MARD: MULTI-LEVEL ADAPTIVE REPRESENTATION DISTILLATION ACROSS HETEROGENEOUS ARCHITECTURES FOR EGOCENTRIC ACTION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deploying powerful pretrained video models for egocentric action recognition is computationally expensive, motivating the transfer of their knowledge to compact students. However, heterogeneous teacher–student architectures organize video representations differently, making direct feature imitation unnecessarily restrictive. Beyond this architectural mismatch, our analysis further shows that higher teacher–student representation similarity does not necessarily lead to better recognition accuracy. Motivated by these observations, we propose Multi-level Adaptive Representation Distillation (MARD), which transfers architecture-agnostic structures from fixed features extracted by a general-purpose teacher while preserving the student's native representation. A lightweight adapter makes these fixed teacher features responsive to the target task. MARD then performs distillation at three complementary levels: Global Similarity Distillation matches the distributions of cross-video relations, Task Conditioned Distillation preserves how relational strength is allocated across verb–noun semantic groups, and Temporal Rhythm Distillation aligns temporal changes across sequences of different lengths. Since these objectives are constructed independently within the native teacher and student spaces, MARD requires neither matched feature dimensions nor aligned temporal tokens. On EPIC-KITCHENS-100, MARD improves action top-1 accuracy over the non-distilled VideoSwin-S and X3D-S baselines by 9.19 and 6.03 percentage points, respectively; on Something-Something V2, the corresponding top-1 gains are 9.54 and 5.30 percentage points. Across both datasets, MARD outperforms representative logit- and feature-based distillation methods with both students and requires no task-specific fine-tuning of the general-purpose teacher.

1 INTRODUCTION

Egocentric action recognition identifies fine-grained human activities from first-person videos and supports applications such as wearable assistants, embodied agents, and human–robot interaction (Damen et al., 2022; Radevski et al., 2023). Practical deployment demands both accurate recognition of compositional verb–noun actions and efficient video processing. Large pretrained video models provide rich object, interaction, and temporal knowledge. Their computational cost, however, makes direct deployment difficult. Knowledge distillation offers a practical alternative by transferring this knowledge to compact students (Hinton et al., 2015).

Two obstacles limit conventional distillation in this setting. First, adapting a large general-purpose teacher to each downstream task introduces substantial training cost. This burden grows with the scale of video foundation models (Qiu et al., 2024; Li et al., 2026b). Second, heterogeneous teacher and student architectures produce representations that differ in dimensionality, tokenization, receptive field, and inductive bias (Lin et al., 2025; Li et al., 2025). Feature-based methods often bridge these spaces with learned projections and pointwise alignment objectives (Romero et al., 2015; Zagoruyko & Komodakis, 2017). Across heterogeneous architectures, however, such alignment implicitly treats the teacher's coordinate system as the target. Our empirical analysis further questions this premise. Figure 1 compares centered kernel alignment (CKA) (Kornblith et al., 2019) with EK100 action accuracy across representative distillation methods for VideoSwin-S and X3D-S. In both panels, the highest CKA does not yield the highest accuracy; the best-performing model

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.

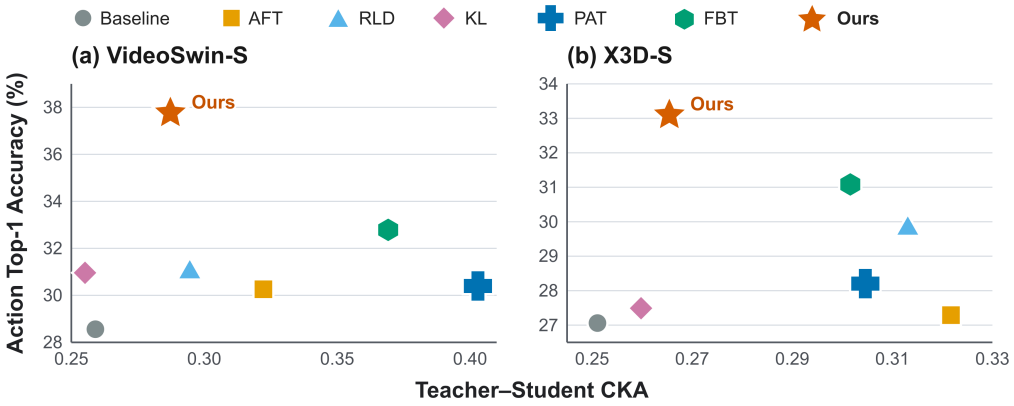

Figure 1: Teacher–student CKA versus EK100 action top-1 accuracy for (a) VideoSwin-S and (b) X3D-S. Across both student architectures, stronger representation alignment does not consistently correspond to higher recognition accuracy. MARD achieves the highest accuracy without maximizing CKA, motivating structure-level transfer that preserves the student's native representation.

instead lies in a moderate-alignment regime. This finding does not diminish the value of teacher knowledge. It shows that reproducing the teacher representation more closely is not, by itself, a reliable distillation objective for heterogeneous students.

To tackle these challenges, we propose a structure-level transfer strategy for heterogeneous video distillation. Rather than forcing teacher and student features into a shared coordinate system, it preserves their native representation spaces and aligns architecture-agnostic structures relevant to recognition. We organize this transfer along three complementary dimensions: global relational distribution, task-conditioned semantics, and temporal evolution. Cross-video similarities characterize the global organization of action instances. Verb–noun-conditioned relations encode task-specific semantic organization. Temporal feature changes describe how an action evolves over time. Each structure is constructed independently within the model's native representation space. This design permits teacher–student alignment without matching feature dimensions or temporal resolutions.

We instantiate this strategy as Multi-level Adaptive Representation Distillation (MARD). MARD reuses fixed features from a frozen general-purpose teacher and employs a lightweight adapter to make them responsive to the downstream task. The adapter is optimized jointly with the student, while a residual constraint keeps the adapted features anchored to the original teacher representation. MARD then transfers complementary knowledge through three objectives. Global Similarity Distillation (GSD) aligns the distributions of cross-video relations to preserve the global organization of task instances. Task Conditioned Distillation (TCD) transfers task-relevant semantic relations under verb–noun supervision. Temporal Rhythm Distillation (TRD) aligns temporal changes between sequences of different lengths. Together, these objectives preserve complementary global, task-semantic, and temporal structures without forcing direct feature correspondence.

Experiments on EK100 and SSv2 validate both the effectiveness and the cross-architecture robustness of MARD. On EK100, MARD improves action top-1 accuracy over training without distillation by 9.19 points for VideoSwin-S and 6.03 points for X3D-S, exceeding the strongest competing results by 4.95 and 2.00 points, respectively. On SSv2, the corresponding top-1 gains are 9.54 and 5.30 points, and MARD again achieves the best student result with both architectures. The general-purpose teacher requires no task-specific fine-tuning, and only the student is retained at inference.

Our main contributions are summarized as follows:

- We introduce a structure-level perspective for heterogeneous video distillation. Since stronger teacher–student representation similarity does not necessarily yield higher accuracy, we align architecture-agnostic relations while preserving each model's native feature space.

- We propose MARD, which uses a constrained residual adapter to make fixed general-purpose teacher features task-responsive, and GSD, TCD, and TRD to transfer complementary global, task-semantic, and temporal structures across heterogeneous representation spaces.
- Experiments on EK100 and SSv2 with VideoSwin-S and X3D-S show consistent gains over representative distillation methods, without task-specific teacher fine-tuning or additional inference cost.

