MG-EATD: MULTI-GRANULARITY ENTROPY-ADAPTIVE TRAJECTORY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Most existing KD methods are primarily designed for homogeneous teacher–student architectures (e.g., CNN→CNN), leaving the potential and flexibility of knowledge transfer across heterogeneous architectures underexplored. Our work focuses on knowledge distillation across heterogeneous architectures. We observe that most existing logit-based methods follow a static endpoint-matching paradigm: they treat the teacher logits as a single deterministic target and constrain the student only to approximate the teacher's final class distribution. In contrast, we reformulate knowledge distillation as dynamic probability-trajectory learning, enabling the student to learn the teacher's probability states across multiple uncertainty levels as well as the local transitions between adjacent trajectory states. Motivated by this dynamic trajectory perspective, we propose Multi-Granularity Entropy-Adaptive Trajectory Distillation (***MG-EATD***), which reformulates final endpoint matching as joint state-and-transition learning along class-probability trajectories. Specifically, *MG-EATD* decomposes teacher and student features into multi-granularity regions and maps them into a unified class-probability space, constructing hierarchical knowledge from global semantics to local discriminative cues. Building on this, we introduce a class-probability trajectory diffusion mechanism that constructs teacher probability states at varying uncertainty levels through discrete forward perturbations. A time-conditioned trajectory predictor estimates the current student probability state, which is then propagated to an adjacent lower-noise state through a local reverse update. To leverage the teacher–student prediction priors naturally available in KD, we estimate sample difficulty using their entropies and jointly adapt both the diffusion-time range and the reverse evolution step size. Consequently, the student learns the teacher's class knowledge across spatial granularities and uncertainty levels, as well as local reverse transitions between adjacent trajectory states. Extensive experiments on multiple benchmark datasets under both homogeneous and heterogeneous models demonstrate the superiority of *MG-EATD*.

1 INTRODUCTION

Knowledge distillation (KD) has attracted growing attention for its ability to transfer knowledge from a strong teacher to a compact student, improving student performance without increasing its parameter count. Most of the existing KD methods (Hinton (2015); Romero et al. (2014); Zagoruyko & Komodakis (2016); Jin et al. (2023); Liu et al. (2023); Chen et al. (2022); Peng et al. (2019); Hossain et al. (2025); Heo et al. (2019); Chen et al. (2021); Tian et al. (2019); Su et al. (2025); Zhao et al. (2022)) have focused on homogeneous teacher—student pairs, such as knowledge transfer between CNNs of different capacities. However, this reliance on architectural homogeneity introduces inherent limitations. First, in specialized tasks or deployment scenarios, teachers that share the student's architectural family and are well matched to its learning characteristics may be limited, making it difficult to form effective teacher–student pairs. Second, with the emergence of diverse vision architectures such as Vision Transformers (Liu et al. (2021); Dosovitskiy et al. (2020); Touvron et al. (2021)) and MLP-based models (Tolstikhin et al. (2021); Touvron et al. (2022)), KD methods tailored to homogeneous pairs may not transfer knowledge effectively across heterogeneous architectures or fully exploit their complementary inductive biases. In contrast, heterogeneous knowledge distillation not only expands the pool of candidate teachers but also exploits complementary induc-

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.

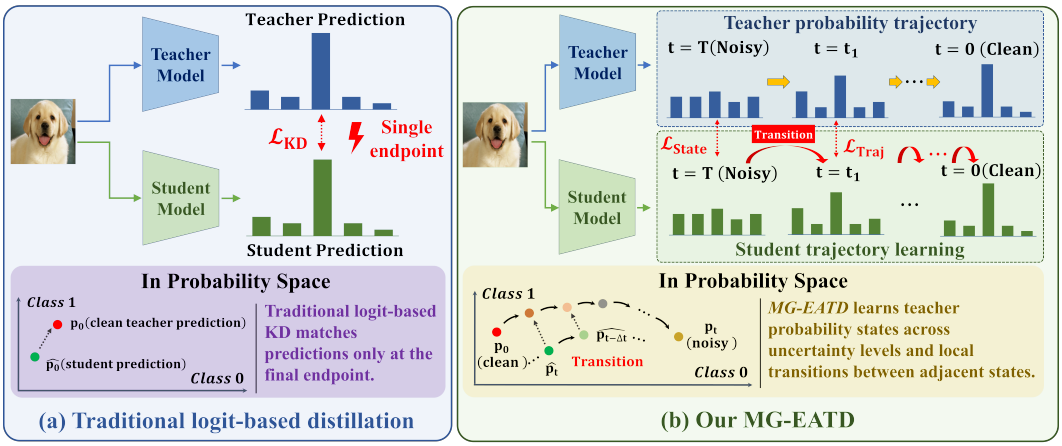


Figure 1: **Comparison between traditional logit-based distillation and *MG-EATD*.** (a) Traditional logit-based distillation aligns teacher and student predictions at a single endpoint. (b) *MG-EATD* extends this endpoint into a probability trajectory, enabling the student to learn both teacher probability states at varying uncertainty levels and local reverse transitions between adjacent states.

tive biases across heterogeneous architectures, providing knowledge that may be unavailable from a homogeneous teacher and potentially improving student performance. For example, with *MG-EATD* on ImageNet-1K, replacing the ResNet-34 teacher with Swin-T raises ResNet-18 accuracy from 72.37% to 72.48%, highlighting the potential of heterogeneous teachers. Motivated by this observation, we develop *MG-EATD* primarily for heterogeneous knowledge distillation and further validate it on homogeneous teacher–student pairs to demonstrate its broader applicability.

Hao et al. (2023) employ Centered Kernel Alignment (CKA) to quantify representational similarity across architectures and reveal a substantial discrepancy between the intermediate representations of heterogeneous models, thereby making it difficult to directly align representations and effectively transfer knowledge across heterogeneous models. This representational discrepancy is closely associated with the distinct inductive biases of different architectures(Raghu et al. (2021); Park & Kim (2022)). Specifically, sliding-window convolutions endow CNNs with strong locality and translation equivariance. In contrast, Vision Transformers and MLP-based models typically tokenize images into patches and capture long-range dependencies through global self-attention or token mixing, respectively, without the explicit locality bias inherent to convolutions.

Early studies have explored improving student performance using heterogeneous teachers. For example, DeiT (Touvron et al. (2021)) introduces a distillation token that enables a Transformer student to learn from a CNN teacher, demonstrating the potential of knowledge transfer across heterogeneous architectures. However, such approaches are typically tied to specific student architectures or specialized designs, limiting their applicability to diverse heterogeneous models. Hao et al. (2023) introduced OFA, the first unified distillation framework applicable to diverse heterogeneous teacher–student pairs. To address the substantial representational discrepancies between heterogeneous features, OFA projects the student's multi-stage intermediate features into the logit space and supervises each output using the teacher logits, thereby avoiding direct alignment between heterogeneous intermediate representations. Subsequently, several logit-based distillation methods have been developed for heterogeneous architectures. Despite their sophisticated designs, these methods essentially follow a static endpoint-matching paradigm: the teacher logits serve as a single deterministic supervision endpoint, and the student is trained solely to match the teacher's final, unperturbed distribution. Consequently, the probability states encoding teacher knowledge at different uncertainty levels, as well as the local transitions between adjacent states, remain explicitly unmodeled.

Prior studies have revealed the difficulty of teacher–student matching in knowledge distillation from different perspectives: Cho & Hariharan (2019); Mirzadeh et al. (2020) show that teacher–student capacity gaps can substantially affect knowledge transfer. Stanton et al. (2021) identify optimization difficulty as a major reason why even students with sufficient representational capacity may fail to faithfully reproduce the teacher's predictive distribution. These findings suggest that relying solely on a single deterministic logit endpoint may be insufficient for effective knowledge transfer, motivating us to extend KD from static endpoint matching to probability-trajectory learning.

Diffusion provides a natural mechanism for constructing trajectory distillation supervision. Existing studies that introduce diffusion models into KD primarily employ them as representation purification

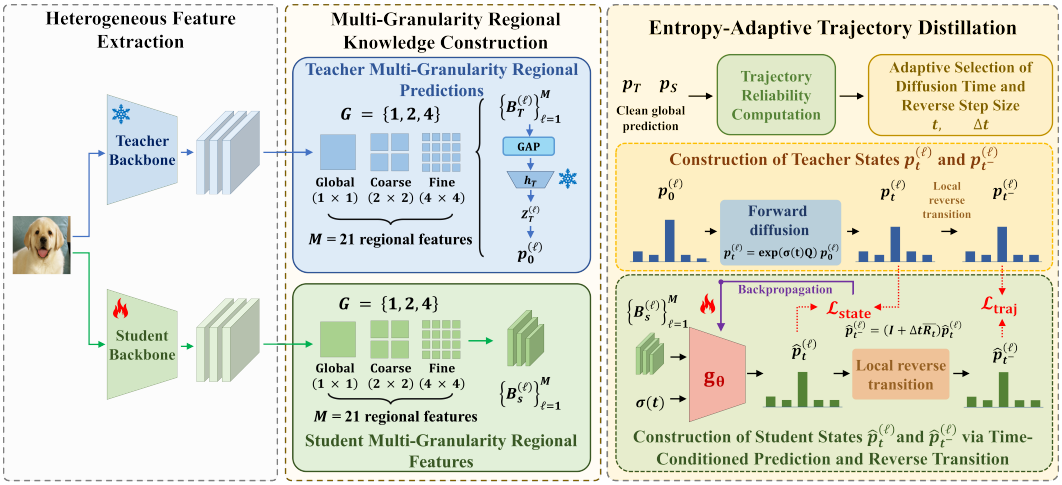


Figure 2: **Overall framework of *MG-EATD*.** Multi-granularity regional knowledge construction decomposes global teacher and student features into regional representations at different spatial granularities. To exploit the sample difficulty priors available in knowledge distillation, we jointly adapt the diffusion-time range and reverse step size based on the entropies of clean global teacher and student predictions. Teacher targets are constructed directly through forward diffusion. A time-conditioned predictor estimates the current student probability state from regional features, and a local reverse transition yields the adjacent lower-noise state.

tools in the feature space (Huang et al. (2023)) or use them to expand and augment the distillation data (Li et al. (2023; 2024)). Meanwhile, these methods still focus mainly on knowledge transfer between homogeneous models, leaving the diffusion process in the logit space and the knowledge encoded by its different trajectory states insufficiently explored. Score Entropy Discrete Diffusion (SEDD) (Lou et al. (2023)) provides a local-transition perspective for modeling reverse evolution in discrete state spaces. Conditioned on the current noisy state and diffusion time, SEDD estimates marginal probability ratios and uses them to construct the reverse transition rate matrix. Simulating these local transitions progressively evolves the marginal distribution toward lower-noise states. This perspective offers a natural foundation for trajectory distillation. As illustrated in Figure 1, discrete diffusion organizes the recovery of a probability distribution from a fully noisy state to its clean endpoint along a temporally ordered trajectory. This extends static endpoint matching to learning probability states and local transitions between adjacent states.

