LEARNING VISUAL RELATIONS WITH ADAPTIVE SEMANTIC CALIBRATION FOR FEW-SHOT ACTION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot action recognition (FSAR) aims to recognize novel action categories from only a few labeled videos. Recent approaches have benefited from set-based video matching and semantic knowledge derived from pretrained vision-language models (VLMs). However, semantic knowledge remains unevenly utilized in FSAR: it is underused in support-query relation learning, insufficiently adapted to video-level recognition, and combined with visual evidence under globally shared supervision that ignores query-specific differences. To address these limitations, we propose **Semantic-Adaptive Relation Fusion (SARF)**, which couples semantic-guided visual relation learning with semantic response recalibration and instance-adaptive semantic–visual supervision. Specifically, the **Text-Guided Feature Alignment (TFA)** module incorporates textual knowledge into support-query relation modeling to strengthen discriminative local correspondences between videos. The **Semantic Logits Recalibration (SLR)** module learns residual corrections for VLM-derived semantic logits, adapting semantic responses to the target FSAR task and alleviating confusion among semantically related categories. Finally, the **Semantic–Visual Fusion (SVF)** module estimates instance-dependent weights from visual relation representations and constructs adaptive semantic–visual predictions that provide query-specific auxiliary supervision during training. Extensive experiments on multiple FSAR benchmarks demonstrate that SARF achieves state-of-the-art performance across most evaluated settings, validating the effectiveness of jointly learning fine-grained visual relations and adaptively coordinating semantic and visual evidence under few-shot supervision.

1 INTRODUCTION

Action recognition is a fundamental task in video understanding, with broad applications in human-computer interaction and intelligent surveillance. Despite the remarkable progress of deep video models on large-scale benchmarks Simonyan & Zisserman (2014); Carreira & Zisserman (2017); Feichtenhofer et al. (2019); Bertasius et al. (2021); Shen et al. (2025); Wu et al. (2024b), their success typically relies on abundant labeled training data. In practical scenarios, however, collecting sufficient labeled videos for newly emerging action categories can be costly or even infeasible. Few-shot action recognition (FSAR) addresses this challenge by recognizing novel action classes from only a few labeled examples, providing a data-efficient paradigm for video understanding.

A central challenge in FSAR is to establish reliable correspondences between support and query videos from extremely limited observations. Existing methods have therefore explored temporal alignment, set-to-set matching, and cross-video relation modeling to capture discriminative action correspondences Cao et al. (2020); Perrett et al. (2021); Wang et al. (2022). Meanwhile, large-scale vision-language models such as CLIP (Contrastive Language-Image Pre-training) Radford et al. (2021) provide transferable semantic knowledge that can complement limited labeled supervision. Recent VLM-based FSAR methods have therefore exploited category-level textual priors and pre-trained vision-language representations to improve generalization to novel action classes Wang et al. (2024a); Wu et al. (2024a); Long et al. (2026). These developments highlight two complementary forms of information for FSAR: visual relation evidence derived from support-query correspondences and pretrained semantic knowledge that provides transferable category-level priors.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

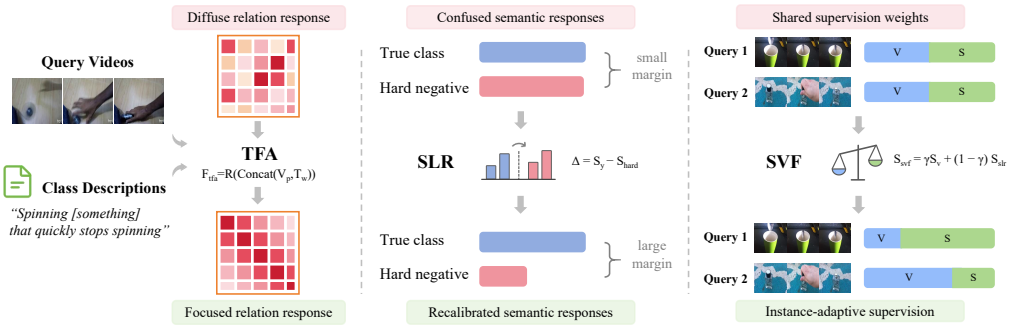


Figure 1: Motivation of SARF. Existing FSAR methods exhibit diffuse visual relation responses, ambiguous semantic predictions, and query-agnostic semantic–visual supervision. SARF addresses these limitations through TFA, SLR, and SVF, respectively.

Despite their complementarity, visual relations and semantic knowledge are often utilized in an imbalanced manner, as illustrated in Figure 1. First, support–query relations are primarily learned from visual features, while textual semantics mainly serve as auxiliary prototypes or prediction cues, limiting their involvement in fine-grained relation modeling. Second, CLIP-derived semantic predictions may remain ambiguous when transferred from image–text pre-training to video-level few-shot recognition, especially for semantically similar action categories. Third, fixed or globally shared semantic–visual supervision applies the same balance across queries, overlooking query-specific variations in the contributions of visual and semantic evidence. These limitations reveal an imbalance in semantic utilization at the relation, response, and supervision levels. This motivates a central question: *how can semantic knowledge guide fine-grained visual relation learning while being adaptively coordinated with visual evidence under few-shot supervision?*

To address this question, we propose **Semantic-Adaptive Relation Fusion (SARF)**, built upon the set-based matching architecture of TS-FSAR Long et al. (2026). SARF organizes visual and semantic evidence into two interacting pathways. Given support and query videos, the visual pathway constructs set-based video representations and introduces category descriptions into support-query relation formation through **Text-Guided Feature Alignment (TFA)**, allowing semantic knowledge to directly participate in fine-grained visual relation learning. In parallel, the semantic pathway obtains CLIP-based class responses and refines them through **Semantic Logits Recalibration (SLR)**, adapting transferred semantic predictions to the target few-shot task and alleviating inter-class ambiguity. The resulting visual relation representations and recalibrated semantic logits are further coupled by **Semantic–Visual Fusion (SVF)**, which predicts query-specific fusion weights from visual relation representations and constructs adaptive semantic–visual predictions for auxiliary supervision during training. Together, these components address semantic utilization imbalance at three complementary levels: textual semantics participate directly in relation formation, CLIP-derived semantic responses are recalibrated under few-shot supervision, while query-specific auxiliary supervision adaptively coordinates visual and semantic evidence. Extensive experiments on multiple standard FSAR benchmarks, including SSv2 Goyal et al. (2017), Kinetics-100 Carreira & Zisserman (2017), HMDB-51 Kuehne et al. (2011), and UCF-101 Soomro et al. (2012), demonstrate consistent improvements over the baseline across different few-shot settings and state-of-the-art performance in most evaluated settings.

2 Related Work

2.1 Few-shot Action Recognition

FSAR aims to recognize novel action categories from only a few labeled videos. Existing methods mainly focus on improving video representations and support-query matching. Early works enhance video features or class prototypes through memory mechanisms and attentive aggregation Zhu & Yang (2018); Bishay et al. (2019); Wang et al. (2021). OTAM Cao et al. (2020) and TRX Perrett et al. (2021) model temporal correspondences between support and query videos, while

HyRSM Wang et al. (2022) and TEAM Lee et al. (2025) further develop fine-grained set-based and temporal-alignment-free matching, respectively. More recently, Trokens Kumar et al. (2025) introduces semantic-aware relational trajectories to improve motion representation.

With the emergence of vision-language models, pretrained visual and textual knowledge has also been introduced into FSAR. CLIP-FSAR Wang et al. (2024a) uses textual priors to modulate visual prototypes, EMP-Net Wu et al. (2024a) exploits multi-level CLIP representations for efficient adaptation, and TS-FSAR Long et al. (2026) combines CLIP adaptation with task-specific relation matching. Despite these advances, category-level textual semantics from pretrained VLMs are still rarely incorporated directly into episodic support-query relation construction, while transferred semantic predictions are often used without explicit adaptation to the target few-shot task.