2 Related Work

2.1 Knowledge Distillation and Heterogeneous Architectures

Knowledge distillation was originally introduced to train a compact student with the softened output distribution of a larger teacher (Hinton et al., 2015). Subsequent work transfers richer intermediate knowledge: FitNets uses hidden representations as training hints (Romero et al., 2015), attention transfer matches spatial attention maps (Zagoruyko & Komodakis, 2017), and similarity-preserving distillation transfers pairwise relations between samples rather than individual feature coordinates (Tung & Mori, 2019). These objectives provide supervision beyond class labels, but feature-level transfer commonly relies on manually selected layers, learned projections, or an implicit correspondence between teacher and student representations. Such assumptions become fragile when the two models differ in inductive bias, feature geometry, or temporal resolution.

Cross-architecture distillation explicitly addresses this mismatch. OFA-KD projects intermediate representations into a shared logit space and suppresses architecture-specific information (Hao et al., 2023), whereas FBT constructs a fused intermediate model from teacher and student modules and replaces pixel-wise matching with a spatial-agnostic objective (Li et al., 2025). A complementary line studies the capacity gap: AID attributes ineffective transfer partly to mismatched intra-class output distributions and fine-tunes the teacher to better match the student (Qian et al., 2025). Together, these studies show that teacher capacity and representation mismatch must be considered when designing the transfer mechanism. MARD targets a different operating point: it keeps the large teacher backbone fixed, introduces no pair-specific feature projector, and compares semantic, temporal, and task-aware structures computed within each model's native representation space.

2.2 Knowledge Transfer from Foundation Models

The increasing cost of foundation models has motivated methods that transfer their knowledge without deploying or fully fine-tuning them. Adaptive Feature Transfer (AFT) regularizes a small downstream model with task-relevant information extracted from one or more frozen foundation models (Qiu et al., 2024). Multi-teacher approaches such as AM-RADIO (Ranzinger et al., 2024) and DUNE (Sarıyıldız et al., 2025) instead consolidate complementary foundation-model representations into a single general-purpose encoder. In video self-supervised learning, SALT decouples teacher and student optimization by using a frozen target encoder for masked latent prediction (Li et al., 2026b). LeVJEPA instead uses empirical characteristic functions to test the normality of random one-dimensional projections and thereby regularize the latent distribution without asymmetric training heuristics (Kuhn et al., 2026). These studies demonstrate the value of frozen teachers and distribution-level constraints, but primarily focus on static-image transfer, universal encoder construction, or representation pretraining. Our goal is downstream egocentric action recognition with heterogeneous lightweight students. Rather than imposing a fixed target distribution on projected embeddings, MARD adapts fixed teacher features to the task and matches architecture-agnostic structures that remain comparable across model families and temporal resolutions.

2.3 Distillation for Video and Egocentric Understanding

Video distillation has been used to transfer supervision across architectures, modalities, and recognition settings. DistInit learns video representations from pretrained image teachers and unlabeled video (Girdhar et al., 2019). FROSTER uses frozen CLIP as a teacher and residual feature distillation to balance video-specific adaptation with open-vocabulary generalization (Huang et al., 2024). Recent work also improves action recognition by dynamically weighting channel-wise feature trans-

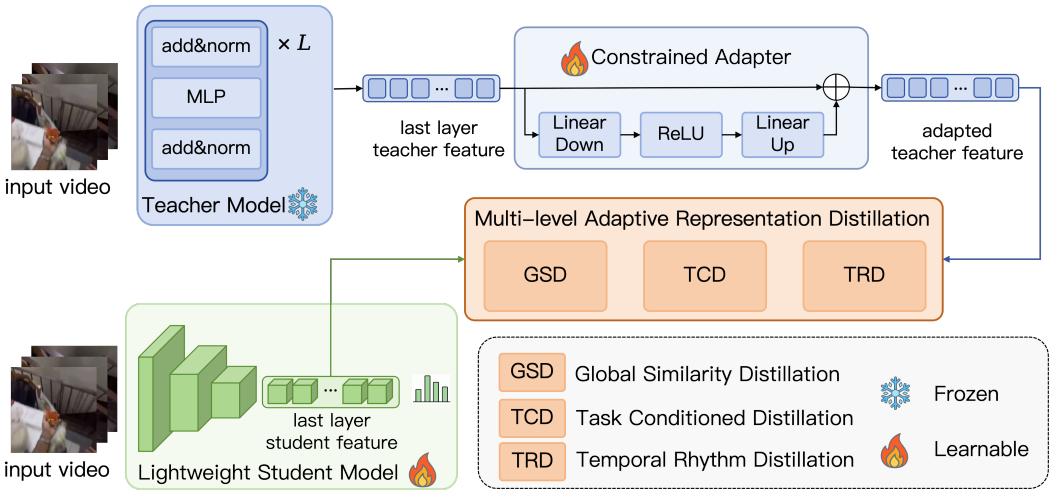

Figure 2: Overall framework of MARD, which transfers knowledge from a frozen general-purpose teacher to a lightweight student through a constrained adapter and three complementary distillation branches.

fer and adapting training samples in the frequency domain (Li et al., 2026a). In egocentric video, multimodal distillation compresses cues from optical flow, audio, and object detections into an RGB student (Radevski et al., 2023), while KARMMA trains a compact multimodal student to remain robust when input modalities are missing (Santos-Villafranca et al., 2025). These methods primarily target representation pretraining, open-vocabulary generalization, feature alignment, or modality compression. MARD instead reuses a frozen video teacher across heterogeneous students by transferring global similarity, task-conditioned, and temporal-rhythm structures without frame-wise or feature-wise correspondence.

3 Method

3.1 Overview

MARD transfers knowledge from a large pretrained video teacher to a lightweight student without assuming compatible architectures or internal representations. As illustrated in Fig. 2, we use a transformer-based teacher and a CNN-based student as the running example throughout this section to make their architectural heterogeneity explicit. The formulation is not limited to this architecture pair. The frozen teacher provides token features, whereas the student produces a spatiotemporal feature map and is optimized from video input. For a mini-batch of B videos, we denote these representations by $\mathbf{F}^t \in \mathbb{R}^{B\times T_t\times D_t}$ and $\mathbf{F}^s \in \mathbb{R}^{B\times T_s\times H_s\times W_s\times C_s}$, respectively, where T_t and T_s are the temporal lengths, H_s and W_s are the student spatial dimensions, and D_t and C_s are the channel dimensions.

A constrained residual adapter makes the fixed teacher knowledge responsive to the downstream task while remaining in the teacher space. It produces

$$\widetilde{\mathbf{F}}^t = \mathbf{F}^t + \mathcal{A}(\mathbf{F}^t). \tag{1}$$

where $\mathcal{A}(\cdot)$ denotes the learnable adapter detailed in Sec. 3.2. The adapted teacher representation and the native student representation are passed to three parallel distillation branches: GSD, TCD, and TRD. Their losses jointly optimize the adapter and student, while the teacher remains frozen. At inference, only the student and its classification heads are retained.

