2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

# Incorporating Scene Graphs into Pre-trained Vision-Language Models for Multimodal Open-vocabulary Action Recognition

Chao Wei and Zhidong Deng*

*Abstract*--- **This paper presents Action-SGFA, a novel action feature alignment approach to learn unified joint embeddings across four action modalities incorporating scene graph (SG) comprehension. A new training paradigm for Action-SGFA is also devised to improve pre-trained VL models using datasets with SG annotation. When learning from image-SG pairs, it captures structure-associated action knowledge for visual and textual encoders. SG supervision generates fine-grained captions based on various graph augmentations highlighting different compositional aspects of action scenes. Furthermore, our research reveals that all combinations of paired data are unnecessary to train such unified embeddings, and only image-paired data is sufficient to bind all action modalities together. Our Action-SGFA can leverage existing large VL models, enhancing their zero-shot capabilities of new modalities due to their natural pairings with images. The open-vocabulary zero-shot performance improves with the strength of the pre-trained VL model and the SG comprehension. We establish a new state-of-the-art in several zero-shot action recognition tasks across modalities, significantly surpassing the vanilla skeleton zero-shot method by 27.0% and 19.7% on NTU-60 and NTU-120, respectively. Additionally, in the context of RGB videos, we surpass the state-of-the-art method on Kinetics-400 by 2.1%.**

## I. INTRODUCTION

Action recognition has been an active research topic due to its diverse applications in human-computer interaction, sports analysis, and entertainment. Previous studies in this area can be categorized into two main approaches: those utilizing visual appearance information extracted from RGB videos [1]--[4] and those focusing on human skeleton data [5]--[8]. Notably, these approaches often treat actions as monolithic events [9]--[12]. Many existing models adopt end-to-end prediction strategies, assigning a single label to an extended video sequence [13]--[17] without explicitly decomposing events into a sequence of interactions between objects.

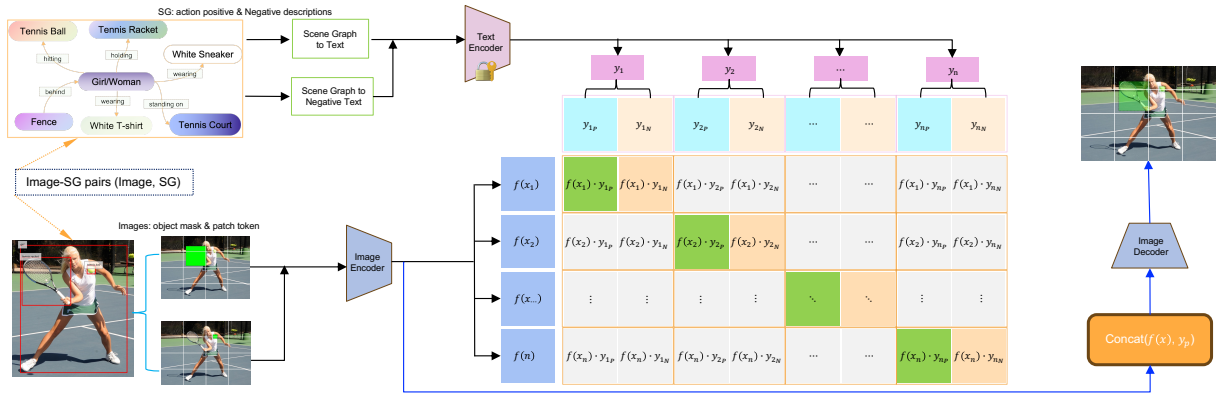
Understanding the structure of visual scenes is a fundamental problem in machine perception, and numerous prior studies [18]--[23] have extensively explored this domain. Datasets containing scene graph (SG) annotations (e.g., Visual Genome [24] and Action Genome [25]) have been collected and employed for training models to capture structural information within visual scenes. While they contribute to scene comprehension, these datasets are relatively small and expansive to collect compared to large-scale image-text pair datasets. Consequently, many large-scale vision and language (VL) models often overlook incorporating such SG data. Our approach demonstrates the feasibility of enhancing VL models by harnessing such specialized data through a tailored model architecture and a novel training paradigm.

In recent years, VL models like CLIP [26] and BLIP [27] have demonstrated remarkable performance across various tasks, showcasing extraordinary zero-shot capabilities in areas such as action recognition, visual question answering, and image captioning. These achievements are attributed to their training on extensive datasets containing image-text pairs, exemplified by LAION 400M [28]. While numerous methods focus on aligning image features with text [26], [28]--[34], audio [35]--[40], and other modalities, they primarily deal with single pairs or, at best, a limited set of visual modalities. Unfortunately, the final embeddings are limited to the pairs of modalities used for training. Consequently, image-skeleton embeddings cannot directly be used for image-text tasks and vice versa. A major obstacle in learning unified joint embedding is the absence of large quantities of multimodal data where all modalities are present together.

