DiA: Directional Adapter

Ashish Kumar Singh[*]
Machine Learning Group
Luleå University of Technology, Sweden
ashish.kumar.singh@associated.ltu.se

Prakash Chandra Chhipa[*]
Scalable Systems
RISE Research Institutes of Sweden, Sweden
prakash.chandra.chhipa@ri.se

[*]***Joint first authors with equal contributions.***

Abstract

Fine-grained action recognition often depends on how an action unfolds over time rather than on appearance alone. Visually similar classes may differ in preparation, execution, or completion phases, making uniformly aggregated temporal context suboptimal. We study this problem under a practical objective: i) improving recognition performance with few tunable parameters, ii) being compute efficient, and iii) using only the readily available video modality. Motivated by this missing temporal observation and the accuracy–compute efficiency–unimodality objective, we propose **Directional Adapter (DiA)**, a parameter-efficient adapter on top of CLIP. DiA formulates temporal specialization into causal and anti-causal directions, allowing past-to-present and future-to-present cues to be modeled differently. To keep this directional modeling parameter- and compute-efficient, both directions use depth-wise temporal convolutions in a compact bottleneck space and are combined through a learnable fusion. DiA achieves state-of-the-art performance among methods using a single video input modality across seven public benchmarks, with only $\sim$3.6M tunable parameters, while also delivering higher throughput and lower inference latency than prior methods. Code is available at https://github.com/prakashchhipa/DiA-Directional-Adapter.

1 Introduction

Fine-grained action recognition requires more than detecting motion. Many action classes are visually similar in appearance and scene context, yet differ in the direction in which motion evidence accumulates over time. In benchmarks such as Diving48 [24] and FineGym [43], class boundaries are often determined by rotation count, transition direction, landing mechanics, body configuration, or short-lived execution stages rather than by objects or background context. The recognition problem is therefore temporal [53] and asymmetric: the *temporal evidence* (discriminative features used in recognition) before an action state and the evidence after it are not interchangeable, and treating the full temporal context uniformly can weaken the cues that separate fine-grained classes.

Further, existing action recognition methods expose two practical challenges. First, full or large-scale fine-tuning of pretrained video models improves adaptation capacity but increases the number of tunable parameters and training cost, as in CLIP-based video methods such as X-CLIP [31] and TimeSformer [2]. Thus, temporal modeling often comes at the cost of either more tunable parameters or higher compute. Second, multi-modal methods can improve fine-action representations, but they add modality-specific data and pre-processing requirements beyond the video modality. Video-text methods such as BDC-CLIP [30] depend on language supervision or prompt construction, while existing fine-grained action recognition methods such as PeVL [54] require pose estimation before recognition. For fine-grained closed-set action recognition, language cues may be weak because class names can be semantically close, whereas the decisive distinction lies in frame-level motion. For

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

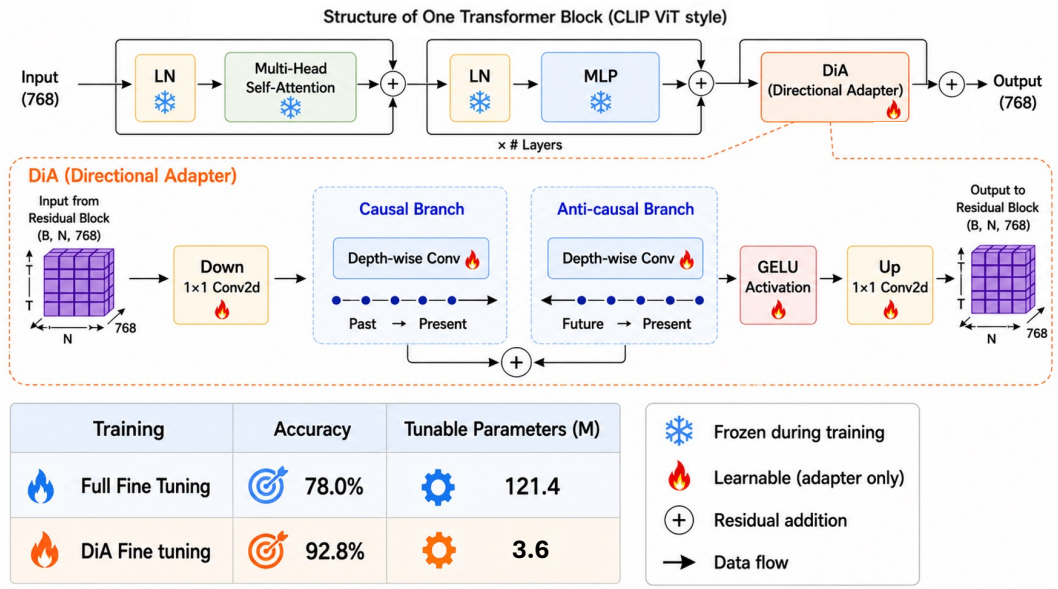

Figure 1: **Overview of DiA.** DiA targets fine-grained action recognition under the accuracy–compute efficiency–unimodality objective. Compared with Full fine-tuning [2] and prior efficient adapters [51], DiA introduces directional temporal specialization to a frozen CLIP visual backbone using lightweight causal and anti-causal adapters with learnable directional fusion.

example, FineGym includes categories such as *switch leap with 0.5 turn* versus *switch leap with 1 turn*, while Diving48 includes categories such as *Forward 25som 2Twis PIKE* versus *Forward 25som 3Twis PIKE*. Additionally, pose estimation can also introduce an additional pre-processing dependency. Thus, multi-modality can add useful information, but it does not remove the need to recover discriminative temporal evidence from the video modality itself.

This leads to our research question: *What temporal formulation is needed to improve fine-grained action recognition using only the video modality, while remaining compute- and parameter-efficient?*

Therefore, we study fine-grained action recognition under the **accuracy–compute efficiency–unimodality objective**: i) improving recognition performance with few tunable parameters, ii) being compute efficient, and iii) using only the readily available video modality. This objective directly follows from the two challenges above. The model must recover asymmetric temporal evidence from video itself, while avoiding the cost of full fine-tuning, heavy video-specific computation, or additional modality pipelines.

We propose **DiA** (**Di**rectional **A**dapter) to address this question (refer Fig. 1). DiA is a parameter-efficient adapter inserted into a CLIP visual backbone. Its design follows directly from the problem structure. Fine-grained actions exhibit temporal asymmetry; therefore, DiA does not treat temporal context uniformly. Instead, it decomposes temporal evidence into causal (past-to-present) and anti-causal (future-to-present) directions. Since the relative importance of these directions can vary across actions, DiA exploits directional fusion [56] with learnable capability. To satisfy the efficiency requirement, temporal specialization is implemented using depth-wise temporal convolutions in a compact bottleneck space and applied as a token-local residual update. This formulation preserves pretrained visual priors while introducing explicit directional temporal specialization for fine-grained action recognition.

Empirical results across seven public benchmarks demonstrate that DiA satisfies the accuracy–compute efficiency–unimodality objective, achieving state-of-the-art performance among single-stream methods using only the video modality. For instance, on Diving48, DiA achieves 92.8% top-1 accuracy with only $\sim$3.6M tunable parameters, improving over prior CLIP-based adaptation while reducing computational cost by 31.24% and increasing throughput by 2.17$\times$ compared to AIM [51]. These results show that directional temporal specialization improves recognition performance while remaining parameter- and compute-efficient using a single video input stream.

Our contributions are:

- We formulate fine-grained action recognition as a **temporal specialization** problem on top of CLIP, targeting the accuracy–compute efficiency–unimodality objective.
- We adapt **directional temporal adaptation**, which decomposes temporal evidence into causal and anti-causal directions to model asymmetric past-to-present and future-to-present cues for **fine-grained action recognition**.
- We propose **learnable directional fusion**, enabling adaptive selection of complementary directional evidence across fine action dynamics.