3.2 Constrained Knowledge Adapter

Fixed teacher features provide stable supervision, but the same immutable representation may not expose equally useful knowledge to students with different inductive biases. We therefore introduce

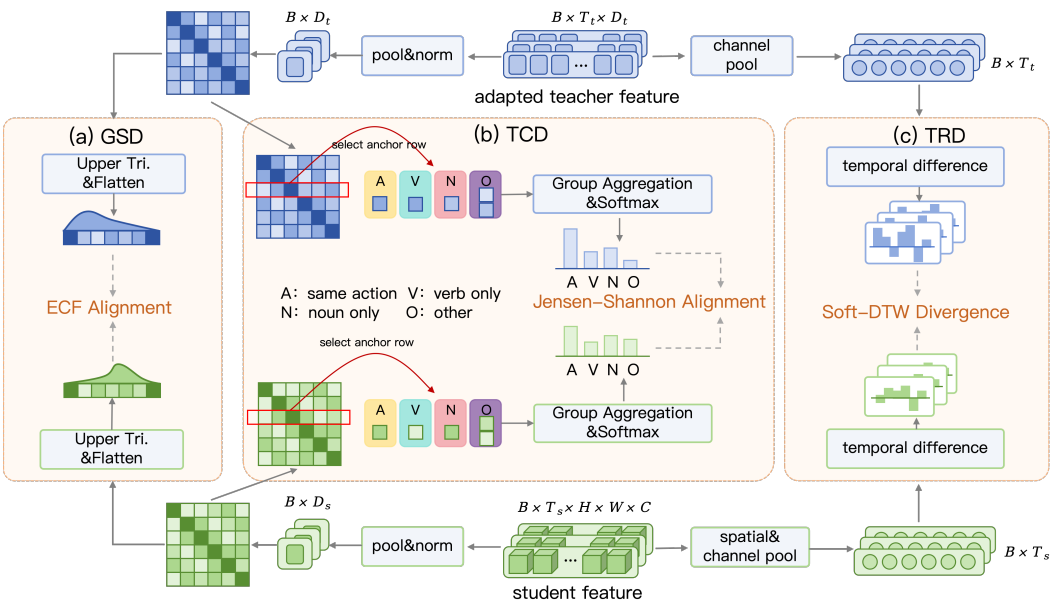

Figure 3: Three complementary branches of MARD: (a) global similarity alignment, (b) task-conditioned relation alignment, and (c) temporal rhythm alignment.

a residual bottleneck adapter to make the teacher features responsive to the downstream task. As shown in Fig. 2, the adapter is

$$\mathcal{A}(\mathbf{F}^t) = \mathrm{ReLU}\big(\mathbf{F}^t \mathbf{W}_{\mathrm{down}}\big)\, \mathbf{W}_{\mathrm{up}}, \tag{2}$$

where $\mathbf{W}_{\mathrm{down}} \in \mathbb{R}^{D_t \times D_a}$, $\mathbf{W}_{\mathrm{up}} \in \mathbb{R}^{D_a \times D_t}$, and $D_a \ll D_t$.

Because the adapter is optimized jointly with the student, it may drift toward an overly student-specific solution and weaken the teacher supervision. We therefore anchor its output to the pretrained teacher representation with

$$\mathcal{L}_{\mathrm{anchor}} = \frac{\|\widetilde{\mathbf{F}}^t - \mathbf{F}^t\|_F^2}{\|\mathbf{F}^t\|_F^2 + \epsilon}. \tag{3}$$

By limiting this deviation, the adapter remains task-responsive while preserving the pretrained teacher as the knowledge source. The adapter is used only during distillation.

3.3 Multi-level Structure Distillation

Direct feature matching is ill-defined when the teacher and student use different dimensions, spatial layouts, or temporal token counts. MARD instead extracts comparable structures independently from $\widetilde{\mathbf{F}}^t$ and $\mathbf{F}^s$. Figure 3 details the three peer-level branches: GSD describes the global distribution of cross-video similarities, TCD describes which task relations receive similarity, and TRD describes how the representation changes over time.

3.3.1 Global Similarity Distillation

GSD transfers the global organization of action instances without forcing any individual student pair to reproduce the corresponding teacher value. As illustrated in Fig. 3(a), GSD pools each video into a global representation, constructs cross-video similarities independently in the teacher and student spaces, and aligns the resulting relation distributions through multi-frequency ECF matching. The normalized global representations are

$$\mathbf{g}_i^t = \mathrm{Norm}\left(\frac{1}{T_t}\sum_{u=1}^{T_t} \widetilde{\mathbf{F}}_{i,u,:}^t\right), \qquad \mathbf{g}_i^s = \mathrm{Norm}\left(\frac{1}{T_s H_s W_s}\sum_{u,h,w} \mathbf{F}_{i,u,h,w,:}^s\right). \tag{4}$$

We then construct separate cosine-similarity matrices in the teacher and student spaces:

$$R^t_{ij} = (\mathbf{g}^t_i)^\top \mathbf{g}^t_j, \qquad R^s_{ij} = (\mathbf{g}^s_i)^\top \mathbf{g}^s_j. \tag{5}$$

Since each matrix is symmetric and its diagonal contains only self-similarities, we retain the $N_p = B(B-1)/2$ upper-triangular entries as the relation samples $\mathcal{S}_m = \{R^m_{ij} \mid i < j\}$ for $m \in \{t, s\}$.

To compare the two relation distributions without sorting or preserving pair identities, we evaluate their empirical characteristic functions (ECFs) at L fixed positive frequencies $\{\omega_\ell\}_{\ell=1}^L$ shared by the teacher and student:

$$\widehat{\phi}_m(\omega_\ell) = \frac{1}{N_p} \sum_{r \in \mathcal{S}_m} \left[\cos(\omega_\ell r) + \mathrm{i}\sin(\omega_\ell r)\right], \qquad \ell = 1, \ldots, L. \tag{6}$$

Each frequency maps the relation samples to averaged sine and cosine responses, providing a different view of the same similarity distribution. Matching these responses encourages the student and adapted teacher to exhibit consistent relation distributions across multiple frequency views. This use of characteristic functions is inspired by the ECF-based normality regularization in LeVJEPA (Kuhn et al., 2026), but GSD requires neither random projection nor a prescribed Gaussian target. Because cosine similarity has already reduced each pair to a bounded scalar, the two ECFs can be compared directly even when $D_t \neq D_s$.

We minimize the squared discrepancy between the two multi-frequency complex signatures:

$$\mathcal{L}_{\mathrm{GSD}} = \frac{1}{L} \sum_{\ell=1}^{L} \left| \widehat{\phi}_t(\omega_\ell) - \widehat{\phi}_s(\omega_\ell) \right|^2. \tag{7}$$

Because the ECF averages are computed over the relation samples $\mathcal{S}_m$, the objective is permutation invariant and requires neither explicit sorting nor pairwise correspondence. Its symmetric form jointly updates the student and adapter, while $\mathcal{L}_{\mathrm{anchor}}$ keeps the adapted teacher representation tied to the pretrained knowledge.

3.3.2 Task Conditioned Distillation

In compositional egocentric action recognition, an action is defined by a verb–noun pair. GSD aligns the overall distribution of cross-video similarities, but it does not specify whether high similarity is assigned to videos sharing the complete action, only the verb, or only the noun. As illustrated in Fig. 3(b), TCD addresses this ambiguity by grouping each anchor's relations into same-action (A), verb-only (V), noun-only (N), and other (O) categories, and aligning the resulting task-conditioned relation distributions.

Given an anchor i with verb label y^v_i and noun label y^n_i, each remaining sample is assigned to one of four mutually exclusive groups: A if both labels match, V if only the verb matches, N if only the noun matches, and O otherwise. We denote the corresponding index sets by $\mathcal{G}^c_i$, where $c \in \{a, v, n, o\}$.