2.2 Semantic Guidance and Logits Calibration

Pretrained vision-language models such as CLIP Radford et al. (2021) provide transferable semantic priors for few-shot recognition, motivating various approaches to adapt such knowledge to downstream tasks. CoOp Zhou et al. (2022b), CoCoOp Zhou et al. (2022a), and MaPLe Khattak et al. (2023) adapt textual or multimodal prompts, while CLIP-Adapter Gao et al. (2024) and Tip-Adapter Zhang et al. (2022) incorporate few-shot visual evidence through lightweight adaptation or cache-based retrieval. More recently, TIMO Li et al. (2025b) explores mutual image-text guidance to jointly refine visual and textual representations.

Beyond adapting prompts and representations, transferred semantic predictions may still suffer from a mismatch with downstream few-shot recognition. Logits DeConfusion Li et al. (2025a) reveals substantial inter-class confusion in CLIP logits and introduces residual correction to alleviate semantic ambiguity. However, existing studies mainly investigate semantic adaptation or logits calibration for image-level recognition, leaving their interaction with structured support-query relation learning in videos largely unexplored.

3 Method

In this section, we first formulate the few-shot action recognition task and present the overall framework of our proposed Semantic-Adaptive Relation Fusion (SARF). We then introduce its three main components: Text-Guided Feature Alignment (TFA), Semantic Logits Recalibration (SLR), and Semantic–Visual Fusion (SVF). Finally, we describe the training objective and inference protocol.

3.1 Problem Formulation and Method Overview

Problem Formulation. We follow the standard episodic formulation of few-shot action recognition Cao et al. (2020); Perrett et al. (2021); Wang et al. (2022). In an N-way K-shot episode, the support set is defined as $\mathcal{S} = \{(x_i^s, y_i^s)\}_{i=1}^{N\times K}$, where each of the N action categories contains K labeled videos, and the query set is denoted as $\mathcal{Q} = \{x_i^q\}_{i=1}^{N\times N_q}$, where N_q denotes the number of query videos per category. Given $\mathcal{S}$, the model predicts the label of each query video within the episode-specific N-way label space.

Overall Architecture. As illustrated in Figure 2(a), SARF is built upon a CLIP-based support–query matching framework. Given the support and query videos, the frozen CLIP visual encoder extracts frame-level visual representations, from which the video representation F_{clip} is obtained. In parallel, the Ladder Side Network (LSN) Long et al. (2026) provides the global token V_{cls} and local patch tokens V_p for relation modeling. Meanwhile, the frozen CLIP text encoder maps the class descriptions to text embeddings T.

Based on these representations, SARF performs relation modeling and semantic refinement during training. For relation modeling, TFA takes V_{cls}, V_p, and T to construct the text-guided relation representation F_{tfa}. In parallel, the dual-α relation modeling produces the main visual relation representation F_{v}. A consistency constraint is introduced between F_{tfa} and F_{v} to incorporate category-level semantic guidance into relation learning. Meanwhile, F_{clip} is matched with T to obtain the CLIP semantic logits S_{clip}, which are further refined by SLR into S_{slr}. The main visual relation representation F_{v} is also temporally aggregated and classified to produce the visual logits S_{v}. Finally,

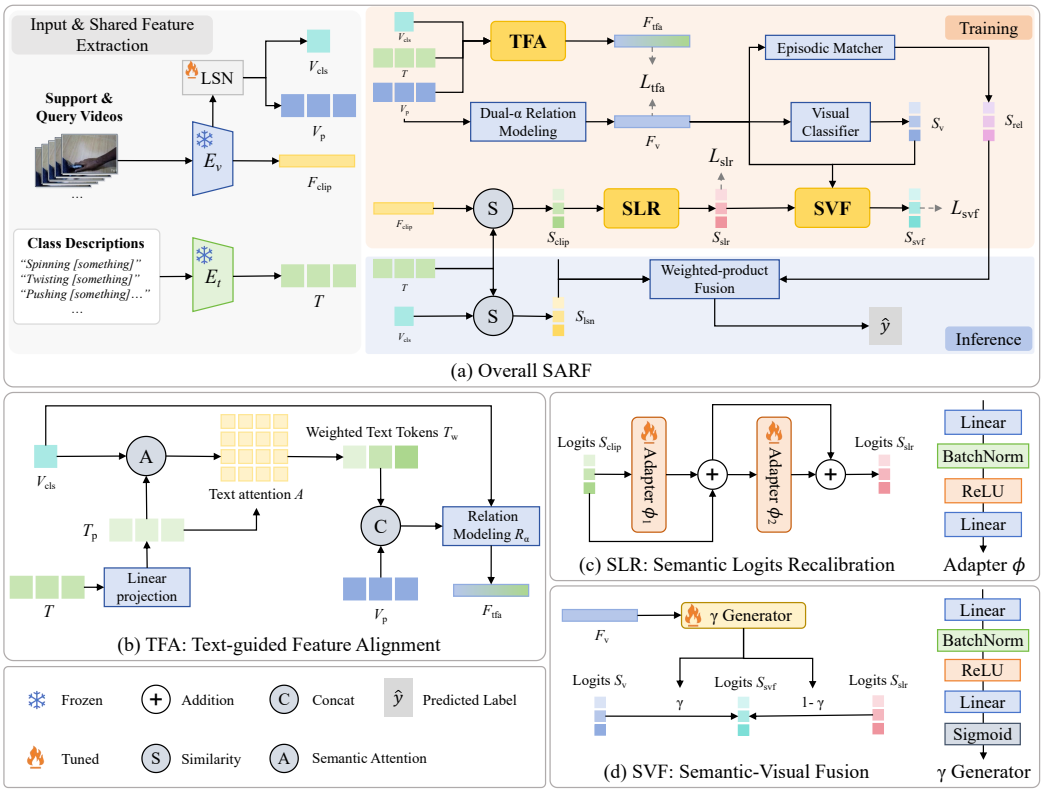


Figure 2: Overview of SARF and its training–inference pipeline. TFA introduces textual semantics into relation modeling, SLR recalibrates CLIP-derived semantic logits, and SVF adaptively fuses visual and semantic logits during training.

SVF adaptively combines S_{v} and S_{slr} to produce S_{svf}, providing additional semantic–visual supervision during training. Through these training objectives, SARF learns improved representations for subsequent episodic prediction.

At inference, SARF directly uses the representations optimized during training to construct the final prediction. The episodic matcher operates on F_{v} to obtain the support–query relation logits S_{rel}. In parallel, F_{clip} and V_{cls} are matched with the class text embeddings to obtain the CLIP semantic logits S_{clip} and the LSN semantic logits S_{lsn}, respectively. The corresponding probabilities of S_{rel}, S_{clip}, and S_{lsn} are then combined for final prediction.

3.2 Text-Guided Feature Alignment

Existing FSAR methods primarily construct support–query relations from visual representations, while the category semantics provided by text are typically exploited only at the prediction stage. This limits their ability to use semantic knowledge when forming fine-grained relation representations. To address this issue, we introduce Text-Guided Feature Alignment (TFA), as illustrated in Figure 2(b). Instead of directly modifying the episodic prediction scores, TFA introduces category-level textual semantics into relation construction during training, yielding a text-guided relation representation that better reflects action-semantic relations.

The text embeddings T are first mapped into the visual feature space through a learnable linear projection, yielding T_p. The same processing is applied to both the support and query branches; for clarity, we describe the support branch below. TFA measures the semantic relevance between $V_{\mathrm{cls}}^{\mathcal{S}}$ and T_p as

$$A^{\mathcal{S}} = \mathrm{Softmax}\big(\tau\ \mathrm{Norm}(V_{\mathrm{cls}}^{\mathcal{S}})\,\mathrm{Norm}(T_p)^{\top}\big)\,, \tag{1}$$

where τ is a similarity scaling coefficient. The resulting attention weights reflect the semantic relevance of each class to the corresponding visual content. During training, T_p contains the projected text embeddings of all training categories, providing broader semantic context for relation learning, while the episodic classifier remains defined over the sampled N classes.

