4$	310	✗	68.5	90.4
MViT [16]	Kinetics-600	MViT-B-24	32×224^2	$0.2 \times 3 \times 1$	53	✗	68.7	91.5
Video Swin [42]	IN21K+K400	Swin-B	32×224^2	$0.3 \times 3 \times 1$	60	✗	69.6	92.7
TAdaConvNeXtV2 [22]	IN1K+K400	ConvNeXt-S	32×224^2	$0.2 \times 3 \times 2$	82	✗	70.0	92.0
EVL❄ [40]	CLIP-400M	ViT-B	32×224^2	$0.68 \times 1 \times 3$	175	✓	62.4	-
ST-Adapter❄ [52]	CLIP-400M	ViT-B	32×224^2	$0.61 \times 1 \times 3$	93	✓	69.5	92.6
DiST❄	CLIP-400M	ViT-B	32×224^2	$0.65 \times 1 \times 3$	105	✓	70.9	92.1
MoTED❄	CLIP-400M	ViT-B	8×224^2	$0.18 \times 1 \times 3$	112	✓	70.1	91.8
MoTED❄	CLIP-400M	ViT-B	16×224^2	$0.34 \times 1 \times 3$	112	✓	71.2	92.4
MoTED❄	CLIP-400M	ViT-B	32×224^2	$0.68 \times 1 \times 3$	112	✓	**71.9**	**92.7**
UniformerV2 [32]	CLIP-400M	ViT-L	32×224^2	$1.73 \times 1 \times 3$	574	✗	73.0	94.5
TAdaFormer [22]	CLIP-400M	ViT-L	32×224^2	$1.70 \times 2 \times 3$	364	✗	73.6	-
EVL❄ [40]	CLIP-400M	ViT-L	32×224^2	$3.21 \times 1 \times 3$	654	✓	66.7	-
EVL❄ [40]	CLIP-400M	ViT-L	32×336^2	$8.08 \times 1 \times 3$	654	✓	68.0	-
ST-Adapter❄ [52]	CLIP-400M	ViT-L	32×224^2	$2.75 \times 1 \times 3$	347	✓	72.3	93.9
DiST❄	CLIP-400M	ViT-L	32×224^2	$2.83 \times 1 \times 3$	336	✓	73.1	93.2
MoTED❄	CLIP-400M	ViT-L	8×224^2	$0.78 \times 1 \times 3$	346	✓	71.5	92.6
MoTED❄	CLIP-400M	ViT-L	16×224^2	$1.49 \times 1 \times 3$	346	✓	73.0	93.4
MoTED❄	CLIP-400M	ViT-L	32×224^2	$2.89 \times 1 \times 3$	346	✓	**73.8**	**93.8**

Table 4. Comparison with the state-of-the-art methods on Something-Something-V2. "Cr." and "Cl." are the abbreviation for "spatial crops" and "temporal clips". "Frozen" indicates freezing the CLIP pre-trained parameters.

Motion Modeling Visualization. In Fig. 4, we also perform the analysis of vision features generated by image and motion encoder to investigate the learned patterns from language supervision. Based on the reasoning tool [8], we depict the attention maps of class token from the final transformer block of the image/motion stream encoder w.r.t. the text encoder. It is observed that, for the case1 in Fig. 4, the CLIP image encoder attends to both motion-relevant foreground and motion-irrelevant background with a major focus on "surface". In contrast, features extracted from the motion modules concentrate on the motion-relevant regions of the moving object, and emphasize the motion-related words "sliding". This phenomenon reveals that the motion features can complement the static spatial semantics of objects in the image features via modeling object motions. More cases can be accessed in the Appendix.

4.3. Fully-supervised Experiments

In the supervised setting, the results on SSv2 and K400 are presented in Tab. 4 and Tab. 6, respectively. Compared with EVL, our proposed MoTED introduces a similar temporal module for CLIP visual encoder, but our method has a significant improvement compared to EVL, with an improvement of 9.5%/2.0% and 5.8%/1.5% on ViT-B and ViT-L respectively on SSv2/K400. The performance gains greatly demonstrate the rationality and effectiveness of using language supervision. Interestingly, the performance gains are particularly pronounced on SSv2. This is because SSv2 is a fine-grained action classification task requiring stronger action discrimination and perception of dynamic information.

