OST: Refining Text Knowledge with Optimal Spatio-Temporal Descriptor for General Video Recognition

Tongjia Chen[1], Hongshan Yu[1✉], Zhengeng Yang[2], Zechuan Li[1], Wei Sun[1], Chen Chen[3]
[1]Hunan University, [2]Hunan Normal University
[3]Center for Research in Computer Vision, University of Central Florida
✉ Corresponding author   Project Page: https://tomchen-ctj.github.io/OST.

Abstract

Due to the resource-intensive nature of training vision-language models on expansive video data, a majority of studies have centered on adapting pre-trained image-language models to the video domain. Dominant pipelines propose to tackle the visual discrepancies with additional temporal learners while overlooking the substantial discrepancy for web-scaled descriptive narratives and concise action category names, leading to less distinct semantic space and potential performance limitations. In this work, we prioritize the refinement of text knowledge to facilitate generalizable video recognition. To address the limitations of the less distinct semantic space of category names, we prompt a large language model (LLM) to augment action class names into Spatio-Temporal Descriptors thus bridging the textual discrepancy and serving as a knowledge base for general recognition. Moreover, to assign the best descriptors with different video instances, we propose Optimal Descriptor Solver, forming the video recognition problem as solving the optimal matching flow across frame-level representations and descriptors. Comprehensive evaluations in zero-shot, few-shot, and fully supervised video recognition highlight the effectiveness of our approach. Our best model achieves a state-of-the-art zero-shot accuracy of 75.1% on Kinetics-600.

1. Introduction

Large-scale contrastive language-image pre-training [25, 46, 65] have shown remarkable performance in various computer vision tasks. The visual-semantic joint space not only serves powerful visual representation but also enables few/zero-shot transferring to downstream tasks with the reference of natural language. However, training a similar model for video recognition can be costly since large-scale video-language datasets are exponentially more massive [57] due to the extra temporal dimension. Hence, a feasible solution is to adapt the pre-trained image-text models for the task of video recognition. As depicted in

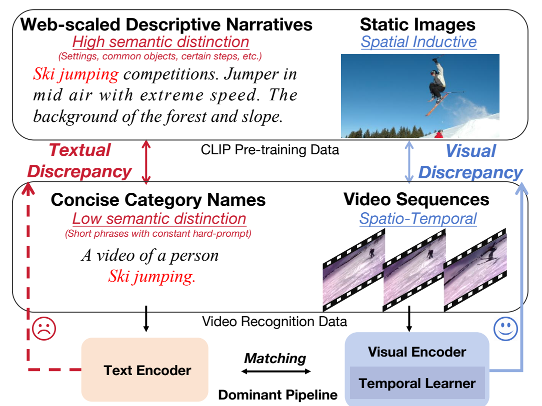

Figure 1. Motivation of our method. Dominant pipelines propose to tackle the visual discrepancies with additional temporal learners while overlooking the textual discrepancy between descriptive narratives and concise category names. This oversight results in a less separable latent space, which may hinder video recognition.

Fig. 1, current methods devise a range of temporal learners to address the visual discrepancy while preserving text-domain knowledge in the semantic space of action category names, often by merging the category name with CLIP-style hard-prompts (*e.g., "a video of a person* {*ski jumping*}*"*) [41, 45, 53, 56, 60]. Despite providing essential inter-class correlations that can benefit general recognition, we speculate this paradigm overlooks the textual discrepancy between web-scaled descriptive narratives in CLIP pre-training and concise category names in downstream video recognition. Given that category names of video datasets generally consist of verbs and nouns, the nouns exhibit variability while the verbs tend to remain consistent. For instance, *playing cello, playing organ & playing violin* are distinct actions related to playing instruments. The sole differentiation between these category names lies in the noun itself, resulting in low discriminative text embeddings. This may lead to a less separable semantic space, potentially introducing ambiguity in recognition tasks [5].

To validate our hypothesis, we perform a sanity check on the semantic distribution of category embeddings across ImageNet [17], Kinetics-400 [7], and Something-Something v2 [20]. Initially, we employ a CLIP-B/16 text encoder to extract semantic embeddings of category names and leverage t-SNE visualization [54] to illustrate embedding clusters across the three datasets. As depicted in Fig. 2 **(Left)**, features from K400 and Sthv2 datasets exhibit denser clustering compared to those from ImageNet, qualitatively indicating the low semantic distinction of video category names. To quantify this distinction and provide further support for our hypothesis, we compute pair-wise cosine similarity within each dataset and determine the average similarity, serving as a measure of semantic density. A higher similarity implies a denser distribution of category embeddings and less separable semantics in the latent space. Fig. 2 **(Right)** visually demonstrates consistently higher mean cosine similarity of category names on video datasets compared to image datasets. This observation suggests that the intrinsic semantic space associated with video category names is less distinct. Since the category embedding serves as a decision plane [60] in cross-modal matching (*i.e.* compute the cosine similarity between category embeddings and visual features), such reduced distinctiveness may potentially diminish its efficacy in recognition tasks.

To mitigate this issue, one could manually craft textual narratives, but this process is labor-intensive. Alternatively, Large Language Models (LLMs) serve as a viable solution, acting as expansive knowledge bases that can generate detailed descriptors efficiently. As shown in Fig. 1, we can substantially refine our comprehension of ski jumping by integrating external contextual information such as the forest, the snow slope, and different action steps performed by the ski jumper. Hence, we propose to prompt LLMs with category names into what we define as *Spatio-Temporal Descriptors* to enrich the semantic space with external knowledge. Where *Spatio Descriptors* should possess the capability to capture static appearances, for instance, the environment and distinct objects included, while *Temporal Descriptors* should focus on describing the temporal evolution of actions. This allows for the disentanglement of the category name into two complementary semantic spaces, thereby enhancing the semantic distinction and providing external knowledge for general recognition.

Based on the obtained descriptors, an intuitive solution is to aggregate these descriptors as a global category embedding via pooling, and match the embedding with corresponding visual features [28, 38]. However, this utilization might be suboptimal due to the following reasons: **1)** Since the descriptors for one action class may not be contained in every video instance in this action category, directly matching the pooled descriptor-level representations with each video is potentially ineffective. **2)** The propensity of LLMs to exhibit hallucinations [69] may bring noises to descriptors. To address this, we need to consider the adaptability of descriptors to individual video instances. In this vein, we propose *Optimal Descriptor Solver* to obtain an optimal transport plan that adaptively aligns features across frame-level tokens and descriptors.