Using the attention weights $A^{\mathcal{S}}$ obtained above, the projected text embeddings T_p are weighted to produce the weighted text tokens $T_w^{\mathcal{S}}$. We then concatenate $T_w^{\mathcal{S}}$ with the local visual tokens $V_p^{\mathcal{S}}$ and feed them into the relation modeling operator R_α:

$$F_{\mathrm{tfa}}^{\mathcal{S}} = R_\alpha\big(V_{\mathrm{cls}}^{\mathcal{S}}, \mathrm{Concat}\big(V_p^{\mathcal{S}}, T_w^{\mathcal{S}}\big)\big)\,. \tag{2}$$

Within R_α, $V_{\mathrm{cls}}^{\mathcal{S}}$ is used to evaluate the relevance of each concatenated token through cosine similarity, and the tokens are reweighted accordingly before relation aggregation. This design emphasizes semantically relevant relation cues while reducing interference from less relevant tokens, yielding a more discriminative text-guided relation representation for fine-grained action matching.

To obtain the main visual relation representation, we employ two relation modeling branches with different α settings. Taking the support branch as an example, the two branches produce $F_{\mathrm{base}}^{\mathcal{S}}$ and $F_{\mathrm{extra}}^{\mathcal{S}}$, respectively, which are fused as

$$F_{\mathrm{v}}^{\mathcal{S}} = \beta F_{\mathrm{extra}}^{\mathcal{S}} + (1-\beta) F_{\mathrm{base}}^{\mathcal{S}}, \tag{3}$$

where β controls the contribution of the additional relation branch. The resulting $F_{\mathrm{v}}^{\mathcal{S}}$ serves as the main visual relation representation for episodic matching. More details of the dual-α design are provided in the supplementary material.

3.3 Semantic Logits Recalibration

Recent CLIP-based FSAR methods leverage textual semantics as complementary cues to visual matching Wang et al. (2024a); Long et al. (2026). In our framework, this semantic information is represented by the CLIP-derived logits S_{clip}. Since these logits are inherited from the pretrained vision–language similarity space and are not explicitly adapted at the logits level to the target FSAR task, their class-wise distinctions can be further refined. Similar observations have also motivated recent studies on CLIP logit deconfusion in few-shot recognition Li et al. (2025a). Accordingly, we introduce Semantic Logits Recalibration (SLR), as illustrated in Figure 2(c), to progressively refine S_{clip} through task-specific residual adaptation.

Specifically, SLR employs two successive residual logits adapters, denoted by ϕ_1 and ϕ_2. The first adapter recalibrates the CLIP-derived semantic logits, while the second further refines the intermediate logits:

$$S_{\mathrm{m}} = S_{\mathrm{clip}} + \phi_1(S_{\mathrm{clip}}), \tag{4}$$

$$S_{\mathrm{slr}} = S_{\mathrm{m}} + \phi_2(S_{\mathrm{m}}), \tag{5}$$

where S_{m} denotes the intermediate logits and S_{slr} denotes the recalibrated semantic logits. Through successive residual refinement, SLR preserves the semantic prior carried by S_{clip} while introducing task-specific corrections, helping adapt the transferred CLIP semantics to the target FSAR task.

3.4 Semantic–Visual Fusion

After obtaining the recalibrated semantic logits, we further exploit their complementarity with visual relation evidence. The main visual relation representation F_{v} is temporally aggregated and passed through a linear classifier to obtain visual logits S_{v}, while SLR provides the recalibrated semantic logits S_{slr}. Both S_{v} and S_{slr} are defined over the training categories. Since the relative importance of visual and semantic cues may vary across video instances, we introduce Semantic–Visual Fusion (SVF), as illustrated in Figure 2(d), to adaptively balance the two types of logits.

Specifically, SVF employs a lightweight γ Generator to predict an instance-specific fusion weight from the temporally aggregated visual relation representation. The generator consists of two linear layers with BatchNorm and ReLU between them, followed by a Sigmoid function:

$$\gamma = G(\mathrm{T_Avg}(F_{\mathrm{v}}))\,, \tag{6}$$

where $\mathrm{T_Avg}$ denotes temporal averaging, and $G(\cdot)$ denotes the γ Generator shown in Figure 2(d).

The fused semantic–visual logits are then computed as

$$S_{\mathrm{svf}} = \gamma S_{\mathrm{v}} + (1-\gamma) S_{\mathrm{slr}}, \tag{7}$$

where γ controls the instance-specific contribution of visual and recalibrated semantic logits.

3.5 Training Objective

SARF introduces three training objectives corresponding to TFA, SLR, and SVF. First, TFA introduces a consistency constraint between the text-guided relation representation F_{tfa} and the main visual relation representation F_{v}:

$$L_{\mathrm{tfa}} = \mathrm{MSE}(\mathrm{Norm}(F_{\mathrm{tfa}}), \mathrm{sg}(\mathrm{Norm}(F_{\mathrm{v}}))), \tag{8}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation. This constraint uses F_{v} as the visual reference to guide F_{tfa} toward the main relation structure while preserving the introduced semantic information.

Second, the recalibrated semantic logits S_{slr} are supervised by the corresponding training-category labels using cross-entropy loss, yielding L_{slr}. This objective directly optimizes the task-specific semantic correction introduced by SLR.

Third, cross-entropy supervision is applied to the fused semantic–visual logits S_{svf}, yielding L_{svf}. This objective encourages SVF to effectively combine the visual logits and recalibrated semantic logits during training.

The three SARF-specific objectives are combined as

$$L_{\mathrm{sem}} = \lambda_{\mathrm{tfa}} L_{\mathrm{tfa}} + \lambda_{\mathrm{slr}} L_{\mathrm{slr}} + \lambda_{\mathrm{svf}} L_{\mathrm{svf}}, \tag{9}$$

where λ_{tfa}, λ_{slr}, and λ_{svf} control the contributions of the corresponding objectives.

In addition, a base training objective L_{base} Long et al. (2026) is incorporated to provide the standard episodic supervision for few-shot action recognition, with its detailed formulation provided in the supplementary material. The overall training objective is therefore

$$L = L_{\mathrm{base}} + L_{\mathrm{sem}}. \tag{10}$$

By jointly optimizing L_{base} and L_{sem}, SARF strengthens visual relation learning with textual guidance while introducing task-specific semantic supervision. The resulting representations are then directly used for episodic inference. Specifically, for each test episode, the learned visual relation representation is used by the episodic matcher to produce the support–query relation logits S_{rel}. In parallel, the CLIP and LSN semantic branches produce the semantic logits S_{clip} and S_{lsn}, respectively. These three predictions provide complementary relation and semantic cues over the candidate classes. Their logits are normalized by Softmax, and the resulting probability distributions are combined through a fixed weighted-product fusion. The class with the highest fused score is taken as the final prediction.

4 Experiments

In this section, we first describe the experimental settings and then compare SARF with existing methods across multiple benchmarks and few-shot settings. We further conduct ablation studies to evaluate the contributions of individual components and analyze the effectiveness of key design choices.

4.1 Experimental Settings

Datasets. Following prior protocols Cao et al. (2020); Perrett et al. (2021); Wang et al. (2022); Long et al. (2026), we evaluate SARF on four widely used datasets: SSv2 Goyal et al. (2017), Kinetics-100 Carreira & Zisserman (2017), HMDB-51 Kuehne et al. (2011), and UCF-101 Soomro et al. (2012). For SSv2, we report results on both the Small and Full splits.

Table 1: Top-1 accuracy (%) comparison of SARF with existing FSAR methods across five benchmarks under the 1-shot and 5-shot settings. Results marked with † are reproduced on our platform.

