

UniDFKD: A Unified Semantic Prior Framework for Architecture-Agnostic Data-Free Knowledge Distillation

Xuewan He¹, Tong Chu¹, Zihan Cheng¹, Yuchen Su¹, Qianxin Xia¹,
Guoming Lu¹, Jielei Wang^{1*}, Wen Li^{1*}

¹University of Electronic Science and Technology of China, Chengdu, China
hershel@std.uestc.edu.cn, jlwang@uestc.edu.cn, liwenbnu@gmail.com

Abstract

Data-Free Knowledge Distillation (DFKD) transfers knowledge from a pretrained teacher model to a compact student model by synthesizing semantically informative data, eliminating the need for access to the original training dataset. Existing DFKD methods rely heavily on architecture-specific statistical priors (e.g., Batch Normalization statistics) to guide data synthesis, however, such architecture-dependent priors are often absent in modern architectures such as Vision Transformers (ViTs), resulting in degraded semantic quality of the synthesized data and consequently catastrophic performance degradation. In this paper, we propose *UniDFKD*, a unified data-free knowledge distillation framework that replaces architecture-specific statistics with explicit, architecture-agnostic semantic priors. *UniDFKD* governs the entire synthesis-distillation pipeline along three dimensions: (1) Categorical Semantic Conditioning (CSC) defines *what* to synthesize by persistently modulating the generator with language-derived embeddings to capture semantic diversity; (2) Spatial Semantic Anchoring (SSA) dictates *where* evidence belongs by anchoring the teacher’s spatial attributions to a Gaussian prior; and (3) Spatial Semantic Distillation (SSD) controls *how* knowledge is transferred by explicitly aligning teacher-student spatial evidence alongside predictions. Extensive experiments across CNNs and ViTs demonstrate that UniDFKD establishes a new state-of-the-art, outperforming existing methods by an average absolute margin of over 20% in both homogeneous and heterogeneous settings.

Introduction

Knowledge distillation is a cornerstone technique for deploying highly performant, large-scale vision models onto resource-constrained edge devices (Hinton, Vinyals, and Dean 2015; Mansourian et al. 2025). In many real-world applications, however, the original training data used to optimize the teacher model is rendered inaccessible due to stringent privacy regulations, proprietary restrictions, or formidable transmission costs (Chen et al. 2019). Data-Free Knowledge Distillation (DFKD) mitigates this bottleneck by synthesizing surrogate training samples directly from the pretrained teacher, effectively hallucinating the missing data distribution (Liu et al. 2021a, 2024; Chen et al. 2025).

Because data synthesis in a high-dimensional space is severely underdetermined (Yin et al. 2020), existing DFKD

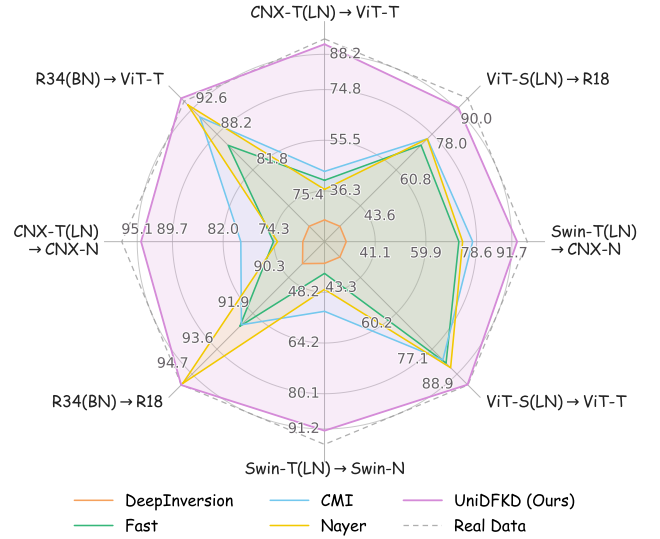


Figure 1: Comparison of student accuracy (%) achieved by different DFKD methods on CIFAR-10 across eight teacher-student distillation settings, including both homogeneous and heterogeneous architectures. BN/LN denotes the normalization type used in the teacher model. The dashed polygon denotes students trained on real data, and numbers indicate axis scales. While existing methods degrade sharply on LN-based teachers, UniDFKD consistently approaches the real-data performance.

methodologies rely on injecting a strong data prior to constrain the generator. The *de facto* standard for this prior has been the running statistics stored within the teacher’s Batch Normalization (BN) layers (Ioffe and Szegedy 2015). By aligning the feature statistics of synthetic batches with these stored real-data statistics, inversion on BN-based CNNs can remarkably approach the accuracy of real-data distillation (cf. R34→R18 and R34→ViT-T settings in Fig. 1). Unfortunately, this BN prior is a mere architectural accident. As the field rapidly pivots toward BN-free architectures such as Vision Transformer (ViT) and ConvNeXt, which rely on Layer Normalization (LN) (Huang et al. 2023), this critical prior simply vanishes. Consequently, state-of-the-art DFKD methods fail to generalize, exhibiting severe degradation when applied to BN-free teachers or heterogeneous architectures

*Corresponding author.

