AQUA: ADAPTIVE QUERYING AND ALIGNMENT FOR INVARIANT PROTOTYPE LEARNING IN FEW-SHOT ACTION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot action recognition (FSAR) aims to recognize novel actions from only a few labeled videos. However, existing methods often struggle to construct reliable and class-consistent prototypes from noisy support sets. We propose AQUA, an Adaptive QUerying and Alignment framework for invariant prototype learning in FSAR. AQUA introduces Invariant Prototype Querying (InvQ), in which learnable queries selectively aggregate class-consistent information across support examples via commonality-biased cross-attention, thereby suppressing sample-level variations. To semantically regularize the raw visual representations, Commonality Alignment (CMAlign) provides an auxiliary training objective that aligns the support representations from InvQ and query representations with class-wise textual anchors. Furthermore, Commonality Adaptation (CMAdapt) dynamically refines textual prototypes through consistency-gated visual conditioning, allowing reliable visual evidence to complement textual semantics. Extensive experiments demonstrate that AQUA consistently improves recognition performance across diverse few-shot action recognition settings, highlighting the effectiveness and generalizability of the proposed framework.

1 INTRODUCTION

Few-shot action recognition (FSAR) aims to recognize novel action classes from only a few labeled videos. Unlike image-based few-shot recognition, action videos exhibit substantial variations in appearance, motion, background, and temporal progression, making it difficult to construct class-representative prototypes from limited support samples. Recent FSAR methods have therefore focused on improving common support representations or query-guided matching (Kong & Fu, 2022; Wanyan et al., 2025; Wang et al., 2022). More recently, pretrained vision-language models such as CLIP (Radford et al., 2021) have further improved FSAR by providing well-aligned semantic priors that complement limited visual supervision (Wang et al., 2024a; 2026).

However, videos from the support set contain not only class-relevant action cues but also sample-specific factors such as background and temporal order, which can bias the resulting prototype. Recent methods (Yang et al., 2026; Wang et al., 2024b) refine class representations using query-support correspondence, but their prototype construction is conditioned on the current query. In contrast, we seek to construct a query-independent class representation directly from the support set by identifying *invariant evidence*: class-representative information that persists across support instances despite subtle variations in the videos.

To construct the invariant prototype despite substantial visual variation across support instances, we selectively extract class-consistent visual evidence from the support set and leverage the shared semantic space of pretrained vision-language models as auxiliary supervision. Specifically, our invariant querying mechanism aggregates visual evidence that consistently recurs across support videos, while class-level textual anchors semantically regularize the support and query representations during training. We further use VLM-generated captions to construct class-wise textual representations that capture richer episode-specific semantics beyond the class label. Since these captions may contain incidental or visually inconsistent descriptions, their contribution is adaptively determined by the constructed invariant visual evidence. As illustrated in Fig. 1, AQUA searches for invariant

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.

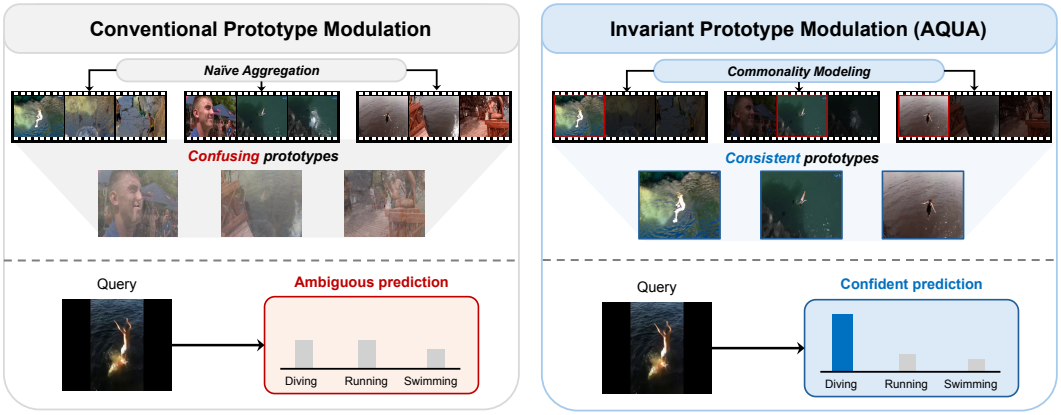


Figure 1: Conceptual comparison between conventional prototype modulation and the proposed invariant prototype modulation. Given support videos with substantial intra-class variation, conventional modulation may retain shot-specific or irrelevant cues, producing confusing prototypes and an ambiguous query prediction. In contrast, AQUA searches for evidence shared across support videos (red boxes) and constructs class-consistent invariant prototypes, resulting in a more confident prediction.

evidence shared across support videos and constructs class-consistent prototypes, thereby reducing shot-specific ambiguity and enabling more confident query predictions.

To this end, we introduce AQUA, a prototype-learning framework built around Invariant Prototype Querying (InvQ) and Commonality Adaptation (CMAdapt), together with an auxiliary Commonality Alignment (CMAlign) objective. InvQ reformulates prototype construction as support-conditioned latent querying, where learnable queries jointly aggregate multiple support videos to emphasize recurring class-consistent evidence while suppressing shot-specific variations. During training, CMAlign regularizes the resulting support and query representations toward their corresponding class-level textual anchors. CMAdapt then uses the invariant visual prototype to refine the caption-derived textual prototype according to visual-text consistency. The resulting visual and textual representations provide complementary class evidence for robust query matching.

Our main contributions are summarized as follows: (1) We propose a novel cross-shot aggregation method, InvQ, which constructs class-representative prototypes by selectively extracting consistent visual evidence from various support videos. (2) We introduce CMAdapt, a visual-to-text prototype adaptation mechanism that uses invariant visual evidence to refine support-dependent textual prototypes according to commonality consistency. (3) We integrate semantic alignment, invariant querying, and prototype adaptation into a unified FSAR framework and demonstrate its effectiveness across four widely used benchmarks.