Method	Pre-train	SSv2-Small		SSv2-Full		Kinetics-100		HMDB-51		UCF-101	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
MatchingNet Vinyals et al. (2016)	INet-RN50	31.3	45.5	-	-	53.3	74.6	-	-	-	-
MAML Finn et al. (2017)	INet-RN50	30.9	41.9	-	-	54.2	75.3	-	-	-	-
ProtoNet Snell et al. (2017)	-	-	-	-	-	64.5	77.9	54.2	68.4	74.0	89.6
TARN Bishay et al. (2019)	INet-RN50	-	-	-	-	64.8	78.5	-	-	-	-
ARN Zhang et al. (2020)	-	-	-	-	-	63.7	82.4	45.5	60.6	66.3	83.1
OTAM Cao et al. (2020)	INet-RN50	36.4	48.0	42.8	52.3	72.2	84.2	54.5	68.0	79.9	88.9
ITANet Zhang et al. (2021)	INet-RN50	39.8	53.7	49.2	62.3	73.6	84.3	-	-	-	-
TRX Perrett et al. (2021)	INet-RN50	36.0	56.7	42.0	60.6	63.6	85.9	53.1	75.6	78.2	96.1
TA^{2}N Li et al. (2022)	INet-RN50	-	-	-	-	72.8	85.8	59.7	73.9	81.9	95.1
STRM Thatipelli et al. (2022)	INet-RN50	37.1	55.3	43.1	68.1	62.9	86.7	52.3	77.3	80.5	96.9
MTFAN Wu et al. (2022)	INet-RN50	45.7	60.4	45.7	60.4	74.6	87.4	59.0	74.6	84.8	95.1
HyRSM Wang et al. (2022)	INet-RN50	40.6	56.1	54.3	69.0	73.7	86.1	60.3	76.0	83.9	94.7
HCL Zheng et al. (2022)	INet-RN50	38.7	55.4	47.3	64.9	73.7	85.8	59.1	76.3	82.5	93.9
Huang et al. Huang et al. (2022)	INet-RN50	38.9	61.6	49.3	66.7	73.3	86.4	60.1	77.0	71.4	91.0
Nguyen et al. Nguyen et al. (2022)	INet-RN50	43.8	61.1	43.8	61.1	74.3	87.4	59.6	76.9	84.9	95.9
SloshNet Xing et al. (2023)	INet-RN50	46.5	68.3	46.5	68.3	70.4	87.0	59.4	77.5	86.0	97.1
MoLo (OTAM) Wang et al. (2023)	INet-RN50	41.9	56.2	56.6	70.6	73.8	85.1	59.8	76.1	85.4	95.1
HM^2 Gao et al. (2025)	INet-RN50	42.5	56.5	-	-	75.5	87.5	61.8	77.9	86.6	96.2
CCLN Wang et al. (2024b)	INet-RN50	46.0	61.3	46.0	61.3	75.8	87.5	65.1	78.8	86.9	96.1
TEAM Lee et al. (2025)	INet-RN50	47.2	63.1	-	-	75.1	88.2	62.8	78.4	87.2	96.2
CLIP-Freeze Radford et al. (2021)	CLIP-B/16	29.5	42.5	30.0	42.4	78.9	91.9	58.2	77.0	89.7	95.7
CapFSAR (OTAM) Wang et al. (2026)	BLIP-B/16	45.9	59.9	51.9	68.2	84.9	93.1	65.2	78.6	93.3	97.8
CLIP-FSAR Wang et al. (2024a)	CLIP-B/16	54.6	61.8	61.9	72.1	94.8	95.4	77.1	87.7	97.0	99.1
EMP-Net Wu et al. (2024a)	CLIP-B/16	57.1	65.7	63.1	73.0	89.1	93.5	76.8	85.8	94.3	98.2
MVP-shot Qu et al. (2025)	CLIP-B/16	55.4	62.0	59.9	64.1	91.0	95.1	77.0	88.1	96.8	99.0
D^{2}ST-Adapter Pei et al. (2025)	CLIP-B/16	55.0	69.3	66.7	81.9	89.3	95.5	77.1	88.2	96.4	99.1
VIM Liu et al. (2026)	CLIP-B/16	57.9	63.8	61.2	72.4	92.8	95.4	77.2	87.4	97.0	98.4
TS-FSAR Long et al. (2026)	CLIP-B/16	60.5	70.3	75.1	83.5	96.3	96.6	85.0	88.9	98.7	**99.3**
TS-FSAR † Long et al. (2026)	CLIP-B/16	60.0	69.9	74.0	83.1	96.2	96.6	85.0	88.7	98.6	99.0
SARF (Ours)	CLIP-B/16	**62.0**	**70.8**	**75.6**	**83.6**	**96.4**	**96.8**	**85.7**	**89.2**	**98.9**	99.1

Implementation Details. We adopt CLIP ViT-B/16 Radford et al. (2021) as the pretrained vision-language backbone. For each video, 8 frames are uniformly sampled as input and resized/cropped following the standard FSAR setting. The model is trained episodically using AdamW Loshchilov & Hutter (2019) with a cosine learning-rate schedule. Unless otherwise specified, experiments are conducted under 5-way 1-shot and 5-way 5-shot settings. During evaluation, we sample 10,000 episodes and report the mean Top-1 accuracy (%).

Additional dataset and implementation details are provided in the supplementary material.

4.2 Comparison with State-of-the-Art Methods

Table 1 compares SARF with representative conventional FSAR methods using ImageNet-pretrained backbones and recent VLM-based approaches under the 5-way 1-shot and 5-way 5-shot settings. Overall, SARF achieves state-of-the-art performance in nine of the ten evaluated settings, demonstrating strong generalization across different action recognition benchmarks. In particular, SARF obtains the best results on both SSv2-Small and SSv2-Full, as well as on Kinetics-100 and HMDB-51 under both shot settings. On UCF-101, SARF achieves the best 1-shot accuracy of 98.9% and remains competitive in the 5-shot setting.

Compared with the reproduced TS-FSAR baseline under identical random seeds and evaluation episodes, SARF consistently improves performance across all settings. The gains are generally more pronounced in the 1-shot setting, with improvements of 2.0% and 1.6% on SSv2-Small and SSv2-Full, respectively, suggesting that the proposed semantic-guided relation learning is particularly beneficial when labeled support examples are highly limited. Improvements are also observed on Kinetics-100, HMDB-51, and UCF-101, indicating that the benefits of SARF are not restricted to a particular dataset.

4.3 Ablation Study

To evaluate the contribution of each component and key design choice in SARF, we conduct ablation studies on SSv2-Small under the 5-way 1-shot setting.

Table 2: Ablation studies of SARF on SSv2-Small under the 5-way 1-shot setting.

(a) Ablation study of the main components.

Method	TFA	SLR	SVF	Acc. (%)
Baseline				59.99
+ TFA	✓			60.89
+ SLR		✓		60.64
+ SLR + SVF		✓	✓	61.06
+ TFA + SLR	✓	✓		61.55
SARF	✓	✓	✓	**61.99**

(b) Ablation study of different TFA designs.

Method	Acc. (%)
w/o TFA	59.99
w/o Semantic Attention	59.85
Direct Concat Text Tokens	59.74
Add Fusion	59.81
TFA (Ours)	**60.89**

(c) Ablation study of different SLR designs.

Method	Acc. (%)
w/o SLR	59.99
Single Adapter	60.17
Two Adapters with Concat	59.09
Original DeConfusion-style	60.01
SLR (Ours)	**60.64**

(d) Ablation study of different fusion strategies in SVF.

Fusion Strategy	Acc. (%)
w/o SVF	61.55
Fixed Average Fusion	60.48
Fixed Weighted Fusion	59.60
Adaptive Fusion (Ours)	**61.99**

Effectiveness of Main Components. As shown in Table 2a, both TFA and SLR improve the baseline when introduced individually, demonstrating the benefits of semantic-guided relation learning and semantic logits recalibration. Combining TFA and SLR further improves the accuracy to 61.55%, outperforming the baseline by 1.56% and indicating that the two mechanisms provide complementary benefits. Adding SVF further improves the accuracy to 61.99%, yielding an overall gain of 2.00%. These results demonstrate the complementary contributions of TFA, SLR, and SVF to the overall performance of SARF.

Analysis of TFA Design. Table 2b compares different strategies for incorporating textual semantics into visual relation learning. Removing semantic attention or directly introducing text tokens through concatenation or additive fusion fails to improve the baseline. In contrast, TFA achieves 60.89%, suggesting that selectively weighting category-level textual semantics according to the current visual content is important for effective semantic-guided relation learning.

Analysis of SLR Design. Table 2c compares different structures for semantic logits recalibration. A single adapter improves the baseline from 59.99% to 60.17%, showing that adapting CLIP-derived semantic responses is beneficial. In contrast, the concatenation-based variant degrades performance, indicating that simply increasing transformation complexity does not necessarily improve semantic adaptation. The original DeConfusion-style design yields only a marginal gain, whereas the proposed two-stage residual SLR achieves the best accuracy of 60.64%. These results support the effectiveness of progressively refining CLIP-derived logits through residual corrections for the target FSAR task.

