Pattern Recognition 162 (2025) 111402

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Semantic matters: A constrained approach for zero-shot video action recognition

Zhenzhen Quan [a,b], Jialei Chen [b], Daisuke Deguchi [b], Jie Sun [d], Chenkai Zhang [b], Yujun Li [a,c,*], Hiroshi Murase [b]

[a] School of Information Science and Engineering, Shandong University, No. 72 Binhai Road, Qingdao, Shandong, China
[b] Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi, Japan
[c] Smart State Governance Lab, Shandong University, No. 72 Binhai Road, Qingdao, Shandong, China
[d] Shandong Provincial Department of Justice, No. 15743, Jingshi Road, Jinan City, Shandong, China

ARTICLE INFO

Keywords:
Visual language model
Video action recognition
Semantic constrained
Zero-shot
Semantic-related

ABSTRACT

Zero-shot video action recognition has advanced significantly due to the adaptation of visual-language models, such as CLIP, to video domains. However, existing methods attempt to adapt CLIP to video tasks by leveraging temporal information, neglecting the semantic information (i.e. the latent categories and their relationships) within videos. In this paper, we propose a Semantic Constrained CLIP (SC-CLIP) approach that leverages semantic information to adjust CLIP for video recognition while ensuring its performance on unseen data. SC-CLIP comprises a semantic-related query generation module and a semantic constrained cross attention module. First, the semantic-related query generation module clusters dense tokens from CLIP to generate semantic-related mask. The semantic-related query is then derived by pooling the adapted CLIP output using the semantic-related mask. Next, the semantic constrained cross attention module feeds the generated semantic-related query back into CLIP to probe semantic-related values, enhancing their ability to leverage the vision-language matching capabilities of CLIP. By generating semantic-related query, the semantic information aids in distinguishing similar actions, thereby improving performance on unseen samples. Experimental results on three zero-shot action recognition benchmarks show improvements of up to 1.9% and 2% in harmonic mean under two settings. Code is available at https://github.com/quanzhenzhen/SC-CLIP.

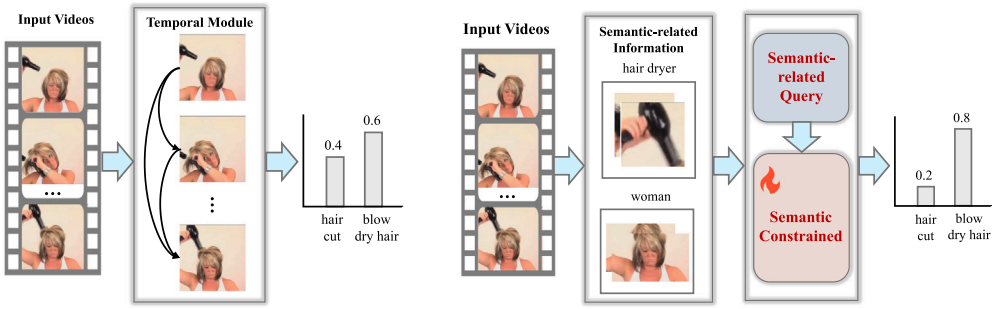
1. Introduction

Research on video action recognition has gained significant attention, driven by advances in video technology [1] and computer vision [2,3]. Existing methods leverage various sensor data [4] for applications in areas such as human–computer interaction and video retrieval. Traditional approaches rely extensively on large-scale manually labeled video datasets to train models. However, with the exponential growth of video data, the cost of manual labeling has risen sharply. These methods often struggle to adapt to unseen action categories, limiting their robustness and scalability in real-world applications. To address these challenges, Zero-Shot Learning (ZSL) [5] has emerged as a promising solution. ZSL aims to enable models to recognize categories absent from the training set by incorporating semantic information to bridge the gap between visual modalities and categories [6], reducing the dependence on large-scale labeled data. Common ZSL methods include embedding-based approaches [7] and generative approaches [8]. However, due to the challenges of low-quality generated samples and poor stability often associated with generative methods, most Zero-Shot Video Action Recognition (ZS-VAR) approaches rely on embedding-based methods.

Recently, the rapid development of vision-language models such as CLIP has prompted many studies to adapt models from the image domain to the video domain, offering a novel solution for ZS-VAR [9]. However, these methods still face significant challenges in ZS-VAR. The first challenge lies in effectively adapting vision-language models from the image domain to the video domain. Video data differs considerably from image data, incorporating temporal information and exhibiting more complex dynamic features and scene semantics. Fully leveraging the multimodal capabilities of vision-language models while adapting to the spatiotemporal characteristics of video data remains a critical research challenge. The second challenge is the insufficient emphasis on

* Corresponding author.
E-mail addresses: quanzhenzhen1991@163.com (Z. Quan), chen.jialei.s6@s.mail.nagoya-u.ac.jp (J. Chen), ddeguchi@nagoya-u.jp (D. Deguchi), 343856@qq.com (J. Sun), zhang1354558057@gmail.com (C. Zhang), liyujun@sdu.edu.cn (Y. Li), murase@nagoya-u.jp (H. Murase).

https://doi.org/10.1016/j.patcog.2025.111402
Received 16 October 2024; Received in revised form 19 December 2024; Accepted 19 January 2025
Available online 25 January 2025



(a) Previous methods (b) SC-CLIP

Fig. 1. Comparison of previous methods and the SC-CLIP approach. Previous methods in (a) process the inter-frame relationship in video through temporal modules or modeling methods. The SC-CLIP approach in (b) distinguishes the potential semantics within video. By leveraging the relationship between actions and potential semantics, SC-CLIP can better recognize actions such as hair cut and blow dry hair.

semantic information in action recognition. Current methods primarily focus on capturing temporal dynamics in video data, either through full fine-tuning of pre-trained vision-language models [10] or by using parameter-efficient transfer learning techniques inspired by natural language processing [11]. However, both approaches prioritize temporal information, either by modeling time or adding temporal modules to capture inter-frame relationships in video data (see Fig. 1(a)). These methods often neglect semantic information in videos, such as objects, actions, and their relationships, which are crucial for accurate action recognition. For similar actions (e.g., "haircut" and "blow-dry hair"), key semantic cues (e.g., scissors and hairdryer) provide crucial information for distinguishing action categories. The lack of semantic modeling significantly limits the adapting capability of existing methods, particularly in complex scenarios. Moreover, although full fine-tuning of vision-language models captures some semantic information, it incurs high computational costs and risks overfitting. In contrast, lightweight transfer learning methods offer only modest performance improvements. Thus, effectively modeling semantic information while ensuring computational efficiency remains a significant challenge.

To address the challenges in adapting vision-language models to the video domain and the underutilization of semantic information in action recognition, we introduce a novel Semantic Constrained CLIP (SC-CLIP) approach, which achieves efficient ZS-VAR by distinguishing potential semantics in the video and utilizing the relationship between actions and potential semantics (See Fig. 1(b)). Specifically, SC-CLIP approach includes a semantic-related query generation module to generate semantics and a semantic constrained cross attention module to utilize the relationship between actions and semantics. First, the semantic-related query generation module clusters the dense tokens of the frozen CLIP through the K-Means and mask fusion methods [12] to generate semantic-related mask encompassing all potential semantics. The CLIP output is then pooled with the semantic-related mask to create query embedded with semantic information. Second, the semantic constrained cross attention module feeds the generated semantic-related queries back into CLIP. By using the semantic-related query as Q and the intermediate token of the second last transformer layer of CLIP as K and V, semantic-related values are probed, enhancing the zero-shot capabilities of CLIP. The probed semantic-related values are then used for classification.

