TAM: TASK-AWARE MEMORY DISTILLATION FOR EFFICIENT SPATIOTEMPORAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge distillation enables efficient spatiotemporal prediction by transferring knowledge from an accurate teacher to a compact student. However, matching outputs or features independently for each sample leaves cross-sample predictive structure underused. Exploiting this structure requires representations and historical references that reflect the dynamics of each task. We propose TAM, a Task-Aware Memory Distillation framework that organizes a frozen teacher's knowledge into a bounded, retrievable history. Memory entries encode latent features, forecast changes, or flow residuals, while task-specific selection rules identify relevant historical references. The student either matches the teacher's similarity distribution over shared references or regresses observation-conditioned residual prototypes. These objectives complement supervised prediction and conventional distillation. The teacher, memory, and auxiliary adapters are used only during training, leaving student inference unchanged. We evaluate TAM on video prediction, weather forecasting, and traffic flow prediction across multiple teacher–student configurations. Averaged over four paired runs, adding TAM improves SSIM on all six video datasets and reduces MSE on five relative to the corresponding KD baselines. Mean paired MSE reductions reach 1.86% on KittiCaltech, 1.93% on WeatherBench with a gSTA teacher, and 1.01% on TaxiBJ. These results demonstrate the utility of historical teacher supervision across distinct forecasting tasks without additional student inference cost. Code and experimental configurations will be released at https://anonymous.4open.science/r/TAM-review-C5D7/.

1 INTRODUCTION

Spatiotemporal prediction supports robotic interaction (Finn et al., 2016; Ebert et al., 2018), weather forecasting (Shi et al., 2015; Ravuri et al., 2021; Gao et al., 2022b), and traffic modeling (Zhang et al., 2017). Recurrent models capture spatiotemporal states (Wang et al., 2017), non-stationary dynamics (Wang et al., 2019b), and motion through historical-state recall (Wang et al., 2019a; Chang et al., 2021). Convolutional, attention-based, and physics-guided predictors offer complementary architectures (Gao et al., 2022a; Tan et al., 2023a; Le Guen & Thome, 2020; Wu et al., 2024; Tan et al., 2023b). Their advances motivate transferring predictive knowledge to compact students for efficient deployment.

Knowledge distillation provides teacher supervision (Hinton et al., 2015) through intermediate hints, attention maps, activation boundaries, and cross-stage features (Romero et al., 2015; Zagoruyko & Komodakis, 2017; Heo et al., 2019; Chen et al., 2021). Structured and channel-wise distillation preserve spatial dependencies in dense prediction (Liu et al., 2019; Shu et al., 2021); forecasting methods also transfer frequency-related knowledge and semantic priors (Li et al., 2025; Wang et al., 2026). Teacher accuracy alone does not guarantee effective transfer when teacher and student capacities differ (Cho & Hariharan, 2019). These findings motivate selecting supervision that the student can exploit.

A second choice concerns the examples defining supervision. Relational and similarity-preserving distillation transfer dependencies among representations (Park et al., 2019; Tung & Mori, 2019), while contrastive distillation transfers their discriminative structure (Tian et al., 2020). Instance-discrimination memory banks, momentum queues, and cross-batch memory retain references beyond the current mini-batch (Wu et al., 2018; He et al., 2020; Wang et al., 2020). CIRKD uses historical

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.

teacher embeddings for cross-image relational distillation (Yang et al., 2022). Unlike temporal state memory within a predictor, these references supply cross-sample training supervision. Forecasting therefore raises a task-specific question: which predictive quantities and historical examples should supervise the current dynamics?

Historical references have different meanings across forecasting tasks. Video dynamics concentrate in moving or occupied regions; weather states depend on geographic location; and future traffic changes depend on the observed flow and its recent trend. A generic feature memory does not explicitly account for these distinctions. We therefore consider two coupled choices: which predictive quantity to store and which historical entries should supervise the current sequence.

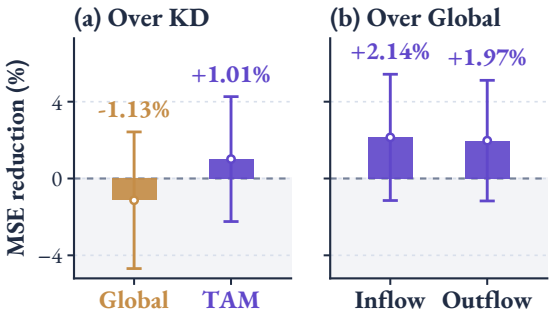


Figure 1: TaxiBJ memory comparison. (a) Overall MSE reduction over KD. (b) TAM's inflow/outflow gains over Global Memory. Bars: four-run paired means; whiskers: population SD. Positive values indicate improvement.

Figure 1 compares memory designs on TaxiBJ with a fixed TAU teacher, U-Net-Base student, and output-plus-feature MSE distillation baseline. Both memory variants use matched training settings, 2,048 historical entries, and 16 neighbors. Relative to the no-memory KD baseline, global residual memory increases test MSE by 1.13%, whereas TAM reduces it by 1.01% on average. Channel-wise analysis further shows that TAM reduces inflow and outflow MSE by 2.14% and 1.97%, respectively, relative to global memory. All percentages average relative changes over four paired runs. These results support organizing historical supervision around observation conditions and prediction horizons. However, gains vary across runs, and this joint comparison does not isolate the individual contributions of conditioned retrieval and horizon grouping.

We introduce TAM, a Task-Aware Memory Distillation framework that makes these choices explicit within a common teacher–student training procedure. A bounded memory stores latent features, forecast changes, or flow residuals from a frozen teacher. Task structure determines the current representation and the eligible historical references. Relational variants match teacher and student similarity distributions over shared references. For traffic flow prediction, observation-conditioned retrieval instead aggregates historical teacher residuals into regression targets for the student.

Memory supervision complements ground-truth prediction, output distillation, and optional feature alignment. The teacher and memory operate only during training, so inference retains the original student architecture and cost. We evaluate the additional effect of memory against matched distillation baselines on video, weather, and traffic prediction.

Our contributions are threefold:

- We formulate task-aware teacher memory as an interface for cross-sample supervision in spatiotemporal prediction, making the stored representation and reference selection explicit.
- We develop shared-reference relation matching and observation-conditioned residual prototypes as task-specific memory objectives that complement prediction and distillation.
- We evaluate teacher–student combinations and memory designs across three prediction domains, using matched comparisons across repeated runs to characterize the effect of historical supervision.

2 Related Work

Spatiotemporal predictive learning. Early recurrent predictors couple spatial processing with temporal state updates. ConvLSTM replaces fully connected transitions with convolutions (Shi et al., 2015), while PredRNN propagates spatiotemporal memory across both time and network depth (Wang et al., 2017). PhyDNet separates a dynamics component constrained by partial differential equations from complementary visual information (Le Guen & Thome, 2020). Recurrent-free models offer

another approach: SimVP uses a convolutional encoder–translator–decoder architecture (Gao et al., 2022a), and TAU combines spatial and temporal attention with a regularizer on frame differences (Tan et al., 2023a). Earthformer organizes space–time attention into cuboids for Earth-system forecasting (Gao et al., 2022b), while Earthfarsser studies a shared architecture for different dynamical systems (Wu et al., 2024). OpenSTL provides a common implementation and evaluation framework for recurrent and recurrent-free predictors (Tan et al., 2023b). Whereas these approaches develop predictive architectures and objectives, TAM augments the supervision of an existing student with task-specific teacher history, without modifying its inference architecture.