Analysis of SVF Strategy. Table 2d compares different strategies for semantic–visual auxiliary supervision, with TFA and SLR enabled in all variants. Fixed Average Fusion uses equal weights and achieves 60.48%, while Fixed Weighted Fusion assigns a fixed weight to balance the visual and semantic logits and obtains 59.60%, both below the 61.55% baseline. This suggests that a globally fixed balance between visual and semantic evidence is suboptimal. In contrast, the proposed instance-adaptive fusion improves the accuracy to 61.99%. This comparison indicates that conditioning the fusion weights on visual relation features provides more effective semantic–visual supervision than using fixed weights.

We further analyze the fusion coefficient γ predicted by SVF. The 5,000 query samples are divided into four groups according to the 20th, 50th, and 80th percentiles of γ. As shown in Table 3, γ is generally large, indicating that visual relation evidence serves as the dominant supervision signal on SSv2, while recalibrated semantic predictions provide complementary guidance. Meanwhile, the learned weights vary across queries. In particular, the lowest-γ group assigns a substantially

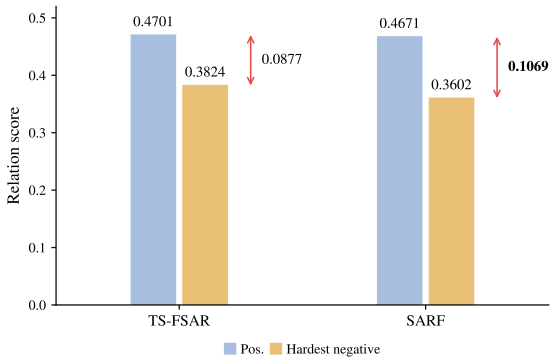


Figure 3: Average ground-truth and hardest-negative relation scores on SSv2-Small.

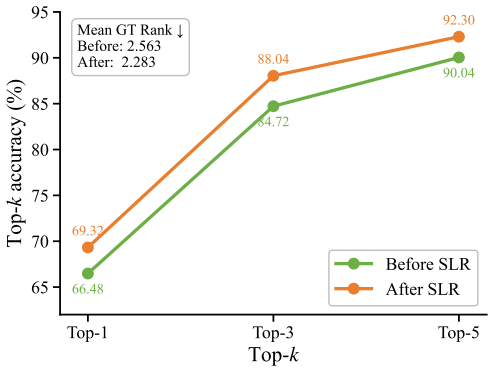


Figure 4: Semantic ranking before and after SLR over all training categories.

larger contribution to semantic evidence and achieves an accuracy of 62.70%. These observations show that SVF does not enforce a fixed semantic–visual balance but adjusts its relative contributions according to individual queries.

Table 3: Distribution of the learned SVF weight γ on SSv2-Small.

γ range	Count	Mean γ	Mean $(1-\gamma)$	Acc. (%)	True-class score
$[0.499, 0.912)$	1000	0.840	0.160	62.70	0.414
$[0.912, 0.967)$	1500	0.944	0.056	60.00	0.384
$[0.967, 0.988)$	1500	0.979	0.021	63.60	0.404
$[0.988, 1.000]$	1000	0.993	0.007	61.10	0.376

4.4 Further Analysis

Visual Relation Margin Analysis. For each query, we compute the relation score of its ground-truth class and the highest score among all incorrect classes, referred to as the hardest-negative score. Figure 3 reports the results averaged over 5,000 queries from 1,000 test episodes. Compared with the reproduced TS-FSAR baseline, SARF enlarges the average relation margin from 0.0877 to 0.1069. While the ground-truth score remains comparable, the hardest-negative score decreases from 0.3824 to 0.3602, indicating that SARF improves relation separability by suppressing confusing negative responses.

Semantic Response Refinement Analysis. To examine the effect of SLR on transferred semantic responses, we compare the semantic predictions before and after recalibration. Here, Top-k accuracy is computed over the full training-category semantic space rather than the episode-specific 5-way label space. Specifically, the ground-truth category is ranked according to the semantic logits for each query, and a prediction is counted as correct if it appears within the top k positions. As shown in Figure 4, SLR improves Top-1, Top-3, and Top-5 accuracy by 2.84%, 3.32%, and 2.26%, respectively, while reducing the mean ground-truth rank by 0.280. These results show that SLR consistently promotes the ground-truth category toward higher positions in the semantic ranking.

5 Conclusion

We propose Semantic-Adaptive Relation Fusion (SARF) for few-shot action recognition. SARF combines text-guided visual relation representation learning, semantic logits recalibration, and instance-adaptive semantic-visual supervision during training. TFA introduces category-level semantics into relation feature construction, SLR adapts transferred semantic responses to the target FSAR task, and SVF provides auxiliary instance-adaptive semantic-visual supervision. Experiments across multiple FSAR benchmarks show the clearest gains on SSv2 and competitive performance on the remaining datasets.

AI USE DISCLOSURE

In this work, we used generative AI tools to assist with literature search and reference verification, manuscript writing, language refinement, and improving the clarity and organization of the paper. All AI-assisted content was reviewed and revised by the authors. We take full responsibility for the final content of this work.

REFERENCES

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

Mina Bishay, Georgios Zoumpourlis, and Ioannis Patras. Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. *arXiv preprint arXiv:1907.09021*, 2019.

Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10618–10627, 2020.

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

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

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Hailiang Gao, Guo-Sen Xie, Rui Yan, Qiongjie Cui, Hongyu Qu, and Xiangbo Shu. Hierarchical motion-enhanced matching framework for few-shot action recognition. *IEEE Transactions on Multimedia*, 27:2450–2462, 2025. doi: 10.1109/TMM.2024.3521712.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International journal of computer vision*, 132(2):581–595, 2024.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pp. 5842–5850, 2017.

Yifei Huang, Lijin Yang, and Yoichi Sato. Compound prototype matching for few-shot action recognition. In *European conference on computer vision*, pp. 351–368. Springer, 2022.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19113–19122. IEEE, 2023.

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

Pulkit Kumar, Shuaiyi Huang, Matthew Walmer, Sai Saketh Rambhatla, and Abhinav Shrivastava. Trokens: Semantic-aware relational trajectory tokens for few-shot action recognition. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13544–13556. IEEE, 2025.

SuBeen Lee, WonJun Moon, Hyun Seok Seong, and Jae-Pil Heo. Temporal alignment-free video matching for few-shot action recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5412–5421, 2025.

Shuo Li, Fang Liu, Zehua Hao, Xinyi Wang, Lingling Li, Xu Liu, Puhua Chen, and Wenping Ma. Logits deconfusion with clip for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25411–25421, 2025a.

Shuyuan Li, Huabin Liu, Rui Qian, Yuxi Li, John See, Mengjuan Fei, Xiaoyuan Yu, and Weiyao Lin. Ta2n: Two-stage action alignment network for few-shot action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

Yayuan Li, Jintao Guo, Lei Qi, Wenbin Li, and Yinghuan Shi. Text and image are mutually beneficial: Enhancing training-free few-shot classification with clip. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5039–5047, 2025b.

Huabin Liu, Tieyuan Chen, Yuxi Li, Shuyuan Li, John See, and Weiyao Lin. Few-shot action recognition via intra-and inter-video information maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Fei Long, Yao Zhang, Jiaming Lv, Jiangtao Xie, and Peihua Li. Task-specific distance correlation matching for few-shot action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7494–7502, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Khoi D Nguyen, Quoc-Huy Tran, Khoi Nguyen, Binh-Son Hua, and Rang Nguyen. Inductive and transductive few-shot video classification via appearance and temporal alignments. In *European Conference on Computer Vision*, pp. 471–487. Springer, 2022.

Wenjie Pei, Qizhong Tan, Guangming Lu, Jiandong Tian, and Jun Yu. D^2st-adapter: Disentangled-and-deformable spatio-temporal adapter for few-shot action recognition. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11317–11326. IEEE, 2025.

Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Temporal-relational crosstransformers for few-shot action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 475–484, 2021.