Unlike existing methods [13] that rely on temporal modules for inter-frame information interaction, SC-CLIP focuses on semantic information in videos. By capturing potential semantics and applying semantic constraints, SC-CLIP highlights features that distinguish similar actions, enabling effective adaptation to the video domain. Additionally, SC-CLIP selectively trains certain CLIP layers while keeping others frozen, ensuring efficiency and reducing overfitting. Our contributions are as follows:

(1) We develop a semantic-related query module that clusters the dense tokens of CLIP to generate the semantic-related mask, which is then used to pool the CLIP output and produce the semantic-related query.

(2) We introduce a semantic-constrained cross attention module that feeds the semantic-related query back into CLIP, where it serves as query to extract classification-related values, thereby constraining the outputs.

(3) We demonstrate superior performance on unseen classes across diverse action recognition datasets.

2. Related works

2.1. Video action recognition

With advancements in deep learning [14], Human Action Recognition (HAR) has emerged as a widely studied field [15]. HAR approaches vary by data modality, encompassing single-modal methods, such as those using RGB images [16] and skeleton data [17], as well as multimodal sensor data methods [4]. HAR methods can also be categorized by action type, including fine-grained action recognition [18], micro-gesture recognition [15], and micro-action analysis [19]. Based on network architecture, HAR methods include Convolutional Neural Network (CNN)-based approaches [19] and transformer-based approaches [20]. CAST [1] demonstrated spatiotemporal understanding of videos using spatiotemporal cross-attention mechanisms. The advent of large-scale vision-language models, such as CLIP, has inspired researchers to adapt these models for HAR [21]. The aforementioned HAR methods require all data to be labeled. Our method employs zero-shot learning for HAR, eliminating the need for labeling all actions. We achieve this by leveraging the zero-shot capabilities of CLIP through semantic-related query generation and semantic constrained cross attention modules. This enables transfer knowledge from seen actions to unseen ones, ultimately facilitating the recognition of unseen actions.

2.2. Zero-shot video action recognition

ZS-VAR is highly valuable in real-world applications as it enables the trained model to recognize unseen actions. Early studies, such as [22], employed action attributes to develop a ZS-VAR models. Similarly, [23] leveraged object-related attributes to model actions. Recently, the widespread adoption of large-scale pre-trained vision-language models for various downstream tasks has inspired researchers to extend these models to zero-shot tasks [24]. Beyond extensive research in zero-shot image classification, zero-shot video action recognition has also been explored [25]. X-CLIP [26] incorporated a cross-attention module into each layer of the transformer to enhance information exchange across the temporal dimension. Existing CLIP-based approaches emphasize inter-frame relationships, achieving ZS-VAR by

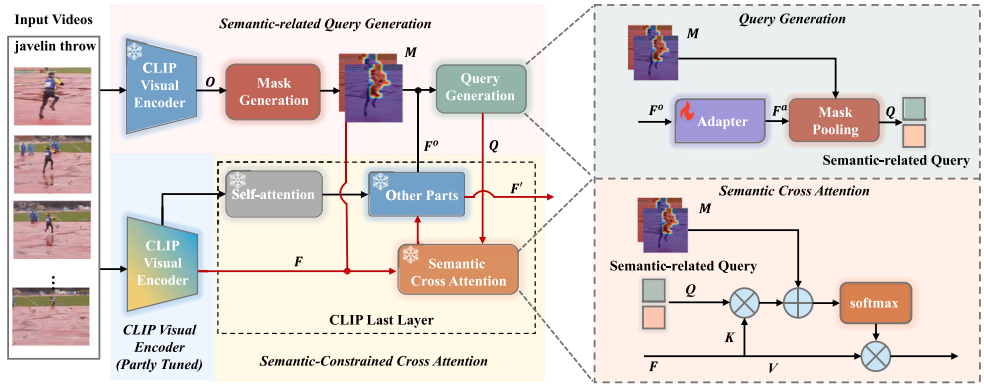


Fig. 2. The overview of the semantic constrained CLIP method is on the left, where the red line represents the semantic constraint process. The detailed view of the query generation and the semantic cross attention are on the top-right and bottom-right.

modeling temporal dynamics or incorporating temporal modules. Additionally, our proposed SC-CLIP approach primarily focuses on semantic information. SC-CLIP leverages CLIP to classify all semantics within videos, utilizing the relationships between actions and semantics to achieve ZS-VAR.

2.3. Pseudo-label generation

When labeled data is scarce, machine learning methods are employed to generate pseudo-labels, enhancing the diversity and richness of the dataset and enhancing the ability of model to adapt to unseen data. Common pseudo-labeling methods include those based on weakly-supervised learning [27], generative adversarial network [8], and clustering [12]. [12] raised a multi-scale K-means algorithm and a mask fusion method to generate pseudo-labels for unknown categories. Unlike existing methods, we do not utilize pseudo-label generation method to create pseudo-labels. Instead, we generate semantic-related mask to provide semantics for video action recognition.

3. Methodology

3.1. Overview

To better distinguish similar actions (e.g., javelin throw and discus throw) utilizing semantic information from videos, we implement a novel semantic constrained CLIP approach. This approach identifies potential semantics in videos and leverages the relationship between actions and these semantics to emphasize those that differentiate similar actions, as shown on the left side of Fig. 2. Firstly, we briefly introduce CLIP in the context of the video domain (See Section 3.1.1). Secondly, we raise a semantic-related query generation module to cluster the dense tokens of the frozen CLIP through the K-Means and mask fusion methods to generate semantic-related mask. The semantic-related mask is utilized to pool the adapted CLIP output to create semantic-related query (See Section 3.2). Then, we construct a semantic constrained cross attention module that feeds the semantic-related query back into CLIP to produce semantic-related embeddings (See Section 3.3). Finally, we give the training objective of SC-CLIP approach (See Section 3.4).

3.1.1. Preliminary

We introduce the manner in which existing methods [25,26] apply CLIP in video recognition, and similarly, we apply this approach to utilize CLIP in ZS-VAR. Consider a video $v_i \in \mathbb{R}^{T\times 3\times H\times W}$ with T frames, where each frame has a resolution of $H\times W$. The visual feature $z_{v,i} \in \mathbb{R}^{T\times D}$ is extracted using the CLIP video encoder f_v, and D is the dimension of the [CLS] token in the ViT encoder. After processing the temporal information using different ways, average pooling is applied to $z_{v,i}$, yielding $\bar{z}_{v,i} \in \mathbb{R}^D$. Our approach treats the video as a collection of frames, which are input into CLIP. For the text processing, the action category is embedded in a predefined template (e.g. "a video of [ ]") as input, where the content inside "[ ]" is the description of action. Given a text description a_j of a video, the text feature $z_{a,j} \in \mathbb{R}^D$ is extracted using CLIP's text encoder f_a. During training, the objective is to maximize the similarity between $\bar{z}_{v,i}$ and $z_{a,j}$ if they belong to the same class.

$$\mathrm{sim}\left(\bar{z}_{v,i}, z_{a,j}\right) = \frac{\left\langle \bar{z}_{v,i}, z_{a,j} \right\rangle}{\left\|\bar{z}_{v,i}\right\| \left\|z_{a,j}\right\|}. \tag{1}$$

3.2. Semantic-related Query Generation (SQG)

To transfer the vision-semantic matching capability of CLIP to video domains, we propose a semantic-related query generation module. Inspired by the pseudo labels generation approach [12] for image, we obtain semantic-related mask from frozen CLIP by the input video. The algorithm proposed demonstrated semantic consistency across different images. Since the variation between frames within the same video is smaller than that between different images, we treat the T frames as a collection of individual frames. We first input a video $V \in \mathbb{R}^{T\times 3\times H\times W}$ into the frozen visual encoder of CLIP. Then we average the dense token $O \in \mathbb{R}^{T\times(H\times W)\times D}$ from CLIP visual encoder to initialize the mask seed C_s through sliding windows of different sizes.

$$C_s = \left\{ \sum_{u=i}^{i+s-1} \sum_{v=j}^{j+s-1} \frac{O[u,v]}{s^2} \,\middle|\, i \in I, j \in J \right\} \tag{2}$$

where $I = \{0, [s/2], \ldots, [H-s]\}$, $J = \{0, [s/2], \ldots, [W-s]\}$, $[\cdot]$ denotes the rounding operation, and $s \in S$. S represents the size of the window. The initialized C_s is then refined using K-means to obtain fine-grained category clusters. To further merge the mask belonging to the same category, we consider mask fusion method by computing the cosine similarity $C \in [-1,1]^{N_c * N_c}$ between C_s and its corresponding mask $M \in \mathbb{R}^{T\times N\times D}$, and merge the mask based on a threshold δ. The value of the threshold δ in mask fusion determines the degree of fusion, with larger values resulting in a higher level of fusion.