Related work	Visual disentangle	Textual disentangle	K400 (Acc / Δ)	SSv2 (Acc / Δ)
Previous SOTA	✗	✗	84.2 / -	69.5 / -
DiST [58]	✓	✗	85.0 / +0.8	70.9 / +1.4
MoTED (Ours)	✓	✓	86.2 / +1.2	71.9 / +1.0

Table 5. Comparison of our method with previous state-of-the-art (SOTA) and the latest related work DiST.

Disentangling spatio-temporal learning is an effective approach to endow the model with temporal capability. As shown in Tab. 5, DiST focuses on temporal disentangling *in the visual side* with gains of +0.8%/+1.4% on K400/SSv2. Our study highlights the effectiveness of temporal disentangling *in the textual side* with detailed descriptions of motions, with further gains of +1.2%/+1.0%. This result reveals that disentangling textual encoder is equally effective w.r.t. disentangling visual encoder for vision-language transfer learning.

4.4. Zero-shot Experiments

Zero-shot generalization is an attractive characteristic of CLIP-extended models, making them more practical in real

Method	Pre-train	Architecture	Input Size	TFLOPs×Cr.×Cl.	Param (M)	Frozen	Top-1	Top-5
SlowFast [17]	-	R101+NL	16×224^2	$0.4 \times 3 \times 10$	60	✗	79.8	93.9
TimeSformer [2]	ImageNet-21K	ViT-L	96×224^2	$8.4 \times 3 \times 1$	430	✗	80.7	94.7
MViT [16]	-	MViT-B	64×224^2	$0.5 \times 1 \times 5$	37	✗	81.2	95.1
ViViT FE [1]	ImageNet-21K	ViT-L	128×224^2	$4.0 \times 3 \times 1$	N/A	✗	81.7	93.8
Video Swin [42]	ImageNet-21K	Swin-L	32×224^2	$0.6 \times 3 \times 4$	197	✗	83.1	95.9
TAdaConvNeXtV2 [22]	ImageNet-21K	ConvNeXt-B	32×224^2	$0.3 \times 3 \times 4$	146	✗	83.7	-
X-CLIP [49]	CLIP-400M	ViT-B	16×224^2	$0.28 \times 3 \times 4$	128	✗	84.7	96.8
ST-Adapter❄ [52]	CLIP-400M	ViT-B	32×224^2	$0.61 \times 1 \times 3$	93	✓	82.7	96.2
EVL❄ [40]	CLIP-400M	ViT-B	32×224^2	$0.59 \times 1 \times 3$	115	✓	84.2	-
DiST❄	CLIP-400M	ViT-B	32×224^2	$0.65 \times 1 \times 3$	112	✓	85.0	97.0
MoTED❄	CLIP-400M	ViT-B	8×224^2	$0.18 \times 1 \times 3$	116	✓	85.1	97.0
MoTED❄	CLIP-400M	ViT-B	16×224^2	$0.34 \times 1 \times 3$	116	✓	85.4	97.2
MoTED❄	CLIP-400M	ViT-B	32×224^2	$0.68 \times 1 \times 3$	116	✓	**86.2**	**97.5**
UnifromerV2 [32]	CLIP-400M+K710	ViT-L	32×224^2	$2.66 \times 2 \times 3$	354	✗	89.3	98.2
TAdaFormer [22]	CLIP-400M+K710	ViT-L	32×224^2	$1.41 \times 4 \times 3$	364	✗	89.5	-
ST-Adapter❄ [52]	CLIP-400M	ViT-L	32×224^2	$2.75 \times 1 \times 3$	347	✓	87.2	97.6
EVL❄ [40]	CLIP-400M	ViT-L	32×224^2	$2.70 \times 1 \times 3$	363	✓	87.3	-
DiST❄	CLIP-400M	ViT-L	32×224^2	$2.83 \times 1 \times 3$	343	✓	88.0	97.9
MoTED❄	CLIP-400M	ViT-L	8×224^2	$0.78 \times 1 \times 3$	349	✓	87.4	97.8
MoTED❄	CLIP-400M	ViT-L	16×224^2	$1.49 \times 1 \times 3$	349	✓	88.0	98.0
MoTED❄	CLIP-400M	ViT-L	32×224^2	$2.89 \times 1 \times 3$	349	✓	**88.8**	**98.2**

Table 6. Comparison with state-of-the-arts on Kinetics-400. "Cr." and "Cl." are the abbreviation for "spatial crops" and "temporal clips". "Frozen" indicates freezing the CLIP pre-trained parameters.