Based on the above analysis, we propose Multi-Granularity Entropy-Adaptive Trajectory Distillation (***MG-EATD***). Standard discrete diffusion models typically sample time steps from a common time range across all examples. While this design is suitable for generative modeling tasks where sample priors are unavailable, it does not fully exploit the sample difficulty priors provided by teacher and student predictions in KD. Ignoring these priors, a fixed diffusion-time range and evolution step size may over-perturb difficult samples or construct overly conservative trajectories for easy ones, thereby making it difficult to provide supervision that is well aligned with each sample's learning difficulty. Entropy (Shannon (1948)), a fundamental information-theoretic measure of uncertainty, provides a natural basis for assessing the reliability of teacher supervision and the student's learning difficulty. Moreover, a single global representation may not fully capture the discriminative information distributed across different local regions. *MG-EATD* therefore decomposes the features into regions at multiple spatial granularities, where coarse-grained representations summarize global semantics and fine-grained representations preserve distinct responses from local regions. Building on these, we construct class-probability trajectories at each spatial granularity and jointly adapt the diffusion-time range and reverse step size for each sample based on the entropies of teacher and student predictions. Thus, *MG-EATD* extends single endpoint matching into a trajectory learning process that spans multiple spatial granularities and adapts to sample difficulty. Extensive experiments across multiple benchmarks demonstrate the superiority of *MG-EATD*. The main contributions of this work are as follows:

1. We identify that single endpoint supervision in traditional logit-based distillation does not explicitly model intermediate probability states at varying uncertainty levels and local transitions between adjacent states. We therefore extend single endpoint matching to joint state-and-transition learning along probability trajectories.

2. We propose Multi-Granularity Entropy-Adaptive Trajectory Distillation (*MG-EATD*), which learns probability trajectories across spatial granularities. An entropy-driven mechanism jointly adapts the diffusion-time range and reverse step size, aligning trajectory supervision with sample difficulty.
3. Experiments on CIFAR-100 and ImageNet-1K validate the effectiveness and generalization of *MG-EATD* across homogeneous and heterogeneous models. Systematic ablation studies further validate the contribution of each component.

2 Related Work

We provide a detailed discussion of **Related Work** to the Appendix A.1.

3 Method

As illustrated in Figure 2, *MG-EATD* extends traditional single endpoint matching to learning teacher probability trajectories at varying uncertainty levels and local transitions between adjacent states. This section details the principles and implementation of *MG-EATD*. We consider a K-class classification dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i is an input image and $y_i \in \{1, \dots, K\}$ is its class label. Let f_T denote a fixed teacher and f_S a trainable student, with feature extractors ϕ_T, ϕ_S and classification heads h_T, h_S, respectively. For an input x, their penultimate features and logits are

$$\begin{aligned} F_T &= \phi_T(x), & F_S &= \phi_S(x), \\ z_T &= h_T(\mathrm{GAP}(F_T)), & z_S &= h_S(\mathrm{GAP}(F_S)), \end{aligned} \tag{1}$$

where GAP denotes global average pooling. Given a temperature $\tau > 0$, the corresponding probability is obtained through temperature-scaled softmax

$$p_T = \mathrm{softmax}(z_T/\tau), \qquad p_S = \mathrm{softmax}(z_S/\tau). \tag{2}$$

3.1 Multi-Granularity Regional Knowledge Construction

Most KD methods typically derive supervision from global representations, overlooking the diverse semantic cues and coupling among local regions within a single feature. Coarse-grained regions primarily capture global semantics, whereas fine-grained regions retain local object information and discriminative details. To extract and decouple knowledge at different spatial granularities, we construct multi-granularity regional knowledge from both teacher and student features.

Architectural differences often lead to discrepancies in feature channel dimensions and spatial organization. For example, CNNs organize visual information as 2D feature maps, whereas ViTs typically represent images as sequences of patch tokens. To facilitate regional partitioning, we first convert F_T and F_S into regular 2D spatial representations, denoted by $\widetilde{F}_T$ and $\widetilde{F}_S$, respectively.

We define the set of spatial granularities as $\mathcal{G} = \{1, 2, 4\}$. For each $g \in \mathcal{G}$, we first pad each feature as needed to make its spatial dimensions divisible by g, then partition it into a regular $g \times g$ grid, yielding g^2 non-overlapping regions at that granularity. Aggregating regions across granularities gives $M = \sum_{g \in \mathcal{G}} g^2 = 1 + 4 + 16 = 21$ regions per sample. The resulting teacher and student local regional feature sets are denoted by

$$\mathcal{B}_T = \left\{ B_T^{(\ell)} \right\}_{\ell=1}^{M}, \qquad \mathcal{B}_S = \left\{ B_S^{(\ell)} \right\}_{\ell=1}^{M}. \tag{3}$$

Here, $B_T^{(\ell)}$ and $B_S^{(\ell)}$ denote the teacher and student features for region ℓ, respectively, where $\ell \in \{1, \dots, M\}$. We then apply global average pooling to each teacher regional feature and reuse the teacher classification head h_T (Liu et al. (2023); Chen et al. (2022)) to map the pooled representation into logit space. The teacher logits for region ℓ are defined as

$$z_T^{(\ell)} = h_T\left(\mathrm{GAP}\left(B_T^{(\ell)}\right)\right). \tag{4}$$

We further convert the regional logits into the corresponding class-probability distribution using softmax with temperature τ:

$$p_T^{(\ell)} = \mathrm{softmax}\left(\frac{z_T^{(\ell)}}{\tau}\right). \tag{5}$$

For the subsequent class-probability diffusion process, we take the unperturbed teacher distribution in region ℓ as the clean initial state of its trajectory, i.e., $p_0^{(\ell)} \triangleq p_T^{(\ell)}$. During trajectory distillation, regional knowledge from both the teacher and student is mapped into class-probability space. The M regional distributions thus form a hierarchical collection of knowledge spanning global semantics and local discriminative details.

3.2 Class-Probability Trajectory Diffusion

To model the evolution of class-probability distributions across different uncertainty states and their local transitions between adjacent states, we introduce Class-Probability Trajectory Diffusion (**CPTD**), the diffusion-based framework underlying our method. CPTD defines a continuous-time Markov chain over the discrete class state space and uses clean regional teacher class-probability distributions $p_0^{(\ell)}$ across spatial granularities as the initial marginal distributions of the corresponding diffusion trajectories.

Forward Process. Let $Q \in \mathbb{R}^{K\times K}$ denote the transition rate matrix governing the forward evolution of class-probability distributions. CPTD adopts the following uniform transition rate matrix:

$$Q = \frac{1}{K}\mathbf{1}\mathbf{1}^\top - I, \tag{6}$$

where $\mathbf{1} \in \mathbb{R}^K$ is the all-ones vector and I is the identity matrix. For $a \neq b$, the off-diagonal entry Q_{ab} specifies the base transition rate between classes a and b before modulation by the noise schedule. Higher rates imply a greater probability of a class transition over a short time interval.

Let $\dot\sigma(t) \geq 0$ denote the instantaneous noise rate, with the time-dependent transition rate matrix given by $\dot\sigma(t)Q$. The corresponding cumulative noise intensity is defined as $\sigma(t) = \int_0^t \dot\sigma(s)\,\mathrm{d}s$. The cumulative noise intensity $\sigma(t)$ is obtained by integrating the instantaneous noise rate $\dot\sigma(t)$. For example, under a log-linear schedule with $\dot\sigma(t) = (1-\epsilon)/[1-(1-\epsilon)t]$, the integral admits the closed-form expression $\sigma(t) = -\log(1-(1-\epsilon)t)$. The class-probability distribution p_t, represented as a column vector, evolves according to

$$\frac{\mathrm{d}p_t}{\mathrm{d}t} = \dot\sigma(t)Qp_t. \tag{7}$$

Since Q is time-independent, this linear system admits the closed-form solution

$$p_t = \exp(\sigma(t)Q)\,p_0. \tag{8}$$

For the uniform transition rate matrix above, this solution further reduces to a convex combination of the clean distribution p_0 and the uniform distribution. A detailed derivation is provided in the Appendix A.2.

Reverse Process. Following prior work (Meng et al. (2022); Lou et al. (2023); Belhasin et al. (2026)), marginal probability ratios between classes play a role analogous of the score function in continuous spaces. For a source class a and a target class b, the concrete score at time t is defined as $s_t(a,b) = \frac{p_t(b)}{p_t(a)}$. The time-reversal formula for continuous-time Markov chains (Anderson (2012); Campbell et al. (2022)) gives the off-diagonal reverse transition rates

$$\overline{R}_t(b,a) = \dot\sigma(t)Q_{ab}s_t(a,b) = \dot\sigma(t)Q_{ab}\frac{p_t(b)}{p_t(a)}, \qquad b \neq a. \tag{9}$$

The diagonal entries are obtained by applying probability mass conservation and rearranging the resulting equations: $\overline{R}_t(a,a) = -\sum_{b\neq a}\overline{R}_t(b,a)$. A detailed derivation is provided in the Appendix A.3.

For a given diffusion time t, let r denote the elapsed reverse time. The reverse evolution of the class-probability distribution is governed by the linear ordinary differential equation

$$\frac{\mathrm{d}p_{t-r}}{\mathrm{d}r} = \overline{R}_{t-r}p_{t-r}. \tag{10}$$

For a sufficiently small reverse step $\Delta t > 0$, the first-order Euler approximation to the local reverse transition matrix is $\overline{\mathcal{T}}_t = I + \Delta t\, \overline{R}_t$. The reverse update from time t to the adjacent lower-noise time $t - \Delta t$ is

$$p_{t-\Delta t} \approx \overline{\mathcal{T}}_t p_t = (I + \Delta t\, \overline{R}_t) p_t. \tag{11}$$

Successive local reverse transitions progressively restore the class-probability distribution from high-noise to lower-noise states.