Next, we feed the video into CLIP visual encoder and obtain the output from the last layer. The output is passed through an adapter with a novel self-attention mechanism, $\mathrm{SA} = \mathrm{softmax}\left(\frac{QK(t-2)\sim(t+2)^T}{\sqrt{d}}\right) V(t-2) \sim (t+2)$, designed to capture temporal information from videos. This process extracts features $F^a \in \mathbb{R}^{T\times(H\times W)\times D}$. The semantic-related query Q is obtained by mask pooling the features F^a through the semantic-related mask M obtained from CLIP on the dimension $T\times H\times W$: (the top-right of Fig. 2)

$$Q = \left[\frac{\sum_{T,H,W} F^a \left[\mathbb{1}(m=1)\right]}{\sum_{T,H,W} \left[\mathbb{1}(m=1)\right]} \,\middle|\, m \in M \right], \tag{3}$$

(a) The attention weights of CLIP (b) The attention weights of SC-CLIP

Fig. 3. The visualization of attention weights of the last layer in CLIP and SC-CLIP.

where $\mathbb{1}$ implies whether the value of mask belongs to 1. For simplicity, the batch size B is omitted. In each epoch, each batch contains different videos, and videos both within and across batches contribute to the generation of semantic-related query.

The semantic-related query Q captures all potential semantic information in every frame. To ensure that the semantic-related query Q aligns more accurately with action-related semantics, we supervise it using the text embeddings a_i from the frozen CLIP text encoder through a cross-entropy:

$$\mathcal{L}_q = -\sum_i \log \frac{\exp\left(\text{sim}\left(q_i, a_i\right)/\tau\right)}{\sum_j \exp\left(\text{sim}\left(q_i, a_j\right)/\tau\right)}, \tag{4}$$

where $\mathcal{L}_q$ is the query alignment loss, $q_i \in Q$ and τ is temperature parameter.

3.3. Semantic Constrained Cross Attention (SCCA)

We propose a semantic constrained cross attention module that retains the semantics of CLIP, enabling it to learn semantic information while leveraging the relationship between actions and semantics to achieve the ZS-VAR task. First, we keep the parameters of the last transformer layer of CLIP unchanged, while making the other layers trainable. All potential semantic information is captured in the semantic-related query generated in Section 3.2. This query is then input into the last transformer layer again, allowing CLIP to learn semantic information. To capture the relationships between actions and potential semantics in videos, we introduce Semantic Cross Attention (SCA) mechanism, replacing the self-attention mechanism in the last transformer layer when inputting semantic-related query, as shown in the bottom-right of Fig. 2. The original query is replaced by the semantic-related query, while the output from CLIP second last layer continues to serve as the key and value. The semantic-related mask is added as a bias matrix to enhance the distinction of semantic-related areas, thereby influencing the semantic cross attention mechanism. SCA is:

$$\text{SCA}\left(Q, F, M\right) = \text{softmax}\left(\frac{\tilde{Q}\tilde{K}^T}{\sqrt{d}} + M\right)\tilde{V}, \tag{5}$$

where $\tilde{Q} = W_q Q$, $\tilde{K} = W_k F$, $\tilde{V} = W_v F$ are parameter-invariant linear projections of the semantic-related query Q and the output F of second last transformer layer.

After these operations, the output F' of the last transformer layer carries enriched semantic information, which we then utilize for action recognition. We propose the semantic constrained loss $\mathcal{L}_{sc}$ to maximize the semantic-related embeddings $f_i' \in F'$ and the corresponding text embeddings a_i from frozen CLIP text encoder via cross-entropy:

$$\mathcal{L}_{sc} = -\sum_i \log \frac{\exp\left(\text{sim}\left(f_i', a_i\right)/\tau\right)}{\sum_j \exp\left(\text{sim}\left(f_i', a_j\right)/\tau\right)}. \tag{6}$$

To assess whether our proposed method captures semantic information, we visualize the attention weights of the last transformer layer from both CLIP and SC-CLIP after passing through the SCCA module in Fig. 3. In the attention visualizations of specific tokens from CLIP (See Fig. 3(a)), the CLS token and other tokens fail to capture meaningful semantic information. However, the SC-CLIP approach effectively captures semantics in the attention weights visualizations, such as hair dryers and body parts (See Fig. 3(b)). This is because [28] demonstrated that clustering effectively captures image semantics. Moreover, CLIP is trained on large-scale image-text pair datasets, enabling text-centered clustering and resulting in superior semantic clustering capabilities. Leveraging this advantage, semantic-related query dynamically learn the semantic clustering features of CLIP through semantic-related mask, capturing detailed, transferable features without predefined category labels. Subsequently, the SCCA module further refines and enhances the completeness and accuracy of the semantic information.

3.4. Training objective

The training objective of SC-CLIP comprises the semantic-constrained loss $\mathcal{L}sc$, the query alignment loss $\mathcal{L}q$, and the adapter alignment loss $\mathcal{L}_a$:

$$\mathcal{L} = \mathcal{L}_{sc} + \alpha\mathcal{L}_a + \beta\mathcal{L}_q, \tag{7}$$

where α and β are hyperparameters. $\mathcal{L}_a$ denotes the cross-entropy loss between the video-level embeddings from the auxiliary adapter and the corresponding text embeddings from the frozen CLIP text encoder like Eqs. (4) and (6), enhancing the performance for unseen categories.

4. Experiment settings

4.1. Datasets

We showcase the effectiveness of the raised SC-CLIP approach by evaluating it on the following datasets.

UCF-101 [29]**:** A HAR dataset, collected from YouTube, consists of 13,320 videos spanning 101 action categories. Each video depicts real-life scenarios. The dataset is officially divided into three parts. Each part allocates 9537 videos for training and 3783 videos for testing.

HHMDB-51 [30]**:** The dataset features 51 action categories, each comprising a minimum of 101 videos, amounting to a total of 6849 videos. Like the UCF-101 dataset, it is split into three segments for training and testing, with each category having 70 videos for training and 30 for testing.

Something Something V2 (SSv2) [31]**:** This dataset features 174 action categories related to human interactions with everyday objects. It focuses on detailed action recognition and introduces more complex temporal dynamics than other datasets. The dataset comprises 168,913 videos for training and 24,777 videos for testing.

Kinetics-400 (K-400) [32] **and Kinetics-600 (K-600)** [33]**:** The K-400 dataset contains 400 action categories, comprising YouTube videos, each approximately 10 s long. The dataset is divided into a training set (240,000 videos) and a test set (20,000 videos). K-600 is an extension of the K-400 dataset, featuring 600 behavioral actions in total, 220 of which are unique to K-600. These new action categories provide valuable resources for zero-shot recognition testing (60 for validation and 160 for testing).

Table 1
The hyperparameters of UCF-101, HMDB-51, SSv2 and K-600 datasets.