Knowledge distillation for efficient prediction. Knowledge distillation transfers information from a trained teacher to a student (Hinton et al., 2015). FitNets uses intermediate hints (Romero et al., 2015); attention transfer aligns attention maps (Zagoruyko & Komodakis, 2017); activation-boundary distillation transfers the partition induced by hidden neurons (Heo et al., 2019). ReviewKD studies connections across feature stages (Chen et al., 2021), and channel-wise distillation aligns spatial probability maps within channels for dense prediction (Shu et al., 2021). Recent forecasting methods further adapt the transferred representations to temporal prediction. Frequency-Aligned KD transfers frequency-related representations (Li et al., 2025), and S^2-KD combines semantic priors from a privileged multimodal teacher with spectral representations (Wang et al., 2026). TAM complements these objectives by organizing supervision across training sequences. A frozen forecasting teacher supplies historical features or forecast changes, and task-specific rules select references for the current sample. Matched output and feature baselines expose the additional effect of historical supervision.

Relational transfer and historical memory. Relational KD transfers distances and angles among examples (Park et al., 2019), and similarity-preserving KD maintains pairwise activation similarities (Tung & Mori, 2019). Contrastive representation distillation uses a contrastive objective to transfer representational information (Tian et al., 2020). More broadly, MoCo demonstrates how a queue can maintain a reference dictionary beyond the current mini-batch, with a momentum encoder controlling representation drift (He et al., 2020). The closest predecessor to our relational objective is CIRKD (Yang et al., 2022): it stores embeddings from a frozen segmentation teacher and aligns pixel-to-pixel and pixel-to-region similarity distributions across images. Our relational objective builds on this shared-reference formulation. We adapt the construction of teacher history to forecasting: motion or occupancy selects video regions, geographic location constrains weather references, and observed traffic flow conditions residual retrieval. The traffic variant further replaces relation matching with residual-prototype regression. Thus, TAM extends historical supervision through task-specific representations and reference organization, rather than introducing a new queue mechanism or relational divergence.

3 METHOD

TAM augments per-sample distillation with supervision from a bounded history of teacher predictions and representations. The task determines both the quantity stored in memory and the historical entries used as references. The student then learns either the teacher's relations to these references or a residual target aggregated from them. Figure 2 summarizes the framework.

3.1 PROBLEM FORMULATION

Given observations $X = (X_1, \dots, X_T)$, spatiotemporal prediction estimates the next K frames or fields, $Y = (Y_1, \dots, Y_K)$. For a training set $\mathcal{D} = \{(X^{(n)}, Y^{(n)})\}_{n=1}^{N}$, we formulate learning as conditional maximum likelihood:

$$\theta^\star = \arg\max_\theta \frac{1}{N} \sum_{n=1}^{N} \log p_\theta\left(Y^{(n)} \mid X^{(n)}\right). \tag{1}$$

Under a fixed-variance isotropic Gaussian observation model with conditional mean $F_\theta(X)$, maximizing this likelihood is equivalent to minimizing the mean squared prediction error:

$$\mathcal{L}_{\text{sup}}(\theta) = \frac{1}{ND} \sum_{n=1}^{N} \left\| Y^{(n)} - F_\theta(X^{(n)}) \right\|_F^2, \tag{2}$$

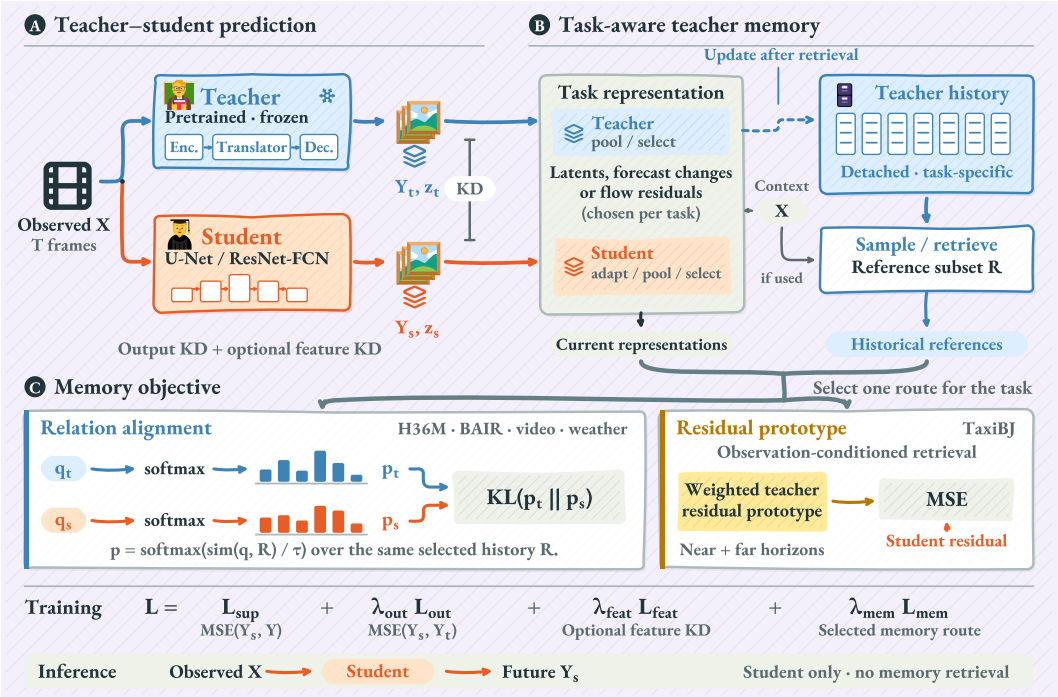


Figure 2: Overview of TAM. **(A)** A frozen teacher provides output and optional feature supervision. **(B)** Task-specific representations enter history after reference selection. **(C)** Memory supervision matches relations over shared references or regresses retrieved residual prototypes. Inference uses only the student. Glyphs and distribution bars are schematic.

where D is the number of scalar elements in each target sequence. The predictor $F_\theta(X)$ provides the deterministic forecast used at inference.

A pretrained teacher F_t and a student F_s produce forecasts $\widehat{Y}_t = F_t(X)$ and $\widehat{Y}_s = F_s(X)$, with intermediate features Z_t and Z_s. Distillation augments the student's supervised objective with output and optional feature matching, while keeping the teacher fixed. We further introduce a memory $\mathcal{M}$ to transfer relationships to previous training samples. Its construction specifies both the predictive representation and the historical references used to supervise it.

3.2 Task-Aware Teacher Memory

Predictive representations. Different tasks expose predictive structure through different quantities: latent features encode spatial states, forecast differences describe temporal changes, and flow residuals measure departures from persistence. We write the corresponding representation as

$$u_i^a = \phi_a(X, Z_a, \widehat{Y}_a)_i, \qquad a \in \{t, s\}, \tag{3}$$

where i indexes a region, a spatial location, or a sequence. The student transformation includes a learned adapter when dimensions differ. Relation matching uses normalized vectors $q_i^a = u_i^a / \|u_i^a\|_2$ and normalized teacher references. Residual regression instead preserves magnitudes to define targets in prediction space.