For model $m \in \{t, s\}$, we aggregate the similarities in anchor row i for group $c \in \{a, v, n, o\}$ using a group-size-normalized log-mean-exp:

$$z^m_{i,c} = \tau_g \log \left[\frac{1}{|\mathcal{G}^c_i|} \sum_{j \in \mathcal{G}^c_i} \exp\left(\frac{R^m_{ij}}{\tau_g} \right) \right]. \tag{8}$$

The normalization prevents the typically larger O group from dominating merely because it contains more samples. Empty groups are masked, and a temperature-scaled softmax with temperature τ_c over the available group scores produces $\mathbf{q}^m_i$, the task-conditioned relation distribution for model m. Unlike a class-prediction logit vector, $\mathbf{q}^m_i$ describes how the anchor's similarity is allocated among the task-defined groups.

We align the teacher and student distributions with Jensen–Shannon divergence:

$$\mathcal{L}_{\mathrm{TCD}} = \frac{1}{B} \sum_{i=1}^{B} D_{\mathrm{JS}}(\mathbf{q}^t_i, \mathbf{q}^s_i). \tag{9}$$

The Jensen–Shannon divergence compares the two distributions symmetrically through their midpoint $\mathbf{m}_i = (\mathbf{q}_i^t + \mathbf{q}_i^s)/2$. The student and constrained adapter can therefore approach a compatible task-conditioned relation structure, while the anchor loss in Eq. equation 3 keeps the adapted teacher representation tied to the pretrained knowledge. TCD aligns the group-level allocation of similarity without constraining individual relations within each group.

3.3.3 Temporal Rhythm Distillation

The temporal evolution of an interaction carries important information that is not fully captured by a global video representation. As illustrated in Fig. 3(c), TRD transfers this temporal structure by summarizing each time step with a scalar response and computing adjacent differences to form a rhythm sequence. For the adapted teacher and native student, respectively, the temporal responses are

$$r_{i,u}^t = \frac{1}{D_t} \sum_{d=1}^{D_t} \widetilde{F}_{i,u,d}^t, \qquad r_{i,u}^s = \frac{1}{H_s W_s C_s} \sum_{h,w,c} F_{i,u,h,w,c}^s. \tag{10}$$

A single adjacent-step difference then defines the rhythm representation

$$\delta_{i,u}^m = r_{i,u+1}^m - r_{i,u}^m, \qquad u = 1, \ldots, T_m - 1, \quad m \in \{t, s\}. \tag{11}$$

The teacher and student rhythms have lengths $T_t - 1$ and $T_s - 1$, so direct pointwise matching is unavailable. We instead apply Soft-DTW with the squared local cost $C_{i,uv} = (\delta_{i,u}^t - \delta_{i,v}^s)^2$. Its monotone alignment paths allow the same action phase to unfold faster or slower in either model without requiring changes at identical absolute time indices. We therefore use its differentiable divergence form (Cuturi & Blondel, 2017; Blondel et al., 2021):

$$\begin{aligned} D_{\mathrm{sDTW}}^{\gamma}(\boldsymbol{\delta}^t, \boldsymbol{\delta}^s) = \mathrm{sDTW}_{\gamma}(\boldsymbol{\delta}^t, \boldsymbol{\delta}^s) \\ - \tfrac{1}{2}\,\mathrm{sDTW}_{\gamma}(\boldsymbol{\delta}^t, \boldsymbol{\delta}^t) - \tfrac{1}{2}\,\mathrm{sDTW}_{\gamma}(\boldsymbol{\delta}^s, \boldsymbol{\delta}^s). \end{aligned} \tag{12}$$

The TRD objective is

$$\mathcal{L}_{\mathrm{TRD}} = \frac{1}{B} \sum_{i=1}^{B} D_{\mathrm{sDTW}}^{\gamma}(\boldsymbol{\delta}_i^t, \boldsymbol{\delta}_i^s). \tag{13}$$

Consequently, TRD tolerates different token counts and modest timing shifts without resampling either feature sequence.

3.4 Training Objective

For compositional egocentric action recognition, the supervised loss is the sum of verb and noun cross-entropies:

$$\mathcal{L}_{\mathrm{cls}} = \mathrm{CE}(\mathbf{p}^v, y^v) + \mathrm{CE}(\mathbf{p}^n, y^n). \tag{14}$$

The complete objective is

$$\mathcal{L}_{\mathrm{MARD}} = \mathcal{L}_{\mathrm{cls}} + \lambda_a \mathcal{L}_{\mathrm{anchor}} + \lambda_g \mathcal{L}_{\mathrm{GSD}} + \lambda_c \mathcal{L}_{\mathrm{TCD}} + \lambda_r \mathcal{L}_{\mathrm{TRD}}. \tag{15}$$

The classification loss updates only the student, the three distillation losses update both the student and adapter, and the anchor loss constrains only the adapter. The teacher backbone and original teacher features remain fixed throughout. Consequently, the general-purpose teacher requires no task-specific fine-tuning, and MARD adds no computation to student inference.

4 Experiments

4.1 Experimental Setup

Datasets and metrics. We evaluate MARD on EPIC-KITCHENS-100 (EK100) (Damen et al., 2022), an egocentric action-recognition benchmark with 97 verb and 300 noun classes. Each action is a verb–noun pair. We report validation top-1 accuracy (%) for actions, verbs, and nouns. We also evaluate both students on Something-Something V2 (SSv2) (Goyal et al., 2017); those results are reported in Appendix A.

Table 1: VideoSwin-S results on the EK100 validation set. Best and second-best student results are bold and underlined.

Method	Type	Teacher fine-tuning	Action ↑	Verb ↑	Noun ↑
Teacher (reference)	–	Yes	46.36	63.99	59.22
VideoSwin-S	None	–	28.56	50.22	44.48
VideoSwin-S+AFT	Feature	No	30.26	52.15	45.84
VideoSwin-S+RLD	Logit	Yes	31.10	53.78	46.02
VideoSwin-S+KL	Logit	Yes	30.96	51.91	47.06
VideoSwin-S+PAT	Feature	Yes	30.40	53.02	45.38
VideoSwin-S+SKD	Logit	Yes	31.30	51.61	46.60
VideoSwin-S+FBT	Feature	Yes	<u>32.80</u>	<u>53.86</u>	<u>47.78</u>
VideoSwin-S+MARD	Structure	No	**37.75**	**59.76**	**50.34**

Teacher, students, and baselines. For both datasets, MARD uses AVION ViT-L (Zhao & Krähenbühl, 2023) pretrained on Ego4D as its frozen teacher. AFT (Qiu et al., 2024) uses the same pretrained checkpoint; the other compared methods use an AVION teacher fine-tuned on the target dataset. The fine-tuned teacher's EK100 accuracy appears in Tables 1 and 2 for reference. We study two students initialized from pretrained checkpoints: the hierarchical transformer VideoSwin-S (Liu et al., 2022) and the convolutional X3D-S (Feichtenhofer, 2020). We compare against training without distillation, conventional logit distillation (KL) (Hinton et al., 2015), and AFT, RLD (Sun et al., 2025), PAT (Lin et al., 2025), SKD (Jeong et al., 2026), and FBT (Li et al., 2025).