Dataset	δ	τ	α	β
UCF-101	0.9	65	0.1	0.1
HMDB-51	0.85	40	0.3	0.3
SSv2	0.85	45	0.2	0.2
K-600	0.8	50	0.0001	0.3

4.2. Evaluation protocols

In line with [13,26], we assess SC-CLIP in both base-to-novel and zero-shot settings. In both the base-to-novel and zero-shot settings, the division of seen and unseen samples in each dataset is the same as in [25].

Base-to-novel setting: We categorize the UCF-101, HMDB-51, and SSv2 datasets into common and novel action categories following the approach outlined in [13]. The common classes are utilized for training, with novel classes designated for testing. Each dataset provides three distinct splits for training and testing. For the SSv2 dataset, validation is performed on the entire test set, whereas for the UCF-101 and HMDB-51 datasets, only the first training-test split is used for validation. The top-1 accuracy is calculated for seen and unseen categories across various test datasets, denoted as $Base$ and $Novel$, respectively, along with their harmonic mean HM.

$$HM = \frac{2 \times Base \times Novel}{Base + Novel}. \tag{8}$$

Zero-shot setting: The SC-CLIP approach begins with training on the K-400 dataset and is then assessed on the UCF-101, HMDB-51, and K-600 datasets. For the UCF-101 and HMDB-51 datasets, we employ the three test splits from the official source and reported the mean and standard deviation of top-1 accuracy across these splits. For the K-600 dataset, we assess performance using 160 action categories that are absent from the K-400 dataset and report the mean and standard deviation of the top-1 accuracy.

4.3. Implementation details

We employ the CLIP architecture with ViT-B/16 across all experiments, with adapters implemented as transformer encoders. Following [25], we utilize GPT-3.5 to enrich action names. For training, we train 11th layer while keeping the other layers frozen. And each video is sampled with 8 frames, and the batch size is 64. For testing, we sample 3 video clips per video, each with 1 crop ("3 × 1" view), and combine the results by averaging. In the base-to-novel setting, training spans 18 epochs per dataset, including 2 warm-up epochs. Initial learning rates are set to 5×10^{-5} for UCF-101, 1×10^{-3} for HMDB-51, and 8×10^{-4} for SSv2, with rates decayed using a cosine scheduler. In the zero-shot setting, training on the K-400 dataset spans 6 epochs, including 2 warm-up epochs, with an initial learning rate of 4×10^{-4}. All experiments are performed using 8 NVIDIA Tesla V100 GPUs. Additional hyperparameter settings shown in Table 1 for all datasets are obtained from the experiments in Section 5.3.

5. Experiment results

5.1. Comparison with state-of-the-art

We primarily compare our method with state-of-the-art CLIP-based ZS-VAR approaches, including Frozen CLIP [24], Action CLIP [34], XCLIP [26], ViFi-CLIP [13], MMA [35], Vita-CLIP [10], GBC [36], M_2-CLIP [37], AIM [9], and ST-Adapter [21], across the UCF-101, HMDB-51, SSv2, and K-600 datasets under both two settings.

5.1.1. Base-to-novel

Table 2 presents the experimental results of base-to-novel comparisons conducted on the UCF-101, HMDB-51, and SSv2 datasets. The results clearly show that the SC-CLIP method significantly outperforms directly applying CLIP [24] to ZS-VAR, achieving notable improvements in recognizing unseen action categories (46.8% to 57.2%). Compared to X-CLIP [26], which modifies each transformer layer of CLIP by adding cross-attention modules for temporal information exchange, SC-CLIP leverages semantic information to constrain CLIP, resulting in improvements of 11.7% (HMDB-51), 21.1% (UCF-101), and 6% (SSv2) on unseen action categories. When compared to ViFi-CLIP [13], which fine-tunes CLIP, SC-CLIP significantly boosts the recognition rate of unseen actions, increasing it from 53.3% to 57.2% on HMDB-51. Furthermore, SC-CLIP demonstrates superior performance over methods that freeze CLIP and employ trainable adapters, such as AIM [9] and ST-Adapter [21], particularly on HMDB-51 and SSv2 datasets. Compared to MMA [35], SC-CLIP also achieves a substantial improvement in the harmonic mean. Notably, SC-CLIP requires only 1/4 of the training parameters used by CLIP, yet achieves over 95% of the performance of fully fine-tuned methods, such as 84.5% compared to 87.0% for FROSTER [25] on the UCF-101 dataset. The effectiveness of SC-CLIP lies in its ability to classify each video frame and leverage the relationship between actions and semantics to achieve ZS-VAR.

5.1.2. Zero-shot

Table 3 presents the zero-shot experimental results on the UCF-101, HMDB-51, and K-600 datasets, including results for the full test set and three split test sets on UCF-101 and HMDB-51 datasets. These results demonstrate that our method surpasses most existing advanced approaches, including uni-modal zero-shot recognition models (ER-ZASR [38]) and CLIP-based adaptation models (such as ActionCLIP [34], Vita-CLIP [10], XCLIP [26], ST-Adapter [21], M_2-CLIP [37], and GBC [36]). SC-CLIP improves accuracy by approximately 28% on the UCF-101 dataset over ER-ZASR [38], by 4.9% compared to the fully fine-tuned approach (Vita-CLIP [10]), and by 2.3% relative to the adapted method (ST-Adapter [21]). Although our approach falls short of GBC [36] on the K-600 dataset, it achieves 96% of performance of GBC method while utilizing only 10% of its training parameters. These improvements highlight the effectiveness of the semantic-related query generation and semantic constrained cross attention modules in SC-CLIP for enhancing CLIP's zero-shot performance in video applications. In terms of GFLOPs, SC-CLIP generates semantic-related masks using the frozen CLIP, creates semantic-related queries from these masks, and processes them through CLIP again. Despite this dual-pass usage, SC-CLIP achieves a GFLOP count of 305, which remains lower than ST-Adapter (455) and GBC (1882), highlighting its greater efficiency.

5.2. Ablation study

We perform ablation studies on various components of the SC-CLIP approach, using the base-to-novel setting on the HMDB-51, UCF-101, and SSv2 datasets, as detailed in Table 4. Fig. 2 demonstrates that the semantic-related query from SQG serve as the input for SCCA. Due to their interdependence, SQG and SCCA are analyzed as an integrated unit. When neither SQG nor SCCA is included (as shown in the 3rd row of Table 4), merely adding adapters to CLIP for the ZS-VAR task leads to poor recognition performance for unseen action categories. This is due to the lack of semantic constraints in CLIP, which prevents it from leveraging the relationships between actions and potential semantics. Excluding L_a, which indicates ignoring the alignment between auxiliary adapter features and text, improves zero-shot performance compared to scenarios where both SQG and SCCA are absent. However, it remains inferior to the SC-CLIP approach. Excluding L_q, which involves disregarding the alignment of semantic-related query and text in SQG, results in suboptimal recognition for both seen and unseen categories. Thus, integrating SQG, SCCA, L_a,

Table 2
The accuracies on the HMDB-51, UCF-101 and SSv2 datasets based on base-to-novel setting. Pub represents publication and the best results are **bolded**. Results are shown in percentage form.

Method	Pub	HMDB-51			UCF-101			SSv2		
		Base	*Novel*	*HM*	*Base*	*Novel*	*HM*	*Base*	*Novel*	*HM*
Frozen CLIP [24]	ICML'21	53.3	46.8	49.8	78.5	63.6	70.3	4.9	5.3	5.1
ActionCLIP [34]	TNNLS'23	69.1	37.3	48.5	90.1	58.1	70.7	13.3	10.1	11.5
XCLIP [26]	ECCV'22	69.4	45.5	55.0	89.9	58.9	71.2	8.5	6.6	7.4
AIM[a] [9]	ICLR'23	64.0	51.6	57.1	89.8	76.4	82.6	8.5	7.9	8.2
ST-Adapter[a] [21]	NIPS'22	65.3	48.9	55.9	85.5	76.8	80.9	9.3	8.4	8.8
ViFi-CLIP [13]	CVPR'23	**73.8**	53.3	61.9	92.9	67.7	78.3	**16.2**	12.1	13.9
MMA [35]	CVPR'24	–	–	–	86.2	**80.0**	82.2	–	–	–
SC-CLIP	–	72.2	**57.2**	**63.8**	**93.5**	77.0	**84.5**	15.8	**12.6**	**14.0**

[a] Indicates that results are generated by our implementation.