2 Related Work

2.1 Few-shot action recognition

Few-shot action recognition aims to recognize novel action categories from only a few labeled video examples. Early FSAR methods focused on temporal matching between support and query videos. OTAM (Cao et al., 2020) performs ordered temporal alignment to account for variations in action execution, while TRX (Perrett et al., 2021) models temporal relations using cross-attention over frame tuples. Subsequent methods further improve video matching by incorporating task-specific relations, motion cues, or structured temporal representations. For example, HyRSM (Wang et al., 2022) jointly models intra- and inter-video relations, while MoLo (Wang et al., 2023) exploits motion-augmented representations to enhance temporal discrimination. More recent studies have increasingly emphasized prototype construction in addition to query–support matching. GgHM (Xing et al., 2023) constructs relation-aware prototypes through graph-guided interactions and combines them with hybrid matching, while TEAM (Lee et al., 2025) represents videos using a fixed number

of pattern tokens for flexible temporal comparison. Most recently, MPL (Yang et al., 2026) tightly couples prototype learning with query–support correspondence, using cross-shot aggregation and match-guided refinement to obtain query-discriminative prototypes. Despite these advances, existing FSAR methods mainly improve support-query matching or aggregate support evidence into discriminative prototypes. Such aggregation may retain shot-specific cues that are not consistently shared across instances, while match-guided prototype construction can make representations dependent on query-specific correspondence. In contrast, we focus on extracting cross-shot commonality solely from the support set before downstream matching.

2.2 Multimodal few-shot action recognition

Recent FSAR studies have increasingly explored multimodal information to complement RGB representations. AMFAR (Wanyan et al., 2023) exploits modality complementarity by adaptively identifying reliable modalities and transferring knowledge across them. With the emergence of vision-language models, textual semantics have also become an important source of auxiliary information. CLIP-FSAR (Wang et al., 2024a) leverages video-text alignment and text-guided prototype modulation to refine visual prototypes, while MORN (Ni et al., 2024) constructs visual and textual prototypes separately and integrates them into multimodal prototypes. More recently, LGA (Qian et al., 2025) goes beyond class-label semantics by decomposing actions into fine-grained textual descriptions and aligning them with visual action phases. CapFSAR (Wang et al., 2026) further employs a pretrained captioning foundation model to automatically generate video-specific captions and integrates the resulting textual representations with visual features for few-shot matching. Despite these advances, existing multimodal FSAR methods mainly focus on modality complementarity, text-guided visual enhancement, or multimodal prototype fusion. In contrast, our method separates invariant visual prototype construction from semantic regularization and adaptation. InvQ directly extracts class-consistent visual evidence from the support set, while class-level textual anchors provide an auxiliary alignment objective that semantically regularizes the learned support and query representations during training. The support-derived invariant visual evidence is then used to adapt complementary caption-based textual prototypes.

3 Methodology

3.1 Overview

Problem Definition. Following the standard metric-based episodic protocol, a dataset is split into class-disjoint $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$. Each N-way K-shot episode consists of a support set $\mathcal{S} = \{(x^s_{c,k}, c)\}^{N,K}_{c=1,k=1}$ with K labeled videos per class and a query set $\mathcal{Q} = \{(x^q_i, y^q_i)\}^{|\mathcal{Q}|}_{i=1}$ from the same classes; each query is classified by comparing it with class representations built from $\mathcal{S}$. Meta-testing applies the same protocol to the unseen classes of $\mathcal{D}_{\text{test}}$, and performance is measured by average episode accuracy.

Overall Architecture. The overall architecture of AQUA is illustrated in Fig. 2. AQUA consists of Invariant Prototype Querying (InvQ), Commonality Adaptation (CMAdapt), and an auxiliary Commonality Alignment (CMAlign) objective. Given an episode $(\mathcal{S}, \mathcal{Q})$, each video x is represented by T uniformly sampled frames, while a pretrained vision-language model generates a natural-language caption $\tilde{x}$ offline. A visual encoder f_{v} extracts frame-level visual features, and a text encoder f_{t} with a learnable projection g_{t} encodes the caption:

$$\mathbf{V}(x) = f_{\text{v}}(x) \in \mathbb{R}^{T\times d}, \qquad \mathbf{e}(x) = g_{\text{t}}\left(f_{\text{t}}(\tilde{x})\right) \in \mathbb{R}^{d}, \tag{1}$$

where d denotes the common feature dimension. We denote the support and query visual features by $\mathbf{V}^s_{c,k} = \mathbf{V}(x^s_{c,k})$ and $\mathbf{V}^q_i = \mathbf{V}(x^q_i)$, respectively.

InvQ builds an invariant visual prototype for each class from the support features (Sec. 3.2), CMAlign regularizes the visual representations with class-level textual anchors during training (Sec. 3.3), and CMAdapt uses the invariant prototype to adapt caption-based textual prototypes (Sec. 3.4). Each query is then scored against both representations of every class,

$$s_{i,c} = \lambda_v s^{\text{vis}}_{i,c} + \lambda_t s^{\text{text}}_{i,c}, \tag{2}$$

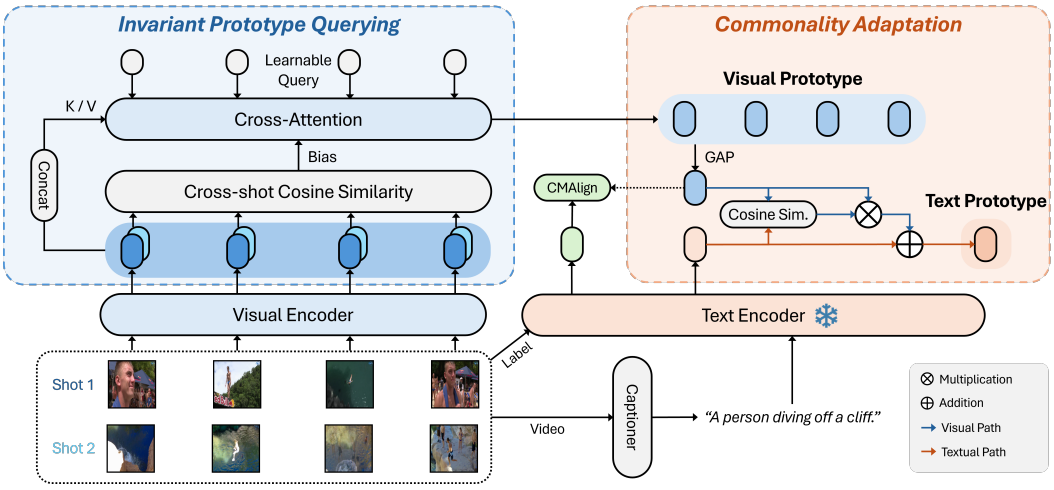


Figure 2: Overview of the proposed AQUA. InvQ (Invariant Prototype Querying) extracts class-consistent visual prototypes through commonality-biased attention, while CMAdapt (Commonality Adaptation) uses invariant visual evidence to adapt textual prototypes. During training, CMAlign (Commonality Alignment) semantically regularizes visual representations using class-level textual anchors. The resulting visual and textual representations are used for query matching.

where $s_{i,c}^{\text{vis}}$ and $s_{i,c}^{\text{text}}$ are the visual and textual matching scores between query x_i^q and class c, and the prediction is $\hat{y}_i^q = \arg\max_{c\in\{1,\dots,N\}} s_{i,c}$.