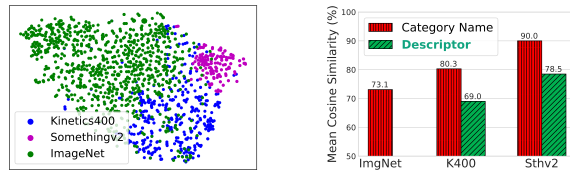

Figure 2. Sanity check on category names. We investigate the semantic distribution of video category names **(Left)** and quantify the semantic density of category names **(Right)**. We observe a higher semantic similarity of category names on K400 and Sthv2 compared to ImageNet. Our proposed *Spatio-Temporal Descriptor* can greatly reduce the semantic similarity in latent space. *Please refer to Sec. 3.2 for comprehensive details.*

In light of the above explorations, we propose Optimal Spatio-Temporal Descriptor (**OST**), a general pipeline for video recognition. Our **OST** comprises two components: We first disentangle the category name into *Spatio-Temporal Descriptors*, which not only bridges the semantic gap between narratives and category names but also serves as a knowledge base for general recognition. Then, we propose *Optimal Descriptor Solver* that adaptively aligns frame-level representations with *Spatio-Temporal Descriptors* to enhance video recognition. To demonstrate the effectiveness of our **OST**, we conduct comprehensive experiments on six benchmarks, including Kinetics-400 [7] & 600 [8], UCF-101 [50], HMDB-51 [31], Something-Something V2 [20], and ActivityNet [6]. The results indicate that our method achieves state-of-the-art performance in open-vocabulary tasks, *e.g.* zero-shot, few-shot, and also consistently improves the performance when combined with existing pipelines in fully-supervised settings. The main contributions of this work are as follows:

- We provide new insights that prior pipelines for adapting vision-language pre-trained models to video recognition are constrained by the semantic space of category names.
- We propose *Spatio-Temporal Descriptors* derived from LLMs to enhance the distinction of semantic space and provide external knowledge for general recognition.
- We introduce *Optimal Descriptor Solver* that forms the video recognition problem as solving the optimal matching flow across frame-level representations and descriptors to fully refine the semantic knowledge.
- Our **OST** presents a new way to utilize external knowledge to adapt pre-trained visual-language models for general video recognition. Experimental results in zero-shot, few-shot, and fully-supervised settings demonstrate the superior performance and generalizability of our method.

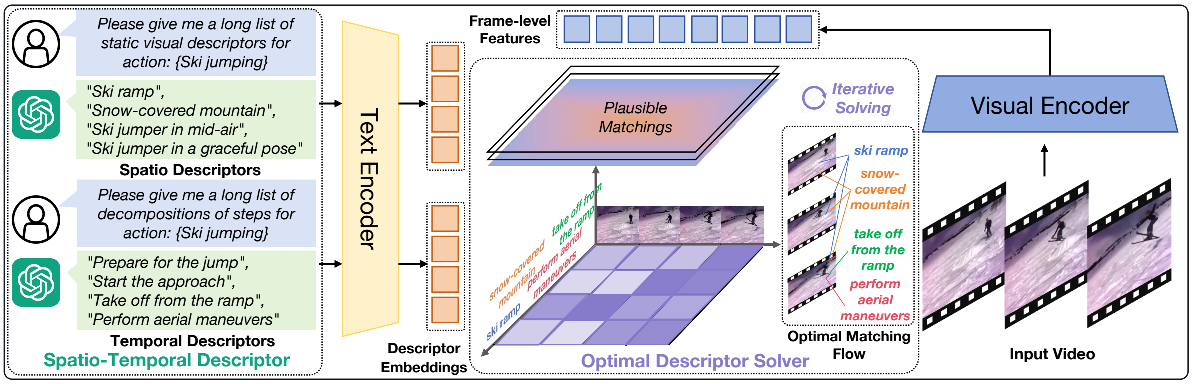

Figure 3. An overview of our pipeline for video recognition. We query the Large Language Model to augment category names to generate corresponding *Category Descriptors*. The descriptors disentangled category names into *Spatio-Temporal Descriptors* for static visual cues and temporal evolution, respectively. To fully refine the textual knowledge, we propose *Optimal Descriptor Solver* that adaptively aligns descriptors with video frames. An optimal matching flow is calculated through the iterative solving of the entropy-regularized OT problem to assign optimal descriptors for each video instance. *Please zoom in for comprehensive details.*

2. Related Work

Video Recognition. As a fundamental component of computer vision, mainstream pipelines have typically explored traditional 2D, 3D CNNs [7, 23, 33, 52, 55, 63] and Transformer-based methods [3, 12, 18, 35, 37, 41, 53, 64]. Additionally, methods modeling action phases [2, 51, 68, 70] have shown promise in video recognition, especially for long-form videos. Recently, cross-modal video recognition [26, 41, 45, 53, 56, 60–62] has benefited a lot from the powerful visual-text joint semantic space of CLIP. This cross-modal paradigm not only fosters strong representations with rich semantics but also achieves great open-vocabulary capacities. However, dominant pipelines [45, 53, 56, 60] focus on the temporal discrepancies between images and videos while maintaining text-domain knowledge constantly. In contrast, our method prioritizes the refinement of text knowledge.

Language for Visual Recognition. Differing from visual signals, natural language contains dense semantic information. Thus, language can serve as a rich source to provide inter-class correlations to benefit visual recognition. CuPL [43] and pipeline proposed by Menon *et al.* [38] utilizes category descriptions from GPT-3 as global category embedding for improved zero-shot image classification. Kaul *et al.* [28] propose to utilize LLM descriptions and visual prototypes to construct a multi-modal classifier for enhanced open-vocabulary object detection. MAXI [34] proposes to construct text bags generated via multiple sources (*e.g.,* captions and descriptions) to perform unsupervised finetuning for robust zero-shot action recognition. ASU [13] utilizes semantic units manually derived from WordNet and Wikipedia for video recognition. In this work, we aim to refine text knowledge by finding the optimal *Spatio-Temporal Descriptors* automatically generated by LLMs to bridge the semantic discrepancy and provide external knowledge to benefit general video recognition.

Optimal Transport. Optimal Transport (OT), also known as Monge Problem [40], is an essential mathematical framework that facilitates the establishment of correspondences between two distinct distributions. Its great characteristics for distribution matching have benefited a variety of machine learning tasks [29], including domain adaptation [14, 16], generative models [1, 21, 48], graph matching [10, 42], image matching [36, 66], and prompt learning [9, 30], *etc*. In this work, we propose to utilize OT distance to solve the cross-modal matching problem. To the best of our knowledge, this is the first work to form the video-text matching problem as solving the OT problem between frame-level representations and textual embeddings.

3. Method

In this section, we first review the preliminaries of optimal transport in Sec. 3.1, then discuss our proposed *Spatio-Temporal Descriptor* and *Optimal Descriptor Solver* scheme in Sec. 3.2 and Sec. 3.3, respectively. Finally, we introduce the training objectives in Sec. 3.4.