Hongyu Qu, Rui Yan, Xiangbo Shu, Hailiang Gao, Peng Huang, and Guo-Sen Xie. Mvp-shot: Multi-velocity progressive-alignment framework for few-shot action recognition. *IEEE Transactions on Multimedia*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Zhongwei Shen, Xiao-Jun Wu, Hui Li, Tianyang Xu, and Cong Wu. I know how you move: Explicit motion estimation for human action recognition. *IEEE Transactions on Multimedia*, 27:1665–1676, 2025. doi: 10.1109/TMM.2022.3211423.

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

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 30, 2017.

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

Anirudh Thatipelli, Sanath Narayan, Salman Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Bernard Ghanem. Spatio-temporal relation modeling for few-shot action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19958–19967, 2022.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

Guangge Wang, Haihui Ye, Xiao Wang, Weirong Ye, and Hanzi Wang. Temporal relation based attentive prototype network for few-shot action recognition. In *Asian Conference on Machine Learning*, pp. 406–421. PMLR, 2021.

Xiang Wang, Shiwei Zhang, Zhiwu Qing, Mingqian Tang, Zhengrong Zuo, Changxin Gao, Rong Jin, and Nong Sang. Hybrid relation guided set matching for few-shot action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19948–19957, 2022.

Xiang Wang, Shiwei Zhang, Zhiwu Qing, Changxin Gao, Yingya Zhang, Deli Zhao, and Nong 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*, pp. 18011–18021, 2023.

Xiang Wang, Shiwei Zhang, Jun Cen, Changxin Gao, Yingya Zhang, Deli Zhao, and Nong Sang. Clip-guided prototype modulating for few-shot action recognition. *International Journal of Computer Vision*, 132(6):1899–1912, 2024a.

Xiang Wang, Shiwei Zhang, Hangjie Yuan, Yingya Zhang, Changxin Gao, Deli Zhao, and Nong Sang. Few-shot action recognition with captioning foundation models. *Computer Vision and Image Understanding*, 271:104894, 2026. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2026.104894. URL https://www.sciencedirect.com/science/article/pii/S1077314226002614.

Xiao Wang, Yan Yan, Hai-Miao Hu, Bo Li, and Hanzi Wang. Cross-modal contrastive learning network for few-shot action recognition. *IEEE Transactions on Image Processing*, 33:1257–1271, 2024b.

Cong Wu, Xiao-Jun Wu, Linze Li, Tianyang Xu, Zhenhua Feng, and Josef Kittler. Efficient few-shot action recognition via multi-level post-reasoning. In *European Conference on Computer Vision*, pp. 38–56. Springer, 2024a.

Cong Wu, Xiao-Jun Wu, Tianyang Xu, and Josef Kittler. Scene adaptive mechanism for action recognition. *Computer Vision and Image Understanding*, 238:103854, 2024b.

Jiamin Wu, Tianzhu Zhang, Zhe Zhang, Feng Wu, and Yongdong Zhang. Motion-modulated temporal fragment alignment network for few-shot action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9151–9160, 2022.

Jiazheng Xing, Mengmeng Wang, Yong Liu, and Boyu Mu. Revisiting the spatial and temporal modeling for few-shot action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 3001–3009, 2023.

Hongguang Zhang, Li Zhang, Xiaojuan Qi, Hongdong Li, Philip HS Torr, and Piotr Koniusz. Few-shot action recognition with permutation-invariant attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 525–542. Springer, 2020.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pp. 493–510. Springer, 2022.

Songyang Zhang, Jiale Zhou, and Xuming He. Learning implicit temporal alignment for few-shot video classification. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 1309–1315. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/181. URL https://doi.org/10.24963/ijcai.2021/181. Main Track.

Sipeng Zheng, Shizhe Chen, and Qin Jin. Few-shot action recognition with hierarchical matching and contrastive learning. In *European conference on computer vision*, pp. 297–313. Springer, 2022.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022b.

Linchao Zhu and Yi Yang. Compound memory networks for few-shot video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 751–766, 2018.

A DATASET DETAILS

We provide additional details of the dataset splits and episodic evaluation protocol used in our experiments. For SSv2-Small, we follow the split adopted by TS-FSAR, which contains 64 meta-training, 12 meta-validation, and 24 meta-testing classes, with 100 videos sampled for each class. SSv2-Full uses the same 64/12/24 class partition but retains all available videos from these classes, providing a larger-scale evaluation setting. Kinetics-100 is also divided into 64/12/24 meta-training, meta-validation, and meta-testing classes. For HMDB-51, the 51 action classes are partitioned into 31 meta-training, 10 meta-validation, and 10 meta-testing classes. For UCF-101, we use 70 meta-training, 10 meta-validation, and 21 meta-testing classes. All class partitions are mutually disjoint and follow the standard splits used in prior few-shot action recognition studies. The validation classes are used for model selection and hyperparameter tuning, while all reported results are computed exclusively on the corresponding meta-testing classes.

All datasets are evaluated under the 5-way 1-shot and 5-way 5-shot protocols. Each test episode samples five classes and contains one or five support videos per class and one query video per class, resulting in five query videos per episode. We report the mean classification accuracy over 10,000 randomly sampled test episodes.

B IMPLEMENTATION DETAILS

Textual Descriptions and Comparison Fairness. We use the textual description set provided in Long et al. (2026), which contains ten fine-grained descriptions for each action class. These descriptions characterize visual appearance, motion patterns, involved objects, human–object interactions, and typical contextual cues. Representative examples are provided in Table 4.

Each description is tokenized and encoded by the frozen CLIP text encoder. For class c, given the ten description embeddings $\{t_c^1, t_c^2, \ldots, t_c^{10}\}$, we obtain the class-level semantic representation by averaging and normalization:

$$t_c = \mathrm{Norm}\left(\frac{1}{10}\sum_{j=1}^{10} t_c^j\right). \tag{11}$$

All SARF variants use exactly the same textual descriptions, frozen CLIP text encoder, and aggregation strategy. The descriptions are fixed before training and are not generated or manually selected specifically for SARF. Therefore, the reported improvements reflect differences in how textual semantics are utilized rather than differences in textual inputs.

Optimization and Training. We optimize all trainable components using AdamW with a weight decay of 0.1 and a cosine learning-rate schedule without warm-up. For the 1-shot setting, training is performed for 5,000 episodic iterations with an initial learning rate of 2.5×10^{-4}. We apply color jitter with brightness, contrast, saturation, and hue factors of 0.2, 0.2, 0.2, and 0.05, respectively, together with random erasing. For the 5-shot setting, training lasts for 10,000 episodic iterations with an initial learning rate of 3×10^{-4}, without color jitter or random erasing. Both settings use one episode per batch and random seed 0. Each video is represented by eight uniformly sampled frames, which are resized to a short side of 256 pixels and cropped to 224×224. When enabled, augmentation is temporally consistent across frames, while independently sampled transformations are applied to support and query videos. During evaluation, we use center cropping without color augmentation. All experiments are conducted using NVIDIA GeForce RTX 4080 SUPER GPUs.

Hyperparameter Settings. We set the semantic similarity scale to $\tau = 50$. For dual-α relation enhancement, the base and auxiliary correlation parameters are set to $\alpha_{\mathrm{base}} = 0.4$ and $\alpha_{\mathrm{extra}} = 1.0$, respectively, with residual weight $\beta = 0.05$. The loss weights are set to $\lambda_{\mathrm{tfa}} = 0.1$, $\lambda_{\mathrm{slr}} = 0.5$, and $\lambda_{\mathrm{svf}} = 0.5$. These hyperparameters are selected on the validation split and kept fixed during evaluation.

Dual-α Relation Enhancement. Figure 5 illustrates the dual-α relation enhancement used to construct the main visual relation representation. Given the same visual tokens, two relation modeling branches are employed with different α settings. The main branch adopts an α-distance correlation

Table 4: Representative textual descriptions used in SARF. For each dataset, one action class and two example descriptions are shown; each class is represented by ten descriptions in our experiments.