Implementation details. We uniformly sample 16 frames per video and use input resolutions of 224×224 for VideoSwin-S and 160×160 for X3D-S. Within each student architecture, all methods share the same student initialization, data processing, batch size, and optimization schedule. Every student is trained for 50 epochs with AdamW, cosine learning-rate decay, a weight decay of 0.05, and mixed precision. The backbone and classification-head learning rates are 5×10^{-5} and 5×10^{-4} for VideoSwin-S, and 10^{-4} and 10^{-3} for X3D-S, respectively. We select the checkpoint with the highest validation action top-1 accuracy on EK100 or classification top-1 accuracy on SSv2. Teacher backbone features are precomputed; methods requiring teacher logits also precompute them from the corresponding teacher checkpoint. Distillation objectives and their method-specific coefficients differ by method. For GSD, we use $L = 17$ fixed frequencies uniformly spaced over $(0, 3]$; teacher and student share these frequencies, and the ECF computation has no learnable parameters. The teacher fine-tuning columns in the result tables make the different teacher preparation requirements explicit.

4.2 Comparison with Representative Distillation Methods

Tables 1 and 2 report EK100 results for VideoSwin-S and X3D-S. The teacher fine-tuning column records whether AVION ViT-L was adapted on EK100 before distillation.

VideoSwin-S results. With VideoSwin-S, MARD achieves the best student results in action, verb, and noun recognition (Table 1). The gain is obtained with a frozen Ego4D-pretrained teacher, without fine-tuning it on EK100. Most competing methods use a task-adapted teacher. This contrast echoes the motivation in Fig. 1: stronger student performance need not come from closer imitation of the teacher's internal features. MARD instead transfers global relations among videos, verb–noun-aware relations, and temporal changes. These complementary cues preserve task-relevant knowledge while allowing the student to learn in its own representation space.

X3D-S results. The same pattern holds for X3D-S. MARD ranks first on all three metrics (Table 2). This setting pairs a convolutional student with the transformer-based teacher. Their feature layouts differ, making direct correspondence difficult. MARD instead compares relations and temporal signals computed within each model's own space. The consistent results across both students highlight the value of structure-level transfer under architectural heterogeneity.

Table 2: X3D-S results on the EK100 validation set. Best and second-best student results are bold and underlined.

Method	Type	Teacher fine-tuning	Action ↑	Verb ↑	Noun ↑
Teacher (reference)	–	Yes	46.36	63.99	59.22
X3D-S	None	–	27.06	56.32	38.99
X3D-S+AFT	Feature	No	27.29	55.91	39.24
X3D-S+RLD	Logit	Yes	29.88	<u>57.36</u>	<u>42.13</u>
X3D-S+KL	Logit	Yes	27.49	56.59	39.32
X3D-S+PAT	Feature	Yes	28.20	56.13	41.73
X3D-S+SKD	Logit	Yes	29.09	56.83	41.78
X3D-S+FBT	Feature	Yes	<u>31.09</u>	56.98	41.78
X3D-S+MARD	Structure	No	**33.09**	**59.45**	**42.33**

4.3 Ablation Studies

Protocol. We ablate the constrained adapter and three distillation objectives with X3D-S on EK100, using the same student training protocol as in the main comparison. Table 3 includes the baseline, an adapter-only variant, each objective alone, all objective pairs, all three objectives without the adapter, and full MARD.

Table 3: Component ablation with X3D-S on the EK100 validation set.

Adapter	GSD	TCD	TRD	Verb ↑	Noun ↑	Action ↑
Baseline				56.32	38.99	27.06
✓	×	×	×	56.37	39.03	27.10
✓	✓	×	×	57.29	40.60	29.49
✓	×	✓	×	57.58	41.15	30.16
✓	×	×	✓	58.18	40.24	29.73
✓	✓	✓	×	58.25	41.88	31.59
✓	✓	×	✓	58.86	41.33	31.36
✓	×	✓	✓	59.04	41.75	32.06
×	✓	✓	✓	59.15	41.97	32.48
✓	✓	✓	✓	**59.45**	**42.33**	**33.09**

Individual component trends. Each objective improves action accuracy when used alone. GSD raises it from 27.06 to 29.49, while TCD achieves the strongest single-objective action and noun results, at 30.16 and 41.15. TRD yields the highest single-objective verb accuracy, at 58.18. These differences are consistent with the global, task-semantic, and temporal emphasis of the three objectives, although a single ablation cannot isolate the representation change responsible for each gain. The adapter-only variant remains close to the baseline, as no distillation objective transfers its adapted teacher features to the student.

Complementarity and adapter effect. Combining objectives produces further gains: TCD with TRD is the strongest pair at 32.06 action accuracy. All three objectives without the constrained adapter reach 32.48, compared with 33.09 for full MARD. Relative to the corresponding two-objective variants, adding GSD, TCD, or TRD in the full configuration increases action accuracy by 1.03, 1.73, or 1.50 points, respectively. These results support the complementarity of the objectives and show an additional gain from the adapter in this setting.

AI use statement

Generative AI tools were used to assist with drafting the initial manuscript and polishing its language. The authors reviewed and revised all AI-assisted text and take full responsibility for the final content of the paper.

REFERENCES

Mathieu Blondel, Arthur Mensch, and Jean-Philippe Vert. Differentiable divergences between time series. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3853–3861, 2021.

Marco Cuturi and Mathieu Blondel. Soft-DTW: A differentiable loss function for time-series. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 894–903, 2017.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision*, 130:33–55, 2022.

Christoph Feichtenhofer. X3D: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 203–213, 2020.

Rohit Girdhar, Du Tran, Lorenzo Torresani, and Deva Ramanan. DistInit: Learning video representations without a single labeled video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 852–861, 2019.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. 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.

Zhiwei Hao, Jianyuan Guo, Kai Han, Yehui Tang, Han Hu, Yunhe Wang, and Chang Xu. One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Xiaohu Huang, Hao Zhou, Kun Yao, and Kai Han. FROSTER: Frozen CLIP is a strong teacher for open-vocabulary action recognition. In *International Conference on Learning Representations*, 2024.

Hyeon-Jin Jeong, Han-Jin Lee, and Seok-Hwan Choi. Streamlined knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, 2019.

Lukas Kuhn, Lucas Maes, Giuseppe Serra, Quentin Le Lidec, Yann LeCun, Randall Balestriero, and Florian Buettner. LeVJEPA: Efficient & scalable video pretraining without the heuristics. *arXiv preprint arXiv:2608.27395*, 2026.

Guopeng Li, Qiang Wang, Ke Yan, Shouhong Ding, Yuan Gao, and Gui-Song Xia. Fuse before transfer: Knowledge fusion for heterogeneous distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3445–3454, 2025.

Ping Li, Chenhao Ping, Jie Song, and Mingli Song. Channel-wise dynamic knowledge distillation via adaptive sample generation for action recognition. *arXiv preprint arXiv:2608.03100*, 2026a.

Xianhang Li, Chen Huang, Chun-Liang Li, Eran Malach, Joshua Susskind, Vimal Thilak, and Etai Littwin. Rethinking JEPA: Compute-efficient video self-supervised learning with frozen teachers. In *International Conference on Learning Representations*, 2026b.

Jhe-Hao Lin, Yi Yao, Chan-Feng Hsu, Hong-Xia Xie, Hong-Han Shuai, and Wen-Huang Cheng. Perspective-aware teaching: Adapting knowledge for heterogeneous distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

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*, pp. 3202–3211, 2022.