Reference selection. Historical entries are useful only when their comparison is meaningful for the task. We express reference selection as

$$\mathcal{R}_i = \mathrm{Select}(\mathcal{M}; X, u_i^t, \gamma_i), \tag{4}$$

where γ_i denotes metadata such as geographic location or episode identity; each instantiation uses the arguments it requires. Observed motion or foreground occupancy selects video anchors, geographic location restricts weather references, and observation keys retrieve sequences with related input conditions. Eligible entries are sampled or ranked by teacher-feature or observation-key similarity. Appendix A specifies the representations and selection rules for each task.

3.3 Distillation over Shared Historical References

Following cross-image relational distillation (Yang et al., 2022), the teacher and student define similarity distributions over a shared set of historical teacher vectors. For normalized teacher references v_j and temperature τ, the distributions are

$$p^a_{i,j} = \frac{\exp((q^a_i)^\top v_j/\tau)}{\sum_{\ell\in\mathcal{R}_i}\exp((q^a_i)^\top v_\ell/\tau)}, \qquad a \in \{t, s\}, \quad j \in \mathcal{R}_i. \tag{5}$$

Using the same references gives corresponding probability components the same meaning. We transfer this relational structure by minimizing

$$\mathcal{L}_{\text{rel}} = \sum_{i\in\mathcal{A}} w_i \, D_{\text{KL}}\big(\text{sg}(p^t_i) \,\|\, p^s_i\big), \tag{6}$$

where $\mathcal{A}$ is the anchor set, sg denotes stop-gradient, and the nonnegative weights satisfy $\sum_{i\in\mathcal{A}} w_i = 1$. Video variants average anchors uniformly; geographic-grid variants can use latitude-dependent weights. This objective supervises relative similarity to historical samples, complementing the direct matching of current outputs and features.

3.4 Observation-Conditioned Residual Prototypes

For traffic flow prediction, we use historical teacher forecasts to construct targets in prediction space. Taking the latest observation X_T as a persistence reference, the residual $\widehat{Y}_{a,h} - X_T$ represents the predicted change at horizon h. We pool and concatenate these residual fields separately over near- and far-horizon groups, yielding vectors r^g_a for $g \in \{\text{near}, \text{far}\}$ and $a \in \{t, s\}$. Keeping the groups separate preserves their horizon-specific targets.

Each memory entry stores the teacher residuals together with an observation key $k(X)$. The key is the normalized concatenation of the spatially pooled latest observation X_T and its recent change $X_T - X_{T-1}$. Retrieval therefore depends only on observed inputs. Let $\mathcal{R}(X)$ contain the top-k historical keys by cosine similarity, and let $\alpha_j(X)$ be the softmax weight of $k(X)^\top k_j/\tau$ over this set. The residual prototype is

$$\bar{r}^g(X) = \sum_{j\in\mathcal{R}(X)} \alpha_j(X)\, r^g_{t,j}, \qquad g \in \{\text{near}, \text{far}\}. \tag{7}$$

Both groups share the retrieved neighbors and weights. The student regresses their respective prototypes through

$$\mathcal{L}_{\text{proto}} = \frac{1}{2} \sum_{g\in\{\text{near},\text{far}\}} \text{MSE}\big(r^g_s, \text{sg}(\bar{r}^g(X))\big). \tag{8}$$

Thus, historical forecasts under similar observed conditions provide an additional target for the student's predicted changes. Unlike relation matching, this objective transfers residual values directly. Appendix A gives the explicit encoding and retrieval definitions.

3.5 Joint Optimization and Memory Updates

We combine historical supervision with ground-truth prediction, output distillation, and optional feature alignment:

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_{\text{out}}\mathcal{L}_{\text{out}} + \lambda_{\text{feat}}\mathcal{L}_{\text{feat}} + \lambda_{\text{mem}}\mathcal{L}_{\text{mem}}. \tag{9}$$

The supervised term is $\mathcal{L}_{\text{sup}} = \text{MSE}(\widehat{Y}_s, Y)$, and output distillation uses $\mathcal{L}_{\text{out}} = \text{MSE}(\widehat{Y}_s, \text{sg}(\widehat{Y}_t))$. MSE averages over tensor elements. The optional feature term aligns dimension-adapted intermediate features, while $\mathcal{L}_{\text{mem}}$ is the task's relation or prototype loss. Training jointly optimizes the student and enabled adapters, with losses aggregated over the batch.

Each iteration retrieves references and evaluates the memory loss before inserting the current detached teacher entries. The bounded memory retains recent training history under the task's location or episode constraints. Before sufficient history is available, optimization uses only prediction and conventional distillation losses. Memory order follows training iterations and does not imply chronological ordering of shuffled sequences.

At inference, $\widehat{Y}_s = F_s(X)$ requires only the student. The teacher, memory, and auxiliary adapters are used exclusively during training and add no student inference cost.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and protocols. We evaluate TAM on video prediction, weather forecasting, and traffic flow prediction. Video benchmarks include Moving MNIST (Srivastava et al., 2015), Moving FMNIST constructed from Fashion-MNIST (Xiao et al., 2017), KTH (Schuldt et al., 2004), Human3.6M (Ionescu et al., 2014), HMDB51 (Kuehne et al., 2011), BAIR (Ebert et al., 2017), and KittiCaltech, which pairs KITTI (Geiger et al., 2013) training data with Caltech Pedestrian (Dollár et al., 2009) test videos. Weather and traffic tasks use WeatherBench two-meter temperature (Rasp et al., 2020) and TaxiBJ inflow/outflow fields (Zhang et al., 2017). KTH and Human3.6M use subject-disjoint splits; HMDB51 follows official split 1; and BAIR uses our episode-balanced window-sampling protocol. WeatherBench uses chronological splits. TaxiBJ reserves the final portion of the official training set for validation and removes overlapping boundary windows. Appendix B specifies our prediction, preprocessing, and evaluation protocols.

Models and comparisons. Teachers use the OpenSTL forecasting implementation (Tan et al., 2023b), with TAU (Tan et al., 2023a), IncepU (Gao et al., 2022a), gSTA (Tan et al., 2025), or UniFormer (Li et al., 2022) translators. Students are forecasting adaptations of U-Net (Ronneberger et al., 2015), with Base and Tiny variants, and a ResNet-18-style encoder (He et al., 2016) with a fully convolutional decoder (ResNet-FCN). We compare supervised-only training, output distillation, output plus standardized feature-MSE distillation, and output plus activation-boundary distillation (Heo et al., 2019). TAM adds historical supervision to the corresponding distillation objective. Memory comparisons retain the same teacher, student, and underlying distillation configuration. Teachers remain frozen during training; inference uses only the student network.