Table 3
The accuracies on UCF-101 and HMDB-51 datasets based on zero-shot setting. UCF* and HMDB* signify assessments on the complete validation set, while UCF and HMDB correspond to evaluations over the three validation splits. Pub and TP represent publication and number of the tunable parameters. The left and right sides of ± correspond to the mean and standard deviation of the top-1 accuracy, respectively.

Method	Pub	TP (M)	GFLOPs	UCF*	UCF	HMDB*	HMDB	K-600
ER-ZASR [38]	ICCV'21	–	–	–	51.8 ± 2.9	–	35.3 ± 4.6	42.1 ± 1.4
ActionCLIP[a] [34]	TNNLS'23	141	141	77.4	77.5 ± 0.8	48.0	48.2 ± 1.5	62.5 ± 1.2
XCLIP [26]	ECCV'22	131.5	145	–	72.0 ± 2.3	–	44.6 ± 5.2	65.2 ± 0.4
ST-Adapter[a] [21]	NIPS'22	7.2	455	77.9	77.6 ± 0.7	50.3	51.1 ± 0.6	60.2 ± 1.8
Vita-CLIP [10]	CVPR'23	124.7	97	–	75.0 ± 0.6	–	48.6 ± 0.6	67.4 ± 0.5
M_2-CLIP [37]	AAAI'24	16	214	–	78.7 ± 1.2	–	47.1 ± 0.4	–
GBC [36]	TCSVT'24	234.3	1882	–	77.7 ± 1.3	–	49.6 ± 1.4	**67.4 ± 0.7**
SC-CLIP	–	21.3	305	**79.9**	**79.9 ± 0.8**	**51.6**	**52.0 ± 0.3**	65.3 ± 1.0

[a] Indicates that results are generated utilizing our implementation.

Table 4
The accuracies on the HMDB-51, UCF-101 and SSv2 datasets based on base-to-novel setting. SQG is semantic-related query generation, SCCA represents semantic constrained cross attention, and L_a and L_q represent adapter alignment loss and query alignment loss.

Method				HMDB-51			UCF-101			SSv2		
SQG	SCCA	L_a	L_q	*Base*	*Novel*	*HM*	*Base*	*Novel*	*HM*	*Base*	*Novel*	*HM*
–	–	✓	✓	66.9	25.7	37.1	91.0	51.8	66.0	13.4	7.4	9.5
✓	✓	–	✓	**73.7**	55.3	63.2	90.9	77.7	83.8	**15.9**	11.6	13.4
✓	✓	✓	–	67.0	56.5	61.3	89.6	76.4	82.5	13.2	10.9	11.9
✓	✓	✓	✓	72.2	**57.2**	**63.8**	**93.5**	**77.0**	**84.5**	15.8	**12.6**	**14.0**

Table 5
The ablation study on similar actions, where Pre represents precision.

Dataset	Method				Category	Pre (%)
	SQG	SCCA	L_a	L_q		
HMDB-51	–	–	✓	✓	Fencing	36.92
	✓	✓	✓	✓		47.37 (↑ 10.45)
	–	–	✓	✓	Hit	28.57
	✓	✓	✓	✓		65.00 (↑ 36.43)
UCF-101	–	–	✓	✓	Javelin throw	28.57
	✓	✓	✓	✓		42.86 (↑ 14.29)
	–	–	✓	✓	Throw discus	50.98
	✓	✓	✓	✓		93.33 (↑ 42.35)

Table 6
The accuracies of different structures of adapter on HMDB-51, UCF-101 and SSv2 datasets based on base-to-novel setting. TF means Transformer.

Adapter		HMDB-51			UCF-101			SSv2		
Structure		*Base*	*Novel*	*HM*	*Base*	*Novel*	*HM*	*Base*	*Novel*	*HM*
MLP	MLP	67.5	57.2	61.9	91.0	77.7	83.8	14.9	11.4	12.9
MLP	TF	68.7	57.3	62.5	91.0	77.6	83.8	15.9	11.2	13.1
TF	MLP	72.0	56.4	63.3	93.4	76.6	84.2	16.3	10.0	12.4
TF	**TF**	**72.2**	**57.2**	**63.8**	**93.5**	**77.0**	**84.5**	**15.8**	**12.6**	**14.0**

and L_q significantly enhances the SC-CLIP approach's performance for both seen and unseen action categories. Additionally, the results for each module in Table 4 are averaged over six experiments. A t-test comparing these results with those of the SC-CLIP approach yields a p-value < 0.05, indicating statistical significance.

Additionally, we conducted ablation experiments on similar actions in Table 5. It presents two sets of similar actions, comparing the method that does not include SQG and SCCA with the SC-CLIP. The results show that with the incorporation of semantic information, there is a significant increase in precision for each action, with the highest improvement being 42.35%. This demonstrates that the SC-CLIP approach enhances the importance of semantics that distinguish similar actions by differentiating potential semantics in the video and leveraging the relationship between actions and these semantics, thereby improving the ability to distinguish similar behaviors.

5.3. Parameter analysis

In this section, we discuss the adapter structure in the top-right of Fig. 2, the parameter τ in Eqs. (4) and (6), the parameters α and β in Eq. (7), the parameter δ in SQG module, as well as the trainable layer of CLIP.

5.3.1. The structure of adapter

Table 6 compares four different combinations of Transformer and MLP structures for the two trainable adapters. The results indicate that the adapter with a transformer structure facilitates greater semantic information interaction during training, enhances the zero-shot capabilities of CLIP, and enables more effective transfer of knowledge from seen to unseen action classes.

Table 7
The accuracies of different values of the threshold δ on HMDB-51, UCF-101 and SSv2 datasets based on base-to-novel setting.

δ	HMDB-51			UCF-101			SSv2		
	Base	*Novel*	*HM*	*Base*	*Novel*	*HM*	*Base*	*Novel*	*HM*
0.7	66.4	52.7	58.8	90.8	76.9	83.3	12.9	10.3	11.5
0.8	70.3	56.3	62.5	93.1	75.8	83.6	15.6	11.5	13.2
0.85	**72.2**	**57.2**	**63.8**	93.4	76.3	84.0	**15.8**	**12.6**	**14.0**
0.9	71.2	54.1	61.5	**93.5**	**77.0**	**84.5**	16.6	11.6	13.7
0.95	71.6	54.7	62.0	93.7	76.8	84.4	16.7	11.6	13.7

Table 8
The accuracies of different values of the parameter τ on HMDB-51, UCF-101 and SSv2 datasets based on base-to-novel setting.