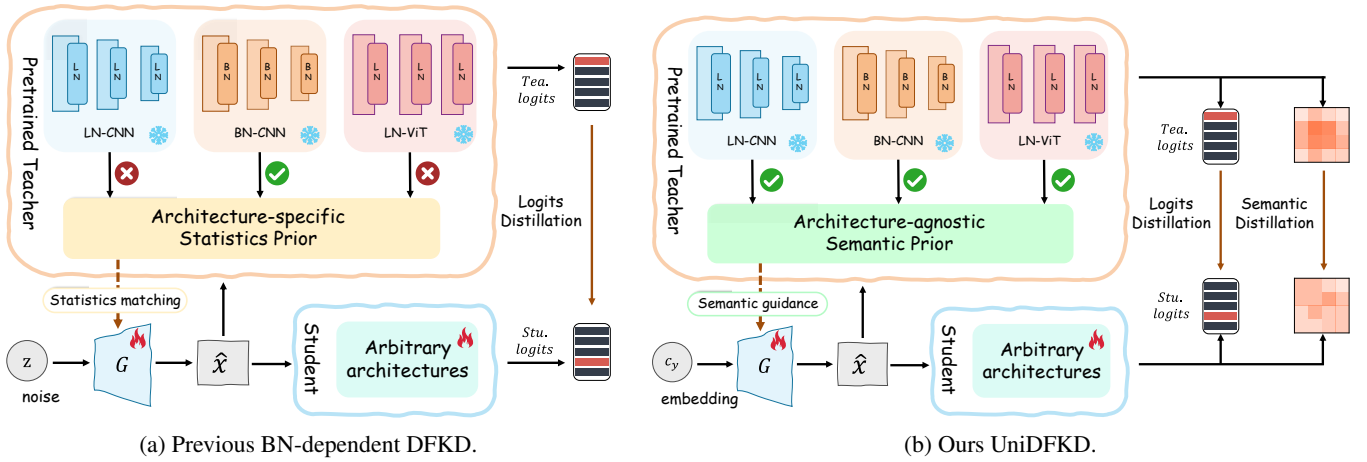


Figure 2: Comparison of DFKD paradigms. (a) Existing methods adopt BN statistics as an implicit semantic prior, confining synthesis to BN-based teachers. (b) UniDFKD replaces this implicit prior with an explicit, architecture-agnostic semantic prior applicable to arbitrary teachers and augments distillation with spatial semantic transfer.

(see the settings with LN-based teachers in Fig. 1).

To formulate an alternative, architecture-agnostic prior, we must first answer a fundamental question (Fig. 3): *What makes BN statistics so effective for DFKD?* We conduct a layer-wise diagnosis across both ResNet and ViT teachers and observe a striking trend: enforcing statistical constraints on deep layers nearly recovers the performance of full-network matching, whereas shallow-layer constraints offer negligible benefits over unconstrained synthesis. This reveals that the true utility of the BN prior does not stem from low-level texture statistics, but rather from its implicit alignment of deep, class-discriminative semantics.

This insight necessitates a paradigm shift in DFKD: from matching architecture-specific feature statistics to enforcing explicit, architecture-agnostic semantic priors (Fig. 2). We propose that an effective semantic prior must systematically address three critical dimensions: *what* semantic content to generate, *where* to spatially localize it, and *how* to transfer it to the student. Currently, no existing DFKD framework successfully unifies these perspectives into a cohesive pipeline.

To this end, we introduce UniDFKD, a unified synthesis and distillation framework driven entirely by explicit semantic priors. During synthesis, UniDFKD answers the *what* via Categorical Semantic Conditioning (CSC). CSC leverages large language models (LLMs) to construct a diverse semantic bank, persistently modulating every block of the generator to ensure synthesized images maintain strict inter-class relations and rich intra-class diversity. To address the *where*, Spatial Semantic Anchoring (SSA) projects the teacher’s deepest responses into a unified Class-Discriminative Spatial Attribution (CDSA) map, aligning it with a center-biased structural Gaussian prior to suppress scattered background artifacts. Finally, to address the *how*, Spatial Semantic Distillation (SSD) transcends traditional logit matching by aligning the per-sample CDSA maps between the teacher and student, guaranteeing that the student learns the exact spatial reasoning behind the teacher’s predictions.

Our contributions are three-fold:

- We reveal that the effectiveness of the BN prior stems primarily from deep semantic representations, fundamentally reframing the challenge of DFKD as the construction of architecture-agnostic semantic priors.
- We propose UniDFKD, which instantiates such priors via CSC for depth-persistent categorical semantics, SSA for class-discriminative spatial semantics, and SSD for transferring the semantics from teacher to student.
- Comprehensive experiments across CNNs and ViTs under homogeneous and heterogeneous settings demonstrate predominantly state-of-the-art performance.

Related Work

Data-Free Knowledge Distillation. Data-free knowledge distillation (DFKD) aims to recover a surrogate training distribution from a pretrained teacher alone (Lopes, Fenu, and Starner 2017; Liu et al. 2021a). Early methods primarily extracted supervision from the teacher’s output space via confidence, class-similarity, or adversarial belief matching (Chen et al. 2019; Nayak et al. 2019; Micaelli and Storkey 2019). However, these output-level signals provide limited guidance for intermediate features, often yielding semantically incomplete samples. A major turning point was DeepInversion (Yin et al. 2020), which established Batch Normalization (BN) running statistics as the de facto data prior. Subsequent studies have improved synthesis diversity, efficiency, scalability, and robustness atop this common anchor (Fang et al. 2021; Yu et al. 2023; Fang et al. 2022; Liu et al. 2024; Tran et al. 2024a; Binici et al. 2022). However, using BN statistics as the primary distributional prior strongly couples data synthesis to the teacher’s architectural design. As a result, such approaches fail to generalize to BN-free teachers (e.g., Vision Transformers and LayerNorm-based models), where this critical prior is absent.