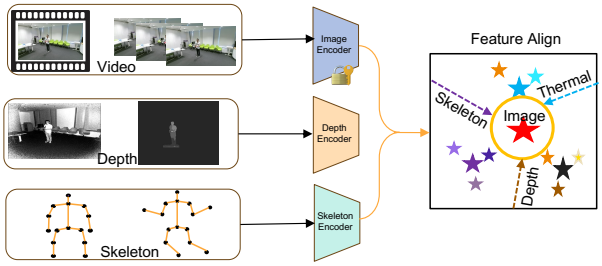
This study presents Action-SGFA, an innovative approach for multimodal open-vocabulary action recognition. Action-SGFA integrates SG comprehension and feature alignment across various action modalities. Due to the constrained ability of pre-trained VL models to comprehend action structures, we focus on fine-tuning these pre-trained models to capture action structure-related knowledge from SG datasets. However, for video, skeleton, depth, and thermal embeddings, which are not directly compatible with pre-trained VL models, Action-SGFA establishes a unified representation space by leveraging diverse image-paired data types. Consequently, Action-SGFA exhibits remarkable efficacy in zero-shot action recognition tasks across a diverse range of modalities, facilitated by its adeptness in visual SG comprehension and multimodal feature alignment (FA) for actions.

In the fine-tuning stage, our approach strategically incorporates components that directly supervise each representation during the learning process from SG paired with images. Initially, we transform image-SG pairs into image-text pairs, creating natural inputs for VL model training. This transformation ensures that the generated text accurately describes the structural nuances present in the SG. A notable advantage of this approach is the production of dense and comprehensive captions, surpassing those found in datasets like LAION. To introduce additional structural information, we leverage the SG to generate hard negatives. For instance, if an SG contains an edge like "men-hitting-ball," we manipulate it by reversing the edge to "ball-hitting-men" or by replacing an irrelevant word like "fence-hitting-ball," thereby creating a corresponding negative caption. A contrastive loss, computed

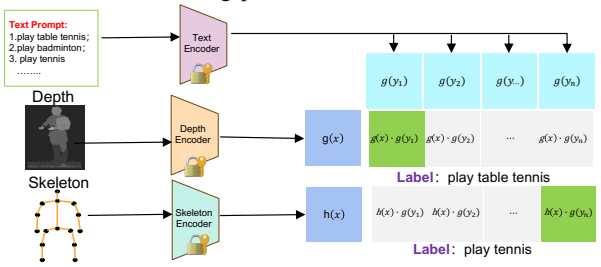
*The corresponding author. Authors are with BNRist, THUAI, Department of Computer Science, Tsinghua University, Beijing 100084, China. {weic18@mails, *michael@mail}.tsinghua.edu.cn




(a) Scene Graph Understanding: The pre-trained VL model underwent a fine-tuning process derived from SG datasets.




(b) Multimodal Action Feature Alignment.




(c) Open-vocabulary Zero-shot Action Recognition.

Fig. 1: The framework of our Action-SGFA: We incorporate SG information into pre-trained VL models, enhancing the capacity to comprehend the contextual aspects of human actions. Subsequently, we endeavor to learn a unified embedding space capable of accommodating diverse human action modalities, including skeleton, depth, and thermal, and leverage image/video-centered modality to align all of them. Ultimately, Action-SGFA possesses excellent zero-shot capabilities of new modalities due to their natural pairings with images.

between the negative and original captions, guides the model to focus on finer structural details. Subsequently, we incorporate these refined structural elements into the visual representation using MAE [41], thus facilitating the training of the visual model. The resulting Action-SGFA model is trained across all modalities using a unified objective function, eliminating the need for additional supervision.

In the multimodal feature alignment stage, Action-SGFA orchestrates the learning of a unified shared representation space by leveraging diverse image-paired data types. Harnessing the binding property of images, we demonstrate that aligning each modality's embedding to image embeddings fosters alignment across all modalities. Action-SGFA strategically combines web-scale (image, text) paired data with naturally occurring paired data, such as (image, skeleton), (image, depth), and more. This fusion enables the learning of a unified joint embedding space. Notably, Action-SGFA implicitly aligns text embeddings with other modalities, including skeleton, depth, etc., paving the way for zero-shot recognition capabilities on these modalities without the necessity for explicit semantic or textual pairing. Furthermore, we showcase the versatility of Action-SGFA by illustrating its initialization with large-scale VL models like CLIP [26], following fine-tuning with SG. This leverages the rich representations of these models in image, text, and image structure. Consequently, Action-SGFA proves applicable to various modalities and tasks with minimal additional training. We leverage extensive image-text paired data alongside naturally paired 'self-supervised' data, encompassing four novel modalities: skeleton, depth, and thermal. Our experiments demonstrate substantial improvements in zero-shot action recognition for each modality. Notably, performance gains increase as the quality of the underlying image representation improves.

To assess our approach's efficacy, we empirically evaluated action recognition tasks and compared them with competitive baselines on three popular benchmark datasets: Kinetics-400 [42], NTU RGB+D 60 [43], and NTU RGB+D 120 [44]. Experimental results show that our model achieves state-of-the-art performance on all three datasets.Our contributions can be summarized as follows: (1) We propose a novel Action-SGFA approach that harnesses structured SG annotations to enrich pre-trained large VL models, which captures structure-associated information for both text and image encoders through direct supervision of visual and textual components when learning from SG labels. (2) A new multi-modality training paradigm for Action-SGFA, which leverages a large fine-tuned VL model to implicitly align text embeddings to other modalities such as skeleton and depth, is also devised to enable zero-shot recognition capabilities on those modalities without explicit semantic or textual pairing. (3) Our Action-SGFA yields state-of-the-art zero-shot performance on three prominent action recognition benchmarks: NTU RGB+D 60, NTU RGB+D 120, and Kinetics-400 achieving accuracy gains of 27.0%, 19.7%, and 2.1%, respectively.