$1/\tau$	HMDB-51			UCF-101			SSv2		
	Base	*Novel*	*HM*	*Base*	*Novel*	*HM*	*Base*	*Novel*	*HM*
1	65.8	38.4	48.5	83.7	69.1	75.7	6.0	5.7	5.8
10	68.5	44.8	54.2	87.6	72.9	79.6	14.2	11.8	13.1
20	69.2	51.3	58.9	90.6	74.3	79.2	15.0	11.6	13.1
40	**72.2**	**57.2**	**63.8**	93.1	76.3	83.9	16.2	11.6	13.5
45	71.2	51.3	59.6	93.3	76.9	84.3	**15.8**	**12.6**	**14.0**
50	71.1	54.5	61.7	93.5	76.9	84.4	16.9	11.0	13.3
60	71.1	54.5	61.7	93.5	76.8	84.3	16.8	11.8	13.9
65	69.4	50.5	58.5	**93.5**	**77.0**	**84.5**	16.4	11.6	13.6
80	70.4	50.0	58.5	93.4	76.7	84.2	16.9	11.8	13.9
100	70.9	48.5	57.6	93.3	76.3	83.9	17.0	11.9	14.0

5.3.2. The threshold δ

In semantic-related query generation module, the number of semantic-related features is determined by the threshold δ in mask fusion. Increasing δ generates more mask seeds. Table 7 shows the performance of different datasets under various thresholds in the base-to-novel setting. The value of δ is selected from $\{0.7, 0.8, 0.85, 0.9, 0.95\}$. As seen in Table 7, UCF-101 dataset requires more semantic-related features than HMDB-51 and SSv2 datasets. This is because the background in UCF-101 is simpler, allowing more semantic information to be captured with a higher threshold δ. For all datasets, a larger threshold δ is not necessarily better. For HMDB-51, increasing the threshold δ results in a decline in zero-shot capability and a reduction in recognition accuracy for seen action categories. In contrast, for UCF-101 and SSv2 datasets, increasing the threshold δ improves recognition of seen action classes but reduces the ability to recognize unseen action samples.

5.3.3. The parameter τ

The τ value in Eqs. (4) and (6) represents the difficulty of training samples across different datasets. Table 8 shows the performance of various datasets in the base-to-novel setting under different τ values. The value of $1/\tau$ is selected from $\{1, 10, 20, 40, 45, 50, 60, 65, 80, 100\}$. As shown in Table 8, The τ value that is either too large or too small hampers the development of zero-shot recognition capability. The HMDB-51 and SSv2 datasets contain more challenging samples compared to UCF-101 dataset, necessitating an increased τ value to achieve better recognition performance.

5.3.4. The hyperparameters α & β

Hyperparameters α and β in Eq. (7) represent the respective proportions of L_a and L_q in the total loss. Fig. 4 displays the results for various datasets in the base-to-novel setting. The values of α and β are selected from $\{0.0001, 0.001, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 1\}$. As shown in Fig. 4, the weight of L_a does not significantly affect overall zero-shot performance. In contrast, the weight of L_q has a greater impact when set too high or too low, especially on the HMDB-51 and UCF-101 datasets. This suggests that aligning the semantic-related query with the action class is essential, but preserving sufficient semantic information within the query is equally important for enhancing performance on unseen categories.

Table 9
The accuracies of training different transformer layer on HMDB-51, UCF-101 and SSv2 datasets based on base-to-novel setting. When the specific layers mentioned in the table are being trained (e.g. $\{9, 10, 11\}$ means from 9th to 11th layer are trained), all other layers remain frozen. "None" indicates that none of the layers are trained.

Trainable	HMDB-51			UCF-101			SSv2		
layer	*Base*	*Novel*	*HM*	*Base*	*Novel*	*HM*	*Base*	*Novel*	*HM*
None	71.8	53.6	61.4	93.4	76.8	84.3	13.4	10.5	11.8
{10}	64.0	24.5	35.4	94.9	69.3	80.1	13.1	7.7	9.7
{11}	**72.2**	**57.2**	**63.8**	**93.5**	**77.0**	**84.5**	**15.8**	**12.6**	**14.0**
{10, 11}	66.9	23.5	34.8	94.9	69.1	80.0	13.1	7.3	9.3
{9, 10, 11}	61.2	19.3	29.3	93.9	63.4	68.1	13.1	7.5	9.5
{1 – 11}	66.2	32.1	43.2	89.9	45.1	60.1	13.3	7.9	9.9

5.3.5. The trainable transformer layer of CLIP

We experiment with training different layers of CLIP in the SC-CLIP approach, and the results are presented in Table 9. It shows that if any transformer layer of CLIP is left untrained (in the 1st row), the recognition performance of the SC-CLIP approach declines for both seen and unseen samples. When multiple layers of CLIP (e.g. from 10th to 11th layer) are trained, the recognition ability for unseen samples significantly decreases, and the inherent zero-shot capability of CLIP is compromised. However, when training a single layer of CLIP (10th or 11th layer), we found that training only the 11th layer of the transformer leads to the strongest retention of zero-shot capability of CLIP. As noted in [39], the last layer of the CLIP visual encoder preserves visual-semantic associations. Freezing its parameters avoids disrupting pre-trained multimodal features and ensures performance on unseen categories.

5.4. Visualization

To extract semantic information from videos, we generate semantic-related mask in SQG and leverage it to create semantic-related query. To evaluate the ability of SC-CLIP to learn semantic information, we visualize the mask seeds alongside the attention weights of the last transformer layer after SCCA, as shown in Figs. 5 and 6. In Fig. 5, the first and fifth columns display three randomly selected frames from a video, while the second to fourth and sixth to eighth columns show the mask seeds. For the selected actions "blow-drying hair" and "playing violin", the semantic-related mask includes three types of mask seeds that capture the primary action semantics. While the content of the seeds varies slightly between frames, they consistently represent the same semantics. Combining frames clarifies semantic information, and the accuracy of the semantic mask matches the results shown in Fig. 5. The attention weights of the last transformer layer after SCCA are shown in the second to fourth columns and the sixth to eighth of Fig. 6. The visualization results indicate that SC-CLIP fully captures the information from the semantic-related mask, including the types and semantics. Additionally, SC-CLIP learns semantics at category level rather than token level, enhancing the completeness of the semantic representation.

In addition to mask visualization, we also utilize t-SNE [40] to visualize the recognition performance of SC-CLIP on unseen action categories, as shown in Fig. 7. By comparing the visualized results of SC-CLIP and Frozen CLIP in zero-shot setting on UCF-101, we observed that SC-CLIP demonstrates significantly stronger capability in distinguishing unseen categories. Specifically, as shown in parts (a) and (b) of Fig. 7, results highlighted by the red circles illustrate the notable advantages of SC-CLIP over Frozen CLIP in terms of clustering performance. SC-CLIP effectively leverages the relationships between actions and latent semantics, achieving a more distinct and well-separated distribution of different category samples. Compared to Frozen CLIP, SC-CLIP forms clearer boundaries between classes, indicating its superior ability to capture characteristic differences across categories.