3.1. Preliminaries

Optimal transport aims to seek the minimal-cost transport plan between two distributions. In this work, we only consider the discrete distribution which is closely related to our framework. Assuming we have two sets of discrete empirical distributions:

$$\boldsymbol{\mu} = \sum_{i=1}^{M} p_i \delta_{x_i}, \quad \boldsymbol{\nu} = \sum_{j=1}^{N} q_j \delta_{y_j}, \tag{1}$$

where p_i and q_j are the probability distribution summing to 1, M and N are number of samples in each empirical distribution, δ denotes the Dirac function. Since each certain distribution is discrete, the optimal transport plan $\boldsymbol{P}$ matching the two distributions is also discrete. In this setting, we can adapt Kantorovich OT formulation [27] and form the optimal transport problem as:

$$\boldsymbol{P}^* = \underset{\boldsymbol{P}\in\mathbb{R}^{M\times N}}{\arg\min} \sum_{i=1}^{M}\sum_{j=1}^{N} \boldsymbol{P}_{ij}\boldsymbol{C}_{ij} \quad \text{s.t.} \quad \boldsymbol{P}\boldsymbol{e} = \boldsymbol{\mu}, \quad \boldsymbol{P}^\top \boldsymbol{e} = \boldsymbol{\nu}. \tag{2}$$

$\boldsymbol{C} \in \mathbb{R}^{M\times N}$ is the cost matrix that represents the distance between the support points x_i and y_j such as $\boldsymbol{C}_{ij} = 1 - sim(x_i, y_j)$. $\boldsymbol{P}^*$ is the optimal transport plan between two empirical distributions to minimize the total distance and $\boldsymbol{e}$ is the vector of ones. Considering the computational and statistical limitations of this original OT formulation, we adopt the Sinkhorn-Knopp [15] algorithm to solve the entropy-regularized OT problem. The regularized OT problem is defined as:

$$\boldsymbol{P}^* = \underset{\boldsymbol{P}\in\mathbb{R}^{M\times N}}{\arg\min} \sum_{i=1}^{M}\sum_{j=1}^{N} \boldsymbol{P}_{ij}\boldsymbol{C}_{ij} - \lambda \boldsymbol{H}(\boldsymbol{P}) \quad \text{s.t.} \quad \boldsymbol{P}\boldsymbol{e} = \boldsymbol{\mu}, \quad \boldsymbol{P}^\top \boldsymbol{e} = \boldsymbol{\nu}, \tag{3}$$

where $\boldsymbol{H}(\cdot)$ is the regularization operator and λ is a regularization coefficient. Eq. 3 is a convex problem and thus can be solved using the Sinkhorn algorithm. With $\boldsymbol{K} = \exp(-\boldsymbol{C}/\lambda)$, the regularized optimal transport can be computed by:

$$\boldsymbol{P}^* = \text{diag}(\boldsymbol{a})\boldsymbol{K}\text{diag}(\boldsymbol{b}), \tag{4}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are marginal constraints:

$$\boldsymbol{a} \leftarrow \boldsymbol{\mu}/\boldsymbol{K}\boldsymbol{b}, \quad \boldsymbol{b} \leftarrow \boldsymbol{\nu}/\boldsymbol{K}^\top \boldsymbol{a}. \tag{5}$$

3.2. Spatio-Temporal Descriptor

In addressing the low semantic distinction of video categories, our objective is to disentangle category names into *Spatio-Temporal Descriptors*. We posit that each type of descriptor yields information that is complementary to the other. *Spatio Descriptors* are intended to capture static visual elements that can be discerned from a single image—such as settings and common objects. For *Temporal Descriptors*, we aim to decompose the action classes in a step-by-step manner to describe the temporal evolution of an action. We use OpenAI's API for GPT-3.5 [4] with a temperature of 0.7 to generate corresponding descriptors.

To generate *Spatio Descriptors*, inspired by [19], we use the following prompt $\mathcal{P}^s(\cdot)$ with category name $\boldsymbol{cls}$ to query LLM: "*Please give me a long list of descriptors for action:* $\{\boldsymbol{cls}\}$*,* N_s *descriptors in total.*"[1]. This prompt enables the LLM to always return a list with N_s descriptors. This process can be formulated as:

$$\boldsymbol{Des^s} = \boldsymbol{LLM}[(\mathcal{P}^s(\boldsymbol{cls}))], \tag{6}$$

[1]For a detailed demonstration of prompts we used, please refer to *Supplementary Material*.

For *Temporal Descriptors*, we utilize the temporal prompt $\mathcal{P}^t(\cdot)$ as "*Please give me a long list of decompositions of steps for action:* $\{\boldsymbol{cls}\}$*,* N_t *steps in total*" and obtain N_t descriptors:

$$\boldsymbol{Des^t} = \boldsymbol{LLM}[(\mathcal{P}^t(\boldsymbol{cls}))]. \tag{7}$$

Nonetheless, our empirical study (*please refer to Sec.4.2*) indicates that the direct application of temporal descriptors $\boldsymbol{Des^t}$, yields only marginal enhancements. As discussed in [22, 34, 39], image-text pre-trained models are less sensitive to verbs. The initial semantic space of the temporal descriptors generated by CLIP might be limited. Thus, we adopt a hard prompt: "*A video of* $\{\boldsymbol{cls}\}$ *usually includes* $\{\boldsymbol{Des^t}\}$" to condition temporal descriptors on the category names. We find this operation brings consistent improvements in recognition tasks.

Through this approach, we can disentangle the category name into two complementary semantic spaces. This disentanglement significantly mitigates the semantic similarity among class names and also serves sufficient knowledge for general recognition.

3.3. Optimal Descriptor Solver

A considerable number of transformer-based video recognition pipelines obtain video-level representation via pooling over image-level [CLS] tokens and then classify the video into a category by calculating the matching score using cosine similarity with category embeddings [47, 53, 56, 60], this pipeline can be formulated as:

$$\boldsymbol{S}_k = cos(\overline{\boldsymbol{V}}, \boldsymbol{Cat_k}), \tag{8}$$

where $cos(\cdot,\cdot)$ is the cosine similarity, $\boldsymbol{V} \in \mathbb{R}^{T\times d}$ is a set of local representations with T frames in total, $\boldsymbol{Cat_k} \in \mathbb{R}^d$ is category embedding for each class. As discussed before, only relying on the understanding of category names may lead to a less distinctive semantic space. After obtaining *Spatio-Temporal Descriptors* introduced in Sec. 3.2, an intuitive operation is to form a global-level descriptor embedding to benefit visual recognition:

$$\boldsymbol{S^s_{k}}_{pool} = cos(\overline{\boldsymbol{V}}, \overline{\boldsymbol{D^s_k}}), \quad \boldsymbol{S^t_{k}}_{pool} = cos(\overline{\boldsymbol{V}}, \overline{\boldsymbol{D^t_k}}), \tag{9}$$

where $\boldsymbol{D_k} \in \mathbb{R}^{N\times d}$ is the embedding of *Spatio-Temporal Descriptors*. By pooling along the N dimension, we can obtain the discriminative global descriptor embedding. However, we find this formation can lead to sub-optimal performances: 1) By averaging the descriptor-level representations, the model treats all of the attributes equally. Since the descriptors are generated by an autoregressive language model without instance-level knowledge, these descriptors may not be contained in every video. 2) The hallucination problem of LLMs may bring noises to the descriptor.