Semantic Priors in DFKD. Beyond the BN prior, early alternatives imposed pixel-space regularizations like total variation and ℓ_2 smoothness (Mahendran and Vedaldi 2015),

which are too weak to independently recover task-relevant distributions. Recent methods introduce semantic knowledge but typically isolate a single facet. To specify *what* to synthesize, NAYER (Tran et al. 2024b) conditions the generator on text embeddings. However, its input-level conditioning leaves categorical fidelity underconstrained in deeper layers and overlooks intra-class diversity. To specify *where* evidence appears, other works extract spatial cues via activation amplitudes (Tran et al. 2024a) or self-attention (Zhong et al. 2025). Yet, these cues rely on architecture-specific operators and merely supplement, rather than replace, the BN prior. UniDFKD bridges these gaps by unifying categorical identity, intra-class diversity, and spatial organization. By persistently injecting diverse categorical semantics throughout the generator and aligning class-discriminative spatial responses via an architecture-agnostic interface, UniDFKD establishes a holistic semantic prior. This eliminates the reliance on BN statistics, enabling effective DFKD and spatial evidence transfer across heterogeneous architectures.

Preliminary Analysis

Reviewing BN Statistics in DFKD

Given a pretrained teacher $\mathcal{T}$ but no access to its training data $\mathcal{D}$, DFKD trains a compact student $\mathcal{S}$ using surrogate images $\hat{x}$ synthesized by a generator $\mathcal{G}$ initialized from scratch. To regularize the synthetic distribution, existing methods align the intermediate feature statistics of $\hat{x}$ with the running statistics stored in the teacher’s BN layers (Yin et al. 2020):

$$\mathcal{L}_{\text{BN}} = \sum_l^L (\|\mu_l(\hat{x}) - \mu_l^{\text{BN}}\|_2 + \|\sigma_l^2(\hat{x}) - (\sigma_l^{\text{BN}})^2\|_2). \quad (1)$$

As Fig. 1 demonstrates, existing DFKD methods degrade substantially under BN-free or heterogeneous settings. This bottleneck directly motivates the design of an architecture-agnostic prior for BN-free models.

What makes BN statistics so effective for DFKD?

To determine what such a prior should preserve, we revisit the information encoded by the BN prior. Existing BN objectives typically aggregate feature statistics from all layers with uniform weighting, implicitly treating constraints at different depths as equally informative. However, teacher representations evolve hierarchically: shallow layers mainly capture local texture and appearance distributions, whereas deeper layers encode more stable and discriminative semantic structures (Gatys, Ecker, and Bethge 2016). The utility of their associated statistical constraints should therefore vary with representation depth. We verify this by isolating each layer in turn, using only its statistics to guide synthesis and measuring the resulting distillation accuracy as that layer’s standalone contribution. Experiments are conducted on both ResNet and ViT teachers, where ViT uses equivalent channel-wise activation moments as the target, serving as an oracle (details are in the supplementary material).

As shown in Fig. 3, both ResNet and ViT teachers reveal that the contribution of individual layers to synthesis quality varies dramatically with depth. Performance rises progressively with the depth of the matched layer, and matching

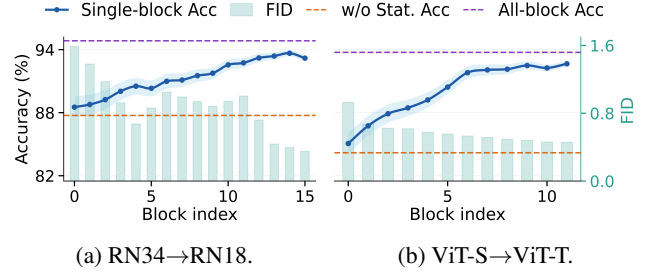


Figure 3: Layer-wise feature-statistic matching on CIFAR-10. Student acc. ($\uparrow$) and FID ($\downarrow$) are reported for analysis.

the deepest layer alone already approaches the performance of full-layer matching. This indicates that *the information governing distillation effectiveness resides predominantly in the teacher’s deep semantic representations*. Consequently, the inherent absence of running statistics in BN-free architectures dictates the paradigm shift: from feature statistic matching to an architecture-agnostic explicit semantic prior.

Methodology

Overall Framework

Building on the above observation, we decompose the construction of such a prior into three perspectives: *what* the semantics span—the class identity, inter-class relations, and intra-class diversity to be encoded in the synthesized distribution; *where* in the image plane they localize—the spatial locations where class-discriminative evidence should concentrate rather than scattering into backgrounds and textures; and *how* they transfer to the student—the explicit distillation of the semantics instilled in the synthetic images beyond logit matching alone. The first two establish the structural and spatial priors for data synthesis, while the third seamlessly bridges this synthesized data with knowledge transfer. Together, they form a closed loop that advances DFKD into a unified, semantics-driven paradigm.

As illustrated in Fig. 4, UniDFKD operationalizes these within a unified two-stage synthesis–distillation pipeline. During synthesis, Categorical Semantic Conditioning (CSC) addresses *what* by conditioning generator $\mathcal{G}$ on language-derived embeddings to preserve inter-class relations and intra-class diversity. Meanwhile, Spatial Semantic Anchoring (SSA) addresses *where* by aligning the class-discriminative attribution with a Gaussian prior, ensuring that discriminative evidence localizes in compact, center-biased regions rather than scattering across backgrounds.