Method	Model	HMDB51	UCF101	K600
ActionCLIP [75]	B/16	40.8±5.4	58.3±3.4	-
X-CLIP [49]	B/16	44.6±5.2	72.0±2.3	65.2±0.4
Vita-CLIP [78]	B/16	48.6±0.6	75.0±0.6	67.4±0.5
DiST❄	B/16	55.4±1.2	72.3±0.6	-
MoTED❄	B/16	58.2±1.1	78.3±0.6	69.9 ± 0.5

Table 7. Comparison of zero-shot accuracy with the state-of-the-art CLIP-based methods on three datasets (*e.g.*, HMDB51 [24], UCF101 [65], and K600 [12]). "❄": frozen backbone.

world applications. For zero-shot settings, we evaluate MoTED on three widely used benchmarks. Following prior work [49], we train the networks on the K400 training set, then conduct the zero-shot evaluation on three unseen datasets (*i.e.*, UCF101, HMDB51 and K600), as shown in Tab. 7. All models share the same architecture of "ViT-B", with 32 frames during inference. Compared with other vison-language methods, MoTED achieves the better zero-shot performances with a significant margin on HMDB51 (+3.8%), UCF101 (+6.0%) and K600 (+2.5%). Different from X-CLIP [49], DiST [58], Vita-CLIP [78] that learns motion representation with the supervision of action category names solely, the proposed MoTED makes full use of conceptual and discriminative descriptions and learn general motion representations with the aid of language supervision. In addition, our method also has a relatively small variance, only about 1%. We assume this is due to the benefits brought by the rich content of the text, as detailed descriptions can promote stable dynamic feature learning.

5. Conclusion

In this study, we aim to overcome the limitations of existing methods in extending large image-text pre-trained models for video understanding. The proposed MoTED framework introduces Motion-Enhanced Descriptions, which are applied to facilitate the extraction of unique temporal features in videos. By generating motion-related descriptions and incorporating perception modules, MoTED aligns visual and textual motion features using a contrastive loss. Experimental results on five benchmarks demonstrate that MoTED provides a strong basis for enhancing CLIP with robust temporal modeling. In future works, we hope to further dedicate to exploring the potential of language supervision and combining it with more powerful dynamic information perception modules to achieve higher performance in video recognition and make it truly practical.

6. Acknowledgement

This work was supported in part by NSFC No. 62222117, the Fundamental Research Funds for the Central Universities under contract WK3490000005, and KY2100000117.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, pages 6836–6846, 2021. 3, 7, 8

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. 3, 8

[3] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *ICLR*, 2023. 5

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020. 2, 3

[5] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "video" in video-language understanding. In *CVPR*, pages 2907–2917, 2022. 2, 4

[6] Adrian Bulat, Juan Manuel Perez Rua, Swathikiran Sudhakaran, Brais Martinez, and Georgios Tzimiropoulos. Space-time mixing attention for video transformer. *NeurIPS*, 34:19594–19607, 2021. 3

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 3

[8] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *ICCV*, pages 387–396, 2021. 6, 7

[9] Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V. Thapliyal, James Bradbury, and Weicheng Kuo. Pali: A jointly-scaled multilingual language-image model. In *ICLR*, 2023. 1, 2

[10] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*, 2021. 2

[11] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113, 2023. 2, 3

[12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision. *arXiv preprint arXiv:2006.13256*, 2020. 2, 5, 8

[13] Ali Diba, Vivek Sharma, Reza Safdari, Dariush Lotfi, Saquib Sarfraz, Rainer Stiefelhagen, and Luc Van Gool. Vi2clr: Video and image for visual contrastive learning of representation. In *ICCV*, pages 1502–1512, 2021. 3