Table 1. Comparisons with state-of-the-art methods for zero-shot video recognition on HMDB51, UCF101 and Kinetics-600. We report Top-1 and Top-5 accuracy using single-view inference.

Method	Venue	Encoder	Frames	HMDB-51	UCF-101	K600 (Top-1)	K600 (Top-5)
Uni-modal zero-shot video recognition models							
ER-ZSAR [11]	ICCV'21	TSM	16	35.3 ± 4.6	51.8 ± 2.9	42.1 ± 1.4	73.1 ± 0.3
JigsawNet [44]	ECCV'22	R(2+1)D	16	38.7 ± 3.7	56.0 ± 3.1	-	-
Adapting pre-trained CLIP							
Vanilla CLIP [46]	ICML'21	ViT-B/16	32	40.8 ± 0.3	63.2 ± 0.2	59.8 ± 0.3	83.5 ± 0.2
ActionCLIP [56]	arXiv'21	ViT-B/16	32	40.8 ± 5.4	58.3 ± 3.4	66.7 ± 1.1	91.6 ± 0.3
Vita-CLIP [58]	CVPR'23	ViT-B/16	8 / 32	48.6 ± 0.6	75.0 ± 0.6	67.4 ± 0.5	-
A5 [26]	ECCV'22	ViT-B/16	32	44.3 ± 2.2	69.3 ± 4.2	55.8 ± 0.7	81.4 ± 0.3
XCLIP [41]	ECCV'22	ViT-B/16	32	44.6 ± 5.2	72.0 ± 2.3	65.2 ± 0.4	86.1 ± 0.8
DiST [45]	ICCV'23	ViT-B/16	32	$\underline{55.4} \pm 1.2$	72.3 ± 0.6	-	-
Tuning pre-trained CLIP							
ViFi-CLIP [47]	CVPR'23	ViT-B/16	32	51.3 ± 0.7	76.8 ± 0.8	71.2 ± 1.0	92.2 ± 0.3
MAXI [34]	ICCV'23	ViT-B/16	16 / 32	52.3 ± 0.6	$\underline{78.2} \pm 0.7$	$\underline{71.5} \pm 0.8$	$\underline{92.5} \pm 0.4$
OST	CVPR'24	ViT-B/16	8	54.9 ± 1.1	77.9 ± 1.3	73.9 ± 0.8	94.1 ± 0.3
			32	$\mathbf{55.9} \pm 1.2$	$\mathbf{79.7} \pm 1.1$	$\mathbf{75.1} \pm 0.6$	$\mathbf{94.6} \pm 0.2$

Hence, a natural question arises: *how can we assign optimal descriptors for each video instance?* In this regard, we introduce *Optimal Descriptor Solver (OD Solver)*, by adapting optimal transport theory, we formulate the video-text matching problem as an optimal matching flow. After obtain a set of frame-level features $\boldsymbol{V} \in \mathbb{R}^{T\times d}$ and descriptor-level embedding for each class $\boldsymbol{D^s_k} \in \mathbb{R}^{N_s\times d}$, $\boldsymbol{D^t_k} \in \mathbb{R}^{N_t\times d}$. The cost matrix for each class can be defined as:

$$\boldsymbol{C^s_k} = 1 - cos(\boldsymbol{V}, \boldsymbol{D^s_k}), \quad \boldsymbol{C^t_k} = 1 - cos(\boldsymbol{V}, \boldsymbol{D^t_k}). \tag{10}$$

According to Eq. 3, the entropy-regularized OT problem can be defined as:

$$\begin{aligned} \boldsymbol{P}^* &= \underset{\boldsymbol{P}\in\mathbb{R}^{T\times N}}{\arg\min} \sum_{i=1}^{T}\sum_{j=1}^{N} \boldsymbol{P}_{ij}\boldsymbol{C}_{ij} - \lambda \boldsymbol{H}(\boldsymbol{P}) \\ \text{s.t.} \quad & \boldsymbol{P}\boldsymbol{e} = \boldsymbol{\mu}, \quad \boldsymbol{P}^\top \boldsymbol{e} = \boldsymbol{\nu}. \end{aligned} \tag{11}$$

We can obtain the optimal transport plan $\boldsymbol{P^{s}_k}^*$ and $\boldsymbol{P^{t}_k}^*$ for *Spatio-Temporal Descriptors* respectively by solving the convex problem in Eq. 11 via the Sinkhorn algorithm as defined in Eq. 4. Here $\boldsymbol{P_k}^* \in \mathbb{R}^{T\times N}$ denotes the optimal matching flow between the video and descriptors. The matching score based on the optimal matching flow can be obtained via Frobenius inner product:

$$\begin{aligned} \boldsymbol{S^s_{k}}_{OT} &= \sum_{i=1}^{T}\sum_{j=1}^{N} \boldsymbol{P^{s*}_{k\,ij}} cos(\boldsymbol{V}_i, \boldsymbol{D^s_{kj}}), \\ \boldsymbol{S^t_{k}}_{OT} &= \sum_{i=1}^{T}\sum_{j=1}^{N} \boldsymbol{P^{t*}_{k\,ij}} cos(\boldsymbol{V}_i, \boldsymbol{D^t_{kj}}). \end{aligned} \tag{12}$$

By fusing the overall matching score in the Euclidean space and Wasserstein space described in Eq. 9 and Eq. 12 respectively, the overall logits can be expressed as:

$$\boldsymbol{S_{k\mathbf{OD}}} = \frac{1}{4}(\boldsymbol{S^s_{k}}_{pool} + \boldsymbol{S^t_{k}}_{pool} + \boldsymbol{S^s_{k}}_{OT} + \boldsymbol{S^t_{k}}_{OT}). \tag{13}$$

Please refer to Supplementary Material for pseudo-codes.