Given the synthesized class-conditional images $\hat{x}$, the generator is updated by minimizing a combined objective comprising the cross-entropy loss $\mathcal{L}_{\text{CE}}$, which optimizes $\hat{x}$ toward the designated class label, and the SSA loss $\mathcal{L}_{\text{SSA}}$:

$$\mathcal{L}_{\mathcal{G}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{SSA}} \mathcal{L}_{\text{SSA}}. \quad (2)$$

During distillation, we address *how* the student learns through a joint knowledge transfer objective. This incorporates the KL divergence $\mathcal{L}_{\text{KL}}$ between the teacher and student predictions, and the Spatial Semantic Distillation (SSD) loss

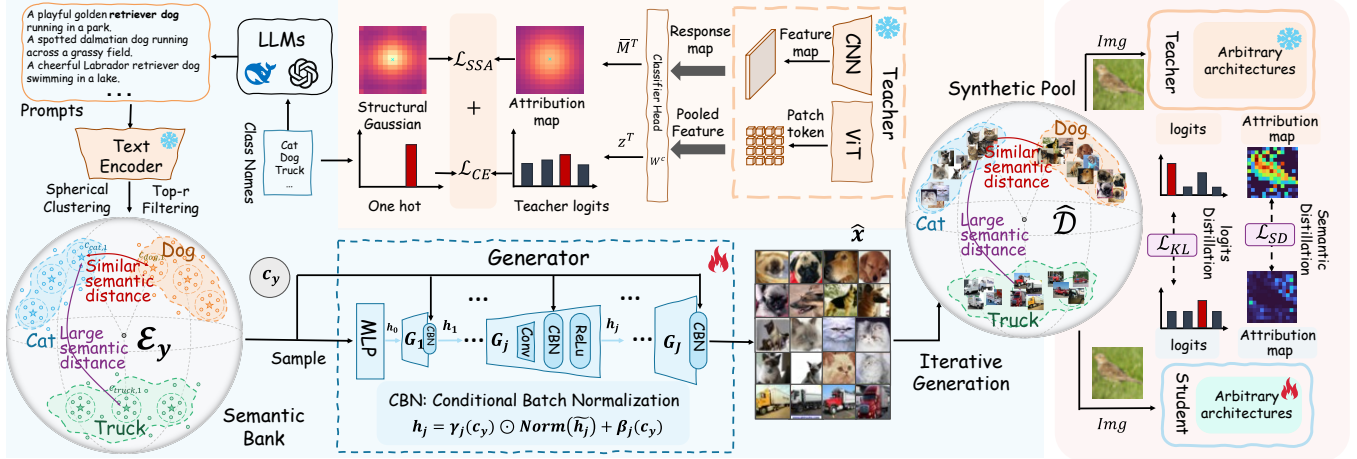


Figure 4: Overview of UniDFKD. **Synthesis (left)**. CSC builds a semantic bank $\mathcal{E}_y$ from LLM-generated class descriptions and modulates every generator block with a sampled embedding $\mathbf{c}_y$ via CBN, while SSA aligns the teacher’s batch-averaged attribution $\bar{\mathbf{M}}^T$ with a Gaussian prior; iterative generation accumulates the surrogate pool $\hat{\mathcal{D}}$. **Distillation (right)**. SSD complements logit matching with attribution alignment, so the student also inherits the teacher’s spatial evidence.

$\mathcal{L}_{SSD}$, which supplements prediction matching by transferring the underlying spatial evidence:

$$\mathcal{L}_{KD} = \mathcal{L}_{KL} + \lambda_{SSD} \mathcal{L}_{SSD}. \quad (3)$$

Categorical Semantic Conditioning

CSC leverages natural-language descriptions in a shared vision-language space as a surrogate for real image geometry, capturing both inter-class relatedness and intra-class diversity (Radford et al. 2021; Pratt et al. 2023). We implement this via a two-stage semantic conditioning pipeline that first refines captions into a compact per-class semantic bank, and then conditions the generator on the sampled embeddings.

Diverse Embedding. For each class y , we first leverage large language models (LLMs) to generate a set of candidate prompts that describe complementary semantic modes such as attributes, viewpoints, and contexts (Pratt et al. 2023). A frozen text encoder (Radford et al. 2021) then maps these prompts into text embeddings. Prompts derived without strict constraints inevitably contain noisy descriptions that act as out-of-distribution conditions for class y . Therefore, we apply spherical k -means (Dhillon and Modha 2001) to partition these embeddings into K clusters $\{\mathcal{C}_{y,k}\}_{k=1}^K$, retaining only the r nearest members to each normalized centroid $\mathbf{v}_{y,k}$ to yield the diverse semantic bank $\mathcal{E}_y$.

$$\mathcal{E}_y = \{\mathbf{e} \in \mathcal{C}_{y,k} \mid 1 \leq k \leq K, \text{rank}_{\mathcal{C}_{y,k}}(\mathbf{e}^\top \mathbf{v}_{y,k}) \leq r\}, \quad (4)$$

where $\text{rank}_{\mathcal{C}_{y,k}}(\cdot)$ denotes the descending rank of the similarity score within $\mathcal{C}_{y,k}$. $\mathcal{E}_y$ keeps $K \times r$ strictly in-distribution embeddings that still comprehensively cover the semantic structures and intra-class diversity of class y . During synthesis, each sample of class y draws its semantic embedding $\mathbf{c}_y$ uniformly from $\mathcal{E}_y$. This sampled embedding is then projected to construct the initial feature $\mathbf{h}_0 = \text{MLP}(\mathbf{c}_y)$, which serves as the foundational input for the generator.

Embedding Injection. Although input-level conditioning establishes the desired semantic mode in $\mathbf{h}_0$, it does not explicitly preserve this information throughout the synthesis hierarchy. In the deeper generators required for 224-scale synthesis, the influence of $\mathbf{c}_y$ progressively weakens across successive upsampling blocks, leaving intermediate features insufficiently constrained by the sampled categorical semantics. To sustain these semantics along the entire synthesis hierarchy, we inject $\mathbf{c}_y$ into every generator block through Conditional Batch Normalization (CBN) (de Vries et al. 2017). Formally, the generator stacks J upsampling blocks, and within each block the static affine of its normalization layer is replaced by linear projections of the sampled embedding:

$$\begin{aligned} \hat{\mathbf{x}} &= (G_J \circ \dots \circ G_1)(\mathbf{h}_0), \\ \mathbf{h}_j &= \gamma_j(\mathbf{c}_y) \odot \text{Norm}(\tilde{\mathbf{h}}_j) + \beta_j(\mathbf{c}_y), \end{aligned} \quad (5)$$

where $\mathbf{h}_j$ is the intermediate feature output by block G_j , $\tilde{\mathbf{h}}_j$ is its pre-normalization activation, and γ_j, β_j are block-specific linear projections broadcast over spatial locations. Persistent conditioning on semantic embeddings encourages the synthesized distribution to retain the inter-class organization of the shared text space, whereas sampling diverse embeddings from $\mathcal{E}_y$ expands the intra-class semantic modes covered for each category (Fig. 8b).