Chengyao Qian, Trung Le, and Mehrtash Harandi. A good teacher adapts their knowledge for distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1239–1248, 2025.

Shikai Qiu, Boran Han, Danielle C. Maddix, Shuai Zhang, Bernie Wang, and Andrew Gordon Wilson. Transferring knowledge from large foundation models to small downstream models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 41644–41657, 2024.

Gorjan Radevski, Dusan Grujicic, Matthew Blaschko, Marie-Francine Moens, and Tinne Tuytelaars. Multimodal distillation for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5213–5224, 2023.

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. AM-RADIO: Agglomerative vision foundation model reduce all domains into one. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12490–12500, 2024.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

Maria Santos-Villafranca, Dustin Carrión-Ojeda, Alejandro Perez-Yus, Jesus Bermudez-Cameo, Jose J. Guerrero, and Simone Schaub-Meyer. Knowledge distillation for multimodal egocentric action recognition robust to missing modalities. *arXiv preprint arXiv:2504.08578*, 2025.

Mert Bülent Sarıyıldız, Philippe Weinzaepfel, Thomas Lucas, Pau de Jorge, Diane Larlus, and Yannis Kalantidis. DUNE: Distilling a universal encoder from heterogeneous 2d and 3d teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30084–30094, 2025.

Wujie Sun, Defang Chen, Siwei Lyu, Genlang Chen, Chun Chen, and Can Wang. Knowledge distillation with refined logits. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1365–1374, 2019.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017.

Yue Zhao and Philipp Krähenbühl. Training a large video model on a single machine in a day. *arXiv preprint arXiv:2309.16669*, 2023.

A Additional Results on Something-Something V2

We additionally evaluate MARD on Something-Something V2 (SSv2) with both student architectures. MARD uses the frozen Ego4D-pretrained AVION ViT-L teacher here as on EK100. Tables 4 and 5 report top-1 and top-5 classification accuracy. The teacher fine-tuning column indicates whether AVION ViT-L is adapted on SSv2 before providing supervision. Best and second-best results are determined among student models; the fine-tuned teacher is shown only as a reference.

Table 4: SSv2 validation results with VideoSwin-S. Best and second-best student results are bold and underlined.

Method	Type	Teacher fine-tuning	Top-1 ↑	Top-5 ↑
Teacher (reference)	–	Yes	72.36	94.01
VideoSwin-S	None	–	56.95	83.31
VideoSwin-S+AFT	Feature	No	58.02	85.60
VideoSwin-S+RLD	Logit	Yes	60.66	84.59
VideoSwin-S+KL	Logit	Yes	63.94	88.89
VideoSwin-S+PAT	Feature	Yes	60.50	84.41
VideoSwin-S+SKD	Logit	Yes	<u>64.20</u>	<u>89.74</u>
VideoSwin-S+FBT	Feature	Yes	63.94	88.89
VideoSwin-S+MARD	Structure	No	**66.49**	**90.94**

With VideoSwin-S, MARD improves the non-distilled baseline by 9.54 points in top-1 accuracy and 7.63 points in top-5 accuracy. It also exceeds the strongest compared student, SKD, by 2.29 and 1.20 points, respectively.

Table 5: SSv2 validation results with X3D-S. Best and second-best student results are bold and underlined.

Method	Type	Teacher fine-tuning	Top-1 ↑	Top-5 ↑
Teacher (reference)	–	Yes	72.36	94.01
X3D-S	None	–	52.29	81.60
X3D-S+AFT	Feature	No	52.97	81.81
X3D-S+RLD	Logit	Yes	<u>57.34</u>	<u>85.36</u>
X3D-S+KL	Logit	Yes	54.63	83.76
X3D-S+PAT	Feature	Yes	52.46	82.16
X3D-S+SKD	Logit	Yes	57.33	85.24
X3D-S+FBT	Feature	Yes	56.82	84.75
X3D-S+MARD	Structure	No	**57.59**	**85.67**

With X3D-S, MARD improves the non-distilled baseline by 5.30 points in top-1 accuracy and 4.07 points in top-5 accuracy. It achieves the highest student top-1 and top-5 results, exceeding RLD by 0.25 and 0.31 points, respectively, and further supports the effectiveness of structure-level transfer across heterogeneous student architectures.

B VISUALIZATION ANALYSIS

We qualitatively compare the spatiotemporal responses of the non-distilled baseline, representative distillation methods, and MARD on both datasets and student architectures. Figures 4–7 each show eight sampled frames from one video. The first row contains the input sequence, while the remaining rows overlay the response maps produced by each model and report its prediction. We deliberately show challenging cases in which MARD is correct and the compared students are incorrect. These examples therefore illustrate characteristic response patterns rather than replace the aggregate comparisons in Tables 1–5.

EPIC-KITCHENS-100. The EK100 examples require recognizing a brief hand–object interaction in a cluttered first-person scene. With VideoSwin-S (Fig. 4), several alternatives distribute substantial response over the static sink or surrounding background and consequently predict actions involving unrelated objects. MARD instead concentrates its strongest responses near the lower interaction region as the manipulated item enters the scene, leading to the correct prediction of *throw napkin*. The X3D-S comparison in Fig. 5 shows a similar trend from a different architectural family. MARD follows the hand-held object through the manipulation, whereas competing responses are more diffuse, intermittent, or displaced toward scene context. The resulting prediction correctly distinguishes *throw skin* from visually related alternatives such as throwing a bag or paper.

Something-Something V2. SSv2 places greater emphasis on temporal relations between objects. In Fig. 6, MARD progressively focuses on the held object and its relation to the supporting object during the discriminative portion of the sequence. Competing methods often respond to the same general area but predict a different interaction, such as putting, hitting, or colliding. In the X3D-S example (Fig. 7), the response of MARD moves with the red container during the tipping motion and weakens after the manipulation ends. This temporally selective pattern supports the correct recognition of *pretending to pour something out of something, but something is empty*, for which object appearance alone is insufficient.

Across the four cases, MARD produces responses that are more consistently associated with the manipulated object, the hand–object interaction, and the discriminative temporal stage. The same qualitative tendency appears for both a transformer-based and a CNN-based student and on both an egocentric and a general action-recognition dataset. These visualizations complement the quantitative and ablation results by illustrating how multi-level structural supervision can promote task-relevant spatiotemporal evidence without imposing direct feature correspondence.