3.3 Training Objective

Multi-granularity regional probability distributions capture knowledge at different spatial granularities, but supervision for each region remains limited to a single endpoint derived from clean logits. CPTD introduces unified class-probability diffusion dynamics for these distributions, extending single endpoint matching to trajectory learning. Conventional diffusion models typically sample diffusion times from a common range and apply the same reverse-time schedule to all samples. In KD, however, teacher and student predictions provide sample difficulty priors before trajectory construction, allowing the schedule to be adapted to each sample. Ignoring this prior information may lead to trajectory supervision that is misaligned with each sample's current learning difficulty. When teacher predictions are unreliable or the student struggles with a sample, excessive perturbation may weaken informative class structure in the teacher distribution, while overly large reverse steps may amplify local discretization errors. Conversely, when teacher predictions are reliable and the student already discriminates well, overly conservative trajectories may limit exploration of the teacher distribution across uncertainty levels. We therefore introduce an entropy-driven mechanism in *MG-EATD* that jointly adapts the diffusion-time range and reverse step size for each sample based on teacher and student entropies.

To adapt trajectory schedules to sample difficulty, we assess the reliability of teacher supervision and the student's difficulty in learning each sample. Entropy (Shannon (1948)) quantifies uncertainty in probability distributions. Su et al. (2025); Kwon et al. (2020); Cheng et al. (2020) have explored its use in KD to assess sample importance and adapt supervision. Motivated by these studies, we combine teacher and student entropies to estimate sample difficulty, which guides the joint adaptation of the diffusion-time range and reverse step size for each sample. Given the temperature-scaled teacher and student class-probability distributions p_T and p_S in the clean state, their entropies are

$$H(p_A) = -\sum_{k=1}^{K} p_A(k) \log p_A(k), \qquad A \in \{T, S\}. \tag{12}$$

For a K-class task, the maximum entropy is $H_{\max} = \log K$. Using these entropies, we define the trajectory reliability of each sample as

$$b = 1 - \frac{H(p_T)}{2H_{\max}} \left(1 + \frac{H(p_S)}{H_{\max}}\right), \qquad b \in [0, 1]. \tag{13}$$

The trajectory reliability b jointly determines the maximum diffusion time and reverse step size through linear mappings:

$$t_{\max} = t_{\max}^{\text{weak}} + b\left(t_{\max}^{\text{strong}} - t_{\max}^{\text{weak}}\right), \quad \Delta t = \Delta t_{\min} + b\left(\Delta t_{\max} - \Delta t_{\min}\right). \tag{14}$$

Here, $t_{\max}^{\text{weak}}$ and $t_{\max}^{\text{strong}}$ denote the maximum diffusion times under weak and strong perturbations, respectively, while $\Delta t_{\min}$ and $\Delta t_{\max}$ bound the reverse step size. This design assigns stronger perturbations and larger reverse steps to samples with higher trajectory reliability, while favoring more conservative supervision for those with lower trajectory reliability. The design rationale is detailed in the Appendix A.4. We sample the current diffusion time within the prescribed range and constrain the reverse step size to keep the adjacent lower-noise time within the same range:

$$t \sim \mathcal{U}(t_{\min}, t_{\max}), \qquad \Delta t \leftarrow \min\{\Delta t,\, t - t_{\min}\}, \qquad t^- = t - \Delta t. \tag{15}$$

The diffusion time and reverse step size are computed per sample and shared across its M regions, ensuring consistent noise intensity and evolution scale across regions at different spatial granularities within the same sample. For any region $\ell \in \{1, \ldots, M\}$, the shared trajectory schedule satisfies

$$t^{(\ell)} = t, \quad \Delta t^{(\ell)} = \Delta t, \quad \sigma^{(\ell)} = \sigma(t), \quad \dot{\sigma}^{(\ell)} = \dot{\sigma}(t). \tag{16}$$

Table 1: **Comparison of Heterogeneous Distillation Results on CIFAR-100.** The best and second-best results are emphasized in **bold** and underlined cases.

Teacher	Student	From Scratch		Feature-based					Logit-based				
		T.	S.	FitNet	CC	RKD	CRD	FBT	KD	DKD	DIST	OFA	**MG-EATD**
Swin-T	ResNet18	89.26	74.01	78.87	74.19	74.11	77.63	81.61	78.74	80.26	77.75	80.54	**83.88**
ViT-S	ResNet18	92.04	74.01	77.71	74.26	73.72	76.60	81.93	77.26	78.10	76.49	80.15	**82.27**
Mixer-B/16	ResNet18	87.29	74.01	77.15	74.26	73.75	76.42	**81.90**	77.79	78.67	76.36	79.39	81.14
Swin-T	MobileNetV2	89.26	73.68	74.28	71.19	69.00	79.80	81.28	74.68	71.07	72.89	80.98	**81.85**
ViT-S	MobileNetV2	92.04	73.68	73.54	70.67	68.46	78.14	82.10	72.77	69.80	72.54	78.45	**82.50**
Mixer-B/16	MobileNetV2	87.29	73.68	73.78	70.73	68.95	78.15	**80.83**	73.33	70.20	73.26	78.78	80.04
ConvNeXt-T	DeiT-T	88.41	68.00	60.78	68.01	69.79	65.94	79.57	72.99	74.60	73.55	75.76	**85.56**
Mixer-B/16	DeiT-T	87.29	68.00	71.05	68.13	69.89	65.35	74.40	71.36	73.44	71.67	73.90	**80.34**
ConvNeXt-T	Swin-P	88.41	72.63	24.06	72.63	71.73	67.09	80.73	76.44	76.80	76.41	78.32	**84.45**
Mixer-B/16	Swin-P	87.29	72.63	75.20	73.32	70.82	67.03	78.44	75.93	76.39	75.85	76.65	**81.47**
ConvNeXt-T	ResMLP-S12	88.41	66.56	45.47	67.70	65.82	63.35	78.03	72.25	73.22	71.93	75.21	**86.43**
Swin-T	ResMLP-S12	89.26	66.56	63.12	68.37	64.66	61.72	77.20	71.89	72.82	11.05	73.58	**85.47**
Average Improvements				↓5.20	↓0.33	↓1.40	↓0.02	↑8.38	↑3.17	↑3.16	↓2.31	↑6.19	**↑11.50**

Let $p_0^{(\ell)}$ denote the clean teacher probability distribution for region ℓ. The teacher targets at time t and the adjacent lower-noise time $t^- = t - \Delta t$ are obtained from the analytical forward solution:

$$p_t^{(\ell)} = \exp(\sigma(t)Q)\, p_0^{(\ell)}, \qquad p_{t^-}^{(\ell)} = \exp\big(\sigma(t^-)Q\big)\, p_0^{(\ell)}. \tag{17}$$

These teacher states supervise current-state prediction and local reverse-transition learning, respectively. For the student, we introduce a time-conditioned trajectory predictor g_θ that estimates the class-probability state at the current diffusion time t from student regional features. The predictor architecture is chosen according to the teacher model to accommodate its representational characteristics. Specifically, g_θ uses a residual convolutional architecture for CNN teachers and a Transformer architecture for Transformer or MLP teachers, both incorporating time modulation. Detailed descriptions and analyses of these architectures are provided in the Appendix A.5. Unlike methods (Lou et al. (2023); Belhasin et al. (2026)) that directly parameterize probability ratios, g_θ outputs class logits conditioned on the cumulative noise intensity $\sigma(t)$. The corresponding probability estimate is obtained through temperature-scaled softmax:

$$\hat{z}_t^{(\ell)} = g_\theta\Big(B_S^{(\ell)}, \sigma(t)\Big), \qquad \hat{p}_t^{(\ell)} = \mathrm{softmax}\left(\frac{\hat{z}_t^{(\ell)}}{\tau}\right). \tag{18}$$

Following Section 3.3, we use ratios between these class probabilities to construct the estimated reverse transition rate matrix:

$$\hat{R}_t^{(\ell)}(b,a) = \dot{\sigma}(t) Q_{ab} \frac{\hat{p}_t^{(\ell)}(b)}{\hat{p}_t^{(\ell)}(a)} \quad (b \neq a), \qquad \hat{R}_t^{(\ell)}(a,a) = -\sum_{c \neq a} \hat{R}_t^{(\ell)}(c,a). \tag{19}$$

We then propagate the current student estimate toward the adjacent lower-noise state using the first-order Euler update operator $\widehat{\mathcal{T}}_t^{(\ell)}$:

$$\widehat{\mathcal{T}}_t^{(\ell)} = I + \Delta t\, \hat{R}_t^{(\ell)}, \qquad \tilde{p}_{t^-}^{(\ell)} = \widehat{\mathcal{T}}_t^{(\ell)} \hat{p}_t^{(\ell)}. \tag{20}$$

Since a finite-step local reverse update may not preserve nonnegativity, we apply a probability correction to the updated student prediction $\tilde{p}_{t^-}^{(\ell)}$, inspired by the numerical stabilization procedure in the implementation of DiDiCM (Belhasin et al. (2026)). Specifically, we clip its entries to a small positive floor and renormalize the resulting vector to ensure a valid probability distribution. During training, the current prediction $\hat{p}_t^{(\ell)}$ and the corrected lower-noise prediction $\tilde{p}_{t^-}^{(\ell)}$ are supervised by their corresponding teacher targets.

For each region, we define a unified distillation weight that combines its uncertainty at the perturbed trajectory state with a difficulty prior derived from the clean global predictions of the corresponding input sample:

$$w_t^{(\ell)} = \underbrace{\frac{H\Big(p_t^{(\ell)}\Big)}{2}\left(1 + \frac{H\Big(\hat{p}_t^{(\ell)}\Big)}{H_{\max}}\right)}_{\text{region-level uncertainty}} \underbrace{\left[1 + \frac{H(p_T)}{2H_{\max}}\left(1 + \frac{H(p_S)}{H_{\max}}\right)\right]}_{\text{sample-level difficulty}}. \tag{21}$$

Table 2: **Comparison of heterogeneous distillation results on ImageNet-1K.** The best and second-best results are emphasized in **bold** and underlined cases.