3.4. Training Objectives

Considering the overall logits calculated by *OD Solver* in Eq. 13 can be described as video-to-text logits $\boldsymbol{S^{v2t}_k}_{\mathbf{OD}} = \boldsymbol{OD}(\boldsymbol{V}, \boldsymbol{D^{s,t}_k})$. A symmetric text-to-video logits can be obtained via a similar way $\boldsymbol{S^{t2v}_k}_{\mathbf{OD}} = \boldsymbol{OD}(\boldsymbol{D^{s,t}_k}, \boldsymbol{V})$. Then, the softmax-normalized similarity scores can be expressed as:

$$\begin{aligned} \boldsymbol{p^{v2t}_i}_{\mathbf{OD}} &= \frac{1}{K}\sum_{k=1}^{K} \frac{exp(\boldsymbol{S^{v2t}_{ki}}_{\mathbf{OD}}/\tau)}{\sum_{j=1}^{B} exp(\boldsymbol{S^{v2t}_{kj}}_{\mathbf{OD}}/\tau)}, \\ \boldsymbol{p^{t2v}_i}_{\mathbf{OD}} &= \frac{1}{K}\sum_{k=1}^{K} \frac{exp(\boldsymbol{S^{t2v}_{ki}}_{\mathbf{OD}}/\tau)}{\sum_{j=1}^{B} exp(\boldsymbol{S^{t2v}_{kj}}_{\mathbf{OD}}/\tau)}, \end{aligned} \tag{14}$$

where τ refers to the temperature hyperparameter for scaling, B is the number of samples in the current mini-batch, and K is the number of classes. Let $\boldsymbol{q^{v2t}}, \boldsymbol{q^{t2v}}$ denotes the ground-truth similarity scores, we can define the Kullback-Leibler (KL) divergence [32] as the overall contrastive loss to optimize the model as:

$$\mathcal{L}_{\mathbf{OD}} = \frac{1}{2}[KL(\boldsymbol{p^{v2t}}_{\mathbf{OD}}, \boldsymbol{q^{v2t}}) + KL(\boldsymbol{p^{t2v}}_{\mathbf{OD}}, \boldsymbol{q^{t2v}})]. \tag{15}$$

4. Experiments

Datasets. We conduct experiments across 6 video benchmarks: Kinetics-400 [7] & 600 [8], UCF-101 [50], HMDB-51 [31], Something-Something V2 [20], and ActivityNet [6]. Our investigation encompasses various settings, including zero-shot, few-shot, and fully-supervised video recognition. *See Supplementary Material for details.*

Implementation Details. We employ a CLIP ViT-B/16 to conduct both zero-shot and few-shot experiments. We generate $N_{s,t} = 4$ descriptors for each category. Following [24, 34, 59], we perform a linear weight-space ensembling between the original CLIP and the finetuned model with a ratio of 0.2. *See Supplementary Material for details.*

Table 2. Comparisons with state-of-the-art methods for few-shot video recognition on HMDB51, UCF101 and Something-Something V2. We scaled up the task to categorize all categories in the dataset with only a few samples per category for training. Here K denotes training samples for each class. We report Top-1 accuracy using single-view inference.

Method	HMDB-51				UCF-101				SSv2			
	K=2	K=4	K=8	K=16	K=2	K=4	K=8	K=16	K=2	K=4	K=8	K=16
Directly tuning on CLIP												
Vanilla CLIP [46]	41.9	41.9	41.9	41.9	63.6	63.6	63.6	63.6	2.7	2.7	2.7	2.7
ActionCLIP [56]	47.5	57.9	57.3	59.1	70.6	71.5	73.0	91.4	4.1	5.8	8.4	11.1
XCLIP [41]	53.0	57.3	62.8	64.0	48.5	75.6	83.7	91.4	3.9	4.5	6.8	10.0
A5 [26]	39.7	50.7	56.0	62.4	71.4	79.9	85.7	89.9	4.4	5.1	6.1	9.7
ViFi-CLIP [47]	<u>57.2</u>	<u>62.7</u>	<u>64.5</u>	<u>66.8</u>	<u>80.7</u>	<u>85.1</u>	<u>90.0</u>	<u>92.7</u>	<u>6.2</u>	<u>7.4</u>	<u>8.5</u>	**12.4**
OST	**59.1**$_{+1.9}$	**62.9**$_{+0.2}$	**64.9**$_{+0.4}$	**68.2**$_{+1.4}$	**82.5**$_{+1.8}$	**87.5**$_{+2.4}$	**91.7**$_{+1.7}$	**93.9**$_{+1.2}$	**7.0** $_{+0.8}$	**7.7** $_{+0.3}$	**8.9** $_{+0.4}$	<u>12.2</u>
Fine-tuned on K400												
ViFi-CLIP [47]	55.8	<u>60.5</u>	64.3	65.4	84.0	86.5	90.3	92.8	6.6	6.8	8.6	11.0
MAXI [34]	<u>58.0</u>	60.1	<u>65.0</u>	<u>66.5</u>	<u>86.8</u>	<u>89.3</u>	<u>92.4</u>	<u>93.5</u>	<u>7.1</u>	<u>8.4</u>	<u>9.3</u>	<u>12.4</u>
OST	**64.8**$_{+6.8}$	**66.7**$_{+6.2}$	**69.2**$_{+4.2}$	**71.6**$_{+5.1}$	**90.3**$_{+3.5}$	**92.6**$_{+3.3}$	**94.4**$_{+2.0}$	**96.2**$_{+2.7}$	**8.0** $_{+0.9}$	**8.9** $_{+0.5}$	**10.5**$_{+1.2}$	**12.6**$_{+0.2}$

Table 3. Fully-supervised video recognition on Kinetics-400, Something-Something V2 and ActivityNet. We report Top-1 accuracy using single-view inference.

Method	Encoder - Frames			
	B/32 - 8	B/32 - 16	B/16 - 8	B/16 - 16
Kinetics-400				
Text4Vis [60]	78.5	79.3	81.4	82.6
OST	**78.7**(+0.2)	**79.8**(+0.5)	**82.0**(+0.6)	**83.2**(+0.6)
Something-Something V2				
Text4Vis [60]	54.3	56.1	57.9	59.9
OST	**54.4**(+0.1)	**56.4**(+0.3)	**58.4**(+0.5)	**60.3**(+0.4)
ActivityNet				
Text4Vis [60]	83.4	85.0	86.4	88.4
OST	**84.0**(+0.6)	**85.8**(+0.8)	**87.1**(+0.7)	**88.7**(+0.3)

4.1. Main Results

Zero-shot video recognition. We present our zero-shot video recognition results and compare our approach with SOTAs in Table 1. The model is first fine-tuned on the Kinetics400 dataset and evaluated directly on downstream datasets to ascertain its generalization capacity with respect to unseen classes. Our approach outperforms regular uni-modal zero-shot video recognition pipelines by a large margin as shown in the upper table. Moreover, we draw comparisons with methods that use K400 to adapt CLIP models for zero-shot recognition. Noteworthy among these are methods that integrate additional temporal learners [26, 41, 45] or employ VL prompting techniques [26, 58]. Contrary to these approaches, our pipeline leverages refined textual knowledge to boost video recognition without altering the underlying architecture. We observe consistent improvements in all datasets with respect to these methods.