3.2 Invariant Prototype Querying

Support videos of the same class vary substantially in appearance and temporal progression, so directly averaging their features preserves sample-specific information rather than patterns shared across the support set. InvQ instead extracts class-consistent information from the K support videos into an invariant prototype.

For each class c, the frame-level features of its K support videos are concatenated into a support token bank:

$$\mathbf{S}_c = \text{Concat}\left(\mathbf{V}_{c,1}^s, \dots, \mathbf{V}_{c,K}^s\right) \in \mathbb{R}^{KT\times d}. \tag{3}$$

InvQ employs T learnable query tokens $\mathbf{Q} \in \mathbb{R}^{T\times d}$ to retrieve representative action patterns from $\mathbf{S}_c$. To bias the queries toward patterns shared across support videos, we compute a cross-shot commonality score. For the t-th frame of support video k, the score is defined by its similarity to the most similar frame in every other support video:

$$r_{c,k,t} = \frac{1}{K-1}\sum_{\substack{k'=1\\k'\neq k}}^{K} \max_{t'} \text{sim}\left(\mathbf{V}_{c,k,t}^s, \mathbf{V}_{c,k',t'}^s\right), \tag{4}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity. The scores are standardized and clipped to form an attention bias $\mathbf{b}_c \in \mathbb{R}^{KT}$, assigning larger values to frames consistently observed across support videos. For the one-shot case ($K = 1$), InvQ operates without the computed scores. The max operation allows semantically corresponding frames to contribute to the commonality score even when their temporal positions differ across support videos.

The learnable queries attend to the support token bank using the commonality scores as an additive attention bias. For the h-th attention head,

$$\mathbf{A}_c^{(h)} = \text{softmax}\left(\frac{\left(\mathbf{Q}\mathbf{W}_Q^{(h)}\right)\left(\mathbf{S}_c\mathbf{W}_K^{(h)}\right)^\top}{\sqrt{d_h}} + \gamma \mathbf{1}_T \mathbf{b}_c^\top\right), \tag{5}$$

where d_h denotes the feature dimension of each head, γ controls the contribution of the commonality bias, and $\mathbf{1}_T\mathbf{b}_c^\top \in \mathbb{R}^{T\times KT}$ broadcasts the bias across query tokens. The remaining computation follows standard multi-head attention: each head outputs $\mathbf{H}_c^{(h)} = \mathbf{A}_c^{(h)}\,\mathbf{S}_c\mathbf{W}_V^{(h)}$, and the H head outputs are concatenated and projected by $\mathbf{W}_O$ to form $\mathbf{H}_c \in \mathbb{R}^{T\times d}$. A Transformer encoder layer then refines $\mathbf{H}_c$, and the result is normalized to obtain the invariant visual prototype

$$\mathbf{P}_c^v = \mathrm{Norm}\big(\mathrm{Transformer}(\mathbf{H}_c)\big) \in \mathbb{R}^{T\times d}. \tag{6}$$

InvQ retains a sequence of invariant tokens rather than collapsing the support set into a single vector, preserving temporally distributed action information for subsequent matching and adaptation.

3.3 Commonality Alignment

Visual consistency alone does not guarantee discriminative class semantics. CMAlign therefore uses class-level textual anchors as auxiliary supervision that aligns the learned visual representations with their action semantics during training.

For each episode class c, we construct a prompt $p(c)$ from its action name and encode it using the text encoder:

$$\mathbf{a}_c = \mathrm{Norm}\left(f_{\mathrm{t}}\left(p(c)\right)\right), \tag{7}$$

where $\mathbf{a}_c$ represents the general semantics of class c, in contrast to the video-specific caption representation $\mathbf{e}(x)$ in Eq. 1.

CMAlign supervises both support and query representations. The invariant support prototype and each query feature sequence are temporally pooled and projected into the text space by a shared projection g_a:

$$\mathbf{z}_c^s = \mathrm{Norm}\Big(g_a\Big(\tfrac{1}{T}\textstyle\sum_{t=1}^{T}\mathbf{P}_{c,t}^v\Big)\Big), \qquad \mathbf{z}_i^q = \mathrm{Norm}\Big(g_a\Big(\tfrac{1}{T}\textstyle\sum_{t=1}^{T}\mathbf{V}_{i,t}^q\Big)\Big). \tag{8}$$

Each is compared with the N anchors $\{\mathbf{a}_c\}$ by cosine similarity, and a cross-entropy loss with the ground-truth class as the positive target is applied at both levels:

$$\mathcal{L}_{\text{CMAlign}} = \mathcal{L}_{\text{align}}^s + \lambda_q\mathcal{L}_{\text{align}}^q, \tag{9}$$

where λ_q weights the query-level term.

3.4 Commonality Adaptation

The caption representations provide complementary semantic information for individual support videos. However, visual and textual cues may not always be equally compatible for a given episode, and indiscriminately injecting visual information into textual prototypes can distort their semantic structure. CMAdapt addresses this issue by using visual-text consistency as a reliability signal to control the amount of visual conditioning applied to the textual prototype.

For each class c, the initial textual prototype is obtained by averaging the caption representations of its K support videos:

$$\mathbf{P}_c^{\mathrm{t}} = \mathrm{Norm}\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_{c,k}^s\right), \tag{10}$$

where $\mathbf{e}_{c,k}^s$ denotes the caption representation of the k-th support video from class c. Meanwhile, the invariant visual prototype produced by InvQ is temporally pooled and projected into the textual embedding space:

$$\bar{\mathbf{v}}_c = \mathrm{Norm}\left(g_{\mathrm{v}}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{P}_{c,t}^{\mathrm{v}}\right)\right), \tag{11}$$

where g_{v} is a learnable visual-to-text projection.

CMAdapt estimates the agreement between the visual and textual representations and uses it to determine the amount of visual information injected into the textual prototype:

$$\alpha_c = \sigma\left(\frac{\bar{\mathbf{v}}_c^\top\mathbf{P}_c^{\mathrm{t}}}{\tau_{\mathrm{g}}}\right), \qquad \widetilde{\mathbf{P}}_c^{\mathrm{t}} = \mathrm{Norm}\left(\mathbf{P}_c^{\mathrm{t}} + \eta\alpha_c\bar{\mathbf{v}}_c\right), \tag{12}$$

where σ is the sigmoid, τ_g controls the sensitivity of the gate, and η scales the residual. High visual-text agreement admits stronger visual conditioning, low agreement preserves the original caption semantics, and the residual form prevents the adaptation from overwriting the initial textual prototype.