Teacher	Student	From Scratch		Feature-based					Logit-based				
		T.	S.	FitNet	CC	RKD	CRD	FBT	KD	DKD	DIST	OFA	**MG-EATD**
DeiT-T	ResNet18	72.17	69.75	70.44	69.77	69.47	69.25	71.22	70.22	69.39	70.64	71.01	**71.45**
Swin-T	ResNet18	81.38	69.75	71.18	70.07	68.89	69.09	72.21	71.14	71.10	70.91	71.76	**72.48**
Mixer-B/16	ResNet18	76.62	69.75	70.78	70.05	69.46	68.40	71.44	70.89	69.89	70.66	71.38	**71.51**
DeiT-T	MobileNetV2	72.17	68.87	70.95	70.69	69.72	69.60	71.78	70.87	70.14	71.08	71.39	**71.81**
Swin-T	MobileNetV2	81.38	68.87	71.75	70.69	67.52	69.58	72.54	72.05	71.71	71.76	72.32	**72.69**
Mixer-B/16	MobileNetV2	76.62	68.87	71.59	70.79	69.86	68.89	72.31	71.92	70.93	71.74	72.12	**72.38**
ResNet50	DeiT-T	80.38	72.17	75.84	72.56	72.06	68.53	75.64	75.10	75.60	75.13	75.73	**75.88**
ConvNeXt-T	DeiT-T	82.05	72.17	70.45	73.12	71.47	69.18	**75.26**	74.00	73.95	74.07	74.41	74.64
Mixer-B/16	DeiT-T	76.62	72.17	74.38	72.82	72.24	68.23	**75.00**	74.16	72.82	74.22	74.46	74.51
ResNet50	Swin-N	80.38	75.53	76.83	76.05	75.90	73.90	77.79	77.58	76.24	77.29	77.76	**78.41**
ConvNeXt-T	Swin-N	82.05	75.53	74.81	75.79	75.48	74.15	77.73	77.15	77.00	77.25	77.50	**78.16**
Mixer-B/16	Swin-N	76.62	75.53	76.17	75.81	75.52	73.38	**76.87**	76.26	75.03	76.54	76.63	76.23
ConvNeXt-T	ResMLP-S12	82.05	76.65	74.69	75.79	75.28	73.57	77.33	76.87	77.23	77.24	77.26	**78.31**
Swin-T	ResMLP-S12	81.38	76.65	76.48	76.15	75.10	73.40	77.42	76.67	76.99	77.25	77.31	**77.67**
Average Improvements				↑1.00	↑0.56	↓0.30	↓1.65	↑2.31	↑1.61	↑1.05	↑1.65	↑2.05	↑**2.42**

Table 3: **Comparison of homogeneous distillation results on ImageNet-1K.** The best and second-best results are emphasized in **bold** and underlined cases.

Teacher	Student	T.	S.	AT	OFD	CRD	ReviewKD	DKD	DIST	FCFD	OFA	FBT	TeKAP	EA-KD	HCD	**MG-EATD**
ResNet34	ResNet18	73.31	69.75	70.69	70.81	71.17	71.61	71.70	72.07	72.24	72.10	72.29	71.35	71.79	72.18	**72.37**
ResNet50	MobileNet	76.61	68.58	69.56	71.25	71.37	72.56	72.05	73.24	73.37	73.28	73.45	72.87	–	73.23	**74.30**

Here, $p_t^{(\ell)}$ and $\hat{p}_t^{(\ell)}$ denote the teacher target and student prediction for region ℓ at time t, respectively, while p_T and p_S are their clean global predictive distributions. The region-level uncertainty factor weights each region using entropies at the current trajectory state, reflecting local uncertainty across spatial locations and granularities. However, these entropies are also affected by the strength of diffusion perturbations. Weighting solely by regional uncertainty may therefore assign insufficient distillation weights to samples assessed as difficult in the clean state. We thus introduce a sample-level difficulty factor that uses the difficulty prior from clean global predictions to adjust the overall distillation strength for each sample.

The current-state prediction loss directly matches the student prediction to the teacher target at the current diffusion time:

$$\mathcal{L}_{\text{state}} = \frac{\tau^2}{M} \sum_{\ell=1}^{M} w_t^{(\ell)} D_{\text{KL}} \left(p_t^{(\ell)} \,\middle\|\, \hat{p}_t^{(\ell)} \right). \tag{22}$$

The trajectory loss further supervises the lower-noise state obtained by applying the local reverse update to the current prediction:

$$\mathcal{L}_{\text{traj}} = \frac{\tau^2}{M} \sum_{\ell=1}^{M} w_{t^-}^{(\ell)} D_{\text{KL}} \left(p_{t^-}^{(\ell)} \,\middle\|\, \tilde{p}_{t^-}^{(\ell)} \right). \tag{23}$$

Together, these losses supervise probability states along the trajectory and local transitions between adjacent states. The overall training objective of *MG-EATD* is

$$\mathcal{L}_{\text{total}} = \lambda_{\text{ce}} \mathcal{L}_{\text{ce}} + \lambda_{\text{kd}} \mathcal{L}_{\text{kd}} + \lambda_{\text{state}} \mathcal{L}_{\text{state}} + \lambda_{\text{traj}} \mathcal{L}_{\text{traj}}, \tag{24}$$

where λ_{ce}, λ_{kd}, λ_{state}, and λ_{traj} control the relative contributions of the corresponding loss terms. Note that during optimization, the trajectory predictor g_θ is updated solely by $\mathcal{L}_{\text{state}}$, while gradients from $\mathcal{L}_{\text{traj}}$ are not propagated to its parameters.

4 Experiments

4.1 Implementation Details

Details of the **models**, **datasets**, **baselines**, and **training protocols** used in our experiments are provided in the Appendix A.6.

4.2 MAIN RESULTS

We conduct systematic experiments across diverse teacher–student pairs. Compared with existing methods, *MG-EATD* achieves superior performance on CIFAR-100 and ImageNet-1K. In heterogeneous distillation, it achieves average Top-1 accuracy gains of 11.50% and 2.42% on CIFAR-100 and ImageNet-1K. It also achieves a gain of 4.17% in homogeneous distillation on ImageNet-1K.

Results on CIFAR-100. Table 1 reports results for 12 heterogeneous teacher–student pairs on CIFAR-100, with *MG-EATD* achieving the best results in most settings. Several methods designed for homogeneous pairs yield limited gains or even degrade student accuracy, highlighting the challenges of transferring knowledge across architectures with different inductive biases. Compared with FBT and OFA, which specifically target heterogeneous distillation, *MG-EATD* improves average accuracy by 3.12% and 5.31%, respectively. The gains are particularly pronounced for Transformer and MLP students. For example, *MG-EATD* achieves accuracies of 85.56% and 86.43% for ConvNeXt-T→DeiT-T and ConvNeXt-T→ResMLP-S12, respectively. These results demonstrate the effectiveness of *MG-EATD* across diverse heterogeneous teacher–student pairs.

Results on ImageNet-1K. Table 2 reports results for 14 heterogeneous teacher–student pairs on ImageNet-1K, with *MG-EATD* also achieving the best results in most settings. Notably, some feature-based methods become more competitive on the larger benchmark, possibly benefiting from increased training data for representation learning. For example, FBT combines feature-level and logit-level supervision, whereas *MG-EATD* performs distillation entirely in the logit space. Furthermore, while logit-based OFA uses features from four student stages, *MG-EATD* constructs multi-granularity knowledge using only penultimate-layer features. With this design,*MG-EATD* also achieves competitive results, demonstrating the effectiveness of probability-trajectory supervision.

Results on homogeneous distillation. To further evaluate the architectural generality and effectiveness of trajectory distillation in *MG-EATD*, we conduct homogeneous distillation experiments on ImageNet-1K. As shown in Table 3, *MG-EATD* achieves the best results in both teacher–student pairs, reaching Top-1 accuracies of 72.37% and 74.30% for ResNet34→ResNet18 and ResNet50→MobileNet, respectively. These results further support the effectiveness of the proposed trajectory distillation strategy for homogeneous knowledge transfer.

4.2.1 FURTHER ANALYSIS

Table 4: **Component ablation of *MG-EATD* on ImageNet-1K.** Results are Top-1 accuracy (%). The first row is baseline KD.

TD	EA	MG	Swin-T →ResNet18	ResNet34 →ResNet18
×	×	×	71.14	70.66
✓	×	×	72.07	72.11
✓	×	✓	72.18	72.23
✓	✓	×	72.25	72.19
✓	✓	✓	72.48	72.37

Ablation Studies. To assess the benefits of trajectory distillation over single endpoint matching in conventional logit-based KD and the contribution of each component, we decompose *MG-EATD* into trajectory distillation (TD), multi-granularity regional knowledge construction (MG), and entropy-driven adaptation (EA). Table 4 presents ablation results on ImageNet-1K using the homogeneous ResNet34→ResNet18 pair and the heterogeneous Swin-T→ResNet18 pair. Compared with baseline KD, TD extends teacher supervision from a single endpoint to a probability trajectory, improving student performance. Adding either MG or EA to TD further improves performance in both settings, while combining all three components yields additional gains, demonstrating their effectiveness and complementarity.

Extended Experiments We present additional experiments in Appendix A.7 to provide a more comprehensive evaluation of *MG-EATD*.

5 CONCLUSION AND FUTURE WORK

We provide a detailed discussion of our **Conclusion and Future Work** in the Appendix A.8.

AI USE STATEMENT

We used generative AI tools to assist with language polishing and improve the readability of the manuscript. The authors reviewed and revised the AI-assisted text and take full responsibility for the final content.

REFERENCES

William J Anderson. *Continuous-time Markov chains: An applications-oriented approach*. Springer Science & Business Media, 2012.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

Omer Belhasin, Shelly Golan, Ran El-Yaniv, and Michael Elad. Advancing image classification with discrete diffusion classification modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 124–134, 2026.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279, 2022.

Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11923–11932. IEEE, 2022.

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.

Xu Cheng, Zhefan Rao, Yilan Chen, and Quanshi Zhang. Explaining knowledge distillation by quantifying the knowledge. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12922–12932. IEEE, 2020.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4793–4801. IEEE, 2019.

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

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. *Advances in Neural Information Processing Systems*, 36:79570–79582, 2023.

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.

Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1921–1930, 2019.

Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Md Imtiaz Hossain, Sharmen Akhter, Choong Seon Hong, and Eui-Nam Huh. Single teacher, multiple perspectives: Teacher knowledge augmentation for enhanced knowledge distillation. In *International Conference on Learning Representations*, volume 2025, pp. 91583–91594, 2025.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *Advances in neural information processing systems*, 35:33716–33727, 2022.