2 Related Work

Conventional video action recognition methods. Prior RGB action-recognition methods establish the importance of temporal modeling. TSN [46], TSM [26], TRN [55], and TQN [53] model temporal evidence using segment aggregation, temporal shifts, relational reasoning, or temporal queries, while I3D [4] and SlowFast [13] use 3D convolutions or multi-pathway designs. These approaches show that temporal structure matters, but they are not designed around the accuracy–compute efficiency–unimodality objective. They either rely on substantial video-specific computation or do not explicitly target the phase-sensitive directional cues that define fine-grained action boundaries.

CLIP-based video recognition methods. Visual foundation models provide a stronger starting point for this objective. CLIP [37] offers transferable visual priors learned from large-scale pretraining, and CLIP-based video methods such as ActionCLIP [47], X-CLIP [31], EVL [27], and ViFi-CLIP [38] adapt these priors to video recognition. However, full or large-scale adaptation increases tunable parameters and computational cost. Parameter-efficient methods such as ST-Adapter [32] and AIM [51] improve this trade-off by preserving the pretrained backbone and adding lightweight temporal adaptation. Yet their temporal specialization are not explicitly formulated around the core fine-grained observation that preparation, execution, and completion phases may contribute different evidence, and that past-to-present and future-to-present context need not be symmetric.

A separate line of work enriches video representations with language, pose, or both, including BDC-CLIP [30], CLAVER [19], and PeVL [54]. These cues can be useful when auxiliary semantics or body structure are reliable. In fine-grained closed-set recognition, however, the main ambiguity often remains temporal: class names can be nearly identical, while the decisive evidence lies in motion order and phase evolution. Pose can also be unreliable under fast motion or occlusion; FineGym reports only 44.7% detectable joints with AlphaPose [43], and PoseConvC3D reaches 54.5% top-1 accuracy on Diving48 [9]. Thus, relying on auxiliary streams can move the method away from the unimodality part of the objective without directly addressing the missing temporal observation.

3 Directional Adapter

We propose **Di**rectional **A**dapter (**DiA**, refer Fig. 1), a parameter-efficient temporal adapter for fully supervised fine-grained action recognition using a CLIP visual backbone. Fine-grained action categories often share similar appearance and scene context, but differ in temporal order, motion evolution, and short-lived action phases. Existing parameter-efficient video adapters mainly model temporal context symmetrically or combine spatial and temporal adaptation, which can obscure these asymmetric cues. DiA addresses this by applying compact temporal specialization in a bottleneck space, with separate *causal* and *anti-causal* directions whose outputs are combined by learnable directional weights.

Let B denote the batch size, T the number of input frames, $H \times W$ the spatial grid of patch tokens per frame, and C the channel dimension. Given the token tensor entering the adapter,

$$\mathbf{X} \in \mathbb{R}^{(HW+1)\times(BT)\times C}, \tag{1}$$

we separate the class token $\mathbf{x}_{\text{cls}} \in \mathbb{R}^{1\times(BT)\times C}$ and the patch tokens $\mathbf{X}_p \in \mathbb{R}^{HW\times(BT)\times C}$ as

$$\mathbf{X} = [\mathbf{x}_{\text{cls}}; \mathbf{X}_p]. \tag{2}$$

We then reshape the patch tokens into a spatio-temporal tensor

$$\mathbf{V} \in \mathbb{R}^{B\times C\times T\times H\times W}. \tag{3}$$

DiA is derived from four modeling hypotheses for fine-grained action recognition, where the discriminative evidence is often temporally directional and phase-sensitive, such as rotation count, transition order, and execution dynamics:

H1: Low-DoF temporal specialization. Effective supervised adaptation should remain capacity-controlled. We therefore adapt features in a bottleneck space of width

$$C_a = \lfloor \alpha C \rfloor, \tag{4}$$

where C_a is the adapter width and $\alpha \in (0, 1]$ is the bottleneck ratio, and perform temporal modeling using depth-wise convolutions, which restrict the update to local temporal structure with low parameter cost.

H2: Directional temporal asymmetry. Fine-grained actions are not temporally symmetric: evidence from the past and evidence from the future need not contribute equally to recognition. We therefore formulate temporal specialization into a *causal* direction and an *anti-causal* direction rather than using a single symmetric temporal specialization.

H3: Adaptive directional evidence selection. The importance of past-oriented versus future-oriented context varies across datasets, layers, and action phases. Thus, their contributions should not be fixed manually; instead, we apply a learnable directional fusion.

H4: Residual identity-preserving adaptation. A pretrained CLIP encoder already provides strong visual priors. To preserve these priors and stabilize optimization, the adapter should act as a residual update that is initialized near identity at the start of training.

Bottleneck temporal adaptation. We first compress the input features using a 1×1 projection:

$$\mathbf{U} = \phi_{\downarrow}(\mathbf{V}), \qquad \phi_{\downarrow} : \mathbb{R}^{C} \to \mathbb{R}^{C_a}, \tag{5}$$

where $\mathbf{U} \in \mathbb{R}^{B \times C_a \times T \times H \times W}$ is the bottleneck feature tensor. To apply temporal operators independently at each spatial location, we flatten spatial positions and obtain

$$\mathbf{U} \to \mathbf{U}_f \in \mathbb{R}^{(BHW) \times C_a \times T}, \tag{6}$$

where the flattened index $m \in \{1, \ldots, BHW\}$ enumerates spatial locations across the batch.

Formulation of temporal specialization. The novel design of DiA is an explicit decomposition of temporal evidence into two asymmetric pathways defined over $\mathbf{U}_f$, so that each spatial token can separately model past-conditioned and future-conditioned motion evolution. This is important in fine-grained action recognition because categories are often distinguished not only by *what* motion occurs, but by *how* it unfolds directionally over time.

The *causal* direction aggregates only past-to-present evidence:

$$\mathbf{Y}^{\rightarrow} = f_{\text{causal}}(\mathbf{U}_f), \tag{7}$$

where $\mathbf{Y}^{\rightarrow} \in \mathbb{R}^{(BHW) \times C_a \times T}$ and $f_{\text{causal}}(\cdot)$ is a depth-wise temporal convolution with kernel size k and left replication padding. More explicitly, for flattened index m, channel $c \in \{1, \ldots, C_a\}$, temporal index $t \in \{1, \ldots, T\}$, and kernel offset $\delta \in \{0, \ldots, k-1\}$,

$$\mathbf{Y}^{\rightarrow}_{m,c,t} = \sum_{\delta=0}^{k-1} w^{\rightarrow}_{c,\delta}\, \mathbf{U}^{\text{pad}}_{f,m,c,t-\delta}, \tag{8}$$

where $w^{\rightarrow}_{c,\delta}$ denotes the causal filter weight for channel c and offset δ, and $\mathbf{U}^{\text{pad}}_{f}$ denotes the left-padded sequence.

The *anti-causal* direction aggregates only present-and-future evidence:

$$\mathbf{Y}^{\leftarrow} = f_{\text{anti}}(\mathbf{U}_f), \tag{9}$$

where $\mathbf{Y}^{\leftarrow} \in \mathbb{R}^{(BHW) \times C_a \times T}$ and $f_{\text{anti}}(\cdot)$ is a depth-wise temporal convolution with right side replication padding:

$$\mathbf{Y}^{\leftarrow}_{m,c,t} = \sum_{\delta=0}^{k-1} w^{\leftarrow}_{c,\delta}\, \mathbf{U}^{\text{pad}}_{f,m,c,t+\delta}, \tag{10}$$

where $w_{c,\delta}^{\leftarrow}$ denotes the anti-causal filter weight for channel c and offset δ.