Spatial Semantic Anchoring

SSA draws on the observation that objects in natural images tend to occupy compact regions near the center (Schauerte and Stiefelhagen 2015), inducing a center-biased distribution of class-discriminative responses. Accordingly, attribution maps (e.g., Grad-CAM (Selvaraju et al. 2017)) aggregated across samples typically exhibit a center-peaked spatial profile (Fig. 5). We approximate this regularity with a Gaussian prior and align the class-discriminative attribution of synthetic images accordingly.

To decouple SSA from architecture-specific internals, we abstract the core principles of existing interpretability techniques (Zhou et al. 2016; Chefer, Gur, and Wolf 2021) to define a unified class-discriminative spatial attribution (CDSA) interface. Given an input $\hat{x}$ and a classifier f over C categories, we first recover a spatial response map $\mathbf{R}^f(\hat{x}) \in \mathbb{R}^{H_p \times W_p \times D_f}$ from its deepest task-related representation before the spatial information is collapsed into a global prediction. We then project each spatial response through the model’s classification-head weights $\mathbf{W}^{C,f} \in \mathbb{R}^{D_f \times C}$:

$$\mathbf{M}^f(\hat{x}) = \mathbf{R}^f(\hat{x})\mathbf{W}^{C,f} \in \mathbb{R}^{H_p \times W_p \times C}, \quad (6)$$

where $\mathbf{M}_c^f[i, j]$ denotes the CDSA score assigned to spatial location (i, j) for category c . Although architectures differ in how $\mathbf{R}^f$ is extracted, all maps consistently yield spatially resolved class-discriminative responses within a common C -dimensional category space. Specifically, for CLS-based models (e.g., ViT), $\mathbf{R}^f$ is assembled from the attention-weighted patch values before CLS token aggregation, whereas for architectures that retain spatial feature maps (e.g., ConvNext), it corresponds directly to the final feature map before pooling. Detailed constructions for each architecture are provided in the supplementary material.

Over a synthetic batch, we aggregate the per-sample maps via ReLU, $\bar{\mathbf{M}} = (\mathbb{E}_{\hat{x}}[\sigma(\mathbf{M}(\hat{x}))]) \in \mathbb{R}^{H_p \times W_p \times C}$, to retain positive attribution responses. We then align it against a Gaussian prior that encodes the center-biased regularity of natural images. Given that the spatial tendency is a broadly observed statistical regularity across natural images, we analytically construct the prior on the $H_p \times W_p$ patch grid, requiring no access to real data:

$$\pi[i, j] = B + A \exp\left(-\frac{1}{2} \left[\frac{(i - \mu_x)^2}{\sigma_x^2} + \frac{(j - \mu_y)^2}{\sigma_y^2} \right]\right), \quad (7)$$

where (μ_x, μ_y) denotes the center coordinates, (σ_x, σ_y) control the expected spatial extent of discriminative evidence, and A, B set the peak strength and the residual background response. The specific values of these prior parameters are provided in the supplementary material. We then align the aggregated CDSA of each category with the Gaussian prior:

$$\mathcal{L}_{\text{SSA}} = \frac{1}{C} \sum_{c=1}^C (1 - \cos(\bar{\mathbf{M}}_c^T, \pi)), \quad (8)$$

The cosine objective aligns only the spatial shape with the prior while remaining agnostic to the teacher-dependent magnitude of $\bar{\mathbf{M}}$, consistent with the design of center bias.

Spatial Semantic Distillation

SSD focuses on effectively transferring both categorical and spatial semantics from synthetic images to the student. Although the KL loss relays the teacher’s categorical knowledge, it discards the class-discriminative spatial semantics behind them. We therefore align the per-sample CDSA maps of teacher and student (Eq. (6)) via channel-wise similarity, upsampling the smaller map when patch resolutions differ:

$$\mathcal{L}_{\text{SSD}} = \mathbb{E}_{\hat{x} \sim \hat{\mathcal{D}}} \left[\frac{1}{C} \sum_{c=1}^C (1 - \cos(\mathbf{M}_c^T, \mathbf{M}_c^S)) \right]. \quad (9)$$

Note that $\mathcal{L}_{\text{SSD}}$ and $\mathcal{L}_{\text{SSA}}$ play complementary roles: the latter constrains the spatial statistics of synthesized images, while the former aligns per-sample attribution during distillation, ensuring the student inherits both the teacher’s class predictions and the spatial discriminative evidence.