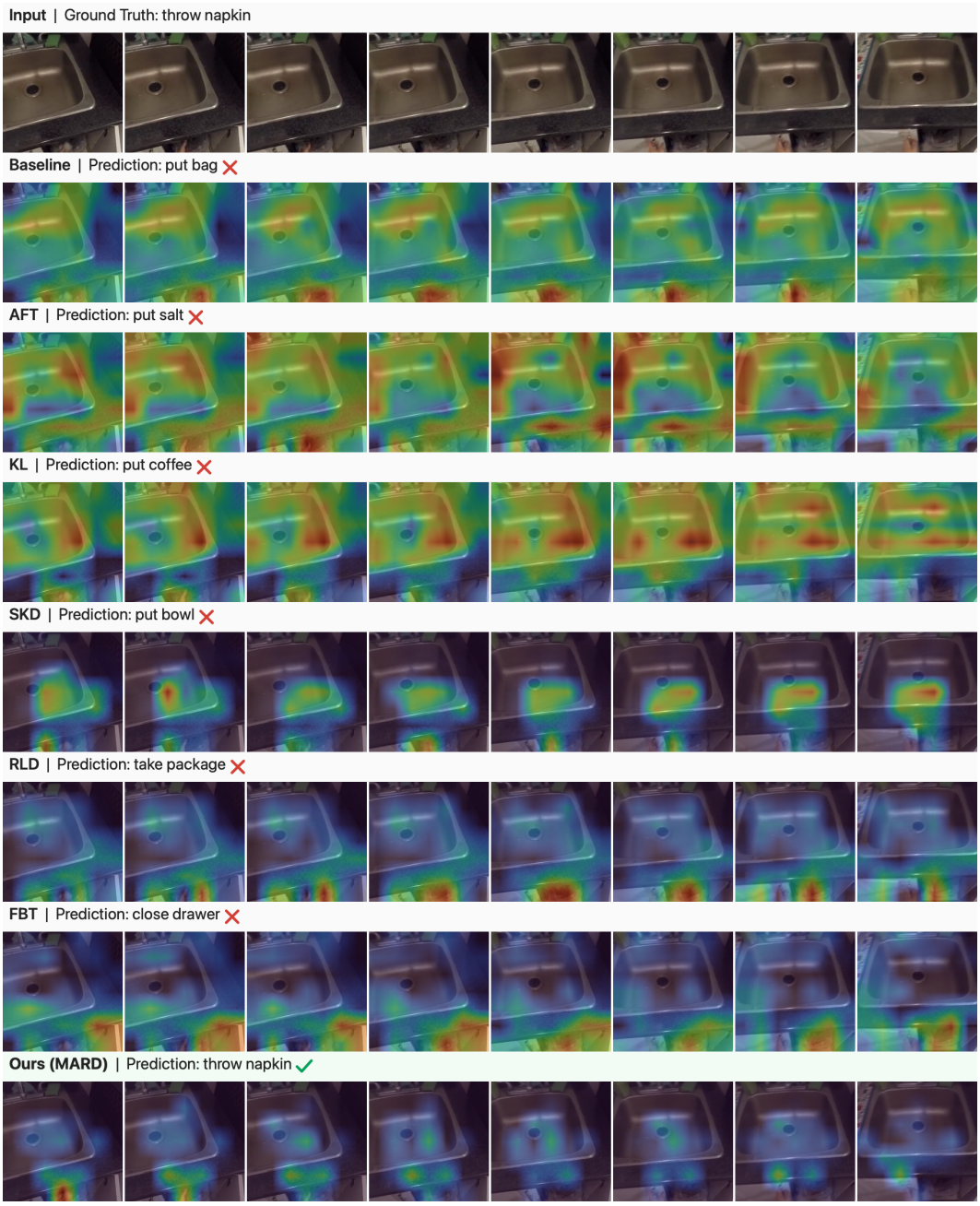

Figure 4: Spatiotemporal response comparison on EK100 with VideoSwin-S. MARD focuses on the hand–object interaction and correctly predicts *throw napkin*.

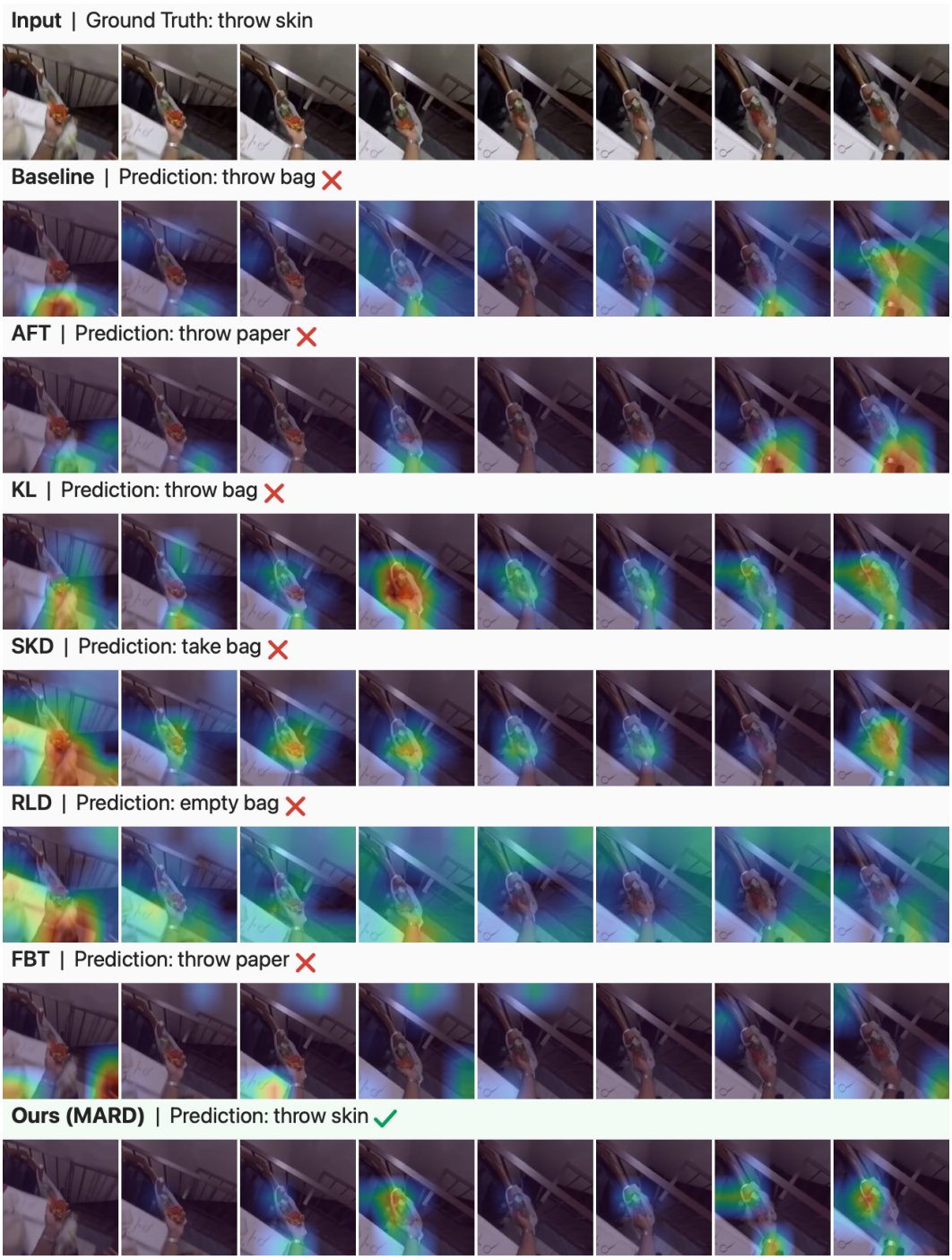

Figure 5: Spatiotemporal response comparison on EK100 with X3D-S. MARD follows the manipulated object and correctly predicts *throw skin*.