Implementation. All configurations use Adam (Kingma & Ba, 2015). Video and traffic tasks use a one-cycle schedule (Smith, 2018) with a peak learning rate of 10^{-3}; WeatherBench uses cosine decay (Loshchilov & Hutter, 2017) with an initial rate of 5×10^{-3}. Training budgets are 200 epochs for Moving MNIST, Moving FMNIST, and BAIR, 50 for WeatherBench, and 100 for the remaining tasks. Batch size is 16, except for KittiCaltech, which uses 8. The output-distillation weight is 1; feature and memory weights depend on the task. Standard relational configurations maintain 4,096 teacher entries, enqueue up to 40 per iteration, and sample 1,024 references. BAIR, WeatherBench, and TaxiBJ instead use episode-aware retrieval, location-specific history, and observation-conditioned residual prototypes, respectively. Appendix C provides the configurations and hyperparameters.

Evaluation. Video metrics are MSE, MAE, PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018). TaxiBJ uses MSE, MAE, and RMSE; WeatherBench uses normalized errors and both ordinary and latitude-weighted RMSE in Kelvin. Tables specify the error reduction and scaling conventions. Multi-seed experiments use seeds 42–45 and report means, standard deviations, and the actual number of runs; relative changes use matched-seed comparisons. Tasks with an independent validation set select checkpoints by validation MSE. The current Moving MNIST, Moving FMNIST, KTH, and Human3.6M implementations share validation and test data; their evaluation-set checkpoint selection is reported separately, alongside available final-epoch results. A dagger in Table 1 denotes this shared-split protocol.

4.2 COMPARISON WITH DISTILLATION BASELINES

Table 1 compares supervised training, conventional knowledge distillation, and the addition of TAM across teacher–student configurations. With the teacher, student architecture, and base distillation objective fixed, TAM improves mean SSIM on all six datasets and reduces mean MSE on five. On KittiCaltech, for example, it reduces MSE from 6.983 to 6.853 (in the table's 10^3 scaling), raises SSIM from 0.6873 to 0.6917, and lowers LPIPS from 0.4199 to 0.4128. These gains indicate that historical teacher information provides useful supervision for cross-domain prediction.

TAM improves all three metrics over the corresponding KD baseline on KittiCaltech, KTH, Moving MNIST, and Moving FMNIST. The gains nevertheless vary across metrics: HMDB51 and BAIR achieve higher SSIM without a corresponding improvement in LPIPS.

Table 1: Video prediction under different teacher–student configurations. Students report mean ± population standard deviation over four runs; teachers use seed 42. Elementwise MSE is scaled by 10^3. Bold denotes the best student mean in each configuration.

Method	Params (M)	MSE ↓	SSIM ↑	LPIPS ↓
KittiCaltech				UniFormer → ResNet-FCN
Teacher	11.787	2.124	0.9099	0.0661
Student	12.811	7.187 ± 0.251	0.6813 ± 0.0066	0.4332 ± 0.0093
Output KD	12.811	6.983 ± 0.087	0.6873 ± 0.0034	0.4199 ± 0.0138
KD + TAM	12.811	**6.853** ± 0.112	**0.6917** ± 0.0019	**0.4128** ± 0.0078
HMDB51				TAU → U-Net-Base
Teacher	10.607	3.285	0.8831	0.0597
Student	7.706	4.331 ± 0.022	0.8371 ± 0.0060	0.0638 ± 0.0017
Feature-MSE KD	7.706	**3.829** ± 0.025	0.8472 ± 0.0031	**0.0625** ± 0.0011
KD + TAM	7.706	3.843 ± 0.036	**0.8525** ± 0.0018	0.0635 ± 0.0008
BAIR				TAU → Tiny U-Net
Teacher	10.607	23.294	0.6706	0.2317
Student	0.119	**24.556** ± 0.332	0.6419 ± 0.0046	0.2454 ± 0.0020
Output KD	0.119	24.773 ± 0.374	0.6439 ± 0.0045	**0.2406** ± 0.0005
KD + TAM	0.119	24.621 ± 0.225	**0.6455** ± 0.0029	0.2423 ± 0.0025
KTH†				IncepU → U-Net-Base
Teacher	12.152	2.544	0.8146	0.2755
Student	7.705	2.673 ± 0.013	0.8120 ± 0.0011	**0.2493** ± 0.0053
Feature-MSE KD	7.705	2.623 ± 0.051	0.8123 ± 0.0024	0.2693 ± 0.0042
KD + TAM	7.705	**2.605** ± 0.024	**0.8132** ± 0.0006	0.2652 ± 0.0011
Moving MNIST†				TAU → ResNet-FCN
Teacher	44.656	6.092	0.9340	0.0557
Student	12.748	15.033 ± 0.087	0.7994 ± 0.0113	0.2382 ± 0.0044
Output KD	12.748	14.845 ± 0.052	0.7990 ± 0.0085	0.2347 ± 0.0035
KD + TAM	12.748	**14.811** ± 0.041	**0.8026** ± 0.0066	**0.2341** ± 0.0022
Moving FMNIST†				TAU → ResNet-FCN
Teacher	44.656	5.999	0.8767	0.1413
Student	12.748	11.981 ± 0.011	0.7455 ± 0.0034	0.4124 ± 0.0024
Output KD	12.748	11.844 ± 0.013	0.7473 ± 0.0022	0.4074 ± 0.0028
KD + TAM	12.748	**11.835** ± 0.017	**0.7483** ± 0.0036	**0.4059** ± 0.0034

These results support teacher memory as a complement to existing distillation objectives. Student parameter counts remain unchanged, and inference requires neither the teacher nor memory retrieval.

4.3 Weather and Traffic Forecasting

Table 2 extends the evaluation to weather and traffic forecasting. On WeatherBench, TAM lowers mean MSE, MAE, and RMSE relative to Feature-MSE KD with both teachers. With gSTA, paired MSE decreases by 1.93% on average and temperature RMSE falls from 1.1191 K to 1.1081 K; TAU yields smaller gains. Geographically organized teacher history thus provides additional supervision, but its benefit depends on the teacher and does not consistently reduce variance.

On TaxiBJ, observation-conditioned residual memory lowers MSE by 1.01% on average over paired Feature-MSE KD runs, while also reducing MAE and RMSE. These gains support historical residual supervision for traffic prediction. Feature-AB KD remains stronger, however, so TAM complements the matched baseline without outperforming every distillation objective. Across video, weather, and traffic prediction, mean gains vary with the underlying distillation configuration.

Table 2: Weather and traffic forecasting (mean ± population SD over four runs). TAM augments Feature-MSE KD; Feature-AB KD is an independent baseline. Bold marks the best student mean per group. MSE/MAE are normalized; RMSE units are specified per group.