This novel formulation imposes an explicit inductive bias toward temporal asymmetry and enables the adapter to preserve distinctions between preparation, execution, and completion phases that are typically collapsed by symmetric temporal updates.

Learnable directional fusion. Rather than enforcing equal contributions from both directions, we learn a global directional weighting. Let

$$\mathbf{a} = [a_{\rightarrow}, a_{\leftarrow}] \in \mathbb{R}^2 \tag{11}$$

be trainable direction logits, where $a_{\rightarrow}$ and $a_{\leftarrow}$ correspond to the causal and anti-causal directions, respectively. We normalize them using softmax:

$$[\lambda_{\rightarrow}, \lambda_{\leftarrow}] = \mathrm{Softmax}(\mathbf{a}), \tag{12}$$

where $\lambda_{\rightarrow}, \lambda_{\leftarrow} \in (0, 1)$ and $\lambda_{\rightarrow} + \lambda_{\leftarrow} = 1$. The fused directional response is then

$$\mathbf{Y}_{\mathrm{dir}} = \lambda_{\rightarrow}\mathbf{Y}^{\rightarrow} + \lambda_{\leftarrow}\mathbf{Y}^{\leftarrow}, \tag{13}$$

where $\mathbf{Y}_{\mathrm{dir}} \in \mathbb{R}^{(BHW)\times C_a \times T}$. This lets the model adaptively determine whether past-oriented or future-oriented evidence is more informative for the task.

Output update. The fused response $\mathbf{Y}_{\mathrm{dir}}$ is reshaped back to $\mathbb{R}^{B\times C_a\times T\times H\times W}$, passed through GELU, projected to the original channel dimension by $\phi_{\uparrow}$, and added residually to $\mathbf{V}$ to produce $\widehat{\mathbf{V}}$, which is then converted back to token form.

Identity-preserving initialization. The output projection $\phi_{\uparrow}$ is zero-initialized so that DiA starts from an identity residual mapping and learns task-specific temporal corrections progressively.

Compact functional form of DiA. The full DiA transformation can be written as

$$\widehat{\mathbf{X}} = [\mathbf{x}_{\mathrm{cls}};\ \mathbf{X}_p + \phi_{\uparrow}(\sigma(\mathcal{F}_{\mathrm{dir}}(\phi_{\downarrow}(\mathbf{X}_p))))], \tag{14}$$

where $\mathcal{F}_{\mathrm{dir}}(\cdot)$ denotes the directional temporal operator defined by

$$\mathcal{F}_{\mathrm{dir}}(\cdot) = \lambda_{\rightarrow} f_{\mathrm{causal}}(\cdot) + \lambda_{\leftarrow} f_{\mathrm{anti}}(\cdot). \tag{15}$$

DiA formulates temporal specialization into **asymmetric causal and anti-causal** decomposition with **learnable directional fusion**, enabling phase-sensitive fine-grained action refinement on frozen CLIP through parameter-efficient, token-local residual adaptation.

Final classification follows the standard CLIP ViT pipeline. Frames are processed by the frozen CLIP transformer, with DiA inserted after the attention and MLP sub-blocks. DiA updates only the patch tokens, while subsequent self-attention propagates the refined information to the $[CLS]$ token. After the final LayerNorm, frame-wise $[CLS]$ features are temporally averaged and passed to a linear classifier for fine-grained action recognition.

4 Results and discussion

4.1 Experimental details

We evaluate DiA on challenging fine-grained action recognition benchmarks: Diving48 [24], FineGym99/288 [43], HAA500 [5], ToyotaSmartHome [6], and MPII Cooking 2 [41], repeated with three random seeds. We use the standard splits for Diving48 , FineGym99/288, HAA500, and ToyotaSmartHome , reporting ToyotaSmartHome under the cross-subject (CS) protocol. For MPII Cooking 2, we use 67 classes with a 70/30 train/test split. Beyond fine-grained action recognition benchmarks, we also report results on two challenging datasets, HMDB51 [21] and Something-Something v2 [15].

All experiments use CLIP ViT-B/16 ($C = 768$) as the visual backbone, with DiA inserted after each of its 12 transformer blocks. DiA uses a bottleneck ratio of $\alpha = 0.25$ ($C_a = 192$) and directional depth-wise temporal convolutions with kernel size $k = 5$ and replication padding. Each layer uses a learnable directional fusion pair initialized as $[0.5, 0.5]$ and normalized with softmax.

Algorithm 1 DiA: Directional Adapter (Forward pass of one DiA module)

Require: Input token tensor $\mathbf{X} \in \mathbb{R}^{(HW+1)\times(BT)\times C}$, frozen CLIP backbone f_θ, DiA parameters $\psi = \{\phi_\downarrow, \phi_\uparrow, f_{\text{causal}}, f_{\text{anti}}, \mathbf{a}\}$, number of frames T, temporal kernel size k

Ensure: Refined token tensor $\widehat{\mathbf{X}}$

Split tokens: $\mathbf{X} \leftarrow [\mathbf{x}_{\text{cls}}; \mathbf{X}_p]$

Reshape patch tokens into video tensor $\mathbf{V} \in \mathbb{R}^{B\times C\times T\times H\times W}$

$\mathbf{U} \leftarrow \phi_\downarrow(\mathbf{V})$

Flatten spatial positions: $\mathbf{U}_f \in \mathbb{R}^{(BHW)\times C_a\times T}$

$\mathbf{Y}^{\rightarrow} \leftarrow f_{\text{causal}}(\mathbf{U}_f)$ ▷ past-to-present temporal modeling

$\mathbf{Y}^{\leftarrow} \leftarrow f_{\text{anti}}(\mathbf{U}_f)$ ▷ present-and-future temporal modeling

$[\lambda_\rightarrow, \lambda_\leftarrow] \leftarrow \text{Softmax}([a_\rightarrow, a_\leftarrow])$

$\mathbf{Y}_{\text{dir}} \leftarrow \lambda_\rightarrow \mathbf{Y}^{\rightarrow} + \lambda_\leftarrow \mathbf{Y}^{\leftarrow}$ ▷ directional fusion

Reshape $\mathbf{Y}_{\text{dir}}$ back to $\mathbb{R}^{B\times C_a\times T\times H\times W}$

$\Delta\mathbf{V} \leftarrow \phi_\uparrow\big(\sigma(\mathbf{Y}_{\text{dir}})\big)$

$\widehat{\mathbf{V}} \leftarrow \mathbf{V} + \Delta\mathbf{V}$

Convert $\widehat{\mathbf{V}}$ back to patch-token layout to obtain $\widehat{\mathbf{X}}_p$

$\widehat{\mathbf{X}} \leftarrow [\mathbf{x}_{\text{cls}}; \widehat{\mathbf{X}}_p]$

return $\widehat{\mathbf{X}}$

Table 1: **Results on Diving48.** DiA achieves the SoTA with substantially fewer tunable parameters. Parameter count 3.6M is for Diving48 and changes slightly depending on the dataset. Total inference parameters: 85.8M (CLIP ViT-B/16) + 3.6M (DiA) = 89.4M. "*" means results reproduced.