The same adaptation is applied to each query video. Specifically, its caption representation $\mathbf{e}_i^q$ is modulated using the temporally pooled query visual feature, producing an adapted query representation $\widetilde{\mathbf{e}}_i^q$. The textual matching score in Eq. 2 is then computed as

$$s_{i,c}^{\text{text}} = \text{sim}\left(\widetilde{\mathbf{e}}_i^q, \widetilde{\mathbf{P}}_c^{\text{t}}\right), \tag{13}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity.

3.5 Training Objective

AQUA is optimized episodically using the query samples in each episode. Given the multimodal matching score $s_{i,c}$ defined in Eq. 2, we minimize the cross-entropy loss over the episode classes:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{|\mathcal{Q}|}\sum_{i=1}^{|\mathcal{Q}|}\log\frac{\exp\left(s_{i,y_i^q}\right)}{\sum_{c=1}^{N}\exp\left(s_{i,c}\right)}. \tag{14}$$

Since $s_{i,c}$ combines both branches, $\mathcal{L}_{\text{cls}}$ optimizes them jointly. Adding the alignment loss of Eq. 9 gives the overall objective as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{a}}\mathcal{L}_{\text{CMAlign}}, \tag{15}$$

where λ_{a} controls the contribution of the commonality alignment objective. During optimization, the trainable components of the visual branch together with the feature projections, InvQ, and CMAdapt are jointly updated, while the captioner and text encoder remain frozen.

4 Experiments

Datasets. We conduct experiments on four widely used benchmarks: HMDB51 (Kuehne et al., 2011), Kinetics (Carreira & Zisserman, 2017), UCF101 (Soomro et al., 2012), and SSv2 (Goyal et al., 2017). For HMDB51 and UCF101, we follow the standard few-shot protocols used in prior work (Wang et al., 2022; 2023). For Kinetics, we use the split from previous work (Yang et al., 2026), with 64 meta-training and 24 meta-testing classes. For SSv2, we follow (Cao et al., 2020), using 64 classes for meta-training and 24 for meta-testing.

Implementation Details. Following prior work (Peng et al., 2023; Wang et al., 2022; 2023), we uniformly sample 8 frames from each video and apply random cropping and color jitter. We employ ImageNet-pretrained ResNet-50 (He et al., 2016), CLIP-RN50 (Radford et al., 2021), and CLIP ViT-B/16 (Dosovitskiy et al., 2020) as visual backbones. Video captions are generated offline using Qwen3-VL-8B (Bai et al., 2025) and encoded with the frozen CLIP text encoder. We use Bi-MHM for visual matching and cosine similarity for textual matching. We train the model for five epochs using Adam (Kingma & Ba, 2014) with a learning rate of 10^{-5}. Following (Wang et al., 2023; 2022), we report the mean accuracy and its 95% confidence interval over the 10,000 test episodes. Further implementation details are provided in the supplementary material.

4.1 Comparison with state-of-the-art methods

Table 1 compares AQUA with prior methods across different backbones and shot settings. With ResNet-50 and CLIP-RN50, AQUA consistently achieves strong performance on HMDB51, Kinetics, and UCF101, outperforming prior methods in most settings. These datasets contain substantial appearance- and scene-specific variation, so suppressing support-specific factors and anchoring prototypes to textual concepts yields clear gains. With CLIP ViT-B/16, AQUA remains highly competitive across datasets and shot settings, achieving leading performance in several settings. The narrower performance differences suggest that highly expressive pretrained features leave less room

Table 1: Comparison with state-of-the-art few-shot action recognition methods on HMDB51, Kinetics, UCF101, and SSv2-Small under 5-way 1/5 shot settings. The best results are **bolded** and the second best results are underlined. † denotes the parameter-efficient fine-tuning of the visual encoder using LoRA (Hu et al., 2021) and ConvLoRA (Zhong et al., 2024).

Method	Reference	Backbone	HMDB51		Kinetics		UCF101		SSv2-Small	
			1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
OTAM	CVPR'20	ResNet-50	54.5	68.0	72.2	84.2	79.9	88.9	36.4	48.0
TRX	CVPR'21	ResNet-50	53.1	75.6	63.6	85.9	78.2	96.1	36.0	56.7
CMPT	ECCV'22	ResNet-50	60.1	77.0	73.3	86.4	71.4	91.0	38.9	61.6
HyRSM	ECCV'22	ResNet-50	60.3	76.0	73.7	86.1	83.9	94.7	40.6	56.1
MoLo	CVPR'23	ResNet-50	60.8	77.4	74.0	85.6	86.0	95.5	42.7	56.4
Huang et al.	IJCV'24	ResNet-50	61.6	77.5	74.0	86.9	74.9	92.5	42.6	**61.8**
HyRSM++	PR'24	ResNet-50	61.5	76.4	74.0	86.4	85.8	95.9	42.8	58.0
TEAM	CVPR'25	ResNet-50	62.8	78.4	75.1	88.2	87.2	96.2	–	–
MPL	CVPR'26	ResNet-50	63.3	78.1	77.4	89.5	90.1	97.5	**43.5**	56.6
AQUA (Ours)	–	ResNet-50	**68.6**±0.3	**81.5**±0.5	**86.8**±0.7	**93.8**±0.5	**95.3**±0.2	**98.4**±0.1	40.8±0.4	56.7±0.4
CLIP-Freeze	ICML'21	CLIP-RN50	51.4	71.0	68.2	85.3	84.6	94.5	26.8	36.3
CLIP-FSAR	IJCV'24	CLIP-RN50	64.2	80.0	82.2	91.2	87.1	97.0	40.3	54.5
MPL	CVPR'26	CLIP-RN50	68.0	**84.9**	83.7	93.8	93.0	98.5	40.1	55.0
AQUA (Ours)	–	CLIP-RN50	71.3±0.3	82.8±0.5	86.4±0.7	94.2±0.5	95.4±0.2	98.2±0.1	40.8±0.4	57.4±0.4
AQUA (Ours)	–	CLIP-RN50†	**71.8**±0.2	84.7±0.5	**88.2**±0.6	**94.6**±0.4	**96.0**±0.2	**98.7**±0.1	**42.5**±0.4	**58.8**±0.4
CLIP-Freeze	ICML'21	CLIP-ViT-B	58.2	77.0	78.9	91.9	89.7	95.7	29.5	42.5
CLIP-FSAR	IJCV'24	CLIP-ViT-B	75.8	**87.7**	89.7	95.0	96.6	99.0	54.5	61.8
MVP-Shot	TMM'25	CLIP-ViT-B	69.2	80.3	91.0	95.1	91.3	97.0	**55.4**	62.0
CapFSAR	CVIU'26	BLIP-ViT-B	70.3	81.3	83.5	92.2	93.1	97.7	45.8	61.1
AQUA (Ours)	–	CLIP-ViT-B	75.9±0.2	84.5±0.4	88.7±0.3	94.9±0.4	96.5±0.2	98.8±0.1	45.1±0.4	62.8±0.3
AQUA (Ours)	–	CLIP-ViT-B†	**76.2**±0.1	86.5±0.3	**89.8**±0.4	**95.4**±0.2	**96.8**±0.3	**99.2**±0.1	48.6±0.3	**65.1**±0.4