Method	MSE $\times 10^3 \downarrow$	MAE $\times 10^3 \downarrow$	RMSE $\downarrow$
WeatherBench · TAU → U-Net-Base			RMSE (K)
Student	2.833 ± 0.012	33.798 ± 0.211	1.1272 ± 0.0024
Feature-MSE KD	2.681 ± 0.025	32.419 ± 0.288	1.0965 ± 0.0051
MSE KD + TAM	**2.679** ± 0.016	**32.326** ± 0.161	**1.0962** ± 0.0034
WeatherBench · gSTA → U-Net-Base			RMSE (K)
Student	2.833 ± 0.012	33.798 ± 0.211	1.1272 ± 0.0024
Feature-MSE KD	2.793 ± 0.098	33.530 ± 0.842	1.1191 ± 0.0197
MSE KD + TAM	**2.740** ± 0.166	**32.961** ± 1.432	**1.1081** ± 0.0330
TaxiBJ · TAU → U-Net-Base			RMSE $\times 10^3$
Student	0.2123 ± 0.0034	8.146 ± 0.137	14.571 ± 0.118
Feature-MSE KD	0.2002 ± 0.0034	8.052 ± 0.145	14.148 ± 0.121
MSE KD + TAM	0.1981 ± 0.0054	7.949 ± 0.061	14.073 ± 0.193
Feature-AB KD	**0.1881** ± 0.0039	**7.830** ± 0.045	**13.716** ± 0.141

4.4 Prediction Horizon Analysis

Figure 3 reports horizon-dependent gains over matched KD on BAIR and TaxiBJ.

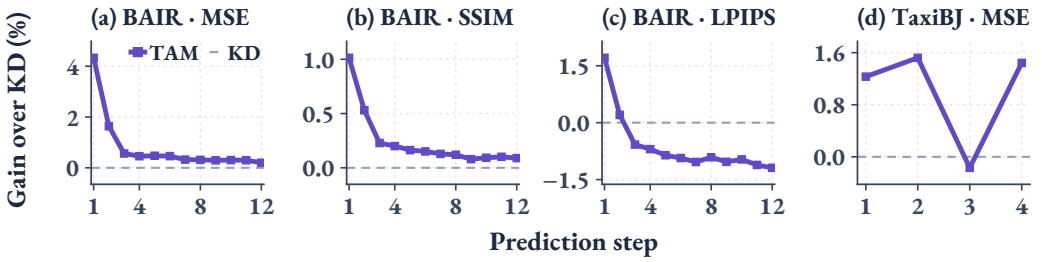


Figure 3: Prediction-horizon gains over matched KD baselines, averaged over four paired runs. Positive values indicate lower MSE/LPIPS or higher SSIM; the dashed line denotes the KD baseline.

On BAIR, TAM improves MSE and SSIM at all 12 prediction steps, but the MSE reduction falls from 4.33% at the first step to 0.20% at the last, indicating stronger near-term benefits. LPIPS improves only at the first two steps, so pixel and structural gains do not imply uniformly better perceptual quality. TaxiBJ exhibits MSE reductions of 1.23%, 1.52%, −0.17%, and 1.44% across its four steps. Thus, gains vary across horizons and are not monotonic.

4.5 Spatial Frequency Analysis

Figure 4 compares U-Net-Base power spectra and frequency-resolved errors for supervised training (S0) and TAM. Results cover all forecast frames from 256 fixed, evenly spaced test windows per dataset and four paired runs. Appendix D defines the transform and frequency bands.

On HMDB51, TAM reduces spectral error relative to S0 by 15.44%, 11.14%, and 6.73% in the low-, mid-, and high-frequency bands, respectively. KTH exhibits smaller reductions of 3.00% and 1.15% in the first two bands, while high-frequency error remains essentially unchanged (a 0.07% increase). Both students retain less high-frequency power than GT. The analysis therefore indicates dataset-dependent improvements in spatial reconstruction, with larger relative gains at low frequencies, rather than uniform recovery of fine detail.

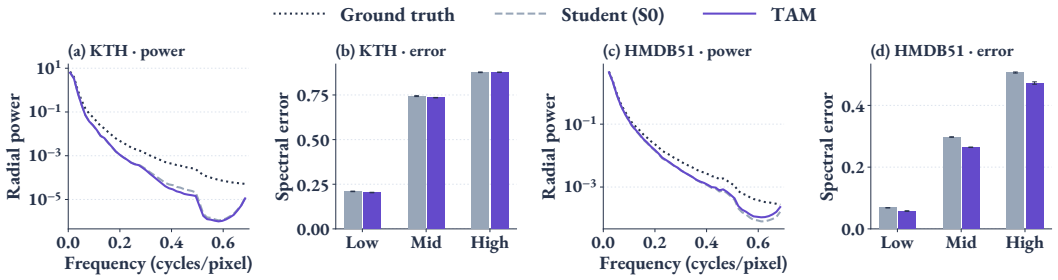


Figure 4: Spatial frequency analysis on KTH and HMDB51. Lower normalized spectral error is better; bars and error bars show means and standard deviations over four training runs.

4.6 Motion-Conditioned Prediction Analysis

Figure 5 compares TAM and S0 across motion conditions on 1,024 fixed windows per dataset and four paired runs. Absolute differences between the final two input frames define shared spatial masks; their spatial means determine motion groups (Appendix E).

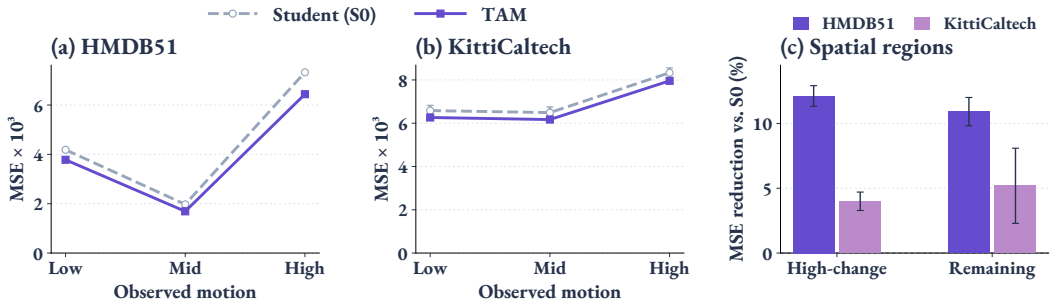


Figure 5: Prediction errors across observed-motion groups (left, middle) and relative MSE reductions in input-defined spatial regions (right). Error bars show standard deviations over four paired runs.

TAM reduces MSE by 9.68%, 14.31%, and 12.12% in the low-, middle-, and high-motion HMDB51 groups, and by 4.78%, 4.85%, and 4.45% on KittiCaltech. High-change regions improve by 12.12% and 4.00%, respectively, although KittiCaltech benefits more in the remaining region (5.19%). Gains thus extend across motion conditions without increasing monotonically with motion strength. These comparisons assess the full framework; Table 1 tests memory against matched KD.

5 Conclusion

We present TAM, a task-aware memory distillation framework that turns a frozen teacher's historical predictions and representations into cross-sample supervision. Task-specific representations and reference selection support relation alignment over shared history or observation-conditioned residual prototype regression, complementing per-sample distillation. Experiments on video, weather, and traffic prediction demonstrate improvements across multiple teacher–student configurations without additional student inference cost. Prediction-horizon, frequency, and motion-conditioned analyses show that gains vary across forecast steps, spatial frequencies, and motion conditions; improvements in pixel and structural metrics do not necessarily yield consistent perceptual gains. These findings support the organization of historical references as an important design choice for predictive distillation. The current framework still relies on task-specific representations and retrieval rules, and the available comparisons do not fully isolate the contributions of individual design choices. Future work will investigate more unified reference selection and controlled ablations to clarify the roles of different memory designs.