Method	Venue	Tunable Params (M)	Frames	Top-1
Non-CLIP-based methods				
ORViT TimeSformer [17]	CVPR'22	160.0	32	88.0
TimeSformer [2]	ICML'21	121.4	16	78.0
BEVT [48]	CVPR'22	88.1	16	86.7
VideoSwin-B [29]	CVPR'22	88.1	32	81.9
SIFAR-B-14 [12]	ICLR'22	87.0	16	87.3
SlowFast R101 [13]	ICCV'19	62.8	64 + 16	77.6
GC-TDN [16]	CVPR'22	27.4	16	87.6
TQN [53]	CVPR'21	–	all	81.8
CTM [35]	CVIU'24	24.4	-	86.5
ART [45]	TNNLS'25	365.0	-	87.9
CLIP-based methods ViT-B/16				
BDC-CLIP [30]*	ICML'25	125.4	16	76.6
EVL [27]*	ECCV'22	32.9	16	51.42
PeVL [54]	CVPR'24	28.0	16	84.5
DiST [36]*	ICCV'23	19.0	16	76.85
AIM [51]	ICLR'23	11.0	32	88.9
DualPath [33]	CVPR'23	10.0	32	88.7
ST Adapter [32]*	NeurIPS'22	7.2	16	84.2
DiA (ours)	-	**3.6**	32	**92.8 ± 0.15**

Following the AIM [51] protocol, we train for 50 epochs using 32-frame clips by default. Unless otherwise specified, we use AdamW with a learning rate of 0.0025, weight decay of 0.05, cosine learning-rate decay, and 2–3 warm-up epochs. These settings are shared across datasets except for the classifier output dimension. Further implementation details are provided in Section E of the Appendix.

Table 2: Results on **FineGym99** and **FineGym288.** † means PeVL [54] results reproduced using video input modality. "*" means results reproduced.

Method	Venue	Frames	Tunable Params (M)	Top-1	
				FineGym99	FineGym288
Non-CLIP-based methods					
TQN [53]	CVPR'21	all	78.0	93.8	89.6
TSM [26]	ICCV'19	16+16	48.6	88.4	83.1
TRNms [55]	ECCV'18	8+8	36.6	87.8	82.0
SlowOnly [13]	ICCV'19	32	32.4	93.46	86.8
I3D [4]	CVPR'17	16	28.0	75.6	77.3
TSN [46]	TPAMI'18	8	24.3	86.0	79.9
ART [45]	TNNLS'25	-	365.0	94.7	90.2
CLIP-based methods ViT-B/16					
PeVL [54]†*	CVPR'24	16	28.0	91.2	85.4
AIM [51]*	ICLR'23	32	11.0	94.9	89.1
DiA (ours)	-	32	3.0	**96.2 ± 0.13**	**91.7 ± 0.10**

Table 3: Results on **HAA500** and **ToyotaSmartHome**. † means PeVL [54] results reproduced using video input modality. "*" means results reproduced.

Method	Tunable Params (M)	HAA500 Top-1
Non-CLIP-based methods		
TPN [50]	33.0	50.5
TSN [46]	24.3	55.3
I3D [4]	28.0	42.80
SlowFast [13]	28.7	25.07
CLIP-based methods ViT-B/16		
P2S + EVL [25]	29.9	80.87
EVL [27]	28.7	76.4
PeVL [54]†*	28.0	78.88
AIM [51]*	11.0	80.2
DiA (ours)	**2.3**	**82.4 ± 0.08**

Method	Tunable Params (M)	ToyotaSH Top-1
Non-CLIP-based methods		
TimeSformer	121.4	68.4
π-ViT [39]	121.4	72.9
MotionFormer [34]	109.1	65.8
Video Swin [29]	88.1	69.8
AssembleNet++ [42]	51.1	63.6
MMNet [3]	34.2	70.1
VPN++ [7]	17.5	71.0
PAAT [40]	–	72.5
CLIP-based methods ViT-B/16		
PeVL [54]†*	28.0	69.3
AIM [51]*	11.0	84.3
DiA (ours)	**3.7**	**86.8 ± 0.12**

4.2 Consistent state-of-the-art performance across benchmarks

DiA delivers consistent state-of-the-art performance among single-stream methods using only the video modality (RGB) across fine-grained actions in sports, atomic human actions, daily activities, cooking actions, and general action-recognition benchmarks.

On Diving48, DiA achieves 92.8% Top-1 accuracy (Table 1) with only 3.6M tunable parameters, improving over AIM from 88.9% while reducing the tunable parameters to 1/4th. The trend remains consistent on FineGym99/288 , where DiA obtains 96.2% and 91.7% Top-1 accuracy (Table 2), compared with AIM at 94.9% and 89.1% under the same 32-frame setting. These substantial gains reiterate that Diving48 and FineGym require resolving phase-sensitive temporal differences, supporting DiA's direction-aware temporal modeling over symmetric aggregation.

DiA also improves over AIM on HAA500 and ToyotaSmartHome (Table 3), as well as MPII Cooking 2 and HMDB51 (Table 4), reaching 82.4%, 86.8%, 78.7%, and 85.6% Top-1 accuracy, respectively. This cross-domain consistency indicates that directional temporal adaptation is not limited to sports actions, but also benefits daily activities, procedural hand-object interactions, and broader action-recognition settings. Additional results on SSv2 [15] in Appendix section B.

Table 4: Results on **HMDB51** and **MPII Cooking 2** . "*" means results reproduced.

HMDB51

Method	Tunable Params (M)	Top-1
Non-CLIP-based methods		
SlowOnly-R101 [10]	62.8	79.0
I3D [4]	28.0	74.8
CLIP-based methods ViT-B/16		
Full-Tuning [8]	86.0	59.4
AIM [51]	11.0	82.4
DUALPATH [33]	10.0	75.6
ST-Adapter [32]	7.2	77.7
VPT [18]	7.0	62.4
VideoPrompt [20]	6.0	66.4
DiA (ours)	**3.6**	**85.6 ± 0.10**

MPII Cooking 2

Method	Tunable Params (M)	Top-1
CLIP-based methods ViT-B/16		
AIM [51]*	11.0	76.1
DualPath [33]*	10.0	74.8
ST [32]*	7.2	75.4
DiA (ours)	**3.7**	**78.7 ± 0.15**

Using only the RGB video stream: no pose, language, or auxiliary cues, DiA matches or surpasses methods using extra modalities. With matched backbones, it outperforms multimodal PeVL (video + pose + text): at ViT-L/14 it achieves 93.8 vs. 92.5 on Diving48 and 92.9 vs. 91.8 on FineGym288, using 12.7M vs. 109M tunable parameters. Under the same RGB-only setting, video-only PeVL scores only 84.5 on Diving48 versus DiA's 92.8, indicating much of PeVL's gain comes from auxiliary streams rather than stronger temporal modeling. DiA therefore remains a parameter- and compute-efficient temporal adapter that achieves state-of-the-art performance using a single input stream.

Few-shot results on UCF101 [44] and HMDB51 [21] are provided in Appendix D.

4.3 Achieving the accuracy–compute efficiency–unimodality objective

The accuracy–compute efficiency–unimodality objective asks whether temporal specialization can improve recognition while remaining compact, fast, and single-stream. We assess this trade-off under the same full-supervised Diving48 protocol on a single NVIDIA A100-SXM4-40GB GPU with FP16 inference. FLOPs are computed with `fvcore`, and latency is measured with CUDA events using batches of 8 clips, each with shape $[3, 32, 224, 224]$.

Table 5 shows that DiA improves accuracy and efficiency simultaneously. Accuracy increases from 88.9% to 92.8%, while per-clip computation decreases from $812.06\,\mathrm{GFLOP}$ to $558.37\,\mathrm{GFLOP}$, giving a 31.24% FLOP reduction. The same reduction appears at the batch level, showing that the gain is not an artifact of per-sample reporting. The runtime measurements confirm the practical impact. DiA increases throughput from $28.15\,\mathrm{clips/s}$ to $61.20\,\mathrm{clips/s}$ and reduces p50 batch latency, i.e., the median wall-clock time for one forward pass over a batch of 8 clips, from $284.14\,\mathrm{ms}$ to $130.73\,\mathrm{ms}$. It also improves realized compute from $22.86\,\mathrm{TFLOP/s}$ to $34.17\,\mathrm{TFLOP/s}$, indicating better use of the same GPU. These results show that DiA achieves the intended objective: stronger recognition with lower FLOPs, lower latency, higher throughput, and higher realized compute, while retaining compact adapter-based training and a single-stream input setting.