## II. Related work

### A. Vision and Language (VL) Models

The incorporation of linguistic signals, such as words or sentences, during image training has proven compelling for zero-shot, open-vocabulary recognition, and text-to-image retrieval [45]--[48]. Leveraging language as supervision not only facilitates these tasks but also contributes to learning robust image representations [49]--[51]. Recent advancements in this domain, including CLIP [26], ALIGN [33], and Florence [32], have curated extensive collections of image-text pairs. These models, trained through contrastive learning, embed image and language inputs into a joint space, showcasing impressive zero-shot performance.

The introduction of CoCa [52] further enhances performance by incorporating an image captioning objective alongside the contrastive loss. LiT [53], adopting contrastive training for fine-tuning, underscores the optimal results achieved by freezing image encoders. It is noteworthy that while prior research primarily concentrated on integrating image and text modalities, our work extends these principles to enable zero-shot recognition across multiple modalities.

### B. Learning Structured Representations

Structured representations have proven instrumental across various computer vision applications, contributing significantly to advancements in video understanding [54]--[56], relational reasoning [57]--[59], vision and language [60], human-object interactions [61], and image-video generation [62]. Over recent years, Scene Graphs (SGs) [20], [63] have emerged as powerful tools, providing semantic representations extensively applied across diverse applications.

Notably, this study underscores a remarkable discovery: even with a modest volume of SG annotations, when contrasted with vast repositories of image-text pair datasets, the infusion of structured knowledge into expansive Visual-Linguistic (VL) models becomes feasible.

## III. Method

This section presents the details of our proposed framework called Action-SGFA. We aim to integrate SG into pre-trained VL models, enhancing the capacity to comprehend the contextual aspects of human actions. Subsequently, we endeavor to learn a unified embedding space capable of accommodating diverse human action modalities, encompassing skeleton, depth, and thermal, by the utilization of image/video modality to align them. We show that the resulting embedding space has a robust zero-shot behavior that automatically associates pairs of modalities without seeing any training data for that specific pair. We illustrate our approach in Figure 1.

### A. Preliminaries

**Scene graph (SG).** A Scene Graph is formally represented as a tuple $G = (V, E)$, where: *(i) Nodes* $V$ Represents the set of $n$ objects in the scene. Each object node encompasses essential information, including a class label, a bounding box, and associated attributes. *-(ii) Edges* $E$ Encompasses the set of $m$ edges, signifying relationships within the scene. These relationships are defined as triplets $(i, e, j)$, where $i$ and $j$ denote object nodes, and $e$ represents the category of the relation between objects $i$ and $j$.

### B. SG for Vision-Language Models

*1) Structural Language Component:* We delve into the Structural Language Component, outlining how SG are transformed into text and subsequently manipulated with graph negatives to enhance the model's structural understanding.

**SG to text.** For a given image $I$ paired with its corresponding SG $G = (V, E)$ from the training dataset, we leverage the graph $G$ to create a textual caption. The process involves systematically traversing connected components within the graph. For each component, a textual caption is constructed by sequentially concatenating class labels extracted from graph edges along a Hamiltonian path. If attributes exist, they are prepended before the object class label. The final step involves generating a single caption by concatenating captions of connected components, separated by a dot.

**Graph negatives (GN).** Recognizing that using SG data solely as image-text pairs with a contrastive loss is insufficient for the model to develop structural understanding, we draw insights from contemporary research [64], [65]. Such studies reveal that conventional contrastive learning tends to overly focus on object labels, overlooking crucial aspects like relations and attributes. To address this, we leverage the SG structure and propose predefined graph-based rules. These rules modify SG, introducing semantic inconsistencies with the corresponding image. The transformed SG then serve as the basis for generating negative textual captions. This set of negatives, paired with a specified loss, motivates the model to prioritize structural aspects.

*2) Structural Visual Component:* In this section, we present the Structural Visual Component, utilizing an encoder-decoder architecture depicted in Figure 1a. The encoder processes the entire input, encompassing both masked and non-masked pixels, while the decoder predicts pixel values solely for the 'masked' subset of the input. The model is trained to minimize reconstruction error specifically for the masked (unseen) part of the input.

**Image/video as spatio-temporal patches.** To facilitate processing, input images or videos are represented as a 4D tensor with shape $T \times H \times W \times C$, where $T$ is the temporal dimension, and $H, W$ represent the spatial dimensions, while $C$ denotes the color channels. Treating images as single-frame videos ($T = 1$), the input is then partitioned into $N$ spatio-temporal patches, each sized $t \times h \times w \times c$ [66].

*3) Losses and training:* We leverage CLIP [26] and OpenCLIP [67], [68], which equipped with an image encoder $F_I$ and a text encoder $F_T$. The similarity computation between an image $I$ and a text $T$ is calculated as follows:

$$\text{Sim}(I, T) = \cos(F_I(I), F_T(T)) \tag{1}$$

Here, cos() denotes the cosine similarity.

Our training process involves two sets: a collection of image-text pairs $< I, T >$ and a set of image-SG pairs $\langle I_G, G \rangle$. Utilizing the latter, we create positive textual

captions $\langle I_G, T^p \rangle$ and negative textual captions $\langle I_G, T^n \rangle$. These pairs serve as inputs to our model during optimization, where the following losses are minimized.