AI USE STATEMENT

We use ChatGPT and Codex to assist with literature search and synthesis, refinement of the research framework and its mathematical presentation, manuscript drafting and translation between Chinese and English, implementation and debugging of inference and analysis scripts, discussion of experimental and post-hoc analysis designs, interpretation of results, and figure creation and typesetting. The authors review and verify the AI-assisted work and take responsibility for the final text, code, experimental results, and scientific conclusions.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix A describe the method and task-specific memory designs. Section 4.1 and Appendices B–C document preprocessing, data splits, checkpoint selection, model configurations, training hyperparameters, and evaluation rules. Appendices D–E specify the sampling and computation procedures for frequency, motion-conditioned, and temporal analyses. We report the number of repeated runs, statistical conventions, and paired comparisons, and explicitly identify settings that share validation and test data. Code and experimental configurations will be released through the anonymous link in the abstract.

REFERENCES

Zheng Chang, Xinfeng Zhang, Shanshe Wang, Siwei Ma, Yan Ye, Xinguang Xiang, and Wen Gao. MAU: A motion-aware unit for video prediction and beyond. In *Advances in Neural Information Processing Systems*, 2021.

Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5008–5017, 2021.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4794–4802, 2019.

Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 304–311, 2009.

Frederik Ebert, Chelsea Finn, Alex X. Lee, and Sergey Levine. Self-supervised visual planning with temporal skip connections. In *Proceedings of the Conference on Robot Learning*, volume 78, pp. 344–356, 2017.

Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, and Sergey Levine. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv preprint arXiv:1812.00568*, 2018.

Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. *Advances in neural information processing systems*, 29, 2016.

Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z. Li. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3170–3180, 2022a.

Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. In *Advances in Neural Information Processing Systems*, volume 35, 2022b.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 2013.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3779–3787, 2019.

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

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. doi: 10.1109/TPAMI.2013.248.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Hilde Kuehne, Hueihan Jhuang, Estibaliz 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*, pp. 2556–2563, 2011. doi: 10.1109/ICCV.2011.6126543.

Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11474–11484, 2020.

Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. UniFormer: Unified transformer for efficient spatiotemporal representation learning. In *International Conference on Learning Representations*, 2022.

Yuqi Li, Chuanguang Yang, Hansheng Zeng, Zeyu Dong, Zhulin An, Yongjun Xu, Yingli Tian, and Hao Wu. Frequency-aligned knowledge distillation for lightweight spatiotemporal forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7262–7272, 2025. doi: 10.1109/ICCV51701.2025.00682.

Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019.

Stephan Rasp, Peter D. Dueben, Sebastian Scher, Jonathan A. Weyn, Soukayna Mouatadid, and Nils Thuerey. WeatherBench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12:e2020MS002203, 2020. doi: 10.1029/2020MS002203.

Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021.

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. URL https://arxiv.org/abs/1412.6550.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, volume 9351, pp. 234–241, 2015.

Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local SVM approach. In *Proceedings of the International Conference on Pattern Recognition*, volume 3, pp. 32–36, 2004.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems*, 2015.

Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5311–5320, 2021.

Leslie N. Smith. A disciplined approach to neural network hyper-parameters: Part 1—learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*, 2018.

Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using LSTMs. In *Proceedings of the International Conference on Machine Learning*, volume 37, pp. 843–852, 2015.

Cheng Tan, Zhangyang Gao, Lirong Wu, Yongjie Xu, Jun Xia, Siyuan Li, and Stan Z. Li. Temporal attention unit: Towards efficient spatiotemporal predictive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18770–18782, 2023a.

Cheng Tan, Siyuan Li, Zhangyang Gao, Wenfei Guan, Zedong Wang, Zicheng Liu, Lirong Wu, and Stan Z. Li. Openstl: A comprehensive benchmark of spatio-temporal predictive learning. In *Advances in Neural Information Processing Systems*, 2023b.

Cheng Tan, Zhangyang Gao, Siyuan Li, and Stan Z. Li. SimVPv2: Towards simple yet powerful spatiotemporal predictive learning. *IEEE Transactions on Multimedia*, 2025. doi: 10.1109/TMM.2025.3543051.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020.

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

Wenshuo Wang, Yaomin Shen, Yingjie Tan, and Yihao Chen. S^2-KD: Semantic-spectral knowledge distillation spatiotemporal forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1195–1203, 2026. doi: 10.1609/aaai.v40i2.37091.

Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R. Scott. Cross-batch memory for embedding learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6388–6397, 2020.

Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S. Yu. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. In *Advances in Neural Information Processing Systems*, 2017.

Yunbo Wang, Lu Jiang, Ming-Hsuan Yang, Li-Jia Li, Mingsheng Long, and Li Fei-Fei. Eidetic 3D LSTM: A model for video prediction and beyond. In *International Conference on Learning Representations*, 2019a.

Yunbo Wang, Jianjin Zhang, Hongyu Zhu, Mingsheng Long, Jianmin Wang, and Philip S. Yu. Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9154–9162, 2019b.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612, 2004.

Hao Wu, Yuxuan Liang, Wei Xiong, Zhengyang Zhou, Wei Huang, Shilong Wang, and Kun Wang. Earthfarsser: Versatile spatio-temporal dynamical systems modeling in one model. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 15906–15914, 2024.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, and Qian Zhang. Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12319–12328, 2022.

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. URL https://openreview.net/forum?id=Sks9_ajex.

Junbo Zhang, Yu Zheng, and Dekang Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

A Task-Specific Memory Instantiations

The implementation instantiates the memory interface with the representations and reference policies summarized in Table A1. These settings specialize the training objective to each task; the prediction network does not access the memory at inference.

Table A1: Memory instantiations. "Region r" denotes non-overlapping $r \times r$ average pooling. Representations and anchor selection are listed separately from reference organization.

Task	Representation / current anchors	Historical references
H36M	Region-2 features; observed-motion mask	4,096-token FIFO; sample 1,024
HMDB51	Features at high-motion locations	4,096-token FIFO; sample 1,024
BAIR	One motion-pooled feature token for each sequence	Episode keys; retrieve 64 from other episodes
KittiCaltech	Region-2 features; all regions are anchors	Sample 1,024; enqueue high- and ordinary-motion tokens
Moving MNIST / FMNIST	Region-2 features; recent-occupancy mask	Sample 1,024; retain top-128 teacher references per anchor
KTH	Regional forecast differences with coordinates; motion anchors	4,096-token FIFO; sample 1,024
WeatherBench	Features at each geographic grid location	64 historical windows per location
TaxiBJ	Pooled residuals from the latest flow; separate near/far horizons	2,048 observation-keyed entries; retrieve 16

Observed motion or occupancy selects video anchors, while the representation and reference policy determine the transferred relation. BAIR uses one motion-pooled token per sequence, retains at most one historical entry per episode, and excludes the current episode from retrieval. KTH augments regional forecast differences with spatial coordinates. Weather memory compares features at corresponding geographic locations and weights the loss by normalized cosine latitude. Appendix C lists pooling sizes, memory capacities, warm-up requirements, and other hyperparameters.