Tao Huang, Yuan Zhang, Mingkai Zheng, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge diffusion for distillation. *Advances in Neural Information Processing Systems*, 36:65299–65316, 2023.

Ying Jin, Jiaqi Wang, and Dahua Lin. Multi-level logit distillation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24276–24285. IEEE, 2023.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Kisoo Kwon, Hwidong Na, Hoshik Lee, and Nam Soo Kim. Adaptive knowledge distillation based on entropy. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7409–7413. IEEE, 2020.

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

Muquan Li, Dongyang Zhang, Tao He, Xiurui Xie, Yuan-Fang Li, and Ke Qin. Towards effective data-free knowledge distillation via diverse diffusion augmentation. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 4416–4425, 2024.

Zheng Li, Yuxuan Li, Penghai Zhao, Renjie Song, Xiang Li, and Jian Yang. Is synthetic data from diffusion models ready for knowledge distillation? *arXiv preprint arXiv:2305.12954*, 2023.

Dongyang Liu, Meina Kan, Shiguang Shan, and Xilin Chen. Function-consistent feature distillation. *arXiv preprint arXiv:2304.11832*, 2023.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2023.

Chenlin Meng, Kristy Choi, Jiaming Song, and Stefano Ermon. Concrete score matching: Generalized score matching for discrete data. *Advances in Neural Information Processing Systems*, 35: 34532–34545, 2022.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5191–5198, 2020.

Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022.

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.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5007–5016, 2019.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew G Wilson. Does knowledge distillation really work? *Advances in neural information processing systems*, 34:6906–6919, 2021.

Chi-Ping Su, Ching-Hsun Tseng, Bin Pu, Lei Zhao, Jiewen Yang, Zhuangzhuang Chen, and Shin-Jye Lee. Ea-kd: Entropy-based adaptive knowledge distillation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 731–740. IEEE, 2025.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):5314–5321, 2022.

Guo-Hua Wang, Yifan Ge, and Jianxin Wu. Distilling knowledge by mimicking features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8183–8195, 2021.

Liuchi Xu, Hao Zheng, Lu Wang, Lisheng Xu, and Jun Cheng. Heterogeneous complementary distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11314–11322, 2026.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 11953–11962, 2022.

Kaixiang Zheng and En-Hui Yang. Knowledge distillation based on transformed teacher matching. In *International Conference on Learning Representations*, volume 2024, pp. 44963–44979, 2024.

Junjie Zhou, Ke Zhu, and Jianxin Wu. All you need in knowledge distillation is a tailored coordinate system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22946–22954, 2025.

A Appendix

A.1 Related Work

A.1.1 Knowledge Distillation

Knowledge distillation (KD) transfers knowledge from a high-capacity teacher to a compact student, aiming to preserve performance while reducing computational and storage costs. Hinton (2015) first proposed KD, transferring knowledge by minimizing the KL divergence between the class-probability distributions derived from teacher and student logits. Subsequent research extends this framework through richer supervision, improved distillation objectives, and other sophisticated strategies.

Distillation between homogeneous architectures. Early research (Hinton (2015); Romero et al. (2014); Zagoruyko & Komodakis (2016); Tian et al. (2019); Park et al. (2019); Peng et al. (2019); Cheng et al. (2020); Kwon et al. (2020); Chen et al. (2021); Wang et al. (2021); Zhao et al. (2022); Chen et al. (2022); Huang et al. (2022); Liu et al. (2023); Jin et al. (2023); Zheng & Yang (2024); Su et al. (2025); Hossain et al. (2025)) primarily considers homogeneous teacher–student architectures, such as CNNs of different depths or widths. Many logit-based methods extend classical KD, primarily by enhancing supervision at the final prediction endpoint (Hossain et al. (2025); Liu et al. (2023); Jin et al. (2023)) or refining distillation objectives to improve the transfer of dark knowledge (Zhao et al. (2022); Su et al. (2025); Huang et al. (2022)). Traditional logit-based distillation relies on final logits for supervision without directly constraining intermediate representations. To exploit the rich knowledge encoded in these representations, Romero et al. (2014) introduce FitNets, which use intermediate teacher features as hints to guide student learning, extending distillation supervision beyond final logits. Subsequent research (Zagoruyko & Komodakis (2016); Tian et al. (2019); Chen et al. (2021)) further develops feature-based distillation to improve the transfer of representational knowledge.

Distillation between heterogeneous architectures. The development of CNNs (He et al. (2016); Howard et al. (2017); Sandler et al. (2018); Liu et al. (2022)), Vision Transformers (Liu et al. (2021); Dosovitskiy et al. (2020); Touvron et al. (2021)), and MLP-based (Tolstikhin et al. (2021); Touvron et al. (2022)) vision models motivates growing interest in heterogeneous KD. These architectures differ substantially in inductive biases, spatial organization, and semantic hierarchies. Consequently, methods developed for homogeneous teacher–student pairs with similar inductive biases and feature hierarchies may limit knowledge transfer or even degrade student performance when directly applied to heterogeneous settings. Hao et al. (2023) identify substantial representational discrepancies through CKA analysis and propose OFA, which maps intermediate student features from multiple stages into logit space for alignment with the teacher's final logits. Subsequent studies further develop general approaches to heterogeneous knowledge distillation. Li et al. (2025) propose FBT, which combines teacher and student modules into a fused model to integrate their inductive biases and processing functions before knowledge transfer, using spatially agnostic contrastive supervision to mitigate spatial mismatches. Zhou et al. (2025) introduce TCS, which tailors a feature coordinate system to transfer dark knowledge from a task-agnostic teacher to task-specific students. Xu et al. (2026) propose HCD, which fuses complementary teacher and student features, projects them into a shared logit space, and applies sub-logit decomposition and orthogonality constraints to promote diverse knowledge transfer.

A.1.2 Diffusion Models

Diffusion models progressively perturb data through a predefined forward process and learn time-conditioned predictions that guide the reverse process for generation. DDPM (Ho et al. (2020)) constructs a diffusion process with Gaussian perturbations in continuous state spaces and connects

its training objective to denoising score matching. D3PM (Austin et al. (2021)) extends this formulation to discrete state spaces, using structured categorical transition matrices for forward corruption and learning the corresponding reverse transitions for image and text generation. SEDD (Lou et al. (2023)) introduces a score entropy objective that extends score matching to discrete spaces by learning marginal probability ratios between states at each diffusion time. These ratios are combined with the forward transition rates to construct the reverse transition rate matrix. DiDiCM (Belhasin et al. (2026)) further applies discrete diffusion to image classification by modeling the posterior distribution over class labels conditioned on the input image. It learns class-probability ratios to construct reverse transition rates and progressively refines predictions through either class-probability evolution or discrete-label sampling, providing an alternative approach to image classification under uncertainty. Together, these studies provide a probabilistic foundation for describing distributional evolution through time-ordered local state transitions, motivating probability-trajectory supervision for knowledge distillation.

Recent studies also incorporate diffusion mechanisms into knowledge distillation. Huang et al. (2023) propose DiffKD, which treats student representations as noisy versions of teacher representations and uses a diffusion model trained on teacher representations to denoise student outputs before matching. DM-KD (Li et al. (2023)) uses diffusion-generated images for distillation without access to real training data, while DDA (Li et al. (2024)) applies diffusion-based augmentation to synthetic samples to increase data diversity.

A.1.3 Summary

These KD approaches that incorporate diffusion models primarily focus on representation denoising or distillation data construction, without jointly supervising teacher class-probability states across noise levels and local transitions between adjacent states. Meanwhile, conventional logit-based distillation methods typically uses a single clean teacher distribution as the supervision endpoint for each input. Note that TeKAP (Hossain et al. (2025)) enriches teacher supervision through random logit perturbations, but we argue that it does not explicitly organize the augmented targets into a time-ordered trajectory or model local transitions between them. *MG-EATD* extends clean endpoint matching to joint state-and-transition learning along discrete diffusion trajectories of class probabilities. To transfer global semantic knowledge and fine-grained regional discriminative cues, it constructs time-ordered teacher probability states across multiple spatial granularities. To exploit the sample-specific difficulty priors available in KD, it uses the entropies of clean global teacher and student predictions to jointly adapt the diffusion-time range and local reverse step size for each sample. Alongside clean endpoint supervision, the student thus learns teacher probability states across spatial granularities and uncertainty levels, together with local reverse transitions toward adjacent lower-noise states.

A.2 Derivation of the Forward Diffusion Process

This section derives the closed-form solution of the forward process used in CPTD. The derivation first establishes the matrix-exponential solution of the forward ordinary differential equation and then simplifies it under the uniform transition rate matrix.

Forward dynamics. Let $p_t \in \mathbb{R}^K$ denote the class-probability distribution at time t, represented as a column vector, where K is the number of classes. The clean distribution p_0 satisfies

$$p_{0,k} \geq 0, \qquad \mathbf{1}^\top p_0 = 1, \tag{25}$$

where $\mathbf{1} \in \mathbb{R}^K$ is the all-ones vector. CPTD adopts the time-independent uniform transition rate matrix

$$Q = \frac{1}{K}\mathbf{1}\mathbf{1}^\top - I, \tag{26}$$

where $I \in \mathbb{R}^{K\times K}$ is the identity matrix. Its entries are

$$Q_{ab} = \begin{cases} \dfrac{1}{K}, & a \neq b, \\ -\dfrac{K-1}{K}, & a = b. \end{cases} \tag{27}$$

Under the column-vector convention, Q_{ab} for $a \neq b$ specifies the base transition rate from class b to class a. The off-diagonal entries are nonnegative, and each column sums to zero, so $\mathbf{1}^\top Q = \mathbf{0}^\top$. Let $\dot{\sigma}(t)$ denote a continuous, nonnegative noise rate on the time interval under consideration. The instantaneous transition rate matrix is $\dot{\sigma}(t)Q$, and the cumulative noise intensity is

$$\sigma(t) = \int_0^t \dot{\sigma}(s)\,\mathrm{d}s. \tag{28}$$

The forward evolution follows

$$\frac{\mathrm{d}p_t}{\mathrm{d}t} = \dot{\sigma}(t)Qp_t, \tag{29}$$

where the initial condition $p_{t=0} = p_0$ specifies that the process starts from the clean class-probability distribution. In ***MG-EATD***, p_0 is the clean teacher class-probability distribution for each region.