for further prototype refinement. Gains are less consistent on SSv2, whose classes are defined by fine-grained motion and object interaction.

Two further trends hold across datasets. AQUA improves more in the 1-shot than in the 5-shot setting; since the commonality bias is inactive for $K = 1$, this gain stems from the textual prior and semantic anchoring compensating for scarce visual evidence, while the 5-shot margins are also bounded by the 94–99% accuracy ceiling. Parameter-efficient fine-tuning (†) helps in almost every setting, most on SSv2, where appearance-centric CLIP features are least aligned with motion-defined categories, and least on saturated UCF101. We hypothesize that lightweight adaptation works because it preserves the pretrained feature space on which both CMAlign (alignment with frozen CLIP text anchors) and InvQ (similarity-based commonality) rely, adapting the backbone without sacrificing the semantic structure needed for class-consistent, shot-invariant prototypes.

4.2 Ablation Studies

Varying Matching Method. We analyze the effect of matching methods and the dependence of AQUA on a specific matching strategy by evaluating our framework with OTAM, TRX, and BiMHM. As shown in Table 2, AQUA maintains consistent performance across different matching methods, indicating that the proposed prototype learning framework generalizes well across matching strategies and can be readily integrated with diverse query-support matching methods.

Varying Modality. To assess the complementarity of the two modalities, we compare visual-only, text-only, and multimodal variants of AQUA (Table 3). The multimodal variant outperforms both single-modality counterparts for every backbone and dataset. Text-only accuracy is nearly identical across backbones, since the caption branch shares the same text encoder, whereas visual-only accuracy improves markedly with the stronger backbone. The gain over the best single modality is largest where the two are most balanced, indicating genuine complementarity rather than dominance of the stronger modality. Moreover, which modality is stronger depends on both dataset and backbone, motivating the episode-adaptive gating of CMAdapt over a fixed fusion rule.

VLM Contribution Analysis. Table 4 compares AQUA with CapFSAR and VLM-based baselines using cosine similarity between query-caption embeddings and mean class-caption embed-

Table 2: Comparison of different matching metrics on Kinetics and SSv2 using an ImageNet-pretrained ResNet-50 backbone.

Method	Backbone	Kinetics		SSv2	
		1-shot	5-shot	1-shot	5-shot
OTAM	ResNet-50	72.2	84.2	36.4	48.0
Ours (OTAM)		**86.7**	**93.5**	**38.5**	**52.3**
TRX		63.6	85.9	36.0	56.7
Ours (TRX)		**86.5**	**93.7**	**36.6**	**56.7**
Bi-MHM		72.3	84.5	40.6	56.1
Ours (Bi-MHM)		**86.8**	**93.8**	**40.8**	**56.7**

Table 3: Comparison of visual-only, text-only, and multimodal approaches on Kinetics and SSv2 with various backbones.

Backbone	Modality	Kinetics		SSv2	
		1-shot	5-shot	1-shot	5-shot
CLIP-RN50	Visual	73.8	82.9	35.7	47.8
	Text	86.3	92.7	38.7	53.0
	Multimodal	**88.2**	**94.6**	**42.5**	**58.8**
CLIP-ViT-B	Visual	81.6	88.6	44.5	58.6
	Text	86.1	92.8	37.3	50.9
	Multimodal	**89.8**	**95.4**	**48.6**	**65.1**

Table 4: Comparison with caption-assisted and VLM-based baselines on Kinetics and SSv2 with CLIP-ViT-B/16 backbone.

Method	Kinetics		SSv2	
	1-shot	5-shot	1-shot	5-shot
CapFSAR	83.5	92.2	<u>45.8</u>	61.1
Video-LLaVA	79.3	92.6	32.7	47.6
Qwen3-VL-2B	83.0	93.1	37.2	57.2
Qwen3-VL-8B	<u>85.6</u>	<u>94.7</u>	39.0	<u>61.5</u>
Ours	**89.8**	**95.4**	**48.6**	**65.1**

Table 5: Ablation study of the proposed key modules of AQUA with CLIP-ViT-B/16 backbone.

Method			Kinetics		SSv2	
CMAlign	InvQ	CMAdapt	1-shot	5-shot	1-shot	5-shot
			87.1	93.3	45.1	58.7
✓			88.4	94.1	46.6	62.4
✓	✓		89.3	94.8	47.6	63.6
✓	✓	✓	**89.8**	**95.4**	**48.6**	**65.1**

dings. Although stronger VLMs generally improve recognition, AQUA consistently outperforms Qwen3-VL-8B on both Kinetics and SSv2. This shows that its gains arise not only from the captioner but also from semantic alignment, invariant prototype construction, and commonality adaptation. The larger gain on SSv2 further demonstrates the complementary role of visual evidence when textual descriptions are insufficient.

Component Ablation. Table 5 adds CMAlign, InvQ, and CMAdapt cumulatively on Kinetics and SSv2 under 5-way 1- and 5-shot settings. The first row denotes the baseline that combines Bi-MHM visual matching with mean-pooled caption prototypes without any proposed module. Each addition improves over the preceding configuration in all four settings, and the full model performs best, indicating complementary rather than redundant benefits. The distribution of the gains is informative. The total improvement is largest on SSv2 5-shot, where the visual branch matters most and multiple shots allow InvQ to exploit cross-shot commonality; CMAlign also yields its largest gain there, suggesting that anchoring visual representations to class semantics is most valuable where captions are least reliable. InvQ still improves the 1-shot settings, although the commonality bias is disabled for $K = 1$, showing that learnable-query refinement is beneficial on its own. CMAdapt contributes more to SSv2 than to Kinetics: on Kinetics, captions already agree well with the visual evidence and the gate saturates, whereas on SSv2, the agreement varies across episodes and selective conditioning matters.