Residual prototype definitions. Let $\mathcal{P}$ apply spatial average pooling to each field and then vectorize the result, preserving the horizon and channel order. With chronological observation notation, the key and residual encodings in Section 3.4 are

$$k(X) = \mathrm{norm}_2(\mathcal{P}([X_T, X_T - X_{T-1}])), \tag{A.1}$$

$$r_a^g = \mathcal{P}\left([\widehat{Y}_{a,h} - X_T]_{h \in \mathcal{H}_g}\right), \qquad a \in \{t, s\}, \quad g \in \{\mathrm{near}, \mathrm{far}\}. \tag{A.2}$$

Here brackets denote concatenation. The TaxiBJ loader stores the latest observation first; the notation above orders observations chronologically. We use $\mathcal{H}_{\mathrm{near}} = \{1, 2\}$ and $\mathcal{H}_{\mathrm{far}} = \{3, 4\}$ for TaxiBJ. Pooling and concatenation retain individual horizons within each group. The normalized observation keys determine the reference weights:

$$\alpha_j(X) = \frac{\exp(k(X)^\top k_j/\tau)}{\sum_{\ell \in \mathcal{R}(X)} \exp(k(X)^\top k_\ell/\tau)}, \qquad j \in \mathcal{R}(X). \tag{A.3}$$

The same weights aggregate the near- and far-horizon residual banks; both targets are detached during student optimization.

The global-residual control in Figure 1 retrieves with the student's normalized predicted residual and combines all future horizons into one vector. Comparing it with observation-conditioned near/far prototypes changes both the retrieval key and horizon organization, without isolating their individual effects.

B Data and Evaluation Protocols

Table A2 summarizes the prediction tasks. A time step denotes a sampled video frame, an hourly temperature field, or a traffic observation, depending on the dataset. Image values are scaled to $[0, 1]$; temperature fields use training-set standardization.

Table A2: Prediction protocols. Channels refer to the predicted field; Human3.6M retains the BGR channel order of its preprocessing pipeline.

Dataset	Input → output	Resolution	Channels
Moving MNIST / FMNIST	$10 \to 10$	64×64	1
KTH	$10 \to 20$	128×128	1
Human3.6M	$4 \to 4$	256×256	3
HMDB51	$4 \to 4$	64×64	3
BAIR	$4 \to 12$	64×64	3
KittiCaltech	$10 \to 1$	128×160	3
WeatherBench	$12 \to 12$	32×64	1
TaxiBJ	$4 \to 4$	32×32	2

Video data. Moving MNIST and Moving FMNIST generate 10,000 training sequences online per epoch, with two moving digit or clothing sprites, and evaluate on fixed sets of 10,000 sequences. KTH uses persons 01–16 for training and 17–25 for testing. Human3.6M uses S1/S5/S6/S7/S8 for training and S9/S11 for testing, sampling every fifth source frame. HMDB51 uses the unstabilized videos of official split 1, with a deterministic, class-stratified 10% validation holdout from the official training videos; action labels are not prediction targets.

BAIR uses the MP4-based episode-balanced protocol: each epoch samples one window per training episode, and a 15-epoch cycle covers all valid starts. Validation and testing enumerate all windows; the test set contains 3,840 windows from 256 episodes. These results use this sampling and preprocessing protocol rather than the separate full-window TFRecord setup. KittiCaltech uses disjoint KITTI recordings for training and validation, and Caltech Pedestrian test recordings for cross-domain evaluation. Caltech preprocessing samples every third decoded frame, resizes to height 128, and center-crops to width 160.

Weather and traffic data. WeatherBench uses two-meter temperature on the 5.625° grid. Training, validation, and testing cover 2010–2015, 2016, and 2017–2018, respectively. A scalar mean and standard deviation computed on the training period normalize all splits; windows do not cross temporal gaps. TaxiBJ maps the OpenSTL input range $[-1, 1]$ to $[0, 1]$. The first 90% of official training windows form the training partition and the final 10% form validation, with a seven-window boundary purge. The official test set provides final evaluation.

Checkpoint selection and metric conventions. HMDB51, BAIR, KittiCaltech, WeatherBench, and TaxiBJ use independent validation data for MSE-based checkpoint selection. Moving MNIST, Moving FMNIST, KTH, and Human3.6M reuse the evaluation data for validation, so their Best results involve evaluation-set selection; available fixed final-epoch results provide a separate checkpoint criterion. Video PSNR and SSIM use a data range of one. Elementwise mean MSE/MAE and OpenSTL spatial/channel-sum errors are distinguished explicitly; the latter multiply the former by CHW for a fixed field shape. Weather RMSE is also computed after denormalization, and latitude weighting uses normalized cosine latitude. Reported multi-run deviations use the population convention. Relative improvements average per-seed relative changes, rather than taking the ratio of aggregate means; each comparison identifies its metric and reference configuration.

C Training Configurations and Hyperparameters

Table A3 lists the teacher–student configurations underlying the task-specific memory experiments. "Base" and "Tiny" denote U-Net-Base and Tiny U-Net; "FCN" denotes ResNet-FCN. Teachers are trained separately and remain in evaluation mode during student optimization. Output distillation has weight $\lambda_{\text{out}} = 1$. A zero feature weight denotes output-only distillation; nonzero entries below use standardized feature MSE. Activation-boundary baselines replace this feature objective while retaining the corresponding task configuration.

Table A3: Training budgets and loss weights for the memory configurations. E: epochs; B: batch size. Budgets describe configured training lengths, not completion of every archived run.

Task	Teacher	Student	E	B	λ_{feat}	λ_{mem}
Moving MNIST	TAU	FCN	200	16	0	0.001
Moving FMNIST	TAU	FCN	200	16	0	0.001
KTH	IncepU	Base/Tiny	100	16	0.03	0.001
Human3.6M	TAU	Base	100	16	0	0.001
HMDB51	TAU	Base	100	16	0.1	0.001
BAIR	TAU	Tiny	200	16	0	0.001
KittiCaltech	UniFormer	FCN	100	8	0	0.003
WeatherBench	TAU/gSTA	Base	50	16	0.1	0.001
WeatherBench	TAU	Tiny	50	16	0.1	0.01
TaxiBJ	TAU	Base	100	16	0.1	0.1

Optimization and representations. Adam and OneCycleLR use a peak rate of 10^{-3} for video and traffic tasks. WeatherBench instead uses Adam with an initial rate of 5×10^{-3} and cosine decay. U-Net-Base uses four scales with base width 64; Tiny U-Net uses three scales with base width 16. ResNet-FCN uses a ResNet-18-style encoder and a progressive upsampling decoder. KittiCaltech uses the UniFormer-hs64 teacher. Feature adapters map student channels to the teacher dimension before per-sample, per-channel spatial standardization and MSE. Feature and memory adapters are distinct unless explicitly shared by the configuration. Activation-boundary losses average over tensor elements; output-only baselines omit feature alignment. Human3.6M uses FP32; the archived BAIR memory and TaxiBJ memory configurations specify BF16. The historical BAIR output-only configuration does not record precision, so a shared precision setting is not assumed.

Reference construction. The standard relational FIFO holds 4,096 entries, enqueues up to 40 per iteration, samples 1,024 references, and starts contributing after the bank fills. Similarity temperature is 0.1 and the auxiliary KD temperature is 1. Moving MNIST/FMNIST use 2×2 region pooling and occupancy from the last two observations with threshold 0.05, retaining the top 128 teacher references per anchor. Human3.6M selects the top 25% of observed-motion locations; HMDB51 likewise uses motion-selected anchors. KittiCaltech enqueues 20 high-motion and 20 ordinary-region tokens while retaining all regions as anchors. KTH uses 16×16 forecast-difference regions, a 25% motion fraction, and coordinates scaled by two.