**Image-text loss:** Our image-text loss compose two components: the contrastive loss and the graph negative loss.

*Contrastive loss:* Training VL models conventionally involves employing the contrastive loss on image-text pairs, as demonstrated in [26]. We adopt this standard approach with both the original pairs $< I, T >$ and those generated from the SG $< I_G, T^p >$. The contrastive loss is formulated as follows:

$$\mathcal{L}_{con} = \text{InfoNCE}(F_I(\hat{I}), F_T(\hat{T})) \tag{2}$$

Here, $\hat{I} = I \cup I_G$ and $\hat{T} = T \cup T^p$.

*Graph negative loss:* For each image associated with a Scene Graph, we possess a positive text $T^p$ that accurately describes it and a negative text $T^n$ that does not. Our loss function aims to encourage $T^p$ to be more akin to $I_G$ than $T^n$ (refer to [64]). The graph-based negative loss is expressed as:

$$\mathcal{L}_{GN} = \sum_{I_G, T^p, T^n} -log\left(\frac{e^{\cos(I_G, T^p)}}{e^{\cos(I_G, T^p)} + e^{\cos(I_G, T^n)}}\right) \tag{3}$$

The image-text loss is a judiciously weighted combination of the contrastive loss and the graph negative loss:

$$\mathcal{L}_{total} = \mathcal{L}_{con} + \alpha \mathcal{L}_{GN} \tag{4}$$

Here, $\alpha$ represents a hyperparameter.

**Image-decoder loss and optimization.** We aim to minimize the reconstruction error between the decoder predictions and the input pixel values. The input pixel values undergo normalization to achieve zero mean and unit variance, as suggested in [69]. This normalized form serves as the target for the loss computation. The $\ell_2$ distance is minimized between predictions and targets over $M$ masked patches. Throughout training, we sample either images or videos in mini-batches, computing the loss based on the decoder predictions.

### C. Action-SGFA Modality Pairing

Action-SGFA employs modality pairs $(\mathcal{I}, \mathcal{M})$, where $\mathcal{I}$ represents image/video, and $\mathcal{M}$ denotes another modality, to facilitate the learning of a unified joint embedding. Our approach leverages an extensively pre-trained VL model, trained on large-scale web datasets with diverse <image, text> pairings covering a wide range of semantic concepts. Additionally, we incorporate the natural, self-supervised pairing of other modalities, such as skeleton, depth, and thermal, with image/video.

Consider the modality pair $(\mathcal{I}, \mathcal{M})$ with aligned observations. For an image $\mathbf{I}_i$ and its corresponding observation in the other modality $\mathbf{M}_i$, we encode them into normalized embeddings: $\mathbf{q}_i = F(\mathbf{I}_i)$ and $\mathbf{k}_i = G(\mathbf{M}_i)$, where $F$ and $G$ are deep networks. The embeddings and encoders are optimized using an InfoNCE loss [70]:

$$L_{con} = -\log \frac{e^{\left(\mathbf{q}_i^\top \mathbf{k}_i / \tau\right)}}{e^{(\mathbf{q}_i^\top \mathbf{k}_i / \tau)} + \sum_{j \neq i} e^{(\mathbf{q}_i^\top \mathbf{k}_j / \tau)}} \tag{5}$$

Here, $\tau$ is a scalar temperature controlling the smoothness of the softmax distribution, and $j$ denotes unrelated observations or "negatives." Following [71], every example $j \neq i$ in the mini-batch is considered a negative. The loss aims to bring the embeddings $\mathbf{q}_i$ and $\mathbf{k}_i$ closer in the joint embedding space, aligning $\mathcal{I}$ and $\mathcal{M}$. In practice, we use a symmetric loss $L_{\mathcal{I},\mathcal{M}} + L_{\mathcal{M},\mathcal{I}}$.

**Alignment of Unseen Modality Pairs.** In Action-SGFA, the alignment process involves modalities paired with images, denoted as $(\mathcal{I}, \mathcal{M})$. This alignment specifically targets each modality's embeddings, aligning them with those obtained from images. Notably, an emergent behavior is observed in the embedding space, facilitating alignment between pairs of modalities, denoted as $(\mathcal{M}_1, \mathcal{M}_2)$, despite the training being conducted solely on pairs $(\mathcal{I}, \mathcal{M}_1)$ and $(\mathcal{I}, \mathcal{M}_2)$. This unique characteristic empowers Action-SGFA to perform diverse zero-shot action recognition tasks without the need for specific training. Remarkably, we achieve state-of-the-art results in zero-shot text-skeleton action recognition, even in the absence of any paired (skeleton, text) samples during training.

## IV. Experiments

### A. Datasets

For the SGs understanding training, we utilize SG data from Visual Genome (VG) [24] and Action Genome (AG) [25] as image-SGs pairs, supplemented with a small subset of the LAION [28]: **(1) LAION** represents a large-scale image-text pair dataset that was automatically curated from the Internet and has been filtered to pre-train the CLIP model. **(2) VG** is annotated with $108,077$ images and SGs. On average, images have 12 entities and seven relations per image. **(3) AG** provides frame-level SG labels for the components of each action, which provide annotations for $234,253$ frames with a total of $476,229$ bounding boxes of 35 object classes and $1,715,568$ instances of 25 relationship classes.