Matrix-exponential solution. Consider the candidate solution

$$\widetilde{p}_t = e^{\sigma(t)Q}p_0. \tag{30}$$

Because Q is time-independent, the time dependence of $e^{\sigma(t)Q}$ enters only through the scalar function $\sigma(t)$. Using the power-series representation of the matrix exponential gives

$$e^{\sigma(t)Q} = \sum_{n=0}^{\infty} \frac{\sigma(t)^n}{n!}Q^n. \tag{31}$$

Since Q and p_0 are independent of t, differentiating this series term by term and applying the scalar chain rule yields

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left[e^{\sigma(t)Q}p_0\right] &= \dot{\sigma}(t)\sum_{n=1}^{\infty}\frac{\sigma(t)^{n-1}}{(n-1)!}Q^n p_0 \\
&= \dot{\sigma}(t)Q\left[\sum_{m=0}^{\infty}\frac{\sigma(t)^m}{m!}Q^m\right]p_0 \\
&= \dot{\sigma}(t)Qe^{\sigma(t)Q}p_0,
\end{aligned} \tag{32}$$

where the second equality follows by substituting $m = n - 1$ and factoring out Q. Therefore, the candidate solution $\widetilde{p}_t = e^{\sigma(t)Q}p_0$ satisfies the forward ordinary differential equation

$$\frac{\mathrm{d}\widetilde{p}_t}{\mathrm{d}t} = \dot{\sigma}(t)Q\widetilde{p}_t. \tag{33}$$

To verify the initial condition, note that the cumulative noise intensity defined in equation 28 vanishes at $t = 0$:

$$\sigma(0) = \int_0^0 \dot{\sigma}(s)\,\mathrm{d}s = 0. \tag{34}$$

Evaluating the candidate solution at $t = 0$ therefore yields

$$\widetilde{p}_{t=0} = e^{\sigma(0)Q}p_0 = e^{0Q}p_0 = Ip_0 = p_0. \tag{35}$$

Thus, $e^{\sigma(t)Q}p_0$ satisfies both the forward ordinary differential equation $\mathrm{d}p_t/\mathrm{d}t = \dot{\sigma}(t)Qp_t$ and the prescribed initial condition $p_{t=0} = p_0$. The continuity of the coefficient matrix $\dot{\sigma}(t)Q$ guarantees the uniqueness of the solution to this linear initial-value problem. Consequently, the forward process admits the unique closed-form solution

$$p_t = e^{\sigma(t)Q}p_0. \tag{36}$$

Convex-combination form under uniform transitions. To obtain an explicit expression for $e^{\sigma(t)Q}$, define the averaging matrix P and the uniform class-probability vector u as

$$P = \frac{1}{K}\mathbf{1}\mathbf{1}^\top, \qquad u = \frac{1}{K}\mathbf{1}. \tag{37}$$

where u is the uniform class-probability distribution. Since $\mathbf{1}^\top\mathbf{1} = K$, the matrix P is idempotent:

$$\begin{aligned} P^2 &= \frac{1}{K^2}\big(\mathbf{1}\mathbf{1}^\top\big)\big(\mathbf{1}\mathbf{1}^\top\big) \\ &= \frac{1}{K^2}\,\mathbf{1}\big(\mathbf{1}^\top\mathbf{1}\big)\mathbf{1}^\top \\ &= \frac{1}{K}\,\mathbf{1}\mathbf{1}^\top = P. \end{aligned} \tag{38}$$

It follows that $P^n = P$ for every integer $n \geq 1$. Using the power-series definition of the matrix exponential,

$$\begin{aligned} e^{\sigma(t)P} &= I + \sum_{n=1}^{\infty} \frac{\sigma(t)^n}{n!} P^n \\ &= I + \left(\sum_{n=1}^{\infty} \frac{\sigma(t)^n}{n!}\right) P \\ &= I + \Big(e^{\sigma(t)} - 1\Big)\, P. \end{aligned} \tag{39}$$

Since $Q = P - I$ and $PI = IP$, the exponential of $\sigma(t)(P - I)$ can be factorized into the product of $e^{-\sigma(t)I}$ and $e^{\sigma(t)P}$. Combining the preceding expansion with $e^{-\sigma(t)I} = e^{-\sigma(t)}I$ yields

$$\begin{aligned} e^{\sigma(t)Q} &= e^{\sigma(t)(P-I)} \\ &= e^{-\sigma(t)I} e^{\sigma(t)P} \\ &= e^{-\sigma(t)} \left[I + \Big(e^{\sigma(t)} - 1\Big)\, P\right] \\ &= e^{-\sigma(t)} I + \Big(1 - e^{-\sigma(t)}\Big)\, P. \end{aligned} \tag{40}$$

Since p_0 is a probability distribution, its entries sum to one, i.e., $\mathbf{1}^\top p_0 = 1$. Consequently,

$$Pp_0 = \frac{1}{K}\mathbf{1}\left(\mathbf{1}^\top p_0\right) = \frac{1}{K}\mathbf{1} = u. \tag{41}$$

Substituting this identity and the matrix-exponential expression in equation 40 into equation 36 yields

$$\boxed{p_t = e^{-\sigma(t)} p_0 + \Big(1 - e^{-\sigma(t)}\Big)\, u}. \tag{42}$$

The forward distribution is therefore a convex combination of the clean prediction p_0 and the uniform distribution u, with their relative contributions controlled by the cumulative noise intensity $\sigma(t)$.

A.3 Derivation of the Reverse Diffusion Process

Notation and basic properties. Let $X_t \in \{1, \ldots, K\}$ denote the class state of the forward continuous-time Markov chain, with marginal probabilities $p_t(a) = \Pr(X_t = a)$. As derived in the Appendix A.2, the marginal distribution evolves under the column-vector convention according to

$$\frac{\mathrm{d}p_t}{\mathrm{d}t} = R_t p_t, \qquad R_t = \dot{\sigma}(t)Q. \tag{43}$$

Accordingly, Q_{ab} specifies the base forward transition rate from class b to class a. Likewise, $R_t(a, b)$ denotes the forward rate from b to a, whereas $\overline{R}_t(b, a)$ denotes the reverse rate from a to b.

Since Q has nonnegative off-diagonal entries and zero column sums, and $\dot{\sigma}(t) \geq 0$, the forward transition rate matrix satisfies

$$R_t(a, b) \geq 0 \quad (a \neq b), \qquad \sum_{a=1}^{K} R_t(a, b) = 0. \tag{44}$$

Under the prescribed noise schedule, $\dot{\sigma}(t)$ is continuous and finite on the time interval under consideration. Since Q is fixed, $R_t = \dot{\sigma}(t)Q$ inherits these properties. Moreover, the initial distribution p_0 is obtained by applying softmax to logits, ensuring that $p_0(a) > 0$ for every class a. Under the uniform forward process, the marginal probability of each class therefore satisfies

$$p_t(a) = e^{-\sigma(t)}p_0(a) + \left(1 - e^{-\sigma(t)}\right) u > 0 \tag{45}$$

Reverse transition rates. For a fixed diffusion time $t > 0$, consider a sufficiently small time increment $\Delta t \in (0, t)$. The forward transition probabilities admit the first-order expansion

$$\Pr(X_t = a \mid X_{t-\Delta t} = b) = \delta_{ab} + \Delta t\, \dot{\sigma}(t) Q_{ab} + o(\Delta t), \tag{46}$$

where δ_{ab} denotes the Kronecker delta and the remainder satisfies $o(\Delta t)/\Delta t \to 0$ as $\Delta t \to 0^+$. By continuity of the forward transition rate matrix, replacing $R_{t-\Delta t}$ with R_t changes the first-order term by $o(\Delta t)$, which is absorbed into the remainder.

To relate the reverse transition probabilities to the forward dynamics, Bayes' rule expresses the probability of transitioning from class a at time t to class b at time $t - \Delta t$ as

$$\Pr(X_{t-\Delta t} = b \mid X_t = a) = \frac{p_{t-\Delta t}(b)\ \Pr(X_t = a \mid X_{t-\Delta t} = b)}{p_t(a)}. \tag{47}$$

Since p_t satisfies the forward ordinary differential equation with continuous coefficients, each marginal probability is differentiable with respect to time, and its first-order Taylor expansion about t gives

$$p_{t-\Delta t}(b) = p_t(b) - \Delta t\, \frac{\mathrm{d}p_t(b)}{\mathrm{d}t} + o(\Delta t). \tag{48}$$

Combining this expansion with the forward transition expansion in equation 46, substituting both into equation 47, and collecting the constant and first-order terms yields

$$\begin{aligned} &\Pr(X_{t-\Delta t} = b \mid X_t = a) \\ &\quad = \delta_{ab} + \Delta t \left[\dot{\sigma}(t) Q_{ab} \frac{p_t(b)}{p_t(a)} - \frac{\delta_{ab}}{p_t(a)} \frac{\mathrm{d}p_t(a)}{\mathrm{d}t}\right] + o(\Delta t). \end{aligned} \tag{49}$$

This simplification uses $\delta_{ab} p_t(b)/p_t(a) = \delta_{ab}$ and $\delta_{ab}\, \mathrm{d}p_t(b)/\mathrm{d}t = \delta_{ab}\, \mathrm{d}p_t(a)/\mathrm{d}t$, since the Kronecker delta vanishes for distinct classes and equals one when the two class indices coincide.

The reverse transition rates can now be identified by comparing equation 49 with the defining infinitesimal relation for the reverse transition rate matrix,

$$\Pr(X_{t-\Delta t} = b \mid X_t = a) = \delta_{ab} + \Delta t\, \overline{R}_t(b, a) + o(\Delta t), \tag{50}$$

Equivalently, subtracting the zero-order term, dividing by Δt, and taking the limit isolates each rate as

$$\overline{R}_t(b, a) = \lim_{\Delta t \to 0^+} \frac{\Pr(X_{t-\Delta t} = b \mid X_t = a) - \delta_{ab}}{\Delta t}. \tag{51}$$

For distinct classes a and b, the Kronecker-delta term in equation 49 vanishes, eliminating the contribution from the marginal derivative. Substituting the remaining first-order term into equation 51 therefore gives

$$\overline{R}_t(b, a) = \dot{\sigma}(t) Q_{ab} \frac{p_t(b)}{p_t(a)} = \dot{\sigma}(t) Q_{ab} s_t(a, b), \qquad b \neq a, \tag{52}$$

where $s_t(a, b) = p_t(b)/p_t(a)$ is the concrete score defined in the main text. This recovers the continuous-time Markov chain time-reversal formula under the column-vector convention (Campbell et al. (2022); Lou et al. (2023)). For distinct classes a and b, the reverse rate from a to b equals the forward rate from b to a multiplied by the marginal probability ratio $p_t(b)/p_t(a)$. In our setting, the uniform forward transition rate matrix Q is symmetric, so $Q_{ab} = Q_{ba}$.