Additionally, DiA consistently outperforms AIM across adaptation depths and frame settings while using substantially fewer tunable parameters, with performance saturating around 10 blocks and improving with increased temporal context (refer section C in Appendix).

Table 5: **DiA compute-efficiency analysis on Diving48**. ↓ indicates lower is better, ↑ indicates higher is better.

Method	Top-1 Acc. ↑	GFLOPs per clip ↓	GFLOPs per batch (B=8) ↓	Throughput (clips/s) ↑	Effective compute (TFLOP/s) ↑	p50 latency (ms) ↓
AIM	88.9	812.06	6496.49	28.15	22.86	284.14
DiA (ours)	**92.8**	**558.37**	**4466.99**	**61.20**	**34.17**	**130.73**
Gain (DiA vs. AIM)	↑ 3.9	↓ 31.24%	↓ 31.24%	↑ 2.17×	↑ 1.49×	↓ 53.98%

Table 6: Component ablation study on Diving48.

Component	Top-1
With Causal	91.3
With Anti-causal	91.5
Causal + Anti-causal (DiA)	92.8

Table 7: Qualitative stats on Diving48 subset ($N = 120$).

Metric	Value
Causal activity ratio	1.0000
Anti-causal activity ratio	1.0000
Mean $\mathrm{corr}(E_c, E_a)$	-0.1884
Mean $\mathrm{corr}(E_c, \mathrm{reverse}(E_a))$	-0.1705
Dominance mean $\mathbb{E}[\bar{D}]$	0.0278
Dominance std $\mathrm{Std}(\bar{D})$	0.0193
Balanced-clip ratio ($\lvert\bar{D}\rvert \leq 0.05$)	0.8833

4.4 How asymmetric directionality benefits DiA?

We evaluate DiA through component ablation and directional behavior. The ablation tests whether causal and anti-causal directions provide complementary temporal evidence, while the qualitative analysis verifies that both directions remain active and non-redundant after training. As shown in Table 6, the full causal+anti-causal DiA achieves 92.8% Top-1 accuracy, outperforming causal-only (91.3%) and anti-causal-only (91.5%) variants, indicating that the two directions capture complementary temporal evidence.

4.4.1 Qualitative evidence for causal and anti-causal modeling

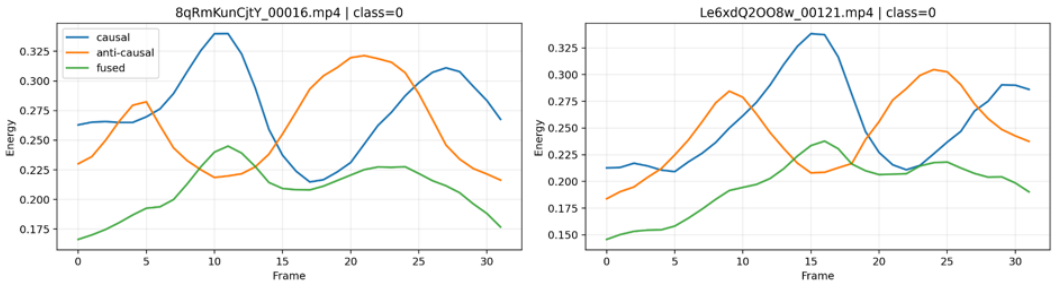

Figure 2: Representative direction-energy trajectories on Diving48 test clips. Causal and anti-causal responses peak at different frames, indicating that the two directions capture different temporal evidence, while the fused response combines them over the clip.

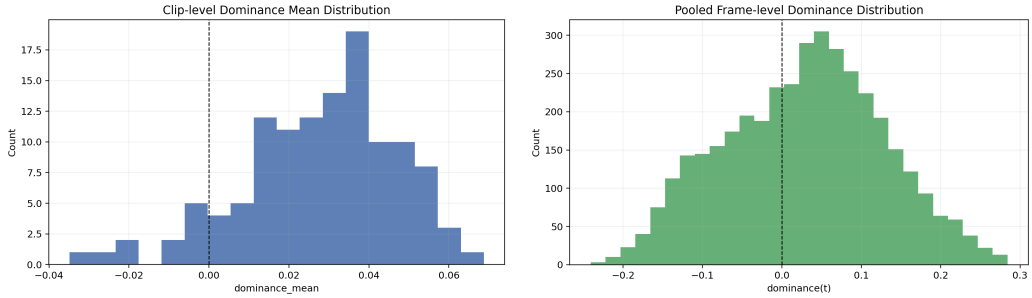

Figure 3: Directional dominance statistics on Diving48. Left: the clip-level mean dominance $\bar{D}$ is concentrated near zero, showing balanced causal and anti-causal use over complete clips. Right: the frame-level dominance $D(t)$ is more spread out, showing that one direction can dominate locally at particular action phases.

We further analyze whether the gain of DiA comes from complementary directional evidence rather than additional capacity alone. Using a held-out Diving48 test subset with $N = 120$ clips, we record causal, anti-causal, and fused activations across adapter layers, spatial tokens, channels, and frames. For each direction, we compute the frame-wise energy

$$E_b(t) = \mathbb{E}_{\ell,s,ch}\left[|z_{b,\ell,s,ch}(t)|\right], \tag{16}$$

where $b \in \{c, a, f\}$ denotes causal, anti-causal, and fused responses, and $z_{b,\ell,s,ch}(t)$ is the activation at frame t, layer ℓ, spatial token s, and channel ch. This energy measures response magnitude over

time; it is not a confidence score. We measure directional dominance as

$$D(t) = \frac{E_c(t) - E_a(t)}{E_c(t) + E_a(t) + \epsilon}. \tag{17}$$

Here, $D(t) > 0$ indicates stronger causal response, $D(t) < 0$ indicates stronger anti-causal response, and $D(t) \approx 0$ indicates balanced use of both directions.

Table 7 reports three diagnostics. Activity ratios test whether each direction remains active. Correlations between E_c and E_a test whether the two directions produce redundant temporal profiles. Dominance statistics test whether DiA collapses toward one direction. Both activity ratios are 1.0000, showing no directional collapse. The negative correlations, including $\mathrm{corr}(E_c, \mathrm{reverse}(E_a))$, indicate that the anti-causal response is not a delayed or reversed copy of the causal response. The dominance statistics show near-balanced use, with $\mathbb{E}[\bar{D}] = 0.0278$ and 88.33% of clips satisfying $|\bar{D}| \leq 0.05$. These results support complementary directional modeling rather than a capacity-only explanation.

Fig. 2 visualizes the temporal responses for representative clips. The causal and anti-causal energies often peak at different frames, showing that the two directions respond to different parts of the action rather than producing the same temporal profile. The fused response combines these directional responses over time. Fig. 3 summarizes this pattern over the test subset. The clip-level dominance is concentrated near zero (shown in blue on left , Fig. 3), showing that DiA uses both directions over a full clip. In contrast, the frame-level dominance is more spread out (shown in green on right , Fig. 3), showing that causal or anti-causal evidence can be stronger at specific frames.

We analyze the learned fusion weights $\lambda_{\rightarrow}$ and $\lambda_{\leftarrow}$, Eq. (13), obtained by softmax over trainable direction logits. Across 10 layers, their means are 0.5030 and 0.4970, with a small mean gap of 0.0342, indicating no collapse to either direction. Together with ablation and dominance results, this shows that DiA uses both past- and future-conditioned evidence for fine-grained action recognition.

Further additional examples and complementary-branch analysis (per-layer, per-clip, per-class) are given in Appendix F.