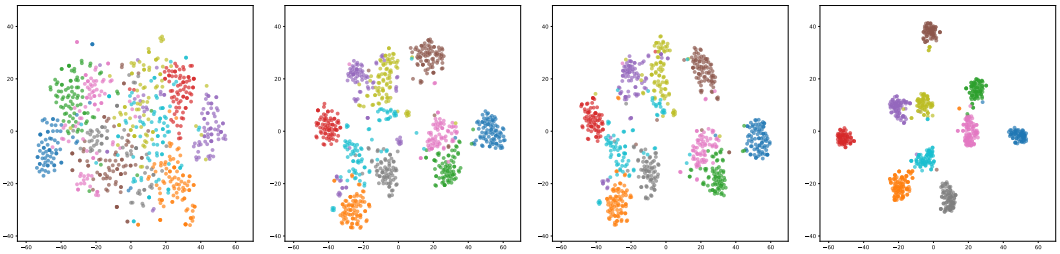
4.3 Visualization results

Effect of Invariant Prototype Querying. Fig. 3 visualizes the cross-attention weights produced by InvQ for support videos from the *Playing Trumpet* class. The red boxes highlight frames receiving peak attention, which consistently capture the characteristic action of playing the instrument despite variations in background and temporal progression. Preceding or background-dominant frames generally receive lower attention weights. These results demonstrate that InvQ selectively identifies class-consistent evidence across support videos without relying on fixed frame-to-frame correspondence. Notably, the attention peaks occur at different temporal positions across the support videos. This behavior illustrates the role of the temporal max operation in Eq. 4, which allows a frame to receive a high commonality score when a semantically similar counterpart appears anywhere in another support video, rather than requiring temporal alignment at the same index.

Feature Distribution Visualization. We qualitatively examine how well the learned representations distinguish action categories by their visual prototypes. Fig. 4 compares AQUA with three



Figure 3: Visualization of cross-attention weights during invariant prototype querying for videos of the 'Playing Trumpet' class.



(a) TRX (b) HyRSM (c) HyRSM++ (d) AQUA

Figure 4: t-SNE visualization of visual prototypes for 10 classes of Kinetics dataset. (5-way 1-shot, 80 episodes), Different colors represent different classes.

representative FSAR approaches Perrett et al. (2021); Wang et al. (2022; 2024b) on the Kinetics test split. The visualization contains 10 action categories collected from 80 episodes under the 5-way 1-shot setting, with each color corresponding to a different category. AQUA forms coherent category-wise groups with limited overlap between classes, indicating that our text-anchor-based aligned visual prototypes retain compact intra-class structures and discriminative inter-class boundaries.

5 Conclusion

We presented AQUA, a prototype-learning framework for few-shot action recognition that extracts class-consistent evidence directly from the support set. InvQ retrieves cross-shot commonality through commonality-biased attention, CMAlign anchors the resulting visual representations to class-level textual semantics during training, and CMAdapt uses the invariant visual prototype to adapt caption-derived textual prototypes according to visual-text agreement. Across four benchmarks and three backbones, AQUA achieves state-of-the-art or competitive accuracy on appearance-driven datasets, and our analyses indicate that these gains arise from the complementarity of the visual and textual branches rather than from the captioner alone. At the same time, AQUA relies on an external captioner at inference, which adds computational cost, and it is less effective on motion-centric categories such as SSv2, where neither frame-level aggregation nor captions captures fine-grained temporal relations; the cross-shot commonality bias also offers little benefit when only a single support video is available. Future work will integrate explicit temporal modeling into invariant prototype learning and replace the captioner with acquiring textual semantics more efficiently.

AI USE STATEMENT

In this work, we used generative AI tools for video caption generation, programming assistance, and manuscript preparation. Specifically, Qwen3-VL-8B-Instruct was used to generate one-sentence action descriptions from videos. These captions were generated offline and used as textual inputs to the proposed method. OpenAI Codex was used to assist with code implementation, debugging, experiment-script preparation, and visualization code. It was also used to improve the organization, grammar, and clarity of the manuscript and to suggest preliminary wording for several sections.

Generative AI tools were not used to generate, modify, or fabricate quantitative experimental results. They were also not used to make final decisions regarding the proposed methodology, experimental design, interpretation of results, or selection of reported findings. All citations and technical claims were manually checked against their original sources. The authors reviewed and, where necessary, revised all generated captions, AI-assisted code, and manuscript text. The code was executed and tested by the authors, and all reported results were obtained from the resulting experiments.

We take full responsibility for the final content of this work, including all text, claims, code, experimental results, and artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

We provide the methodological and experimental details necessary to reproduce our results throughout the paper and the appendix. Section 3 specifies the proposed components, matching procedure, and training objective, while Section 4 describes the datasets, evaluation protocols, baselines, and ablation settings. Appendix A further documents the video and text preprocessing pipeline, including the caption-generation prompt, parameter-efficient backbone adaptation, module configurations and hyperparameters, optimization settings, and evaluation procedure. It also provides additional analyses of caption–label leakage, prototype aggregation, and generalization across episode configurations and backbone capacities. Upon acceptance, we intend to publicly release the implementation, configuration files, and instructions for reproducing the reported experiments together with the camera-ready version.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022.

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.

SuBeen Lee, WonJun Moon, Hyun Seok Seong, and Jae-Pil Heo. Temporal alignment-free video matching for few-shot action recognition. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5412–5421. IEEE, 2025.

Xinzhe Ni, Yong Liu, Hao Wen, Yatai Ji, Jing Xiao, and Yujiu Yang. Multimodal prototype-enhanced network for few-shot action recognition. In *Proceedings of the 2024 international conference on multimedia retrieval*, pp. 1–10, 2024.

Kunyu Peng, Alina Roitberg, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhagen. Delving deep into one-shot skeleton-based action recognition with diverse occlusions. *IEEE Transactions on Multimedia*, 25:1489–1504, 2023.

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.

Zefeng Qian, Xincheng Yao, Yifei Huang, Chongyang Zhang, Jiangyong Ying, and Hong Sun. Beyond label semantics: Language-guided action anatomy for few-shot action recognition. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10421–10431. IEEE, 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.

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.

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, Zhiwu Qing, Zhengrong Zuo, Changxin Gao, Rong Jin, and Nong Sang. Hyrsm++: Hybrid relation guided temporal set matching for few-shot action recognition. *Pattern Recognition*, 147:110110, 2024b.

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*, pp. 104894, 2026.