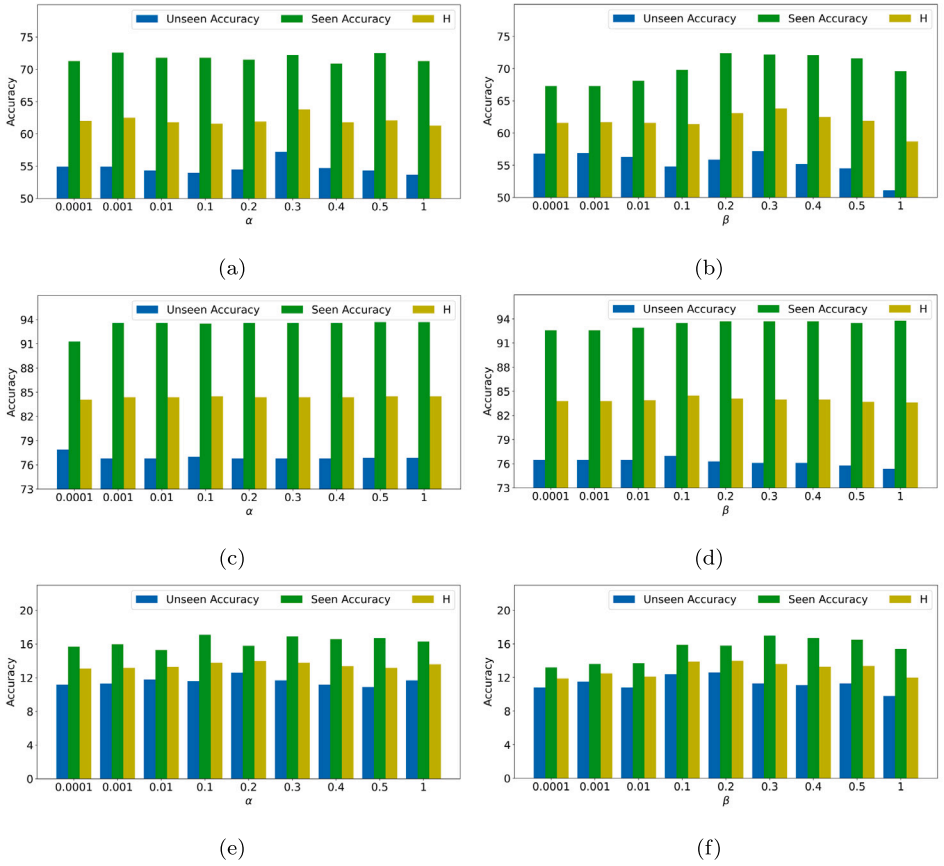


Fig. 4. The results of different values of parameters α and β on different datasets based on base-to-novel setting. (a) and (b) are the results on the HMDB-51 dataset, (c) and (d) are the results on the UCF-101 dataset, and (e) and (f) are the results on the SSv2 dataset.

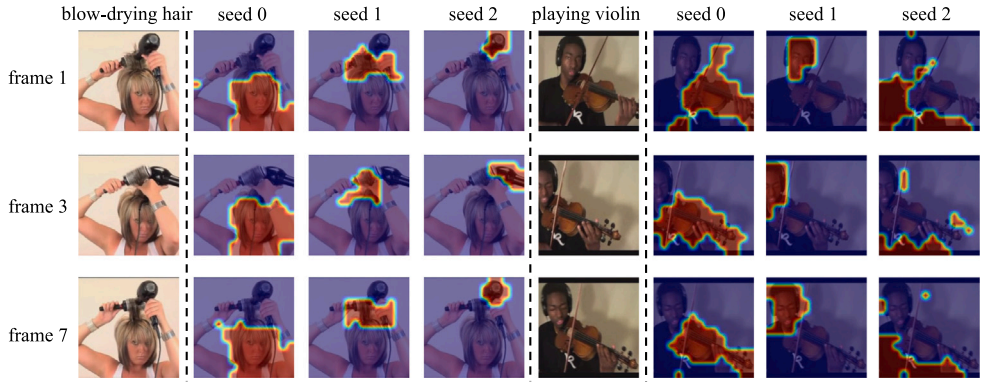


Fig. 5. The visualization of different mask seeds which represent different semantics.

6. Conclusion

In this work, we propose SC-CLIP, a novel approach to enhance the performance of CLIP in zero-shot video action recognition by integrating semantic information through the Semantic Query Generation (SQG) and Semantic Constrained Cross Attention (SCCA) modules. SQG extracts dense tokens from CLIP and clusters them to generate the semantic-related query to provide richer semantic context. Additionally, SCCA introduces a semantic cross-attention mechanism, feeding the generated semantic-related query back into CLIP to better capture semantic information and the relationships between actions and semantics. By focusing on semantic-related information, SC-CLIP sharpens its ability to distinguish between visually similar actions, ultimately improving its performance on unseen action categories. Evaluations across multiple benchmarks consistently demonstrate improved performance.

Limitations and Future Directions: Despite its demonstrated success, SC-CLIP approach is currently evaluated using the CLIP ViT-B architecture, which may limit insights into how it performs with other model variants. In the future, we aim to apply our method to a wider variety of architectures to evaluate its adaptability and effectiveness. Additionally, while this work primarily centers around the integration of semantic information for improving action recognition, temporal dynamics within video data have not yet been fully explored. Future efforts will aim to incorporate temporal modeling, potentially by developing modules that jointly learn both semantic and temporal relationships. Furthermore, the performance of SC-CLIP is highly dependent on the quality and size of the training dataset. We plan to explore its scalability to more diverse and large-scale datasets to assess

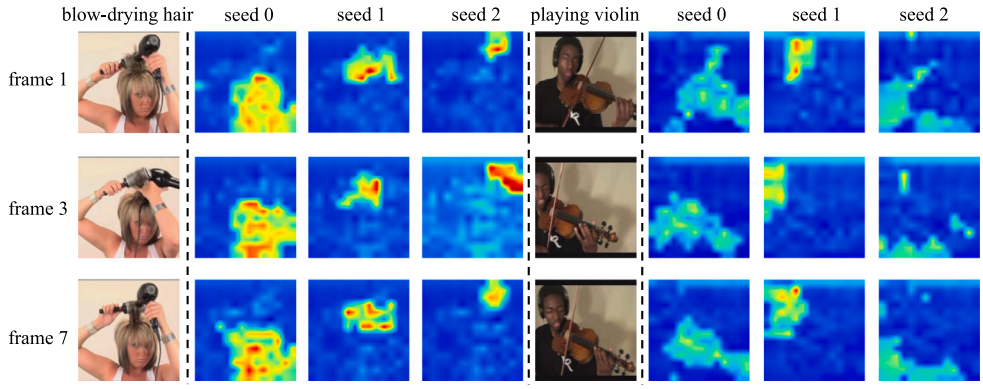
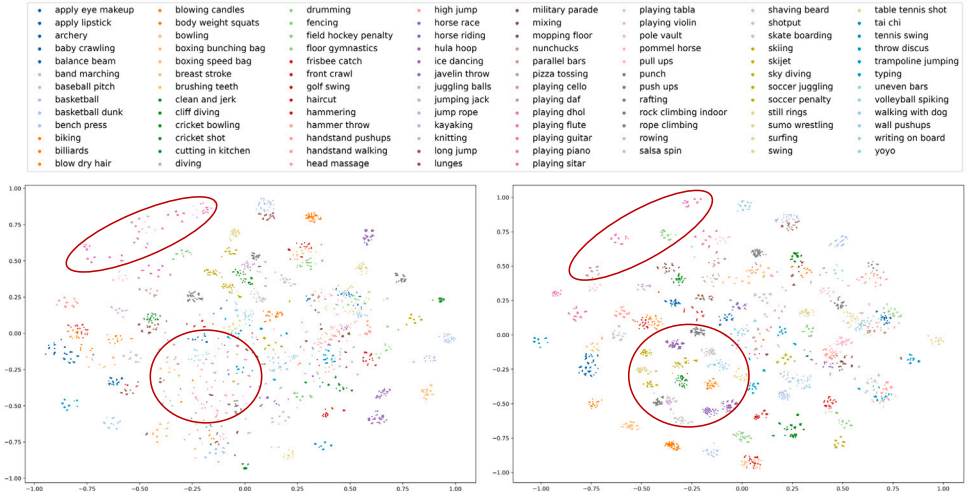


Fig. 6. The visualization of the attention weights after SCCA.



(a) The visualization of Frozen CLIP

(b) The visualization of SC-CLIP

Fig. 7. t-SNE visualization on UCF-101 dataset based on zero-shot setting. (a) is the Frozen CLIP method, and (b) is the SC-CLIP approach.

its robustness across different domains. Although SC-CLIP has been evaluated in action recognition tasks, its performance in cross-domain applications, such as healthcare or autonomous driving, remains to be explored.