For the diagonal case $b = a$, the marginal-derivative term must be retained because $\delta_{aa} = 1$, and the same coefficient comparison yields

$$\overline{R}_t(a, a) = \dot{\sigma}(t) Q_{aa} - \frac{1}{p_t(a)} \frac{\mathrm{d}p_t(a)}{\mathrm{d}t}. \tag{53}$$

To express the time derivative in terms of the forward transition rates, take the a-th component of the forward evolution equation and expand the matrix-vector product:

$$\begin{aligned}\frac{\mathrm{d}p_t(a)}{\mathrm{d}t} &= \left[\frac{\mathrm{d}p_t}{\mathrm{d}t}\right]_a \\ &= \dot{\sigma}(t)\left[Qp_t\right]_a \\ &= \dot{\sigma}(t)\sum_{b=1}^{K} Q_{ab}p_t(b) \\ &= \dot{\sigma}(t)Q_{aa}p_t(a) + \dot{\sigma}(t)\sum_{b\neq a} Q_{ab}p_t(b).\end{aligned} \tag{54}$$

Here, $[\cdot]_a$ denotes the a-th component of a vector, and the final equality separates the diagonal term $b = a$ from the contributions of the remaining classes. Substituting this decomposition into equation 53 cancels the terms involving Q_{aa}, while the remaining sum can be expressed using the off-diagonal reverse rates in equation 52, giving

$$\begin{aligned}\overline{R}_t(a,a) &= -\dot{\sigma}(t)\sum_{b\neq a} Q_{ab}\frac{p_t(b)}{p_t(a)} \\ &= -\sum_{b\neq a}\overline{R}_t(b,a).\end{aligned} \tag{55}$$

Consequently, the time-reversal relation and the forward dynamics determine each diagonal entry as the negative total reverse transition rate out of its source class. The diagonal entries are therefore uniquely specified by the off-diagonal rates, ensuring that every column of $\overline{R}_t$ sums to zero.

A.4 Interpretation of Entropy-Adaptive Trajectory Scheduling

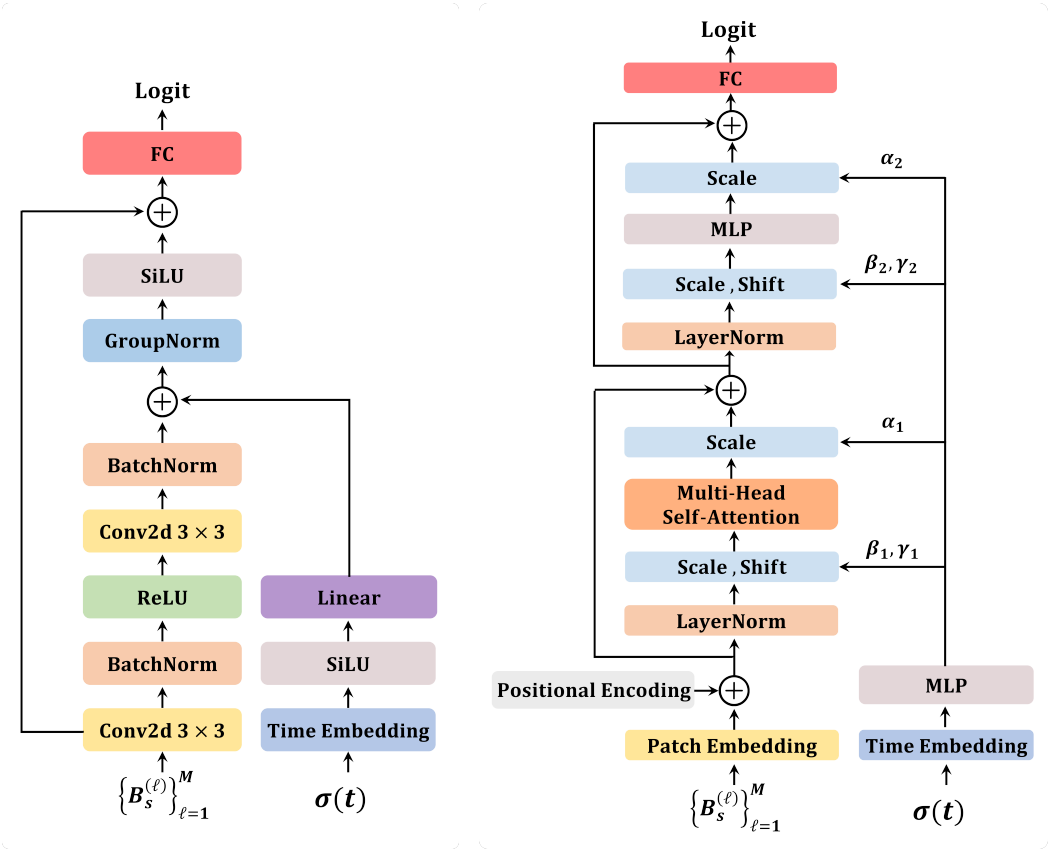
We compute trajectory reliability from the clean global teacher and student predictions as

$$b = 1 - \frac{H(p_T)}{2H_{\max}}\left(1 + \frac{H(p_S)}{H_{\max}}\right). \tag{56}$$

Here, $H(\cdot)$ denotes entropy, and $H_{\max} = \log K$ is its theoretical upper bound for a K-class task. Higher entropy indicates greater predictive uncertainty. This formulation uses teacher uncertainty as the primary signal, with student uncertainty modulating its contribution to trajectory reliability. When either entropy increases while the other remains fixed, the trajectory reliability b decreases or remains unchanged. The following limiting cases illustrate this behavior:

$$b \longrightarrow \begin{cases} 0, & H(p_T) \to H_{\max}, \quad H(p_S) \to H_{\max}, \\ \dfrac{1}{2}, & H(p_T) \to H_{\max}, \quad H(p_S) \to 0, \\ 1, & H(p_T) \to 0, \quad H(p_S) \to 0. \end{cases} \tag{57}$$

When both predictions are highly uncertain, b approaches zero. Under the linear scheduling mappings, the maximum diffusion time and reverse step size consequently approach $t_{\max}^{\text{weak}}$ and $\Delta t_{\min}$, respectively, yielding conservative trajectory supervision. When the teacher remains highly uncertain but the student prediction becomes confident, b approaches $1/2$, placing both scheduling parameters near the midpoints of their prescribed ranges. This expands the range of supervised probability states compared with the jointly uncertain case, while maintaining a more conservative schedule to account for teacher uncertainty. When both teacher and student predictions become nearly deterministic, b approaches one, permitting a broader perturbation range and larger reverse steps. The mechanism incorporates uncertainty from both models, using teacher entropy as the primary signal and student entropy to modulate the trajectory range and local reverse step size.



(a) CNN-based time-conditioned predictor **(b)** Transformer-based time-conditioned predictor

Figure 3: **Architectures of the time-conditioned trajectory predictors in *MG-EATD*.** (a) A residual convolutional predictor for CNN teachers. (b) A Transformer predictor for Transformer or MLP teachers. Both predictors take student regional features and the cumulative noise intensity $\sigma(t)$ as inputs and output class logits at the corresponding diffusion time.

A.5 Architecture of the Time-Conditioned Trajectory Predictor

FCFD (Liu et al. (2023)) emphasizes the functional consistency of teacher and student features under subsequent network processing, while FBT (Li et al. (2025)) combines their inductive biases and feature-processing functions through architectural fusion. Inspired by these studies, we design time-conditioned trajectory predictors according to the teacher's architectural characteristics, allowing student regional features to be transformed in a manner better suited to the teacher's representations. Specifically, we use a residual convolutional predictor for CNN teachers and a Transformer predictor for Transformer or MLP teachers. Both predictors take student regional features and the cumulative noise intensity as inputs and output class logits at the corresponding diffusion time.

CNN-based time-conditioned predictor. As illustrated in Figure 3a, this predictor uses two 3×3 convolutional layers to extract local spatial information, together with BatchNorm and ReLU for feature transformation. The temporal branch encodes the cumulative noise intensity $\sigma(t)$ into a time embedding, which is processed by SiLU and a linear projection to match the feature channels. The resulting conditioning vector is added to the normalized features. The fused features are processed by GroupNorm and SiLU, combined with the residual features, and mapped to class logits through a FC module. This architecture incorporates the local modeling properties of convolution to accommodate CNN teacher representations, while time conditioning supports probability-state prediction at different noise levels.

Table 5: **Training hyperparameters on CIFAR-100 and ImageNet-1K.** Our training configurations follow OFA (Hao et al. (2023)). Values in parentheses indicate the corresponding settings in FBT (Li et al. (2025)) where they differ from ours.

	ImageNet-1K		CIFAR-100	
	CNN	Transformer/MLP	CNN	Transformer/MLP
Epochs	100	300	300	300
Image resolution	224×224	224×224	224×224	224×224
Batch size	64 (512)	128 (1024)	128 (1024)	128 (512)
Initial LR	0.1	5×10^{-4}	0.05 (0.1)	5×10^{-4}
Minimum LR	10^{-6}	10^{-6}	10^{-3}	10^{-5}
Optimizer	SGD	AdamW	SGD	AdamW
Weight decay	10^{-4}	5×10^{-2}	2×10^{-3}	5×10^{-2}
LR schedule	$\times 0.1$ at [30, 60, 90]	Cosine	Cosine	Cosine
Warmup epochs	3	20	3	20
EMA decay	–	0.99996	–	–
RandAugment	–	9/0.5	–	9/0.5
Mixup	–	0.8	–	0.8
CutMix	–	1.0	–	1.0
Random erasing prob.	–	0.25	–	0.25

Transformer-based time-conditioned predictor. As shown in Figure 3b, student regional features are first converted into token representations through patch embedding, and positional encodings are then added to incorporate spatial information. The tokens are then processed by two residual branches comprising multi-head self-attention and a feed-forward MLP, respectively. Inspired by the conditioning mechanism in DiT (Peebles & Xie (2023)), we use an MLP to map the time embedding to branch-specific modulation parameters. Specifically, γ_i and β_i scale and shift the features after LayerNorm, respectively, while α_i scales the branch output before residual addition. Here, $i \in \{1, 2\}$ denotes the attention and feed-forward branches, respectively. The resulting representations are mapped to class logits through a a FC module. This design uses self-attention to model relationships among spatial locations within each region, while temporal modulation adapts the feature transformations to the current diffusion time.