Yuyang Wanyan, Xiaoshan Yang, Chaofan Chen, and Changsheng Xu. Active exploration of multimodal complementarity for few-shot action recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6492–6502. IEEE, 2023.

Yuyang Wanyan, Xiaoshan Yang, Weiming Dong, and Changsheng Xu. A comprehensive review of few-shot action recognition. *International Journal of Computer Vision*, 133(10):6832–6859, 2025.

Jiazheng Xing, Mengmeng Wang, Yudi Ruan, Bofan Chen, Yaowei Guo, Boyu Mu, Guang Dai, Jingdong Wang, and Yong Liu. Boosting few-shot action recognition with graph-guided hybrid matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1740–1750, 2023.

Feng Yang, Jie Zhao, Fulin Luo, Anyong Qin, Tiecheng Song, Yue Zhao, Chenqiang Gao, and Junwei Han. Mpl: Match-guided prototype learning for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34439–34448, 2026.

Zihan Zhong, Zhiqiang Tang, Tong He, Haoyang Fang, and Chun Yuan. Convolution meets lora: Parameter efficient finetuning for segment anything model. In *International Conference on Learning Representations*, volume 2024, pp. 26755–26779, 2024.

A Additional Implementation Details

A.1 Video and Text Preprocessing

Each video is represented by $T = 8$ uniformly sampled frames. During training, we apply a spatially consistent random crop and color jitter to all frames of a video. At test time, center cropping is used. The frames are resized to 256×256 and cropped to 224×224. We use the normalization statistics associated with each pretrained visual backbone.

Video captions are generated offline using Qwen3-VL-8B-Instruct (Bai et al., 2025). The model samples the input video at 2 frames per second and is prompted to produce a single concise sentence describing the actor, action, interacting object, and scene context. We use the following instruction:

> Describe this action video in exactly one short sentence using this pattern: "The person is [action verb ending in -ing] [the main object if visible] [in/on/at the relevant location or scene if visible]." The sentence must combine the actor, action, and object/location context. Do not output bullet points, labels, JSON, explanations, or multiple sentences. If no clear object is visible, omit the object and describe only the action and location.

Captions are generated for all videos, including support and query videos, before episodic training or evaluation. We sample each input video at 2 frames per second, allow at most 2,048 input frames, and generate at most 128 new tokens. Caption generation is not performed during AQUA training or inference.

The generated captions are encoded using the frozen CLIP ViT-B/16 text encoder (Radford et al., 2021). For the class-level textual anchors used by CMAlign, underscores in each class name are replaced with spaces, and the following template is employed:

$$p(c) = \text{“a photo of } \{\text{class name}\}\text{”} \tag{16}$$

All visual and textual features are projected to a common dimension of $d = 512$ and ℓ_2-normalized.

A.2 Parameter-Efficient Backbone Adaptation

We evaluate AQUA using an ImageNet-pretrained ResNet-50 (He et al., 2016), CLIP-RN50, and CLIP ViT-B/16. For CLIP-based backbones, the pretrained parameters are frozen and only lightweight adapters and the task-specific modules are optimized.

For CLIP ViT-B/16, LoRA (Hu et al., 2021) is applied to the attention and MLP projections of the last four Transformer blocks. Specifically, the adapters are inserted into the input and output

projections of multi-head attention and the two linear layers of the MLP. We use a LoRA rank of 8 and a scaling factor of 16. The second low-rank matrix is initialized to zero so that the initial output remains identical to that of the frozen backbone.

For CLIP-RN50, ConvLoRA (Zhong et al., 2024) is applied to the 1×1 convolutions in the last three residual blocks of `layer4`. LoRA is additionally applied to the query, key, value, and output projections of the attention-pooling layer. Both ConvLoRA and attention-pooling LoRA use rank 8 and scaling factor 16. The batch-normalization layers of the frozen CLIP-RN50 backbone remain in evaluation mode during training. The task-specific projection from the CLIP-RN50 output to the common 512-dimensional space is learnable.

For the ImageNet-pretrained ResNet-50 variant, the complete visual backbone and its 2048-to-512 projection are fine-tuned, whereas the CLIP-based visual backbones remain frozen except for their LoRA parameters and task-specific projection layers.

A.3 Module Configuration

InvQ uses $Q = 8$ learnable query tokens. Its cross-attention module and the subsequent Transformer encoder contain eight attention heads, one Transformer layer, and a dropout rate of 0.1. The commonality scores are standardized over all KT support frames of each class and clipped to $[-3, 3]$. They are added to the cross-attention logits with weight $\gamma = 1.0$. If the score standard deviation is smaller than 10^{-4}, the standardized bias is set to zero. For $K = 1$, the commonality bias is disabled, while the learnable-query attention and Transformer refinement remain active.

CMAdapt uses a gate temperature of $\tau_g = 0.2$ and a residual scale of $\eta = 0.1$. The visual and textual matching scores are combined using $\lambda_v = 0.5$ and $\lambda_t = 0.5$, respectively. We set the weight of CMAlign to $\lambda_a = 0.1$ and the weight of its query-level alignment term to $\lambda_q = 0.25$. The temperature used for classification and text-anchor alignment is 0.07.

Unless otherwise specified, a soft variant of Bi-MHM (Wang et al., 2024b) is used for frame-level visual matching. Bidirectional matching weights are computed using softmax with a temperature of 0.05.

A.4 Optimization and Evaluation

All models are trained episodically for five epochs using Adam (Kingma & Ba, 2014) with a learning rate of 1×10^{-5}, a weight decay of 5×10^{-4}, and an episode batch size of four. The captioner and CLIP text encoder remain frozen throughout training. The visual adapters, feature projections, InvQ, CMAdapt, and the remaining task-specific layers are jointly optimized.

Following prior FSAR evaluation protocols, the main experiments use 5-way 1-shot and 5-way 5-shot episodes with one query video per class at test time. For each dataset and shot setting, we evaluate the final checkpoint on 10,000 randomly sampled episodes from the test split and report the mean classification accuracy. All experiments are implemented in PyTorch.

B Additional Experimental Results

B.1 Caption-Label Leakage Analysis

To examine whether the generated captions directly reveal the ground-truth action categories, we measure the proportion of test captions that explicitly contain their corresponding class labels. We normalize capitalization, punctuation, underscores, and CamelCase class names before matching. A caption is considered an exact match only when the complete normalized class label appears as a contiguous phrase.

Table 6 summarizes the results. Exact class-label occurrences are relatively limited, appearing in 9.04% of Kinetics captions, 18.69% of UCF101 captions, and 4.41% of HMDB51 captions. Across all four benchmarks, 787 out of 8,837 test captions contain the complete class label, corresponding to an overall exact-match rate of 8.91%.