5 Conclusion and limitations

The central finding is that fine-grained action recognition benefits from directional temporal specialization under a tight efficiency budget. The gains over causal-only and anti-causal-only variants show that past- and future-conditioned cues are complementary, while the cross-dataset improvements show that this benefit is not obtained by full model adaptation or auxiliary modalities. DiA therefore demonstrates that a frozen visual backbone can be effectively specialized for phase- and order-sensitive actions with only a compact set of tunable parameters and lower inference cost. Two directions remain open for the future work which are currently not addressed. First, studying longer-range fine-grained action structure will require curating long-duration fine-grained action benchmarks, as existing datasets contain relatively short clips and may not capture extended temporal dependencies. Second, DiA uses both causal and anti-causal directions within the observed clip, which is well suited to offline recognition but leaves online recognition as a natural extension, where future frames are not available at prediction time.

6 Acknowledgment

This publication has emanated from research supported by the Kempe Foundations (Kempestiftelserna), JCSMK23-0109, Sweden.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6836–6846, 2021.

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 2, page 4, 2021.

[3] XB Bruce, Yan Liu, Xiang Zhang, Sheng-hua Zhong, and Keith CC Chan. MMNet: A model-based multimodal network for human action recognition in rgb-d videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3522–3538, 2022.

[4] 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 (CVPR)*, pages 21–26, 2017.

[5] Jihoon Chung, Cheng-hsin Wuu, Hsuan-ru Yang, Yu-Wing Tai, and Chi-Keung Tang. HAA500: Human-centric atomic action dataset with curated videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13465–13474, 2021.

[6] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome: Real-world activities of daily living. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 833–842, 2019.

[7] Srijan Das, Rui Dai, Di Yang, and Francois Bremond. VPN++: Rethinking video-pose embeddings for understanding activities of daily living. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9703–9717, 2021.

[8] 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.

[9] H. Duan et al. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[10] Haodong Duan, Yue Zhao, Yuanjun Xiong, Wentao Liu, and Dahua Lin. Omni-sourced webly-supervised learning for video recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 670–688, 2020.

[11] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6824–6835, 2021.

[12] Quanfu Fan, Rameswar Panda, et al. Can an image classifier suffice for action recognition? *arXiv preprint arXiv:2106.14104*, 2021.

[13] 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 (ICCV)*, pages 6202–6211, 2019.

[14] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens Van Der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16102–16112, 2022.

[15] 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 (ICCV)*, pages 5842–5850, 2017.

[16] Yanbin Hao, Hao Zhang, Chong-Wah Ngo, and Xiangnan He. Group contextualization for video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 928–938, 2022.

[17] Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Object-region video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3148–3159, 2022.

[18] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 709–727, 2022.

[19] Yang JingYi, YU Zitong, He Jia, Hui Li, et al. Kronecker mask and interpretive prompts are language-action video learners. In *International Conference on Learning Representations (ICLR)*, 2025.

[20] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 105–124, 2022.

[21] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2556–2563, 2011.

[22] Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. UniFormer: Unified transformer for efficient spatiotemporal representation learning. *arXiv preprint arXiv:2201.04676*, 2022.

[23] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. MViTv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4804–4814, 2022.

[24] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528, 2018.

[25] Yong-Lu Li, Xiaoqian Wu, Xinpeng Liu, Zehao Wang, Yiming Dou, Yikun Ji, Junyi Zhang, Yixing Li, Xudong Lu, Jingru Tan, et al. From isolated islands to pangea: Unifying semantic space for human action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16582–16592, 2024.

[26] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7083–7093, 2019.

[27] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard De Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 388–404, 2022.

[28] Enqi Liu, Liyuan Pan, Yan Yang, Yiran Zhong, Zhijing Wu, Xinxiao Wu, and Liu Liu. Storyboard-guided alignment for fine-grained video action recognition. *Advances in Neural Information Processing Systems*, 38:57070–57096, 2026.

[29] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3202–3211, 2022.

[30] Fei Long, Xiaoou Li, Jiaming Lv, Haoyuan Yang, Xianjun Cheng, and Peihua Li. BDC-CLIP: Brownian distance covariance for adapting CLIP to action recognition. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

[31] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–18, 2022.

[32] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. ST-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35:26462–26477, 2022.

[33] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Dual-path adaptation from image to video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2203–2213, 2023.

[34] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. *Advances in Neural Information Processing Systems*, 34:12493–12506, 2021.

[35] Huifang Qian, Jialun Zhang, Jianping Yi, Zhenyu Shi, and Yimin Zhang. Ctm: Cross-time temporal module for fine-grained action recognition. *Computer Vision and Image Understanding*, 244:104013, 2024.

[36] Zhiwu Qing, Shiwei Zhang, Ziyuan Huang, Yingya Zhang, Changxin Gao, Deli Zhao, and Nong Sang. Disentangling spatial and temporal learning for efficient image-to-video transfer learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13934–13944, 2023.

[37] 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 *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

[38] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6545–6554, 2023.

[39] Dominick Reilly and Srijan Das. Just add π! pose induced video transformers for understanding activities of daily living. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18340–18350, 2024.

[40] Dominick Reilly, Aman Chadha, and Srijan Das. Seeing the pose in the pixels: Learning pose-aware representations in vision transformers. *arXiv preprint arXiv:2306.09331*, 2023.

[41] Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision*, pages 1–28, 2015. doi: 10.1007/s11263-015-0851-8. URL `http://dx.doi.org/10.1007/s11263-015-0851-8`.

[42] Michael S Ryoo, AJ Piergiovanni, Juhana Kangaspunta, and Anelia Angelova. Assemblenet++: Assembling modality representations via attention connections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 654–671, 2020.

[43] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2616–2625, 2020.

[44] 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.

[45] Baoli Sun, Yihan Wang, Xinzhu Ma, Zhihui Wang, Kun Lu, and Zhiyong Wang. Toward an effective action-region tracking framework for fine-grained video action recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[46] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11):2740–2755, 2018.

[47] Mengmeng Wang, Jiazheng Xing, and Yong Liu. ActionCLIP: a new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021.

[48] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. BEVT: Bert pretraining of video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14733–14743, 2022.

[49] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3333–3343, 2022.

[50] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 591–600, 2020.

[51] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. AIM: Adapting image models for efficient video action recognition. In *International Conference on Learning Representations (ICLR)*, 2023.

[52] Bowen Zhang, Jiahui Yu, Christopher Fifty, Wei Han, Andrew M Dai, Ruoming Pang, and Fei Sha. Co-training transformer with videos and images improves action recognition. *arXiv preprint arXiv:2112.07175*, 2021.

[53] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Temporal query networks for fine-grained video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4486–4496, 2021.

[54] Haosong Zhang, Mei Chee Leong, Liyuan Li, and Weisi Lin. Pevl: Pose-enhanced vision-language model for fine-grained human action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18857–18867, 2024.

[55] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018.

[56] Kevin Zhu, Alexander Wong, and John McPhee. Fencenet: Fine-grained footwork recognition in fencing. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3588–3597. IEEE, 2022.

Appendix

A Analysis of Video-Only and Multimodal Recognition

To isolate the contribution of the video modality, we compare DiA with PeVL [54] under video-only and multimodal settings. PeVL combines video, pose, and text, whereas DiA uses only the video stream. Since the official PeVL code is not publicly available, we re-implemented its video-only setting. On Diving48, our reproduced result matches the video-only accuracy reported by PeVL (84.5%); for FineGym99/288, where video-only results are not reported, our implementation obtains 91.2% and 85.4%, respectively.

Table 8: **Video-only and multimodal comparison with PeVL.** Top-1 accuracy (%) is reported on Diving48 and FineGym99/288. "Reported" denotes results from the original PeVL paper [54]; the remaining PeVL results are reproduced using our implementation. DiA uses only the video modality.