[14] Sixun Dong, Huazhang Hu, Dongze Lian, Weixin Luo, Yicheng Qian, and Shenghua Gao. Weakly supervised video representation learning with unaligned text for sequential videos. In *CVPR*, pages 2437–2447. IEEE, 2023. 3

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 4

[16] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, pages 6824–6835, 2021. 3, 7, 8

[17] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019. 3, 5, 7, 8

[18] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *CVPR*, pages 3299–3309, 2021. 3

[19] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *NeurIPS*, 2022. 3

[20] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, pages 5842–5850, 2017. 2, 5

[21] Ziyuan Huang, Shiwei Zhang, Liang Pan, Zhiwu Qing, Mingqian Tang, Ziwei Liu, and Marcelo H Ang Jr. Tada! temporally-adaptive convolutions for video understanding. In *ICLR*, 2022. 3

[22] Ziyuan Huang, Shiwei Zhang, Liang Pan, Zhiwu Qing, Yingya Zhang, Ziwei Liu, and Marcelo H Ang Jr.

Temporally-adaptive models for efficient video understanding. *arXiv preprint arXiv:2308.05787*, 2023. 7, 8

[23] Simon Jenni and Hailin Jin. Time-equivariant contrastive video representation learning. In *ICCV*, pages 9970–9980, 2021. 3

[24] H Jhuang, H Garrote, E Poggio, T Serre, and T Hmdb. A large video database for human motion recognition. In *ICCV*, page 6, 2011. 2, 5, 8

[25] 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 *ICML*, pages 4904–4916. PMLR, 2021. 1, 2

[26] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *ECCV*, pages 105–124. Springer, 2022. 1, 2

[27] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *ECCV*, pages 105–124. Springer, 2022. 1

[28] 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. 2, 3, 5

[29] Jan Kocon, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydlo, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocon, Bartlomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Milkowski, Marcin Oleksy, Maciej Piasecki, Lukasz Radlinski, Konrad Wojtasik, Stanislaw Wozniak, and Przemyslaw Kazienko. Chatgpt: Jack of all trades, master of none. *Inf. Fusion*, 99: 101861, 2023. 3

[30] Jie Lei, Tamara L. Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In *ACL*, pages 487–507, 2023. 2, 4

[31] Chunyuan Li, Haotian Liu, Liunian Harold Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, and Jianfeng Gao. ELEVATER: A benchmark and toolkit for evaluating language-augmented visual models. In *NeurIPS*, 2022. 2

[32] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *arXiv preprint arXiv:2211.09552*, 2022. 3, 5, 7, 8

[33] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *TPAMI*, 2023. 3, 5

[34] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. HERO: hierarchical encoder for video+language omni-representation pre-training. In *EMNLP*, pages 2046–2065, 2020. 2

[35] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, pages 10955–10965, 2022. 2

[36] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. *arXiv preprint arXiv:2212.00794*, 2022. 1, 2

[37] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *ICLR*, 2022. 2

[38] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, pages 4804–4814, 2022. 3

[39] Xudong Lin, Fabio Petroni, Gedas Bertasius, Marcus Rohrbach, Shih-Fu Chang, and Lorenzo Torresani. Learning to recognize procedural activities with distant supervision. In *CVPR*, pages 13843–13853. IEEE, 2022. 3

[40] 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 *ECCV*, pages 388–404. Springer, 2022. 1, 2, 5, 7, 8

[41] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 3

[42] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, pages 3202–3211, 2022. 3, 7, 8

[43] Timo Lüddecke and Alexander S. Ecker. Image segmentation using text and image prompts. In *CVPR*, pages 7076–7086, 2022. 2

[44] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *ICLR*, 2023. 3

[45] 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 *ICCV*, pages 2630–2640, 2019. 2

[46] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, pages 9879–9889, 2020. 2

[47] Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. Verbs in action: Improving verb understanding in video-language models. *ICCV*, 2023. 2, 3, 4

[48] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *ICCV*, pages 3163–3172, 2021. 3

[49] 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 *ECCV*, pages 1–18. Springer, 2022. 1, 2, 5, 8