Table 6: Exact class-label occurrence in VLM-generated captions from the test splits. "Matched" denotes captions containing the complete normalized class label as a contiguous phrase.

Dataset	Test captions	Matched	Ratio (%)
Kinetics	2,400	217	9.04
UCF101	2,745	513	18.69
HMDB51	1,292	57	4.41
SSv2-Small	2,400	0	0.00
Overall	8,837	787	8.91

This exact-match criterion provides a conservative estimate of caption–label overlap, as it does not capture morphological variations, paraphrases, or semantically equivalent expressions. For example, "the person is blowing out candles" is not counted as an exact occurrence of the UCF101 class *BlowingCandles*. Similarly, SSv2 class names contain placeholders such as `[something]`, which are typically instantiated with concrete objects in generated captions. Therefore, the 0% exact-match rate on SSv2 should not be interpreted as the complete absence of class-relevant semantic information. These results indicate that verbatim class-label occurrence is not ubiquitous, although it remains non-negligible on UCF101. This analysis measures only direct lexical overlap and does not rule out semantic overlap between captions and action categories.

B.2 Comparison with Mean Aggregation

To further examine the effectiveness of InvQ as a prototype construction mechanism, we compare it with conventional mean aggregation on Kinetics. Mean aggregation assigns equal importance to all support representations and may therefore retain sample-specific or irrelevant information. In contrast, InvQ employs learnable queries to selectively retrieve representative visual evidence and construct a refined sequence-level prototype. As shown in Figure 5, InvQ consistently outperforms mean aggregation in multi-shot settings, with the performance gap becoming larger as more support examples are available.

We further examine the 1-shot setting, where cross-shot commonality is unavailable. Since there is only a single support video, no cross-shot aggregation is performed; the baseline directly uses its frame-level feature sequence as the class representation. InvQ instead refines the same support sequence through learnable-query attention while disabling the cross-shot commonality bias. Under this setting, InvQ achieves performance comparable to the direct support representation, indicating that the learnable-query refinement does not degrade the information contained in a single support example. These results suggest that InvQ provides a unified prototype construction mechanism across different shot settings, while its cross-shot commonality modeling becomes particularly beneficial when multiple support examples are available.

B.3 Generalization across Ways and Backbone Capacities

We further evaluate the robustness of AQUA under varying episode complexity and backbone capacity. Specifically, we compare AQUA with HyRSM++ (Wang et al., 2024b) by increasing the number of classes per episode from 5-way to 10-way and by using visual backbones of different capacities.

As shown in Figure 6, recognition accuracy naturally decreases as the number of episode classes increases, since distinguishing among more candidate categories makes the few-shot task increasingly challenging. Nevertheless, AQUA maintains consistent performance advantages over HyRSM++ across different N-way settings, demonstrating that the proposed prototype learning framework remains effective as episode complexity increases.

We further observe consistent performance across different backbone capacities. The improvements are maintained when using both smaller and larger visual encoders, suggesting that AQUA does not rely on a particular backbone capacity and can effectively complement representations of varying

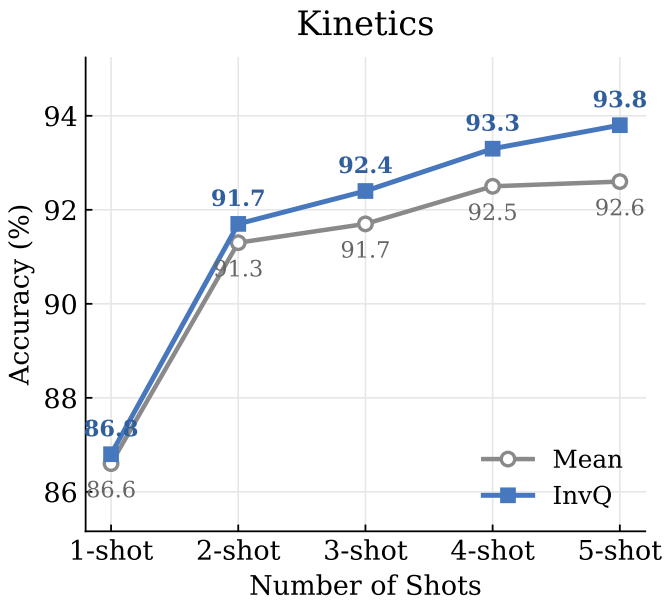


Figure 5: Mean frame-wise aggregation and InvQ on Kinetics under 5-way 1 to 5-shot evaluation.

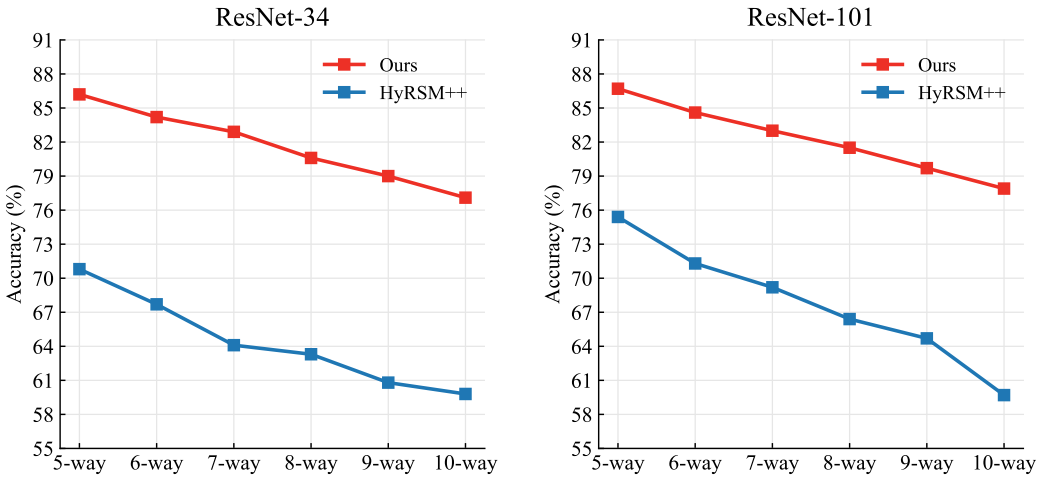


Figure 6: Comparison of AQUA and HyRSM++ (Wang et al., 2024b) under different backbone depths on Kinetics (N-way 1-shot). Performance is evaluated from 5-way to 10-way settings.

strengths. Overall, these results demonstrate the robustness and scalability of AQUA across both task complexity and backbone architecture.