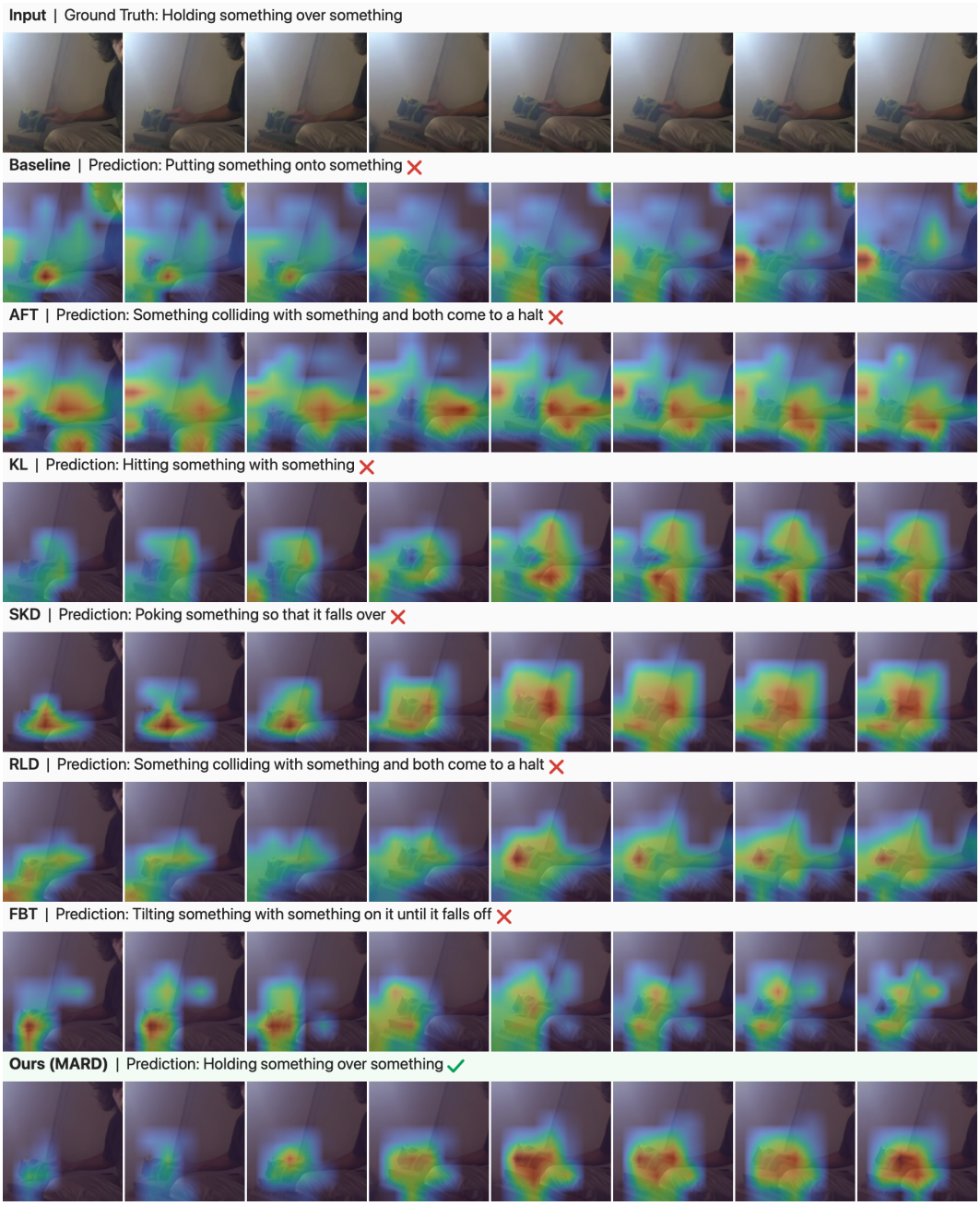

Figure 6: Spatiotemporal response comparison on SSv2 with VideoSwin-S. MARD captures the relation between the held and supporting objects and predicts the correct action.

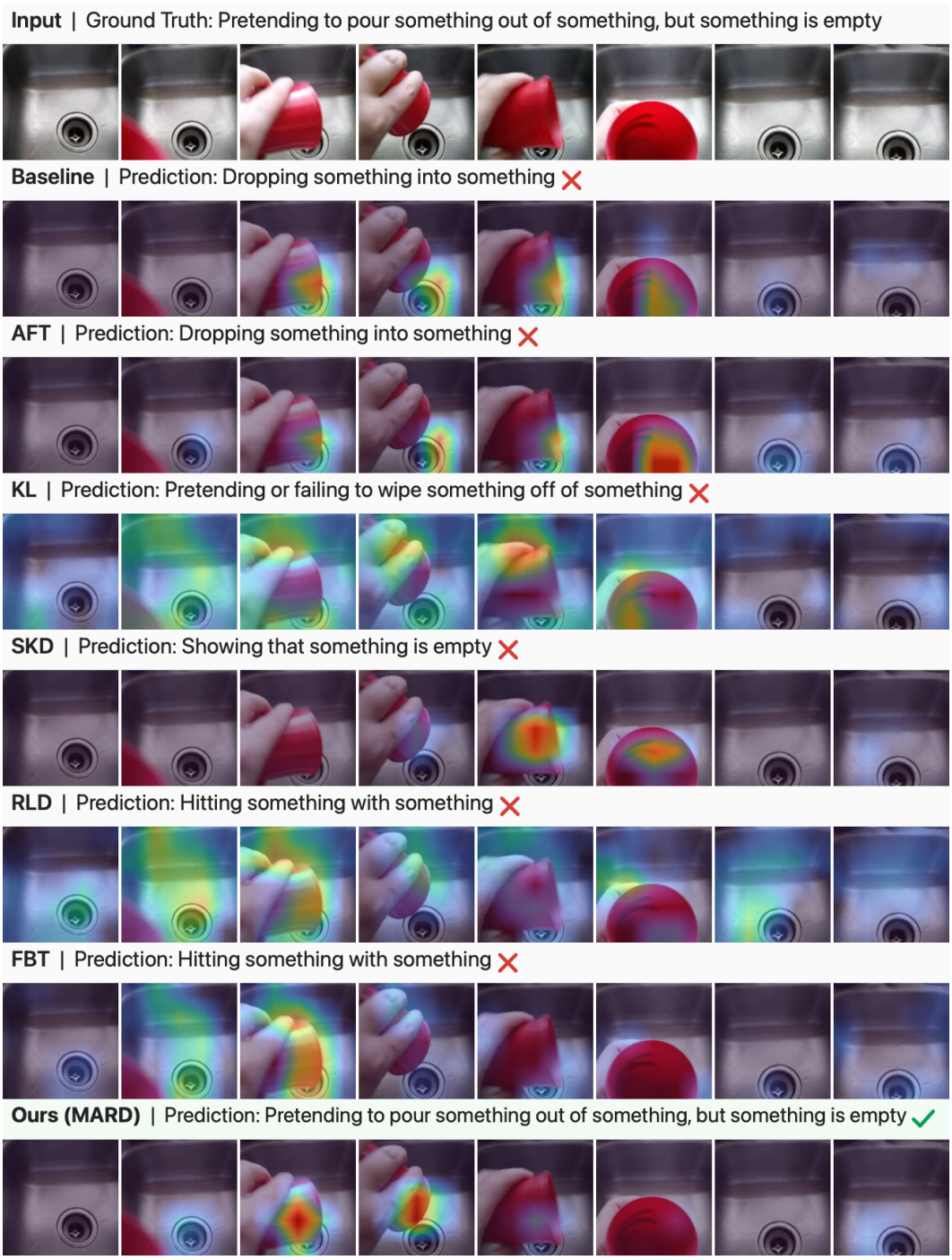

Figure 7: Spatiotemporal response comparison on SSv2 with X3D-S. MARD tracks the container during the discriminative tipping motion and predicts the correct action.