We further compare our method with other fully finetuning paradigms [34, 47]. Serving as a baseline to our method, ViFi-CLIP [47] relies on the direct utilization of category names to fine-tune the CLIP model. Notably, utilizing only 8 frames for training and validation, our method demonstrates competitive performance, surpassing our baseline by a large margin. Upon scaling up the input frames to 32, our method consistently exhibits improvements across all datasets in comparison to prior SOTAs. Even against MAXI [34] which leverages more diverse textual knowledge, such as frame-level captions, our approach showcases superior accuracy with a 3.6% improvement on HMDB, 1.5% on UCF, and 3.6% on K600.

Few-shot video recognition. We demonstrate our method's learning capacity and generalizability under the challenging all-way few-shot regime. The Top-1 accuracy on three datasets is reported in Table 2. We conduct experiments in two different aspects. We first conduct an experiment that directly tunes CLIP for few-shot recognition. Our method shows consistent improvement over our baseline [47] on HMDB-51, UCF101, and even temporal-heavy dataset SSv2.

Following [34], we adopt our best model in zero-shot settings to further verify our method's generalization capacity. As a comparison, ViFi-CLIP shows degraded performance in this fashion (*e.g.* $K = 4$ on UCF, $K = 16$ on SSv2). In this regime, our method outperforms the unsupervised contrastive training framework MAXI [34] in different shot settings by an average of $\sim$5% on HMDB, $\sim$3% on UCF, and $\sim$1% on SSv2. This indicates the generalizability of our pipeline in the extremely low-shot settings.

Fully-supervised video recognition. We also conduct fully-supervised experiments on three large-scale video benchmarks Kinetics-400, Something-Something V2, and ActivityNet to validate the effectiveness of our method in supervised settings. Serving as a standard pipeline to adapt pre-trained vision-language models for supervised video recognition, we choose Text4Vis [60] as our baseline and vary different encoders ViT-B/32, and ViT-B/16 with 8, and 16 frames, respectively. As shown in Table 3, we find our method improves upon our corresponding baseline for all different architectures on all datasets. We can see that the

Table 4. Ablation studies. We utilize ViT-B/16 as the backbone and use 8 frames for training/validation unless otherwise specified. All of the performances are top-1 accuracy (%) in the zero-shot setting using single-view inference and spatial size of 224×224.

Method	HMDB	UCF	K600
Category Name [47]	50.9	75.5	70.8
Descriptors*	53.3 (+2.4)	76.6 (+1.1)	69.3
OD Solver	**54.5** (+3.6)	**77.9** (+2.4)	**72.3** (+1.5)

(a) Study on cross-modal matching mechanisms. Here we apply the number of descriptors $N_{s,t} = 4$. * denotes pooling descriptors along with category names.

Spatio	Temporal	HMDB	UCF	K600
✓	✗	46.7	65.3	56.3
✗	✓	53.1	77.5	71.6
✓	✓	**54.5**	**77.9**	**72.3**

(b) The impact of different descriptors. Here ✓ means applying corresponding Spatio/Temporal descriptors.

N	HMDB	UCF	K600
2	53.8	77.3	72.1
4	**54.5**	**77.9**	72.3
8	53.0	77.5	**72.6**

(c) Comparisons between different number of descriptors N.

Spatio	Temporal	HMDB	UCF	K600
✗	✗	49.8	74.1	64.2
✓	✗	53.5	**79.0**	71.8
✓	✓	53.5	78.9	72.1
✗	✓	**54.5**	77.9	**72.3**

(d) Study on category conditioning operation. ✓ means conditioning corresponding descriptors on category names.

Ensemble	HMDB	UCF	K600
✗	55.4	80.1	72.9
✓	**55.9**	79.7	**75.1**

(e) The effects of weight-space ensembling. ✓ means perform ensemble with a ratio of 0.2. 32 frames are used during training/validation.

Table 5. Additional cost analysis of our method, we report step latency during training, and throughput (TP) during inference. We refer to Top-1 as zero-shot accuracy on Kinetics-600.

Method	Top-1 (%)	Latency (*s*)	TP (*video/s*)
ViFi-CLIP [47]	71.2	**0.40** (1.0×)	**40.9** (1.00×)
OST	**75.1**	0.44 (1.1×)	40.0 (0.98×)

performance on K400 and SSv2 is about 0.5% higher than Text4Vis [60]. For ActivityNet, the accuracy is even 0.8% higher than our counterparts.

4.2. Ablation Studies

We conduct ablation studies on zero-shot settings in Table 4 to investigate our **OST**'s learning capacity and generalizability in different instantiations.

Different cross-modal matching mechanisms. Table 4a shows the effects of different cross-modal matching mechanisms. For a fair comparison, we start with a baseline that uses the category name during matching as [47]. By simply aggregating the *Descriptors* along with the category name via mean pooling, the accuracy on HMDB and UCF improved by 2.4% and 1.1%, respectively. However, on the K600 dataset, we observe a 1.5% performance drop. This validates our hypothesis that the enhanced distinction brought by pooling operation can benefit downstream recognition, but might not be optimal. We then introduce our *OD Solver* to solve the optimal matching flow, we find that our approach can further boost the performance on HMDB and UCF, and achieve a remarkable improvement of 1.5% on the large-scale dataset K600. Notably, the categories in the K600 validation set are more complicated compared to HMDB and UCF. This validates our *OD Solver*'s effectiveness, especially in complicated open-vocabulary settings.

The impact of different descriptors. We investigate the impact of *Spatio-Temporal Descriptors* on the performance of our proposed method. The results shown in Table 4b demonstrate that each descriptor is complementary to others. Indicating that both *Spatio* and *Temporal Descriptors* provide crucial information for recognition tasks. We also observe that the effect of temporal descriptors is more convincing compared to *Spatio Descriptors*.

Numbers of descriptors. We investigate the influences of varying the number of descriptors N in Table 4c. We conducted experiments with 2, 4, and 8 *Spatio-Temporal Descriptors*. We can observe that the performance reaches its peak at $N_{s,t} = 4$. We've further checked the quality[2] of descriptors when varying N. We find that 2 descriptors can not afford enough information to supply cross-modal matching. When the number of descriptors reaches 8, the hallucination problem of LLM becomes more severe, resulting in a significant amount of noisy descriptors. In this case, we set N as 4 in our basic settings.