[50] OpenAI. Gpt-4 technical report, 2023. 1, 2, 4

[51] TB OpenAI. Chatgpt: Optimizing language models for dialogue. *OpenAI*, 2022. 3

[52] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. Parameter-efficient image-to-video transfer learning. *arXiv e-prints*, pages arXiv–2206, 2022. 1, 2, 7, 8

[53] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *CVPR*, pages 11205–11214, 2021. 3

[54] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Dual-path adaptation from image to video transformers. *CVPR*, 2023. 3

[55] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. *NeurIPS*, 34:12493–12506, 2021. 3

[56] Sarah M. Pratt, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. *ICCV*, 2022. 3

[57] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *CVPR*, pages 6964–6974, 2021. 3

[58] Zhiwu Qing, Shiwei Zhang, Ziyuan Huang, Yingya Zhang, Changxin Gao, Deli Zhao, and Nong Sang. Disentangling spatial and temporal learning for efficient image-to-video transfer learning. *ICCV*, 2023. 1, 2, 5, 7, 8

[59] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *ICCV*, pages 5533–5541, 2017. 3

[60] 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 *ICML*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 4, 5, 6

[61] Kanchana Ranasinghe and Michael S. Ryoo. Language-based action concept spaces improve video self-supervised learning. *CoRR*, abs/2307.10922, 2023. 3

[62] Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. *NeurIPS*, 34:12786–12797, 2021. 3

[63] Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, and Lorenzo Torresani. Only time can tell: Discovering temporal data for temporal modeling. In *WACV*, pages 535–544, 2021. 2, 4

[64] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *NeurIPS*, 27, 2014. 3

[65] 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. 2, 5, 8

[66] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *NeurIPS*, 2022. 3, 5

[67] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015. 3

[68] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, pages 6450–6459, 2018.

[69] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *ICCV*, pages 5552–5561, 2019. 3

[70] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 4

[71] Heng Wang, Du Tran, Lorenzo Torresani, and Matt Feiszli. Video modeling with correlation networks. In *CVPR*, pages 352–361, 2020. 3

[72] Junke Wang, Dongdong Chen, Chong Luo, Xiyang Dai, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. Chatvideo: A tracklet-centric multimodal and versatile video understanding system. *CoRR*, abs/2304.14407, 2023. 3

[73] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *TPAMI*, 41(11): 2740–2755, 2018. 3

[74] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae V2: scaling video masked autoencoders with dual masking. In *CVPR*, pages 14549–14560, 2023. 3

[75] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 1, 2, 8

[76] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, pages 23318–23340, 2022. 2

[77] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. In *CVPR*, pages 14733–14743, 2022. 3

[78] Syed Talal Wasim, Muzammal Naseer, Salman H. Khan, Fahad Shahbaz Khan, and Mubarak Shah. Vita-clip: Video and text adaptive CLIP via multimodal prompting. *CVPR*, 2023. 1, 2, 8

[79] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, pages 14668–14678, 2022. 3

[80] Wenhao Wu, Zhun Sun, and Wanli Ouyang. Revisiting classifier: Transferring vision-language models for video recognition. In *AAAI*, pages 2847–2855, 2023. 1

[81] 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. In *EMNLP*, pages 6787–6800, 2021. 2

[82] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas M. Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *CVPR*, pages 18113–18123, 2022. 2

[83] An Yan, Yu Wang, Yiwu Zhong, Chengyu Dong, Zexue He, Yujie Lu, William Wang, Jingbo Shang, and Julian J. McAuley. Learning concise and descriptive attributes for visual recognition. *ICCV*, 2023. 3

[84] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *CVPR*, pages 3333–3343, 2022. 7

[85] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *CVPR*, pages 19187–19197, 2023. 3

[86] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. MM-REACT: prompting chatgpt for multimodal reasoning and action. *CoRR*, abs/2303.11381, 2023. 3

[87] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 1, 2

[88] 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. 1, 2

[89] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *CVPR*, pages 6586–6597. IEEE, 2023. 3

[90] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. Regionclip: Region-based language-image pretraining. In *CVPR*, pages 16772–16782, 2022. 2

[91] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592, 2023. 3