Method	Tunable Params (M)	Diving48	FineGym99	FineGym288
ViT-B/16 backbone				
PeVL (video + pose + text), reported	42.0	91.9	**96.5**	90.5
PeVL (video)	28.0	84.5	91.2	85.4
DiA (video)	**3.7**	**92.8**	96.2	**91.7**
ViT-L/14 backbone				
PeVL (video + pose + text), reported	109.0	92.5	**97.0**	91.8
DiA (video)	**12.7**	**93.8**	–	**92.9**

Table 8 shows a clear modality–parameter trade-off. With ViT-B/16, removing pose and text from PeVL reduces accuracy from 91.9% to 84.5% on Diving48, indicating a substantial contribution from its additional modalities. In contrast, DiA reaches 92.8% using only video and 3.7M tunable parameters. The trend remains at the larger ViT-L/14 backbone: DiA achieves 93.8% versus 92.5% on Diving48 and 92.9% versus 91.8% on FineGym288, while using 12.7M rather than 109M tunable parameters. Thus, DiA remains competitive with or exceeds the multimodal PeVL results under matched backbones, while operating with a single video input stream and substantially fewer tunable parameters.

B Additional results on Something-Something v2

On SSv2 (Table 9), methods with the highest accuracy rely on large-capacity models, e.g., MViTv2-L (213M, 73.3%), ViViT-L (311M, 65.4%), and MTV-B (310M, 67.6%), indicating that strong performance is often achieved through substantial parameter scaling. In contrast, parameter-efficient methods operate in a much smaller regime: AIM (14M, 68.1%) and ST (7.2M, 69.3%). DiA uses only 3.0M tunable parameters and achieves 69.2% top-1 accuracy, remaining competitive within this low-parameter regime.

This highlights a clear trade-off: while larger models achieve higher absolute accuracy, they require one to two orders of magnitude more tunable parameters. DiA prioritizes parameter and compute efficiency, achieving comparable performance to existing efficient methods with significantly fewer parameters. The gap to large-scale models reflects the increased temporal complexity of SSv2, where higher capacity can be beneficial, but DiA maintains a favorable accuracy–efficiency balance under strict parameter constraints.

C Ablation on number of blocks and frames

We ablate the number of adapted blocks and input frames to evaluate whether DiA remains effective under different capacity and temporal-context budgets.

Table 9: Comparison on **Something-Something V2 (SSv2)**. We report Top-1 accuracy on a challenging temporal reasoning benchmark. DiA maintains competitive performance while using a compact set of tunable parameters.

Method	Tunable Params (M)	Top-1
Non-CLIP-based methods		
ViViT-L/16×2 [1]	311	65.4
MTV-B [49]	310	67.6
MViTv2-L (312↑) [23]	213	**73.3**
TimeSformer-L [2]	121	62.4
VideoSwin-B [29]	89	69.6
MViTv2-B [23]	51	70.5
UniFormer-B [22]	50	71.2
MViT-B [11]	37	67.1
CTM [35]	24.4	65.1
Omnivore [14]	-	71.4
CoVeR [52]	-	70.9
CLIP-based methods		
EVL [27]	86	62.4
AIM [51]	14	68.1
ST [32]	7.2	69.3
DiA (ours)	**3.0**	69.2

Table 10 shows that DiA achieves strong performance even with limited adaptation depth. With only 6 adapted blocks, DiA already outperforms AIM while using substantially fewer tunable parameters. Performance improves further when increasing to 10 blocks, but saturates at 12 blocks. This indicates that DiA does not require dense adaptation across the full backbone; most of the benefit comes from a compact number of directionally specialized temporal updates. We therefore use 10 blocks as the best accuracy–parameter trade-off.

Table 10: Impact of adaptation depth on Diving48. DiA achieves higher accuracy than AIM with fewer tunable parameters across different numbers of adapted CLIP blocks.

Method	# Blocks	Tunable Params (M)	Top-1
AIM	6	5.3	87.3
DiA	6	**1.8**	**90.6**
AIM	10	8.9	87.7
DiA	10	**3.0**	**92.7**
AIM	12	11.0	88.9
DiA	12	**3.7**	**92.8**

Table 11: Impact of temporal context on Diving48. DiA consistently outperforms AIM across frame budgets, showing that directional temporal adaptation uses available temporal evidence more effectively.

Method	# Frames	Top-1
AIM	8	81.8
DiA	8	**83.0**
AIM	16	84.7
DiA	16	**90.5**
AIM	32	88.9
DiA	32	**92.7**

Table 11 shows that DiA benefits from additional temporal context. The largest improvement appears when increasing from 8 to 16 frames, suggesting that directional temporal modeling needs sufficient temporal evidence to capture action phases. The gain from 16 to 32 frames is smaller but remains positive, indicating that longer clips still provide useful phase-completion cues. DiA also outperforms AIM at every frame setting, showing that the causal/anti-causal design uses available frames more effectively rather than simply relying on more input.

Overall, these ablations show that DiA is robust across both adaptation depth and temporal input length, while preserving the paper's accuracy–efficiency objective.

D Few-shot comparison on UCF101 and HMDB51

To assess whether DiA remains effective when labelled data is scarce, we evaluate it in the K-shot setting on UCF101 [44] and HMDB51 [21], with $K \in \{2, 4\}$ training videos per class. We compare with Storyboard [28] under the same setting.

Table 12: Few-shot Top-1 accuracy on UCF101 and HMDB51 with K training videos per class.

Method	UCF101		HMDB51	
	$K = 2$	$K = 4$	$K = 2$	$K = 4$
Storyboard [28]	92.1	93.0	64.0	66.1
DiA (ours)	**93.7**	**96.9**	**77.1**	**85.5**

As shown in Table 12, DiA outperforms Storyboard at both shot counts on both datasets. The gains on UCF101 are 1.6 and 3.9 points. The gains on HMDB51 are larger, at 13.1 and 19.4 points. These results indicate that the compact directional adapter does not depend on large amounts of labelled data and transfers well to the low-data regime.

E Implementation details

We use CLIP ViT-B/16 as the visual backbone and train with a batch size of 8. For Diving48, FineGym99, HMDB51, HAA500, ToyotaSH and MPII Cooking 2, we sample 32 input frames with segment-based sampling, resize the short side to 256, and apply 224×224 crops. For SSv2, we use 16 input frames with the same sampling and crop pipeline. Optimization follows AdamW with a default learning rate of 0.0025 and cosine learning-rate decay, warmup for 2–3 epochs, and weight decay in the range 0.03–0.05. We train for 50 epochs on Diving48, MPII Cooking 2, FineGym99, HAA500, ToyotaSH and HMDB51, and 45 epochs on SSv2. We follow standard video data processing, including random cropping, resizing, and horizontal flipping for augmentation, along with multi-threaded video decoding.

F Qualitative evidence for causal and anti-causal modeling: additional examples and complementary branches analysis

F.1 Complementary branches analysis

We extend the analysis of Sec. 4.4.1 on the same Diving48 test subset ($N = 120$ clips, 20 classes), using the energy and dominance measures of Eqs. (16)–(17). The component ablation (Table 6) already shows that each direction is necessary: causal-only and anti-causal-only variants reach 91.3% and 91.5%, respectively, compared with 92.8% for the fused model. Below, we report per-layer, per-clip, and per-class statistics of the trained model.

Table 13 shows the learned fusion weights (Eq. (12)) for each adapted block. The split remains close to $50/50$ at every depth (mean $0.503/0.497$), with a maximum imbalance of 0.139 at block 3. Although the softmax could concentrate the weight on one direction, both directions are retained throughout the network.