Experiments

We evaluate UniDFKD on CIFAR-10, CIFAR-100 (Krizhevsky and Hinton 2009), and ImageNet-100 (a 100-class subset of ILSVRC-2012 (Deng et al. 2009)). Our experiments span four architecture families: the BN-based ResNet (RN) (He et al. 2016), alongside the LN-based ConvNeXt (CNX) (Liu et al. 2022), Vision Transformer (ViT) (Dosovitskiy et al. 2021), and Swin Transformer (Swin) (Liu et al. 2021b). To comprehensively assess our method, we construct distillation pairs under two regimes: homogeneous (e.g., within CNNs or ViTs) and heterogeneous (e.g., ViT $\rightarrow$ ResNet). Since modern architectures natively expect 224×224 inputs, we upsample CIFAR images to this resolution accordingly except for ResNet. We compare our approach against the representative DeepInversion (Yin et al. 2020), Fast (Fang et al. 2022), CMI (Fang et al. 2021), and NAYER (Tran et al. 2024b). For fair benchmarking, we adopt their official implementations but optimize their configurations for BN-free architectures. Beyond discarding the inapplicable $\mathcal{L}_{\text{BN}}$, we finely tune the learning rate, optimizer, data augmentation, and synthesis iterations to establish strong baselines. Implementation details are provided in the supplementary material.

Main Results

As shown in Tab. 1, UniDFKD achieves state-of-the-art accuracy across almost all 24 evaluated settings, attaining averages of 82.50% and 78.33% in homogeneous and heterogeneous scenarios and outperforming the previous best (NAYER) by 21.23% and 23.66%, respectively. When the teacher contains BN, whether in homogeneous (RN34 $\rightarrow$ RN18) or heterogeneous (RN34 $\rightarrow$ ViT-T) settings, existing methods already approach the real-data reference on CIFAR, as $\mathcal{L}_{\text{BN}}$ provides effective distributional guidance. UniDFKD matches this performance while achieving consistent gains at least 3% over NAYER on the more challenging ImageNet-100, confirming that semantic priors complement BN statistics during high-resolution synthesis.

The advantage grows substantially with BN-free teachers, where the absence of $\mathcal{L}_{\text{BN}}$ removes the dominant distributional prior. While baselines maintain moderate accuracy on CIFAR (e.g., NAYER yields 85.86% for ViT), they degrade sharply on ImageNet-100, dropping to 39.30% for ViT and collapsing to a mere 29.76% for Swin (vs. 88.30% real-data). In stark contrast, UniDFKD consistently bridges this gap, reaching 93.91% on CIFAR-10 and over 71% on ImageNet-100 across both architectures. Without distributional priors, high-resolution synthesis proves untenable for existing methods, whereas UniDFKD recovers to 71.00%.

The same pattern holds in heterogeneous settings with BN-free teachers, where baselines similarly collapse on ImageNet-100 (e.g., NAYER drops to 11.74% for CNX-

	CIFAR-10				CIFAR-100				ImageNet-100				
<i>Homogeneous distillation</i>													
<i>Teacher</i>	RN34	CNX-T	ViT-S	Swin-T	RN34	CNX-T	ViT-S	Swin-T	RN34	CNX-T	ViT-S	Swin-T	Avg.
<i>Student</i>	RN18	CNX-N	ViT-T	Swin-N	RN18	CNX-N	ViT-T	Swin-N	RN18	CNX-N	ViT-T	Swin-N	
T. acc.	95.70	97.54	95.25	96.43	78.05	83.10	77.93	81.24	85.92	88.98	85.06	90.83	88.00
S. acc.	95.20	97.41	94.01	96.08	77.10	82.50	73.80	79.22	84.72	86.27	83.20	88.30	86.48
DeepInversion	89.69	69.84	53.34	39.09	61.32	10.70	14.51	17.77	55.98	19.36	3.60	3.28	36.54
Fast	94.05	73.31	83.71	32.24	74.34	29.85	23.65	56.17	69.84	34.53	13.34	22.81	50.65
CMI	94.84	79.28	82.19	54.18	77.04	45.38	42.42	68.41	70.42	39.50	23.56	18.30	57.96
NAYER	95.21	73.69	85.86	47.26	77.54	45.07	57.51	66.32	78.60	39.10	39.30	29.76	61.27
UniDFKD	95.33	94.47	93.91	91.75	77.37	80.23	70.41	77.94	83.08	83.40	71.16	71.00	82.50
<i>Heterogeneous distillation</i>													
<i>Teacher</i>	RN34	CNX-T	ViT-S	Swin-T	RN34	CNX-T	ViT-S	Swin-T	RN34	CNX-T	ViT-S	Swin-T	Avg.
<i>Student</i>	ViT-T	ViT-T	RN18	CNX-N	ViT-T	ViT-T	RN18	CNX-N	ViT-T	ViT-T	RN18	CNX-N	
T. acc.	95.70	97.54	95.25	96.43	78.05	83.10	77.93	81.24	85.92	88.98	85.06	90.83	88.00
S. acc.	94.01	94.01	95.20	97.41	73.80	73.80	77.10	82.50	83.20	83.20	84.72	86.27	85.44
DeepInversion	71.72	25.31	33.81	30.36	52.53	7.02	11.86	33.02	15.58	5.06	6.07	10.99	25.28
Fast	86.15	40.35	72.73	72.07	63.15	11.79	15.70	41.98	49.83	7.65	23.73	48.91	44.50
CMI	91.19	43.72	75.79	77.18	69.83	25.30	35.55	65.70	57.30	9.16	24.95	37.83	51.13
NAYER	93.38	36.90	75.73	73.46	70.92	34.46	44.66	61.88	66.54	11.74	28.52	57.86	54.67
UniDFKD	94.54	91.99	90.64	93.64	74.59	73.15	71.67	75.58	70.32	68.36	56.72	78.72	78.33

Table 1: Student accuracy (%) under homogeneous and heterogeneous architectures. *T.* and *S.* acc. denote from-scratch on real data. “Avg.” is the mean across the datasets for homogeneous or heterogeneous setting. Best in **bold**, second best underlined.

	Base	Single			Pair			Full
CSC	×	✓	×	×	✓	✓	×	✓
SSA	×	×	✓	×	✓	×	✓	✓
SSD	×	×	×	✓	×	✓	✓	✓
Acc.	82.45	89.77	87.51	88.29	90.51	92.53	92.69	93.91

Table 2: Ablation study of the proposed CSC, SSA, and SSD modules (ViT-S→ViT-T, CIFAR-10). Combining all three components yields the best performance.