BAIR stores at most one entry per episode, starts after 1,024 distinct episodes, and retrieves 64 references from other episodes. Its 384-dimensional key concatenates the latest RGB observation and signed first-to-last change after 8×8 pooling. WeatherBench Base configurations retain 64 historical entries per location and use latitude-weighted aggregation; the Tiny configuration instead uses a global FIFO with 1,024 sampled references. TaxiBJ stores 2,048 entries, retrieves 16 neighbors at temperature 0.1, and uses 4×4 pooling with horizon groups $\{1, 2\}$ and $\{3, 4\}$. All memory losses access history before the current detached teacher entries are inserted.

D Spatial Frequency Analysis Protocol

Sampling and predictions. For each of KTH and HMDB51, we select 256 windows at evenly spaced indices in the ordered test set before inspecting their predictions. The same windows are used for all models and four paired training runs (42–45), without filtering for positive gains. We evaluate every predicted frame: 20 for KTH and four for HMDB51. Both comparisons use U-Net-Base, with the checkpoint criterion described in Appendix B. This is a descriptive analysis of fixed test subsets; overlapping windows are not treated as independent evidence of statistical significance.

Spectral transform. For a predicted or target frame channel $z \in \mathbb{R}^{H\times W}$, let w denote a separable two-dimensional Hann window. We remove its weighted spatial mean $\mu_w(z)$ and compute an orthonormal two-dimensional Fourier transform:

$$F(z) = \mathrm{FFT}_2\left(\frac{w \odot (z - \mu_w(z))}{\sqrt{\langle w^2\rangle}}\right), \qquad \mu_w(z) = \frac{\sum_{u,v} w_{uv} z_{uv}}{\sum_{u,v} w_{uv}}. \tag{D.1}$$

Here $\langle w^2\rangle$ is the spatial mean squared window value. We use raw floating-point predictions without clipping, sharpening, or contrast adjustment. DC coefficients are excluded. Consequently, the spectral analysis omits per-frame mean-brightness discrepancies and is not identical to pixel MSE. Power is averaged over native image channels and all prediction times.

Radial spectra and band errors. Radial power averages $|F(z)|^2$ within 48 uniformly spaced frequency annuli. Radial spatial frequency is measured in cycles per pixel; its maximum is $\sqrt{1/2}$ at the corners of the two-dimensional frequency grid. Low, middle, and high bands are $(0, 1/8]$, $(1/8, 1/4]$, and $(1/4, \sqrt{1/2}]$, respectively. For model m, training run s, and band b, we compute

$$E^{(b)}_{m,s} = \frac{\sum_{i,t,c} \sum_{\omega\in\Omega_b} |F(\widehat{Y}^{i,t,c}_{m,s})(\omega) - F(Y^{i,t,c})(\omega)|^2}{\sum_{i,t,c} \sum_{\omega\in\Omega_b} |F(Y^{i,t,c})(\omega)|^2}. \tag{D.2}$$

The indices i, t, c cover sampled windows, future times, and channels. This complex-valued residual retains phase discrepancies; greater power alone need not indicate more accurate detail. Each band is normalized by its own GT energy, so band heights do not measure their contributions to total pixel MSE. Figure 4 reports means and population standard deviations across training runs. Relative reductions average $100(E^{(b)}_{\mathrm{S0},s} - E^{(b)}_{\mathrm{TAM},s})/E^{(b)}_{\mathrm{S0},s}$ over paired runs. Numerical checks verify Parseval energy conservation for the windowed fields.

E Motion and Temporal Analysis Protocol

Test subsets and aggregation. We select 1,024 evenly spaced windows from each ordered test set before inspecting predictions. HMDB51 uses U-Net-Base and KittiCaltech uses ResNet-FCN, with the existing validation-selected checkpoints and four paired runs (42–45). Predictions are evaluated in floating point without clipping or post-processing. The sampled windows span 882 HMDB51 clips and 66 KittiCaltech recordings; overlapping windows are not treated as independent statistical replicates. For each run, MSE is averaged over windows after averaging the relevant pixels, channels, and future times within each window. Relative reductions are computed from paired run-level errors and then averaged across runs. Error bars use population standard deviations.

Input-defined motion groups and regions. For window i, we define a spatial change map and its mean strength as

$$d_i(u,v) = \frac{1}{C}\sum_{c=1}^{C}\left|X^i_{T,c}(u,v) - X^i_{T-1,c}(u,v)\right|, \qquad a_i = \frac{1}{HW}\sum_{u,v} d_i(u,v). \tag{E.1}$$

Dataset-specific tertiles of a_i define the low-, middle-, and high-motion groups. The high-change mask retains the top 20% of pixels ranked by d_i, subject to $d_i > 1/255$; stable pixel-index order breaks ties. Its complement defines the remaining region. Both masks are fixed from the input and shared across models, runs, and future frames. This proxy also captures camera and illumination changes and does not identify semantic foreground or track moving objects. Group analysis uses all 1,024 windows. Both regional metrics use the same windows with nonempty high-change masks: 983 for HMDB51 and 1,024 for KittiCaltech. Regional errors weight windows equally.

Reconstruction of temporal changes. For each transition between future frames, we compare predicted and target frame differences:

$$D_{m,s,h} = \frac{1}{NCHW}\sum_{i=1}^{N}\left\|\left(\widehat{Y}^i_{m,s,h} - \widehat{Y}^i_{m,s,h-1}\right) - \left(Y^i_h - Y^i_{h-1}\right)\right\|_F^2, \quad h = 2,3,4. \tag{E.2}$$

Here m and s index the model and training run. On the same 1,024 HMDB51 windows, TAM reduces this error relative to S0 by 4.74%, 1.68%, and 0.96% across the three transitions (Figure 6). Gains diminish at later transitions. The metric measures agreement with observed temporal changes rather than visual smoothness alone. KittiCaltech predicts one future frame and therefore does not support this within-forecast analysis.

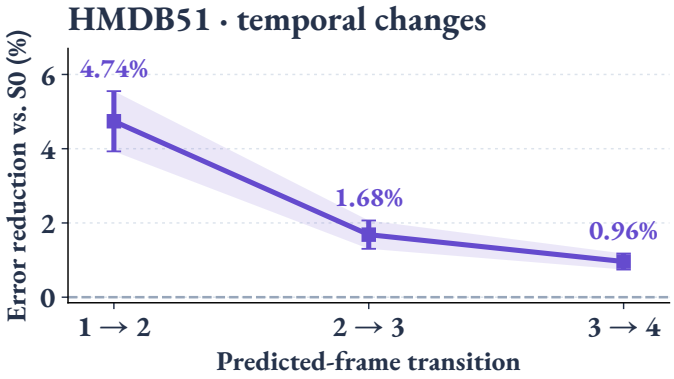


Figure 6: HMDB51 frame-difference MSE reduction relative to S0. Points and error bars show the mean and standard deviation over four paired runs.