Table 13: Learned directional fusion weights per adapted block on Diving48. The larger weight in each block is shown in bold.

Block	2	3	4	5	6	7	8	9	10	11
Causal $\lambda_{\rightarrow}$	**0.502**	**0.570**	0.486	0.482	0.481	0.480	**0.516**	**0.511**	**0.502**	0.500
Anti-causal $\lambda_{\leftarrow}$	0.498	0.430	**0.514**	**0.518**	**0.519**	**0.520**	0.484	0.489	0.498	0.500
$\lvert\lambda_{\rightarrow} - \lambda_{\leftarrow}\rvert$	0.004	0.139	0.028	0.036	0.037	0.040	0.031	0.022	0.005	0.000

At the clip level, both branches are active on all clips, yet their energy trajectories are anti-correlated (mean $\mathrm{corr}(E_c, E_a) = -0.19$), indicating that the two directions respond at different points of an action. Table 14 shows that this correlation is negative for every representative clip, for both correct

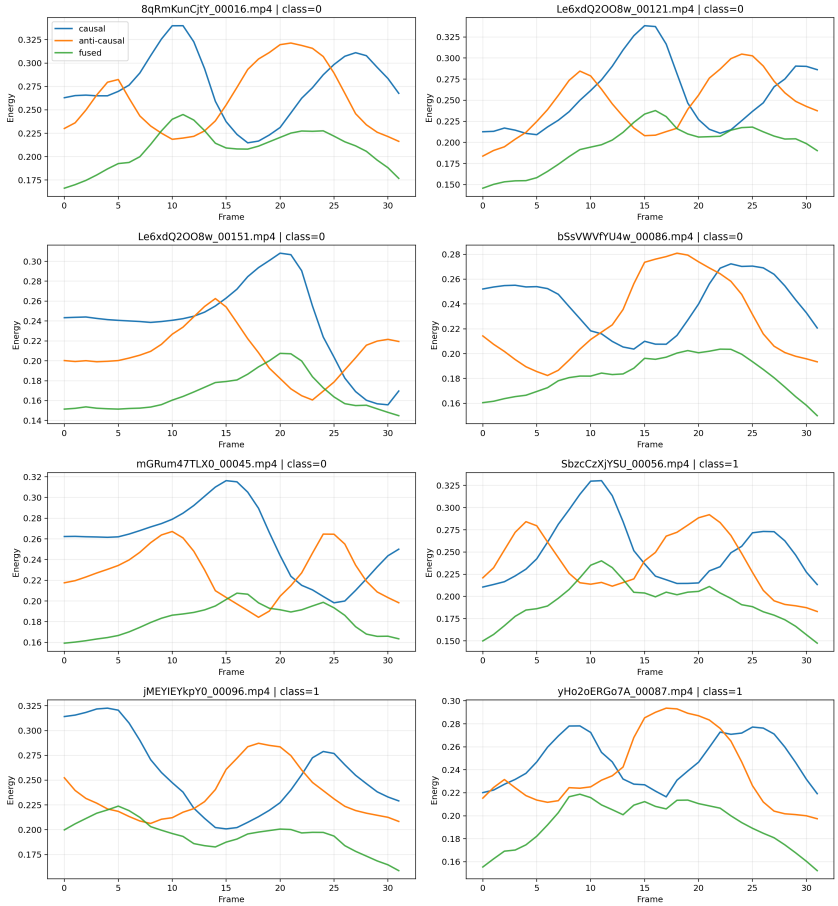

Figure 4: Representative direction-energy trajectories on Diving48 test clips. Each plot shows how causal, anti-causal, and fused responses vary over frames. Peaks at different temporal locations indicate that the two directions respond to different action phases, while the fused response combines their evidence over the clip.

and incorrect predictions. The fused energy is below both directional energies, which implies partial cancellation between the two responses. The correlation with the time-reversed anti-causal profile varies in sign across clips but is negative on average (-0.17, Table 7), so the anti-causal branch is not a reversed copy of the causal one.

Table 14: Per-clip statistics on representative Diving48 test clips. E_c, E_a, and E_f denote the clip-averaged causal, anti-causal, and fused energies; $\mathrm{rev}(\cdot)$ denotes time reversal.

Video	Correct	Conf.	E_c	E_a	E_f	$\mathrm{corr}(E_c, E_a)$	$\mathrm{corr}(E_c, \mathrm{rev}(E_a))$
8qRmKunCjtY_00016	✓	0.90	0.28	0.26	0.21	−0.62	0.71
Le6xdQ2OO8w_00121	✓	0.97	0.26	0.25	0.20	−0.21	−0.72
Le6xdQ2OO8w_00151	×	0.61	0.24	0.21	0.17	−0.13	0.49
mGRum47TLX0_00045	✓	0.77	0.26	0.23	0.18	−0.32	−0.59
SbzcCzXjYSU_00056	✓	0.86	0.25	0.24	0.19	−0.43	0.86
jMEYIEYkpY0_00096	×	0.95	0.26	0.24	0.19	−0.40	−0.59
yHo2oERGo7A_00087	✓	0.75	0.25	0.24	0.19	−0.23	−0.03
Le6xdQ2OO8w_00201	✓	0.81	0.26	0.24	0.19	−0.51	−0.55
roFeEJPgJD8_00038	✓	0.83	0.27	0.24	0.20	−0.27	−0.72

Finally, to verify that the global balance does not hide classes relying on a single branch, we compute the statistics per class (Table 15). All 20 classes are balanced ($|\bar{D}| \leq 0.05$), and 18 of 20 have a

negative mean correlation, reaching -0.65 for class 24 and -0.63 for class 8. The remaining two classes are close to zero ($+0.01$ and $+0.07$). Complementary use of both directions therefore holds within individual classes.

Table 15: Per-class statistics on the Diving48 test subset. A class is balanced if $|\bar{D}| \leq 0.05$.

Class	#Clips	Mean $\bar{D}$	Mean $\mathrm{corr}(E_c, E_a)$	Balanced
0	5	0.04	-0.32	✓
1	3	0.03	-0.35	✓
5	3	0.01	0.01	✓
8	3	0.04	-0.63	✓
21	3	0.03	-0.27	✓
26	5	0.01	-0.21	✓
34	6	0.04	-0.35	✓

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction are supported by the proposed DiA method in section 3 and the experimental results in section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations and future directions in section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not present theoretical results requiring formal assumptions or proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer:[Yes]

Justification: Paper describes complete details on architecture, training setup, evaluation protocols, public datasets, measurement, and metrics needed to reproduce the all the results which are described in section 4, along with complete source code released anonymously for implementation details.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provides an anonymized code repository for DiA, and the experiments use publicly available benchmark datasets with reproduction details described in section 4.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The training and evaluation setup, including datasets, splits, backbone, number of frames, optimizer, training schedule, and default hyperparameters, is described in section 4, with additional implementation details provided in the released code.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports the standard deviation of the proposed method's performance, with all main results obtained from experiments repeated three times using different random seeds.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The experimental setup and compute-efficiency evaluation are described in section 4, and the released code provides the remaining implementation details needed to reproduce the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All sections.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

Justification: The paper focuses on open research problem, fine-action recognition and uses public dataset which means the stated work has no remotely relation with any societal impact.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The work does not release high-risk assets such as pretrained language models, image generators, or scraped datasets, so special safeguards are not applicable.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Existing datasets and models are credited in section 2 and section 4, as they are open sourced licenses and terms of use.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces and releases the DiA code as a new asset, with an anonymized repository link provided in the abstract.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing experiments or new research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: This work does not involve new crowdsourcing, user studies, or direct interaction with human subjects. The experiments use existing public video action-recognition benchmarks, so IRB approval or equivalent human-subjects review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method does not use LLMs as an important, original, or non-standard component.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.