T→ViT-T) while UniDFKD maintains 68.36%. These results validate that architecture-agnostic semantic priors effectively decouple synthesis quality from normalization-specific statistics, enabling robust data-free transfer across modern teacher–student configurations.

Analytical Experiments

Effectiveness of the proposed modules. Tab. 2 reports the ablation of CSC, SSA, and SSD. Each module individually improves the baseline, confirming that categorical and spatial semantics both benefit data-free synthesis. Pairwise combinations consistently surpass their single counterparts, and the full model achieves the highest accuracy, validating that the three modules address complementary aspects of the data-free synthesis bottleneck.

Design Choices of CSC. As shown in Tab. 3, replacing noise with a single class prototype confirms that categorical semantics provide a strong generation prior. Diversifying intra-class embeddings and filtering OOD samples further maximizes accuracy to 89.77%, aligning the synthetic

	Sem.	Div.	Filt.	Acc.	Mechanism	Acc.
Noise	×	×	×	82.45	w/o inject	87.61
Proto.	✓	×	×	87.56	Additive	87.58
Diverse	✓	✓	×	89.09	SE attn.	89.13
Filtered	✓	✓	✓	89.77	CBN	89.77

(a) Conditioning type

(b) Injection mechanism

Table 3: Analysis of the proposed CSC module (ViT-S→ViT-T, CIFAR-10). (a) Conditioning types include Gaussian noise, single-class prototype embeddings, diverse class embeddings, and filtered embeddings. (b) Injection mechanisms include intermediate feature addition, SE attention, and our conditional batch normalization (CBN) style injection.

distribution with the ground-truth categorical structure. For feature injection, direct additive fusion falls below the no-injection baseline by collapsing generation diversity. While SE attention (Hu, Shen, and Sun 2018) helps via channel-wise reweighting, our CBN performs best by modulating normalization statistics, achieving the optimal trade-off between semantic consistency and generation diversity.

Design Choices of SSA. We justify CDSA as the optimal alignment target in SSA by comparing it with raw attention (*Attn.*) and feature maps (*Feat.*) from a ViT-S teacher (Fig. 5). Instance-wise (Fig. 5a), both *Attn.* and *Feat.* suffer from severe background activation, whereas CDSA precisely localizes the target-class subject to provide compact discriminative evidence. Distribution-wise under real data (Fig. 5b), *Feat.* collapses into a near-uniform profile. While *Attn.* exhibits a Gaussian-like center bias, its reliance on

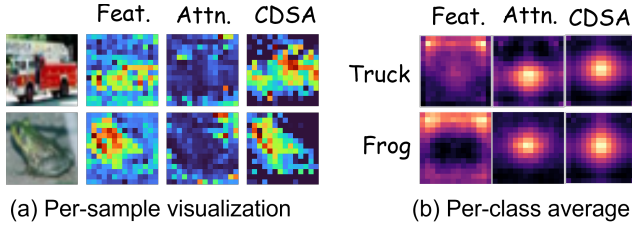


Figure 5: Visual comparison of our class-discriminative spatial attribution map (CDSA) against alternative attention mechanisms. *Feat.* and *Attn.* denote the L_2 norm of patch tokens and the CLS attention of the last layer, respectively.

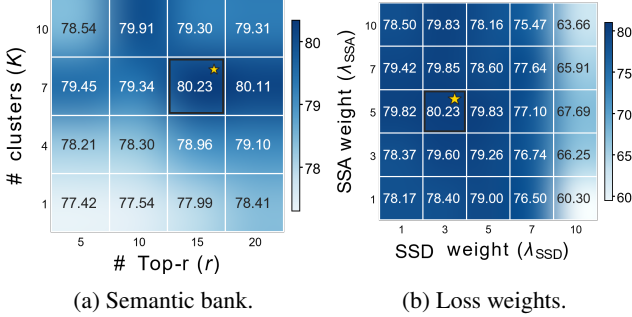


Figure 6: Hyperparameter analysis on CIFAR-100 (CNX-T $\rightarrow$ CNX-N). (a) Joint effects of cluster number and top- r in the diverse semantic bank. (b) Joint effects of synthesis weight λ_{SSA} and distillation weight λ_{SSD} .

self-attention precludes generalization to CNNs. Ultimately, CDSA uniquely produces a smooth, Gaussian profile that naturally conforms to our spatial prior (Eq. (7)).

Hyperparameter Analysis. UniDFKD introduces two hyperparameter pairs: (K , r) for semantic bank sizing and (λ_{SSA} , λ_{SSD}) for loss weighting, both exhibiting broad stability (Fig. 6, CNX-T $\rightarrow$ CNX-N). For the semantic bank (Fig. 6a), filtering an initial pool of 200 LLM-generated prompts per class down to roughly half ($K=7$, $r=15$) is optimal, as too few clusters fail to capture intra-class diversity, while more clusters or a larger r admit noisy prompts. The loss weights in (Fig. 6b) exhibit a distinct structural asymmetry. Since $\mathcal{L}_{SSA}$ imposes constraints only on batch-level statistics, λ_{SSA} remains robust across its entire range. In contrast, the sample-wise $\mathcal{L}_{SSD}$ competes directly with $\mathcal{L}_{KL}$, meaning an excessive λ_{SSD} (> 5) over-prioritizes spatial layout at the expense of class predictions.