Dataset	Action Class	Generated Descriptions
SSv2	Pouring [something] into [something]	**(1)** In the video, the action of pouring involves taking a container filled with a liquid or substance and tilting it to allow the contents to flow into another container. This motion requires careful control to ensure that the [something] is directed accurately into [something], preventing spills and ensuring that the desired amount is transferred effectively. **(2)** In the video, the action of pouring involves taking a container filled with a liquid or substance and tilting it to allow the contents to flow into another container. This process showcases the precise movement and control needed to ensure that the desired amount is transferred without spilling, emphasizing the action of pouring effectively.
Kinetics-100	air drumming	**(1)** In the video, the individual is performing air drumming, mimicking the movements of a drummer without actual drums. They use their hands to tap out rhythms in the air, utilizing gestures like drumming on imaginary surfaces and striking invisible cymbals, all while embodying the energy and enthusiasm of a live performance. **(2)** In the video, the person performs air drumming by mimicking the motion of playing drums without actual instruments. They use their hands to simulate hitting drumheads, while their feet tap the imaginary bass drum. The enthusiastic gestures capture the rhythm and energy, showcasing their passion for music and drumming.
HMDB-51	brush hair	**(1)** The person gently lifts a section of hair and runs the brush through it from roots to tips. They repeat this motion smoothly, detangling and smoothing the strands. The brushing helps make the hair neat and tidy, removing any knots and giving it a polished look. **(2)** The person gently lifts a section of hair and runs the brush through it from roots to tips. They repeat this brushing action smoothly, ensuring the hair is untangled and neat. The motion is steady and careful, maintaining hair health while styling.
UCF-101	Blowing Candles	**(1)** In the video, the action of blowing candles involves a person leaning forward toward a cake, taking a deep breath, and forcefully exhaling to extinguish the flames on the candles. This celebratory gesture often follows singing a birthday song, marking a special moment during a birthday celebration. The act of blowing candles symbolizes making a wish. **(2)** In the video, the person begins by making a wish before leaning forward. They take a deep breath and then blow out the candles, creating a puff of air that extinguishes the flames one by one. The surrounding atmosphere is filled with excitement and joy as onlookers celebrate this moment.

setting with $\alpha = 0.4$ and produces the base relation feature F_{base}, which provides the primary relation representation. In parallel, an additional branch with $\alpha = 1.0$ produces F_{extra}, providing a complementary relation response under a different correlation setting.

Rather than directly averaging the two branches, we incorporate the additional relation information through the residual formulation defined in Eq. 3. The coefficient β controls the contribution of the additional branch. This formulation preserves the main relation information in F_{base} while incorporating complementary cues captured by F_{extra}. The resulting F_{v} serves as the main visual relation representation for subsequent episodic matching.

Inherited Training Objectives. For reproducibility, Table 5 summarizes the loss weights adopted from the base training objective Long et al. (2026) under the SSv2-Small 5-way 1-shot setting. We retain the episodic classification, visual and semantic classification, and KL consistency objectives, collectively denoted by L_{base}. The KL consistency term uses a distillation temperature of 4. SARF

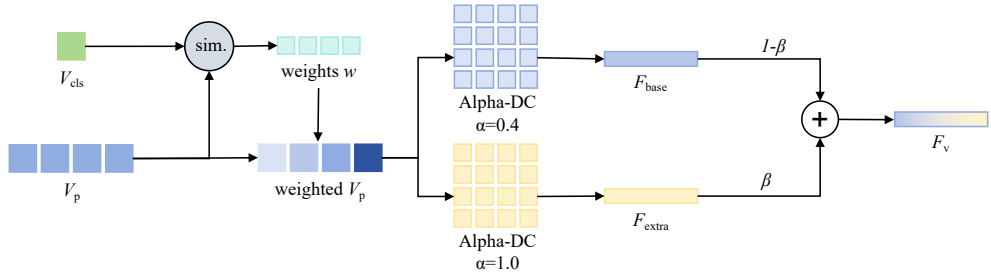
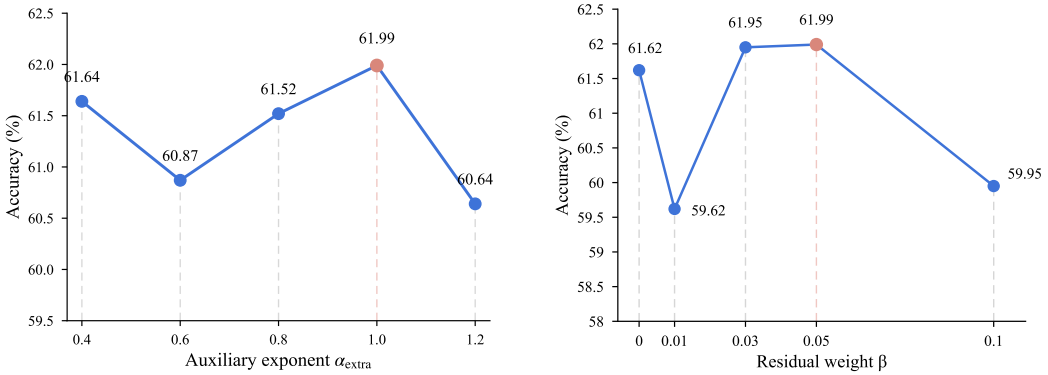


Figure 5: Illustration of the dual-α relation enhancement mechanism. The main branch computes F_{base} with the default α-distance correlation setting, while the auxiliary branch computes F_{extra} with $\alpha = 1.0$. The two relation features are combined through a residual formulation controlled by β, providing complementary local relation cues without replacing the original relation representation.

Table 5: Weights of the inherited training objectives on SSv2-Small under the 5-way 1-shot setting.

Training objective	Weight
Episodic support-query classification	1.0
Visual classification	2.0
CLIP semantic classification	0.1
LSN semantic classification	0.1
KL consistency regularization	0.1

further augments L_{base} with the semantic enhancement objective L_{sem}, which jointly promotes semantic-aware relation learning, task-specific refinement of transferred semantics, and adaptive coordination between visual and semantic cues.



(a) Auxiliary exponent α_{extra}.

(b) Residual weight β.

Figure 6: Sensitivity to the dual-α parameters on SSv2-Small. (a) α_{extra} is varied with $\alpha_{\text{base}} = 0.4$ and $\beta = 0.05$. (b) β is varied with $\alpha_{\text{base}} = 0.4$ and $\alpha_{\text{extra}} = 1.0$.

C ADDITIONAL EXPERIMENTS AND ANALYSES

Sensitivity to Dual-α Parameters. Figure 6 evaluates the sensitivity of the dual-α relation enhancement to α_{extra} and β on SSv2-Small. As shown in Figure 6(a), $\alpha_{\text{extra}} = 1.0$ achieves the best performance among the tested settings, improving the accuracy by 0.35% over $\alpha_{\text{extra}} = \alpha_{\text{base}} = 0.4$. Both smaller and larger values lead to inferior performance, showing that an appropriately different α setting is important for providing complementary relation information.

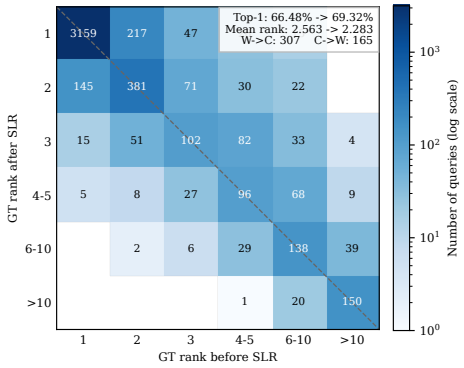


Figure 7: Ground-truth rank transitions before and after SLR on SSv2-Small. Numbers denote the corresponding query counts.

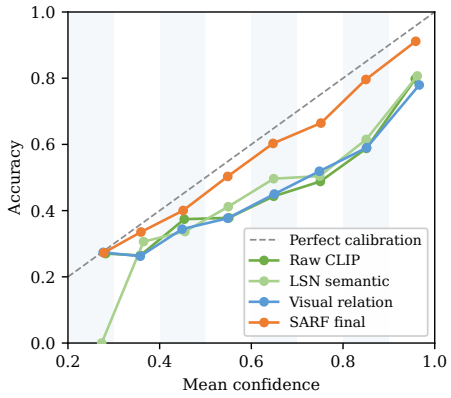


Figure 8: Reliability diagram of different prediction branches on SSv2-Small. The dashed diagonal denotes perfect calibration.