For multimodal feature alignment and the evaluation of zero-shot action recognition, we utilize the following datasets: **(1) Kinetics-400** [42]: This dataset, curated by DeepMind, comprises approximately 300,000 video clips sourced from YouTube, encompassing a wide array of 400 human action classes. To align with the skeleton modality, we adapt the Kinetics-Skeleton-400 dataset, derived from the Kinetics-400 video dataset through the application of the OpenPose [72] pose estimation toolbox. The Kinetics-Skeleton-400 dataset comprises 240,436 training and 19,796 testing skeleton sequences across 400 classes. **(2) NTU RGB+D 60** [43]: This action recognition dataset contains 56,578 RGB videos, skeleton sequences, depth and thermal video. Over 60 action classes are captured from 40 distinct subjects and three different camera view angles. Each skeleton graph has N = 25 body joints as nodes for the skeleton modality, with their 3D locations in space as initial features. **(3) NTU RGB+D 120** [44]: An extension of NTU RGB+D 60, this dataset includes an additional 57,367 skeleton sequences spanning 60 supplementary action classes. In total, NTU RGB+D 120 comprises 113,945 samples distributed across 120 classes.

The data is collected from 106 subjects and recorded using 32 unique camera setups.

### B. Results

We assess the performance of Action-SGFA in the context of open-vocabulary zero-shot action recognition tasks, employing text prompt templates as outlined in [26]. Specifically, we evaluate Action-SGFA on the RGB video modality using the Kinetics-400 dataset, while for the human skeleton, depth, and thermal modalities, we conduct evaluations on both NTU RGB+D 60 and NTU RGB+D 120 datasets. Across all three datasets, Action-SGFA demonstrates notable improvements across all modalities. Detailed results are presented in Table I.

Each modality is a testbed for evaluating Action-SGFA's capability to establish associations between text embeddings and the image modality, even without joint training exposure. Given the novelty of our problem setting, it is essential to note that no conventional "fair" baselines are available for direct comparison with Action-SGFA. However, our evaluation compares prior work that incorporates text paired with specific modalities, such as skeleton-based approaches [80], [81]. Additionally, we employ the CLIP model directly for certain ‘‘visual-like’’ modalities such as depth and thermal. Moreover, we provide results for each benchmark's best-reported supervised upper bound. In the context of the NTU-60 and NTU-120 zero-shot learning (ZSL) settings with splits of 48/12 and 96/24, respectively, SynSE [79] achieves accuracy levels of 33.3% and 38.7%. SynSE represents the state-of-the-art among classic ZSL methods, as indicated in Table I. Notably, Action-SGFA outperforms SynSE by a significant margin, achieving accuracy gains of 27.0% and 19.7% on NTU-60 and NTU-120, respectively. We also report the standard RGB video (Kinetics-400 [42]) zero-shot recognition tasks for completeness. Action-SGFA achieved 74.1% performance, surpassing the SOTA method VideoCoCa by 2.1%.

Skeleton-DGCFA performs outstanding open-vocabulary zero-shot action recognition, registering significant improvements across all benchmark datasets. These findings underscore the effectiveness of Action-SGFA in aligning various action modalities by leveraging the integration of SG. This alignment process implicitly extends the text-based supervision associated with images to other modalities. Notably, Action-SGFA shows strong alignment for non-visual modalities such as the skeleton, highlighting the substantial value of their inherent pairing with images as a potent source of supervision.

### C. Ablation Study

Utilizing the CLIP model, we conducted an ablation study on the NTU RGB+D 120 dataset using our Action-SGFA approach (Refer to Table II).

**SG knowledge.** To assess the optimal utilization of SG knowledge, Table IIa illustrates how the visual and textual components contribute to the approach. When incorporating solely the textual descriptions generated from the SGs (CLIP + Graph Text (GT) in the table), we observe an improvement in action recognition scores over the CLIP baseline, specifically, an increase of 2.7%. Including the graph-based negatives (GN) prompt results in an even more substantial enhancement, with a 4.5% improvement compared to the baseline, demonstrating the effectiveness of the generated negatives. Furthermore, introducing the SG MAE-decoder leads to a remarkable improvement of 5.7% in skeleton scores, while the text score remains comparable. This emphasizes the mutual benefits of both visual and textual components when leveraging SG labels for learning.

**Visual encoder components.** In this ablation, we justify our design choices within the visual encoder while employing consistently generated captions and negatives across all variants. We report performance in Table IIb on the skeleton (NTU RGB+D 120) modality. To assess the contribution of the visual encoder, we introduce an alternative variant, *CLIP + Decoder*, which excludes masked image tokens and predicts the same-scale image using a visual decoder. Notably, *CLIP + Decoder + Masking* outperforms *CLIP + Decoder*, demonstrating the advantages of incorporating masked image tokens. Finally, our "SG MAE" variant, which exclusively applies random masking to the action component in the input images/videos, outperforms all other models, demonstrating that our proposed adaptation technique tailored to the visual tokens allows better learning of the action recognition task.

**Scaling the image encoder.** The core idea in Action-SGFA is aligning the embeddings of all modalities to image embeddings, leveraging the comprehension of action SGs. Consequently, image embeddings play a pivotal role in aligning unseen modalities, and we study their effect on zero-shot performance. We vary the size of the image encoder and train an encoder for the skeleton modalities to match the image representation. To isolate the influence of the image representation, we fix the size of the skeleton encoders. In this experiment, we employ pre-trained CLIP (ViT-B/32, ViT-B/16, and ViT-L/16) and OpenCLIP (ViT-H/16) image and text encoders. Our results in Table IIc show that Action-SGFA's zero-shot action recognition performance on all modalities improves with better visual features. For skeleton action recognition, the stronger ViT-H/16 $vs$ the ViT-B/32 image encoder provides a gain of 9.2%. Thus, more substantial visual features can improve recognition performance even on non-visual modalities.