The impact of conditioning descriptors on category names. We study the effect of conditioning descriptors on category names on the final zero-shot accuracy. Table 4d shows that conditioning temporal descriptors on category names can achieve the best performances while conditioning both descriptors may lead to performance degradation. This further indicates the points framed in [22, 34, 39] that visual-language pre-trained models are less sensitive to verbs. As a result, the category conditioning technique can ensure the semantic distribution of the *Temporal Descriptors* clustered well, making the optimization process smoother.

The effects of weight-space ensembling. We investigate the effects of the linear weight-space ensembling technique. As shown in Table 4e, the ensembling technique greatly mitigates the catastrophic forgetting problem, especially on the large-scale Kinetics-600 dataset, where the zero-shot accuracy is improved by 2.2%.

[2]Please refer to *Supplementary Material* for examples of descriptors.

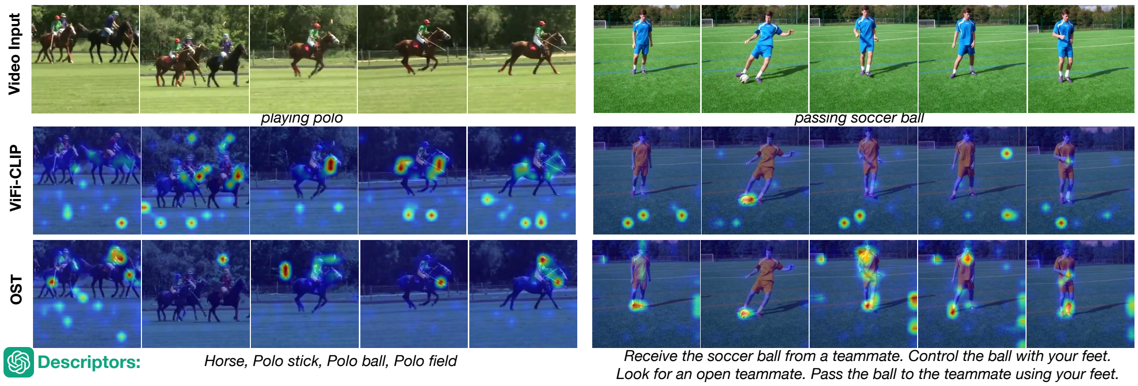

Figure 4. Attention map on K600 validation set. We demonstrate *Spatio Descriptors* and *Temporal Descriptors* on the left and right, respectively. **(Left):** For videos that can be recognized via static frames, our **OST** attends to the certain object more while ViFi-CLIP [47] is often distracted by the backgrounds. **(Right):** For classes that require more temporal clues, ViFi-CLIP [47] attends to appearance (*e.g.* soccer ball and soccer field) more, while our **OST** shows consistent attention to the body's temporal salient parts such as the player's feet.

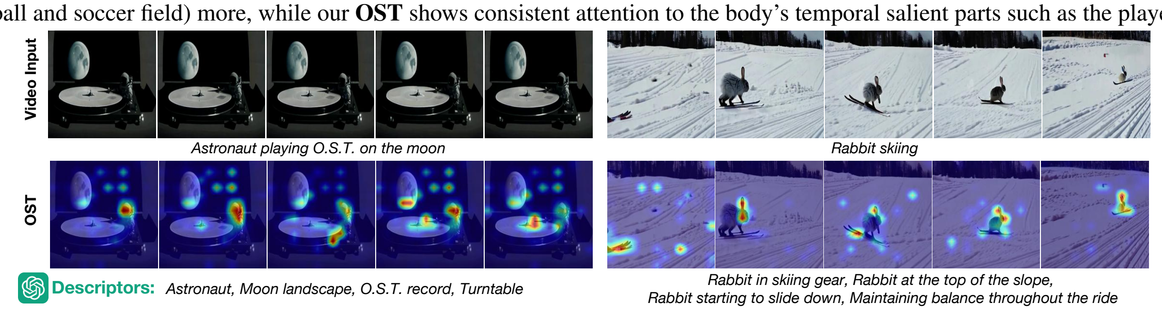

Figure 5. Generalization on extreme outliers. We utilize the text-to-video diffusion model Show-1 [67] to generate synthetic videos with a semantic distribution distinct from the fine-tuning data in Kinetics-400 to further demonstrate the generalizability of our method. Attention map for *Spatio Descriptors* and *Temporal Descriptors* are visualized on the left and right, respectively.

4.3. Cost Analysis

We analyze the additional cost of our method during training and inference in Table 5. Latency is measured in our basic training setting and throughput is measured using the largest possible batch size before running out of memory with a single NVIDIA 4090-24G. Notably, the original implementation of ViFi-CLIP [47] utilizes cross-entropy loss and maintains the logits for all categories in every mini-batch during training, leading to a larger latency. For a fair comparison, we re-implement ViFi-CLIP with local infoNCE-styled loss [56] to analyze the training cost. Our pipeline only requires an extra $0.1\times$ training time and reduces the throughput by about 2%, which is acceptable given the improvement in performance.

4.4. Visualizations

We conduct a qualitative study on the attention map of our **OST** in the zero-shot setting. As depicted in Fig. 4, compared to our baseline ViFi-CLIP [47] our method can not only focus on varied spatial cues but also consistently attend to temporal salient elements (*e.g.* the player's feet) for videos that include more scene dynamics. Additionally, we investigate the attention map of our method on extreme outlier samples in Fig. 5. Our empirical findings indicate that out **OST** upholds robust generalization capabilities, even in extreme out-of-distribution examples. *Please refer to Supplementary Material for more qualitative results.*

5. Conclusion

In this work, we introduce a novel general video recognition pipeline **OST**. We prompt an LLM to augment category names into *Spatio-Temporal Descriptors* and refine the semantic knowledge via *Optimal Descriptor Solver*. Comprehensive evaluations in six datasets and three different tasks demonstrate the effectiveness of our approach.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant U2013203, 62373140, U21A20487, U1913202, 62103137, and 61973106, the National Key Research and Development Program of China(2023YFB4704503), the Project of Science Fund for Distinguished Young Scholars of Hunan Province (2021JJ10024), the Leading Talents in Science and Technology Innovation of Hunan Province (2023RC1040), the Natural Science Fund of Hunan Province (2022JJ40100, 2022JJ30024), the Project of Talent Innovation and Sharing Alliance of Quanzhou City(2021C062L), the Key Research and Development Project of Science and Technology Plan of Hunan Province (2022GK2014). The work was done while Tongjia was a research intern mentored by Chen Chen. We thank Ming Li and Yong He for proofreading and discussion.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017. 3