Table 6: Calibration results of different prediction branches on SSv2-Small under the 5-way 1-shot setting.

Logits	Acc. (%)	ECE ↓	MCE ↓	NLL ↓	Brier ↓
Raw CLIP	49.20	0.175	0.263	1.422	0.682
LSN semantic	55.56	0.168	0.277	1.282	0.618
Visual relation	51.94	0.177	0.262	1.387	0.656
SARF final	**61.84**	**0.050**	**0.087**	**0.979**	**0.506**

Table 7: Loss-weight sensitivity on SSv2-Small (5-way 1-shot).

λ_{tfa}	λ_{slr}	λ_{svf}	Acc. (%)
0.05	0.5	0.5	60.7
0.10	0.5	0.5	**62.0**
0.20	0.5	0.5	60.8
0.10	0.3	0.7	60.3
0.10	0.7	0.3	60.5

Figure 6(b) further studies the residual weight β with $\alpha_{\text{base}} = 0.4$ and $\alpha_{\text{extra}} = 1.0$. The best performance is obtained at $\beta = 0.05$, which improves the accuracy by 0.37% points over $\beta = 0$. A larger $\beta = 0.1$ leads to degraded performance, indicating that the additional relation branch is most effective when used as a moderate complement to the main branch. These results support our final choice of $\alpha_{\text{extra}} = 1.0$ and $\beta = 0.05$.

Semantic Rank Transition Analysis. To further investigate how SLR refines the transferred semantic responses, we visualize the sample-level transitions of the ground-truth (GT) rank before and after recalibration. As shown in Figure 7, the horizontal and vertical axes denote the GT ranks before and after SLR, respectively, where smaller ranks indicate better semantic predictions. For clarity, lower-frequency ranks are grouped into 4–5, 6–10, and > 10. Cells above the diagonal correspond to improved GT ranks, whereas cells below the diagonal indicate degraded ranks. A considerable number of samples move toward higher-ranked positions after SLR, particularly from rank 2 or 3 to rank 1. Overall, SLR corrects 307 initially incorrect predictions while introducing 165 new errors, increasing Top-1 accuracy from 66.48% to 69.32% and reducing the mean GT rank from 2.563 to 2.283. These results provide sample-level evidence that SLR effectively refines the semantic response ordering rather than merely changing the logit magnitude.

Calibration and Semantic Response Analysis. Table 6 evaluates the calibration of different prediction branches over 1,000 test episodes. Before evaluation, the final weighted product is normal-

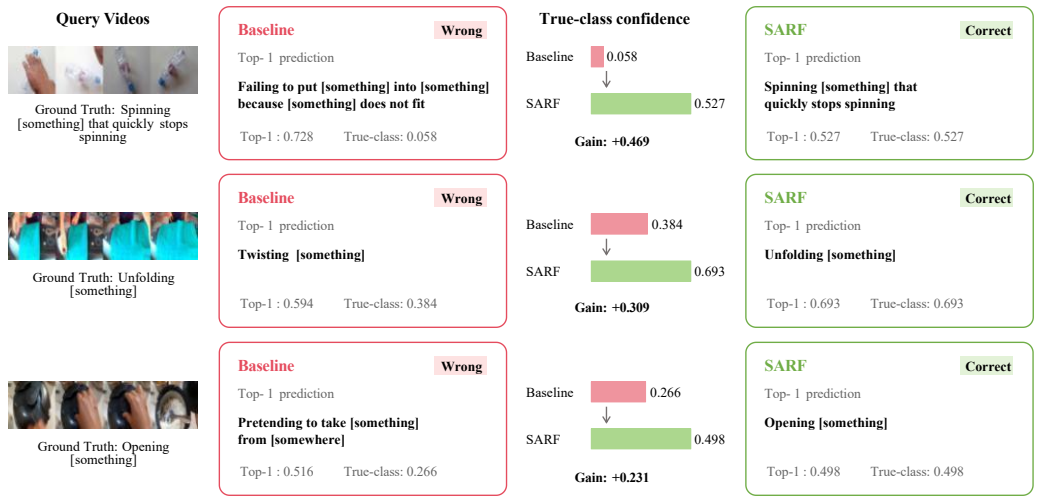


Figure 9: Qualitative examples on SSv2-Small. Compared with the baseline, SARF increases the confidence of the ground-truth class and corrects confusing predictions.

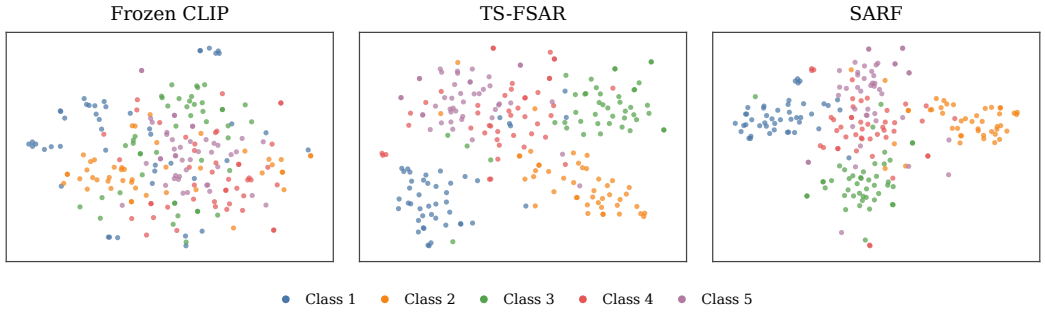


Figure 10: t-SNE visualization of frozen CLIP and visual relation features on SSv2-Small. Colors denote five action classes, and all panels use the same query videos. The Davies–Bouldin Indices (DBI) of frozen CLIP, TS-FSAR, and SARF are 5.2277, 3.6846, and 3.5728, respectively, computed on the original ℓ_2-normalized features (lower is better).

ized into a valid probability distribution. SARF final achieves the lowest ECE, MCE, NLL, and Brier score among the evaluated branches, indicating better agreement between predictive confidence and empirical accuracy. As further illustrated in Figure 8, the SARF final prediction remains closer to the perfect-calibration diagonal across most confidence intervals, while the individual semantic and visual branches show larger deviations. These results demonstrate the effectiveness of SARF in producing better-calibrated predictions with more reliable confidence estimates.

Qualitative Visualization. Figure 9 compares the predictions of the reproduced TS-FSAR Long et al. (2026) baseline and SARF on several ambiguous examples. All confidence values are computed from normalized final probability distributions. While the baseline may favor visually or semantically similar categories, SARF assigns higher probabilities to the ground-truth classes and corrects these predictions. These examples illustrate that SARF learns more discriminative predictions for challenging action categories.

Sensitivity to Loss Weights. We further analyze the sensitivity of SARF to the loss weights in the semantic-enhancement objective. As shown in Table 7, SARF achieves the best performance when $\lambda_{\text{tfa}} = 0.10$, $\lambda_{\text{slr}} = 0.5$, and $\lambda_{\text{svf}} = 0.5$. Reducing or increasing λ_{tfa} leads to lower accuracy, indicating that text-guided feature alignment should provide moderate semantic guidance rather than dominate visual relation learning. When changing the relative weights of SLR and SVF, the performance remains above the reproduced baseline but is lower than the balanced setting. This suggests

that semantic logits recalibration and semantic–visual fusion supervision are complementary and benefit from balanced optimization.

Feature Distribution Visualization. Figure 10 compares frozen CLIP features with the visual relation features of TS-FSAR Long et al. (2026) and SARF. CLIP is projected separately because its feature space differs from that of the relation features; TS-FSAR and SARF are projected jointly. SARF shows clearer separation for several classes, although overlap remains. On the original ℓ_2-normalized features, the Davies–Bouldin Index decreases from 5.2277 (CLIP) to 3.6846 (TS-FSAR) and 3.5728 (SARF), indicating a modest further improvement over the baseline.

Classes 1–5 correspond to Poking a stack of [something] so the stack collapses", Twisting [something]", Dropping [something] into [something]", Pushing [something] with [something]", and "Spinning [something] that quickly stops spinning", respectively.