### D. Analysis of Various Attributes

We conduct experiments on training design choices to evaluate the influences of text prompt, action encoder, and contrastive loss temperature selection. These results are presented in Table III. We focus on the human skeleton action modality, which is non-visual and has a temporal component. We found that studying this modality led to robust and transferable design decisions

**Influence of text prompts.** The text prompt design has an enormous impact on the model performance. We show the influences of different text prompts in Table IIIa. By directly using label name (with prefix or suffix) as the text prompt in Skeleton-DGCFA for the baseline, the modest improvements

| | Kinetics-400 | NTU RGB+D 60 | | | NTU RGB+D 120 | | |
|---|---|---|---|---|---|---|---|
| Model | RGB video | skeleton | depth | thermal | skeleton | depth | thermal |
| Random Chance | 0.25 | 1.6 | 1.6 | 1.6 | 0.8 | 0.8 | 0.8 |
| Absolute SOTA | 90.6 | 89.8 | - | - | 85.1 | - | - |
| CLIP [73] | 62.3 | - | - | - | - | - | - |
| X-CLIP [74] | 65.2 | - | - | - | - | - | - |
| Text4Vis [75] | 68.9 | - | - | - | - | - | - |
| VideoCoCa [76] | 72.0 | - | - | - | - | - | - |
| ReViSE [77] | - | 17.4 | - | - | 32.3 | - | - |
| JPoSE [78] | - | 28.7 | - | - | 32.4 | - | - |
| SynSE [79] | - | 33.3 | - | - | 38.7 | - | - |
| Action-FA (ours) | 70.6 | 55.8 | 56.2 | 56.9 | 54.7 | 58.1 | 60.6 |
| Action-SGFA (ours) | **74.1** | **60.3** | **65.5** | **66.4** | **58.4** | **63.3** | **64.1** |

TABLE I: **Open-vocabulary zero-shot action recognition results**.

(a) Scene graph knowledge

| Model | Acc.(%) |
|---|---|
| CLIP | 51.7 |
| CLIP + GT | 54.4(↑ 2.7) |
| CLIP + GT + GN | 56.2(↑ 4.5) |
| CLIP + GT + GN + SG MAE | 58.4(↑ 5.7) |

(b) Visual encoder components

| Model | Acc.(%) |
|---|---|
| CLIP | 53.7 |
| CLIP + Decoder | 54.3(↑ 0.6) |
| CLIP + Decoder + Masking | 55.2(↑ 1.5) |
| CLIP + Decoder + SG MAE | 58.4(↑ 4.7) |

(c) Scaling the image encoder

| Model | Acc.(%) |
|---|---|
| ViT-B/32 | 49.2 |
| ViT-B/16 | 50.5(↑ 1.3) |
| ViT-L/16 | 53.8(↑ 4.6) |
| ViT-H/16 | 58.4(↑ 9.2) |

TABLE II: **Ablation study of the NTU RGB+D 120 dataset.**

(a) Text prompt type

| Prompt type | Acc(%) |
|---|---|
| Label name | 56.3 |
| Synonym | 56.9(↑ 0.6) |
| Body parts | 57.6(↑ 1.3) |
| Paragraph | 58.1(↑ 1.9) |
| SG Text | **58.4**(↑ 2.1) |

(b) Different skeleton encoders

| Backbone | Acc(%) | |
|---|---|---|
| | w./o. SGs | w. SGs |
| ST-GCN | 52.2 | 54.6(↑ 2.4) |
| MS-G3D | 53.9 | 57.8(↑ 3.9) |
| MST-GCN | 53.7 | 58.1(↑ 4.4) |
| CTR-GCN | 54.3 | **58.4**(↑ 5.1) |

(c) Temperature for loss

| $\lambda$ | Acc(%) |
|---|---|
| 1.0 | 58.1 |
| 0.8 | **58.4** |
| 0.7 | 58.3 |
| 0.5 | 58.1 |
| 0.2 | 57.8 |

TABLE III: **Analysis of different components of Action-SGFA, including text prompt, skeleton encoder and $\lambda$ selection.**

(0.6%) over the baseline are observed by incorporating a synonym list for label names. Utilizing prompts based on descriptions of body parts further makes enhancement of performance $1.3\%$, enriching the semantic context associated with each action class. The best performance is yielded by combining SG text descriptions for prompts, resulting in an accuracy of $58.4\%$.

**Different modality encoders.** The proposed Action-SGFA offers a network architecture-agnostic solution that can enhance encoders of various modalities. Taking skeleton modality as an example, we show experimental results of applying Action-SGFA to ST-GCN, MS-G3D, MST-GCN, and CTR-GCN in Table IIIb. Action-SGFA brings consistent improvements of (2.4-5.1%) over original feature alignment without sense graph understanding learning at inference, demonstrating the effectiveness and generalization ability of Action-SGFA for downstream tasks.

**Contrastive loss temperature.** We investigate the influence of the temperature parameter, denoted as $\tau$ (as described in Eq. 5), and present our findings in Table IIIc. Our experimentation uses a learnable temperature initialized to 0.08 (parametrized in the log-scale), a technique inspired by [26]. This is contrasted with the application of various fixed temperature values. Interestingly, our results diverge from those presented in [26]. We find that a fixed temperature yields superior performance for skeleton action recognition. Furthermore, our observation reveals that higher temperature settings prove more effective in training the skeleton encoders, whereas a lower temperature is better for training the depth and thermal encoders.