[2] Federico Becattini, Tiberio Uricchio, Lorenzo Seidenari, Lamberto Ballan, and Alberto Del Bimbo. Am i done? predicting action progress in videos. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(4):1–24, 2020. 3

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

[4] 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. *NeurIPS*, 2020. 4

[5] Adrian Bulat and Georgios Tzimiropoulos. Lasp: Text-to-text optimization for language-aware soft prompting of vision & language models. In *CVPR*, 2023. 1

[6] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 2, 5

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 2, 3, 5

[8] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 2, 5

[9] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Prompt learning with optimal transport for vision-language models. In *ICLR*, 2023. 3

[10] Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. Graph optimal transport for cross-domain alignment. In *ICML*, 2020. 3

[11] Shizhe Chen and Dong Huang. Elaborative rehearsal for zero-shot action recognition. In *ICCV*. 5, 2

[12] Tom Tongjia Chen, Hongshan Yu, Zhengeng Yang, Ming Li, Zechuan Li, Jingwen Wang, Wei Miao, Wei Sun, and Chen Chen. First place solution to the cvpr'2023 aqtc challenge: A function-interaction centric approach with spatiotemporal visual-language alignment. *arXiv preprint arXiv:2306.13380*, 2023. 3

[13] Yifei Chen, Dapeng Chen, Ruijin Liu, Hao Li, and Wei Peng. Video action recognition with attentive semantic units. In *ICCV*, 2023. 3

[14] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *NeurIPS*, 2017. 3

[15] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *NeurIPS*, 2013. 4, 1, 2

[16] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *ECCV*, 2018. 3

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2

[18] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, 2021. 3

[19] Zhili Feng, Anna Bair, and J Zico Kolter. Leveraging multiple descriptive features for robust few-shot image learning. *arXiv preprint arXiv:2307.04317*, 2023. 4

[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*, 2017. 2, 5

[21] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *NeurIPS*, 2017. 3

[22] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*, 2021. 4, 7

[23] Rui Hou, Chen Chen, and Mubarak Shah. Tube convolutional neural network (t-cnn) for action detection in videos. In *ICCV*, 2017. 3

[24] Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights. *NeurIPS*, 2022. 5, 2

[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*, 2021. 1

[26] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *ECCV*, 2022. 3, 5, 6

[27] Leonid V Kantorovich. On the translocation of masses. *Journal of mathematical sciences*, 133(4):1381–1382, 2006. 3

[28] Prannay Kaul, Weidi Xie, and Andrew Zisserman. Multi-modal classifiers for open-vocabulary object detection. *arXiv preprint arXiv:2306.05493*, 2023. 2, 3

[29] Abdelwahed Khamis, Russell Tsuchida, Mohamed Tarek, Vivien Rolland, and Lars Petersson. Earth movers in the big data era: A review of optimal transport in machine learning. *arXiv preprint arXiv:2305.05080*, 2023. 3

[30] Kwanyoung Kim, Yujin Oh, and Jong Chul Ye. Zegot: Zero-shot segmentation through optimal transport of text prompts. *arXiv preprint arXiv:2301.12171*, 2023. 3

[31] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, 2011. 2, 5

[32] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86, 1951. 5

[33] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, 2019. 3

[34] 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. *arXiv preprint arXiv:2303.08914*, 2023. 3, 4, 5, 6, 7, 2

[35] 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*, 2022. 3

[36] Benlin Liu, Yongming Rao, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Multi-proxy wasserstein classifier for image classification. In *AAAI*, 2021. 3

[37] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, 2022. 3

[38] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *ICLR*, 2023. 2, 3

[39] Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. Verbs in action: Improving verb understanding in video-language models. In *ICCV*, 2023. 4, 7

[40] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781. 3

[41] 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*, 2022. 1, 3, 5, 6, 2

[42] Hermina Petric Maretic, Mireille El Gheche, Giovanni Chierchia, and Pascal Frossard. Got: an optimal transport framework for graph comparison. *NeurIPS*, 2019. 3

[43] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *CVPR*, 2023. 3

[44] Yijun Qian, Lijun Yu, Wenhe Liu, and Alexander G Hauptmann. Rethinking zero-shot action recognition: Learning from latent atomic actions. In *ECCV*, 2022. 5

[45] 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. In *ICCV*, 2023. 1, 3, 5, 6

[46] 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*, 2021. 1, 5, 6

[47] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *CVPR*, 2023. 4, 5, 6, 7, 8, 2

[48] Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving gans using optimal transport. *arXiv preprint arXiv:1803.05573*, 2018. 3

[49] Richard Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967. 1, 2

[50] 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

[51] Ombretta Strafforello, Xin Liu, Klamer Schutte, and Jan van Gemert. Video bagnet: short temporal receptive fields increase robustness in long-term action recognition. In *ICCV*, 2023. 3

[52] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018. 3

[53] Shuyuan Tu, Qi Dai, Zuxuan Wu, Zhi-Qi Cheng, Han Hu, and Yu-Gang Jiang. Implicit temporal modeling with learnable alignment for video recognition. *arXiv preprint arXiv:2304.10465*, 2023. 1, 3, 4

[54] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008. 2

[55] 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*, 2016. 3

[56] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 1, 3, 4, 5, 6, 8

[57] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 1

[58] Syed Talal Wasim, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Vita-clip: Video and text adaptive clip via multimodal prompting. In *CVPR*, 2023. 5, 6

[59] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *CVPR*, 2022. 5, 2

[60] Wenhao Wu, Zhun Sun, and Wanli Ouyang. Revisiting classifier: Transferring vision-language models for video recognition. In *AAAI*, 2023. 1, 2, 3, 4, 6, 7

[61] Wenhao Wu, Xiaohan Wang, Haipeng Luo, Jingdong Wang, Yi Yang, and Wanli Ouyang. Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. In *CVPR*, 2023.

[62] 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 representation alignment. *arXiv preprint arXiv:2209.06430*, 2022. 3

[63] Taojiannan Yang, Sijie Zhu, Matias Mendieta, Pu Wang, Ravikumar Balakrishnan, Minwoo Lee, Tao Han, Mubarak Shah, and Chen Chen. Mutualnet: Adaptive convnet via mutual learning from different model configurations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[64] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. Aim: Adapting image models for efficient video action recognition. *ICLR*, 2023. 3

[65] 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

[66] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *CVPR*, 2020. 3

[67] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023. 8

[68] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *CVPR*, 2017. 3

[69] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023. 2

[70] Jiaming Zhou, Kun-Yu Lin, Yu-Kun Qiu, and Wei-Shi Zheng. Twinformer: Fine-to-coarse temporal modeling for long-term action recognition. *IEEE Transactions on Multimedia*, 2023. 3