Comparison of time-conditioned predictors with different architectures. To evaluate how predictor architecture affects heterogeneous knowledge transfer, we compare the residual convolutional and Transformer predictors across different heterogeneous teacher–student pairs in Appendix A.7.3.

A.6 Implementation Details

A.6.1 Models

Following the teacher–student configurations in OFA (Hao et al. (2023)), we evaluate both heterogeneous and homogeneous distillation across three architectural families: CNNs, Transformers, and MLPs. Specifically, the CNN models include ResNet (He et al. (2016)), MobileNet (Howard et al. (2017)), MobileNetV2 (Sandler et al. (2018)), and ConvNeXt (Liu et al. (2022)). The Transformer models include ViT (Dosovitskiy et al. (2020)), DeiT (Touvron et al. (2021)), and Swin (Liu et al. (2021)). The MLP models include MLP-Mixer (Tolstikhin et al. (2021)) and ResMLP (Touvron et al. (2022)).

A.6.2 Datasets

We conduct experiments on CIFAR-100 and ImageNet-1K. CIFAR-100 (Krizhevsky et al. (2009)) comprises 50,000 training and 10,000 test images across 100 classes, with an image resolution of 32×32 pixels. Compared with CIFAR-100, ImageNet-1K (Russakovsky et al. (2015)) provides substantially more training images and a broader range of categories, enabling evaluation in a larger-scale classification setting. Specifically, ImageNet-1K (Russakovsky et al. (2015)) contains approximately 1.28 million training images and 50,000 validation images across 1,000 classes.

Table 6: **Effect of spatial granularity selection on ImageNet-1K.** Results are Top-1 accuracy (%) on ImageNet-1K for two heterogeneous teacher–student pairs. $\mathcal{G}$ specifies the spatial partition grids, where each $g \in \mathcal{G}$ corresponds to a $g \times g$ partition of the feature map.

$\mathcal{G}$	Swin →ResNet18	DeiT →MobileNetV2
$\{1\}$	72.25	71.61
$\{1, 2\}$	72.31	71.67
$\{1, 2, 4\}$	72.48	71.81

Table 7: **Ablation of loss terms and local reverse-transition learning.** Results are Top-1 accuracy (%) on ImageNet-1K using Swin-T→ResNet18. The clean-state CE and KD losses are retained in all configurations.

Configuration	$\mathcal{L}_{\text{state}}$	$\mathcal{L}_{\text{traj}}$	Top-1 (%)
Baseline KD	×	×	71.14
State only	✓	×	72.06
Trajectory only	×	✓	71.88
Dual-time prediction	✓	✓	72.16
MG-EATD	✓	✓	**72.48**

A.6.3 Baselines

Following FBT (Li et al. (2025)), we compare our method with representative knowledge distillation approaches with publicly available implementations. The logit-based baselines include KD (Hinton (2015)), DKD (Zhao et al. (2022)), DIST (Huang et al. (2022)), OFA (Hao et al. (2023)), TeKAP (Hossain et al. (2025)), HCD (Xu et al. (2026)) and EA-KD (Su et al. (2025)). The feature-based baselines include FitNet (Romero et al. (2014)), ReviewKD (Chen et al. (2021)), CC (Peng et al. (2019)), RKD (Park et al. (2019)), CRD (Tian et al. (2019)), FCFD (Liu et al. (2023)), AT (Zagoruyko & Komodakis (2016)), OFD (Heo et al. (2019)), and FBT (Li et al. (2025)). The baseline results are adopted from those reported in FBT.

A.6.4 Training Protocols

Our method and the compared baselines generally follow the training protocols of OFA (Hao et al. (2023)), while FBT (Li et al. (2025)) differs primarily in batch size, using batches 4–8 times as large as those of the other methods. On CIFAR-100, all student models are trained for 300 epochs, using SGD for CNNs and AdamW for Transformer and MLP architectures. On ImageNet-1K, CNN students are trained for 100 epochs using SGD, while Transformer and MLP students are trained for 300 epochs using AdamW. Experiments on CIFAR-100 use a single NVIDIA RTX 3090 GPU, whereas those on ImageNet-1K use four NVIDIA RTX 3090 GPUs. We repeat each experiment three times and report the average results. Detailed configurations are provided in Table 5.

A.7 Extended Experiments

A.7.1 Comparison of Different Spatial Granularity Settings

We evaluate the effect of spatial granularity on ImageNet-1K using two heterogeneous teacher–student pairs: Swin→ResNet18 and DeiT→MobileNetV2. We compare three granularity configurations, $\mathcal{G} = \{1\}$, $\{1, 2\}$, and $\{1, 2, 4\}$, which yield 1, 5, and 21 regions from each feature map, respectively. As shown in Table 6, incorporating progressively finer regions improves accuracy for both students. Compared with global-only supervision using $\{1\}$, the $\{1, 2, 4\}$ configuration increases accuracy from 72.25% to 72.48% and from 71.61% to 71.81%, yielding gains of 0.23% and 0.20%, respectively. These results support the effectiveness of complementing global supervision with local regional knowledge at multiple granularities.

Table 8: **Comparison of time-conditioned predictor architectures.** Results are Top-1 accuracy (%) on ImageNet-1K for two heterogeneous teacher–student pairs. CNN-based and Transformer-based denote the two predictor architectures.

Predictor	DeiT-T →MobileNetV2	ResNet50 →Swin-N
CNN-based	71.64	78.41
Transformer-based	71.81	78.07

A.7.2 Ablation of Local Reverse-Transition Learning and Loss Components

We evaluate the two loss terms and local reverse-transition learning on ImageNet-1K using Swin-T→ResNet18, retaining the $\mathcal{L}_{\mathrm{ce}}$ and $\mathcal{L}_{\mathrm{kd}}$ in all configurations. As shown in Table 7, using only $\mathcal{L}_{\mathrm{state}}$ or $\mathcal{L}_{\mathrm{traj}}$ yields accuracies of 72.06% and 71.88%, respectively, both exceeding the baseline KD accuracy of 71.14%. Note that in the trajectory-only configuration, gradients from $\mathcal{L}_{\mathrm{traj}}$ are allowed to update the time-conditioned predictor g_θ. Combining both losses yields the best performance, highlighting the benefits of jointly learning probability states at different uncertainty levels and local reverse transitions between adjacent states for effective knowledge transfer.

In *MG-EATD*, the adjacent lower-noise student state is obtained by applying a local reverse update to the current probability estimate. To further evaluate the role of local reverse-transition learning, we introduce a dual-time direct-prediction variant that replaces the local reverse update with direct predictions from the same time-conditioned predictor g_θ at both t and t^-, each supervised by its corresponding teacher target. In this variant, $\mathcal{L}_{\mathrm{state}}$ and $\mathcal{L}_{\mathrm{traj}}$ supervise the predictions at t and t^-, respectively, and both losses contribute gradients to update the parameters of g_θ. Full *MG-EATD* achieves 72.48% accuracy, outperforming the dual-time prediction variant at 72.16%. These results highlight the benefits of explicitly linking adjacent probability states through local reverse updates for effective knowledge transfer.

A.7.3 Comparison of Time-Conditioned Predictor Architectures

To evaluate how predictor architecture affects heterogeneous knowledge transfer, we compare CNN-based and Transformer-based time-conditioned predictors on ImageNet-1K using DeiT-T→MobileNetV2 and ResNet50→Swin-N. As shown in Table 8, the Transformer-based predictor achieves 71.81% accuracy with the DeiT-T teacher, outperforming the CNN-based predictor at 71.64%. Conversely, with the ResNet50 teacher, the CNN-based predictor achieves 78.41%, exceeding the 78.07% obtained by the Transformer-based predictor. Across both pairs, the predictor corresponding to the teacher's architectural family performs better, despite receiving features from a student with a different architecture. These results support designing the time-conditioned predictor according to the teacher's architectural characteristics to better accommodate its representations when transforming student regional features.

A.8 Conclusion and Future Work

In this work, we present *MG-EATD*, which extends single-endpoint supervision to jointly learning probability states at different uncertainty levels and local transitions between adjacent states along class-probability trajectories. First, *MG-EATD* partitions features of the teacher and student into regions at multiple spatial granularities and maps regional knowledge into a shared class-probability space, enabling distillation to exploit both global semantics and local discriminative information. Second, during trajectory distillation across granularities, an entropy-driven adaptation mechanism uses the entropies of clean global teacher and student predictions as difficulty priors to jointly adjust each sample's diffusion-time range and reverse step size. We then design a time-conditioned predictor aligned with the teacher's architectural characteristics to better accommodate its representational biases. The predictor estimates the student's class-probability state at the current diffusion time, followed by a local reverse update that yields the adjacent lower-noise state for joint state-and-transition learning. Experiments on CIFAR-100 and ImageNet-1K demonstrate the effectiveness of *MG-EATD* across heterogeneous and homogeneous teacher–student pairs, with ablation studies further supporting the contributions of its components.

Our work provides an initial exploration of integrating discrete diffusion with logit-based knowledge distillation. Unlike prior work that confines diffusion models to feature denoising or data augmentation, we directly exploit diffusion dynamics to reformulate the conventional single-endpoint matching in knowledge distillation as state-and-trajectory learning. Future work may build on this formulation to explore more effective strategies for integrating diffusion mechanisms with knowledge distillation. More flexible noise schedules and refined selection of diffusion times and reverse step sizes could better accommodate the knowledge transfer requirements of different teacher–student pairs. The design of the time-conditioned predictor offers another direction for improvement. While the current design is guided primarily by the teacher's architectural characteristics, jointly considering the properties of teacher representations and the spatial organization of student features could facilitate more effective knowledge transfer. These extensions may further improve the effectiveness and training efficiency of trajectory distillation across diverse teacher–student architectures.