## V. Conclusion

This study investigates the problem of open-vocabulary zero-shot action recognition. We propose Action-SGFA, a novel action feature alignment approach that learns unified joint embeddings across four action modalities incorporating SG comprehension. Experimental results demonstrate that the enhancement of pre-trained VL models is attributed to incorporating the comprehension of action SGS and adopting a new training paradigm as well, enhancing their zero-shot capabilities of new modalities due to their natural pairing with images. We also establish a new state-of-the-art in zero-shot action recognition tasks across modalities, outperforming the vanilla skeleton zero-shot method by 27.0%, 19.7%, and 2.1% on NTU-60, NTU-120, and Kinetics-400, respectively.

## ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation of China (NSFC) under Grant No. 62176134, by research and application on AI technologies for smart mobility funded by SAIC Motor, and by a grant from the Institute Guo Qiang (2019GQG0002), Tsinghua University.

## REFERENCES


[1] M. Cheng, K. Cai, and M. Li, "RWF-2000: An Open Large Scale Video Database for Violence Detection," in *ICPR*, 2021.

[2] Z. Islam, M. Rukonuzzaman, R. Ahmed, M. H. Kabir, and M. Farazi, "Efficient Two-Stream Network for Violence Detection Using Separable Convolutional LSTM," in *IJCNN*, 2021.

[3] X. Wang, Z. Che, K. Yang, B. Jiang, J.-B. Tang, J. Ye, J. Wang, and Q. Qi, "Robust Unsupervised Video Anomaly Detection by Multipath Frame Prediction," *Neural Networks and Learning Systems*, vol. 33, pp. 2301--2312, 2022.

[4] M. Z. Zaheer, A. Mahmood, M. H. Khan, M. Segu, F. Yu, and S.-I. Lee, "Generative Cooperative Learning for Unsupervised Video Anomaly Detection," in *CVPR*, 2022.

[5] Y. Su, G. Lin, J. Zhu, and Q. Wu, "Human Interaction Learning on 3D Skeleton Point Clouds for Video Violence Recognition," in *ECCV*, 2020.

[6] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh, "Learning Regularity in Skeleton Trajectories for Anomaly Detection in Videos," in *CVPR*, 2019.

[7] A. Markovitz, G. Sharir, I. Friedman, L. Zelnik-Manor, and S. Avidan, "Graph Embedded Pose Clustering for Anomaly Detection," in *CVPR*, 2020.

[8] C. Liu, R. Fu, Y. Li, Y. Gao, L. Shi, and W. Li, "A Self-Attention Augmented Graph Convolutional Clustering Networks for Skeleton-Based Video Anomaly Behavior Detection," *Applied Sciences*, vol. 12, no. 1, 2022.

[9] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 961--970.

[10] H. Zhao, A. Torralba, L. Torresani, and Z. Yan, "Hacs: Human action clips and segments dataset for recognition and temporal localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8668--8678.

[11] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *European Conference on Computer Vision*. Springer, 2016, pp. 510--526.

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

[13] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 4489--4497.

[14] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision*. Springer, 2016, pp. 20--36.

[15] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299--6308.

[16] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6202--6211.

[17] N. Hussein, E. Gavves, and A. W. Smeulders, "Timeception for complex action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 254--263.

[18] R. Herzig, A. Bar, H. Xu, G. Chechik, T. Darrell, and A. Globerson, "Learning canonical representations for scene graph to image generation," in *European Conference on Computer Vision*, 2020.

[19] R. Li, S. Zhang, and X. He, "Sgtr: End-to-end scene graph generation with transformer," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19 464--19 474, 2021.

[20] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene Graph Generation by Iterative Message Passing," 2017, pp. 3097--3106.

[21] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Conference on Computer Vision and Pattern Recognition*, 2018.

[22] J. Yang, Y. Z. Ang, Z. Guo, K. Zhou, W. Zhang, and Z. Liu, "Panoptic scene graph generation," in *European Conference on Computer Vision*, 2022.

[23] S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid, and I. Laptev, "Language conditioned spatial relation reasoning for 3d object grounding," *ArXiv*, vol. abs/2211.09646, 2022.

[24] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32--73, 2017.

[25] J. Ji, R. Krishna, L. Fei-Fei, and J. C. Niebles, "Action genome: Actions as compositions of spatio-temporal scene graphs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 236--10 247.

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748--8763.

[27] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 888--12 900.

[28] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," *arXiv preprint arXiv:2111.02114*, 2021.

[29] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 278--25 294, 2022.

[30] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. Van Der Maaten, "Exploring the limits of weakly supervised pretraining," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 181--196.

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

[32] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li, *et al.*, "Florence: A new foundation model for computer vision," *arXiv preprint arXiv:2111.11432*, 2021.

[33] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*. PMLR, 2021, pp. 4904--4916.

[34] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, *et al.*, "Flamingo: a visual language model for few-shot learning, 2022," *URL https://arxiv.org/abs/2204.14198*.

[35] P. Morgado, N. Vasconcelos, and I. Misra, "Audio-visual instance discrimination with cross-modal agreement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 475--12 486.