CRediT authorship contribution statement

Zhenzhen Quan: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis. **Jialei Chen:** Writing – review & editing, Software. **Daisuke Deguchi:** Writing – review & editing, Funding acquisition. **Jie Sun:** Writing – review & editing. **Chenkai Zhang:** Writing – review & editing. **Yujun Li:** Writing – review & editing, Supervision, Funding acquisition. **Hiroshi Murase:** Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the annual project funding of the Smart State Governance Laboratory, Shandong University, and JSPS KAKENHI Grant Number 23K28164 and 24H00733, and JST CREST Grant Number JPMJCR22D1 . All computations of this paper are carried out on the supercomputer "Flow" at the Information Technology Center, Nagoya University. The author Zhenzhen Quan is sponsored by the China Scholarship Council.

Data availability

Data will be made available on request.

References

[1] D. Lee, J. Lee, J. Choi, CAST: cross-attention in space and time for video action recognition, in: Advances in Neural Information Processing Systems, 2024.

[2] H. Li, M. Li, Q. Peng, S. Wang, H. Yu, Z. Wang, Correlation-guided semantic consistency network for visible-infrared person re-identification, IEEE Trans. Circuits Syst. Video Technol. 34 (6) (2024) 4503–4515.

[3] X. Wang, S. Zhang, Z. Qing, C. Gao, Y. Zhang, D. Zhao, N. Sang, Molo: Motion-augmented long-short contrastive learning for few-shot action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 18011–18021.

[4] Z. Quan, Q. Chen, M. Zhang, W. Hu, Q. Zhao, J. Hou, Y. Li, Z. Liu, MAWKDN: A multimodal fusion wavelet knowledge distillation approach based on cross-view attention for action recognition, IEEE Trans. Circuits Syst. Video Technol. 33 (10) (2023) 5734–5749.

[5] S. Wang, J. Chang, Z. Wang, H. Li, W. Ouyang, Q. Tian, Content-aware rectified activation for zero-shot fine-grained image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 46 (6) (2024) 4366–4380.

[6] Y. Liu, X. Gao, J. Han, L. Shao, A discriminative cross-aligned variational autoencoder for zero-shot learning, IEEE Trans. Cybern. 53 (6) (2022) 3794–3805.
[7] Y. Liu, Y. Dang, X. Gao, J. Han, L. Shao, Zero-shot learning with attentive region embedding and enhanced semantics, IEEE Trans. Neural Netw. Learn. Syst. 35 (3) (2022) 4220–4231.
[8] Y. Liu, K. Tao, T. Tian, X. Gao, J. Han, L. Shao, Transductive zero-shot learning with generative model-driven structure alignment, Pattern Recognit. 153 (2024) 110561.
[9] T. Yang, Y. Zhu, Y. Xie, A. Zhang, C. Chen, M. Li, Aim: Adapting image models for efficient video action recognition, in: Proceedings of the International Conference on Learning Representations, 2023.
[10] S.T. Wasim, M. Naseer, S. Khan, F.S. Khan, M. Shah, Vita-clip: Video and text adaptive clip via multimodal prompting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 23034–23044.
[11] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, in: Proceedings of the International Conference on Learning Representations, 2021.
[12] J. Chen, D. Deguchi, C. Zhang, X. Zheng, H. Murase, Clip is also a good teacher: A new learning framework for inductive zero-shot semantic segmentation, 2023, arXiv preprint arXiv:2310.02296.
[13] H. Rasheed, M.U. Khattak, M. Maaz, S. Khan, F.S. Khan, Fine-tuned clip models are efficient video learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 6545–6554.
[14] J. Chen, D. Deguchi, C. Zhang, X. Zheng, H. Murase, Frozen is better than learning: A new design of prototype-based classifier for semantic segmentation, Pattern Recognit. 152 (2024) 110431.
[15] H. Chen, H. Shi, X. Liu, X. Li, G. Zhao, Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis, Int. J. Comput. Vis. 131 (6) (2023) 1346–1366.
[16] Y. Zhang, Z. Chen, T. Xu, J. Zhao, S. Mi, X. Geng, M.-L. Zhang, Temporal segment dropout for human action video recognition, Pattern Recognit. 146 (2024) 109985.
[17] K. Gedamu, Y. Ji, L. Gao, Y. Yang, H.T. Shen, Relation-mining self-attention network for skeleton-based human action recognition, Pattern Recognit. 139 (2023) 109455.
[18] D. Shao, Y. Zhao, B. Dai, D. Lin, Finegym: A hierarchical video dataset for fine-grained action understanding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2616–2625.
[19] D. Guo, K. Li, B. Hu, Y. Zhang, M. Wang, Benchmarking micro-action recognition: Dataset, methods, and applications, IEEE Trans. Circuits Syst. Video Technol. 34 (7) (2024) 6238–6252.
[20] V. Mazzia, S. Angarano, F. Salvetti, F. Angelini, M. Chiaberge, Action transformer: A self-attention model for short-time pose-based human action recognition, Pattern Recognit. 124 (2022) 108487.
[21] J. Pan, Z. Lin, X. Zhu, J. Shao, H. Li, St-adapter: Parameter-efficient image-to-video transfer learning, in: Advances in Neural Information Processing Systems, 2022, pp. 26462–26477.
[22] R. Zellers, Y. Choi, Zero-shot activity recognition with verb attribute induction, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 946–958.
[23] J. Gao, T. Zhang, C. Xu, I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 8303–8311.
[24] 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: Proceedings of the International Conference on Machine Learning, 2021, pp. 8748–8763.
[25] X. Huang, H. Zhou, K. Yao, K. Han, Froster: Frozen clip is a strong teacher for open-vocabulary action recognition, in: Proceedings of the International Conference on Learning Representations, 2024.
[26] B. Ni, H. Peng, M. Chen, S. Zhang, G. Meng, J. Fu, S. Xiang, H. Ling, Expanding language-image pretrained models for general video recognition, in: Proceedings of the European Conference on Computer Vision, ECCV, 2022, pp. 1–18.
[27] Z. Wang, S. Wang, H. Li, Z. Dou, J. Li, Graph-propagation based correlation learning for weakly supervised fine-grained image classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 12289–12296.
[28] S. Gao, Z.-Y. Li, M.-H. Yang, M.-M. Cheng, J. Han, P. Torr, Large-scale unsupervised semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 45 (6) (2022) 7457–7476.
[29] K. Soomro, A.R. Zamir, M. Shah, UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild, CRCV- TR- 12- 01, 2012.
[30] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, HMDB: a large video database for human motion recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2011, pp. 2556–2563.
[31] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, et al., The “something something" video database for learning and evaluating visual common sense, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2017, pp. 5842–5850.
[32] J. Carreira, 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.
[33] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, A. Zisserman, A short note about kinetics-600, 2018, arXiv preprint arXiv:1808.01340.
[34] M. Wang, J. Xing, J. Mei, Y. Liu, Y. Jiang, ActionCLIP: Adapting language-image pretrained models for video action recognition, IEEE Trans. Neural Netw. Learn. Syst. (2023) 1–13.
[35] L. Yang, R.-Y. Zhang, Y. Wang, X. Xie, MMA: Multi-modal adapter for vision-language models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 23826–23837.
[36] Z. Yang, G. An, Z. Zheng, S. Cao, Q. Ruan, GBC: Guided alignment and adaptive boosting CLIP bridging vision and language for robust action recognition, IEEE Trans. Circuits Syst. Video Technol. 34 (9) (2024) 8172–8187.
[37] M. Wang, J. Xing, B. Jiang, J. Chen, J. Mei, X. Zuo, G. Dai, J. Wang, Y. Liu, A multimodal, multi-task adapting framework for video action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2024, pp. 5517–5525.
[38] S. Chen, D. Huang, Elaborative rehearsal for zero-shot action recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13638–13647.
[39] C. Zhou, C.C. Loy, B. Dai, Extract free dense labels from clip, in: Proceedings of the European Conference on Computer Vision, ECCV, 2022, pp. 696–712.
[40] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008).