[36] M. Patrick, Y. Asano, P. Kuznetsova, R. Fong, J. F. Henriques, G. Zweig, and A. Vedaldi, "Multi-modal self-supervision from generalized data transformations," 2020.

[37] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 631--648.

[38] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Computer Vision--ECCV 2020: 16th European Conference, Glasgow, UK, August 23--28, 2020, Proceedings, Part XI 16*. Springer, 2020, pp. 776--794.

[39] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 609--617.

[40] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir, "Multimae: Multi-modal multi-task masked autoencoders," in *European Conference on Computer Vision*. Springer, 2022, pp. 348--367.

[41] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000--16 009.

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

[43] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010--1019.

[44] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2684--2701, 2019.

[45] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," *Advances in neural information processing systems*, vol. 26, 2013.

[46] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 207--218, 2014.

[47] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.

[48] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," *arXiv preprint arXiv:1707.05612*, 2017.

[49] J.-B. Alayrac, A. Recasens, R. Schneider, R. Arandjelović, J. Ramapuram, J. De Fauw, L. Smaira, S. Dieleman, and A. Zisserman, "Self-supervised multimodal versatile networks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 25--37, 2020.

[50] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2630--2640.

[51] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, "End-to-end learning of visual representations from uncurated instructional videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9879--9889.

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

[53] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer, "Lit: Zero-shot transfer with locked-image text tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 123--18 133.

[54] C.-Y. Wu and P. Krähenbühl, "Towards Long-Form Video Understanding," in *CVPR*, 2021.

[55] Y. Zhang, P. Tokmakov, M. Hebert, and C. Schmid, "A structured model for action detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9967--9976.

[56] E. B. Avraham, R. Herzig, K. Mangalam, A. Bar, A. Rohrbach, L. Karlinsky, T. Darrell, and A. Globerson, "Bringing image scene structure to video via frame-clip consistency of object tokens," in *Thirty-Sixth Conference on Neural Information Processing Systems*, 2022.

[57] K. Xu, J. Li, M. Zhang, S. S. Du, K. ichi Kawarabayashi, and S. Jegelka, "What can neural networks reason about?" in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=rJxbJeHFPS

[58] V. Zambaldi, D. Raposo, A. Santoro, V. Bapst, Y. Li, I. Babuschkin, K. Tuyls, D. Reichert, T. Lillicrap, E. Lockhart, *et al.*, "Relational deep reinforcement learning," *arXiv preprint arXiv:1806.01830*, 2018.

[59] A. Jerbi, R. Herzig, J. Berant, G. Chechik, and A. Globerson, "Learning object detection from captions via textual scene attributes," *ArXiv*, vol. abs/2009.14558, 2020.

[60] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *ECCV*, 2020.

[61] C. Gao, J. Xu, Y. Zou, and J.-B. Huang, "Drg: Dual relation graph for human-object interaction detection," *ArXiv*, vol. abs/2008.11714, 2020.

[62] A. Bar, R. Herzig, X. Wang, A. Rohrbach, G. Chechik, T. Darrell, and A. Globerson, "Compositional video synthesis with action graphs," in *ICML*, 2021.

[63] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3668--3678.

[64] S. Doveh, A. Arbelle, S. Harary, R. Panda, R. Herzig, E. Schwartz, D. Kim, R. Giryes, R. Feris, S. Ullman, *et al.*, "Teaching structured vision&language concepts to vision&language models," *arXiv preprint arXiv:2211.11733*, 2022.

[65] M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou, "When and why vision-language models behave like bags-of-words, and what to do about it?" in *International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=KRLUvxh8uaX

[66] R. Girdhar, M. Singh, N. Ravi, L. van der Maaten, A. Joulin, and I. Misra, "Omnivore: A single model for many visual modalities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 102--16 112.

[67] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, "Reproducible scaling laws for contrastive language-image learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2818--2829.

[68] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, *et al.*, "Openclip," *Zenodo*, vol. 4, p. 5, 2021.

[69] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. pmlr, 2015, pp. 448--456.

[70] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[71] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3733--3742.

[72] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291--7299.

[73] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers." Association for Computational Linguistics, 2019, pp. 5099--5110.

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

[75] W. Wu, Z. Sun, and W. Ouyang, "Transferring textual knowledge for visual recognition," *arXiv preprint arXiv:2207.01297*, 2022.

[76] S. Yan, T. Zhu, Z. Wang, Y. Cao, M. Zhang, S. Ghosh, Y. Wu, and J. Yu, "Video-text modeling with zero-shot transfer from contrastive captioners," *arXiv preprint arXiv:2212.04979*, 2022.

[77] Y.-H. Hubert Tsai, L.-K. Huang, and R. Salakhutdinov, "Learning robust visual-semantic embeddings," in *Proceedings of the IEEE International conference on Computer Vision*, 2017, pp. 3571--3580.

[78] M. Wray, D. Larlus, G. Csurka, and D. Damen, "Fine-grained action retrieval through multiple parts-of-speech embeddings," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 450--459.

[79] P. Gupta, D. Sharma, and R. K. Sarvadevabhatla, "Syntactically guided generative embeddings for zero-shot skeleton action recognition," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 439--443.

[80] A. Nagrani, P. H. Seo, B. Seybold, A. Hauth, S. Manen, C. Sun, and C. Schmid, "Learning audio-video modalities from image captions," in *European Conference on Computer Vision*. Springer, 2022, pp. 407--426.

[81] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending clip to image, text and audio," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 976--980.