Synthetic Distribution Quality. To verify that our gains stem from a better-aligned synthetic distribution, we evaluate the real and synthetic CIFAR-100 features across visual, manifold, and hierarchical dimensions (Fig. 7 and 8). Visually, t-SNE projections demonstrate that our synthetic clusters closely overlap with and broadly cover the real distribution, correcting the noticeable deviations seen in NAYER. At the manifold level, UniDFKD consistently improves all four PRDC metrics (Naeem et al. 2020), confirming higher fidelity and broader spatial coverage. At the hierarchical level,

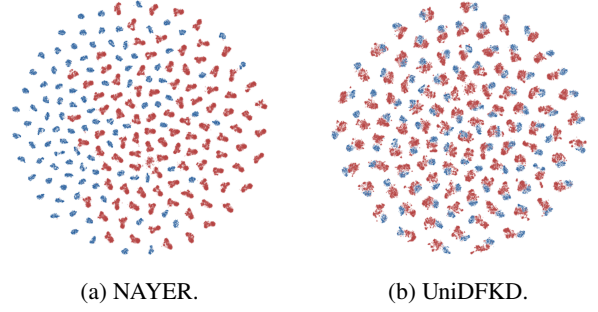


Figure 7: Coverage of synthetic distribution on CIFAR-100 (CNX-T $\rightarrow$ ViT-T), visualized by t-SNE of teacher features on training (blue) and synthetic (red) data.

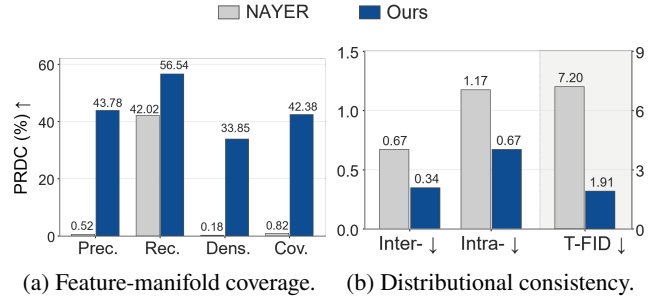


Figure 8: Synthetic distribution quality on CIFAR-100 (CNX-T $\rightarrow$ ViT-T). (a): PRDC (precision, recall, density and coverage); (b): inter-class MAE between the inter-class correlation matrices computed from real and synthetic data, intra-class Vendi MAE between their per-class Vendi scores, and T-FID between their teacher-feature distributions.

Conclusion

This paper addresses the architecture dependence of data-free knowledge distillation by shifting from implicit BN statistics to explicit, architecture-agnostic semantic priors. Our layer-wise analysis reveals that the effective content of the BN prior resides in deep semantic representations, motivating the design of *UniDFKD*. *UniDFKD* leverages categorical semantics to dictate *what* to synthesize and spatial semantics to pinpoint *where* the discriminative evidence is located. Consequently, semantic distillation resolves *how* to effectively transfer this preserved knowledge from the teacher to the student. Experiments across CNNs and ViTs under homogeneous and heterogeneous settings demonstrate substantial improvements over existing methods.

References

- Binici, K.; Aggarwal, S.; Pham, N. T.; Leman, K.; and Mitra, T. 2022. Robust and resource-efficient data-free knowledge distillation by generative pseudo replay. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 6089–6096.
- Chefer, H.; Gur, S.; and Wolf, L. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 782–791.
- Chen, H.; Wang, Y.; Xu, C.; Yang, Z.; Liu, C.; Shi, B.; Xu, C.; Xu, C.; and Tian, Q. 2019. Data-free learning of student networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3514–3522.
- Chen, X.; Li, Y.; Han, Y.; Xu, G.; and Shen, J. 2025. Coupling the generator with teacher for effective data-free knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2152–2160.
- Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2818–2829.
- Choi, Y.; Choi, J.; El-Khamy, M.; and Lee, J. 2020. Data-Free Network Quantization With Adversarial Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.
- de Vries, H.; Strub, F.; Mary, J.; Larochelle, H.; Pietquin, O.; and Courville, A. 2017. Modulating early visual processing by language. In *Advances in Neural Information Processing Systems*, volume 30.
- DeepSeek-AI. 2026. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence. *arXiv preprint arXiv:2606.19348*.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. IEEE.
- Dhillon, I. S.; and Modha, D. S. 2001. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42: 143–175.
- Do, K.; Le, H.; Nguyen, D.; Nguyen, D.; Harikumar, H.; Tran, T.; Rana, S.; and Venkatesh, S. 2022. Momentum Adversarial Distillation: Handling Large Distribution Shifts in Data-Free Knowledge Distillation. In *Advances in Neural Information Processing Systems*, volume 35.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Hounsby, N. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.
- Fang, G.; Mo, K.; Wang, X.; Song, J.; Bei, S.; Zhang, H.; and Song, M. 2022. Up to 100x faster data-free knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 6597–6604.
- Fang, G.; Song, J.; Wang, X.; Shen, C.; Wang, X.; and Song, M. 2021. Contrastive model inversion for data-free knowledge distillation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 2374–2380.
- Friedman, D.; and Dieng, A. B. 2023. The Vendi Score: A Diversity Evaluation Metric for Machine Learning. *Transactions on Machine Learning Research*.
- Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2414–2423.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.
- He, X.; Wang, J.; Cheng, Z.; Su, Y.; Huang, S.; and Lu, G. 2025. PRISM: Precision-recall informed data-free knowledge distillation via generative diffusion. *arXiv preprint arXiv:2509.16897*.
- Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 6626–6637.
- Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141.
- Huang, L.; Qin, J.; Zhou, Y.; Zhu, F.; Liu, L.; and Shao, L. 2023. Normalization techniques in training DNNs: Methodology, analysis and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 10173–10196.
- Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, 448–456. PMLR.
- Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. *Technical report, University of Toronto*.
- Li, M.; Zhang, D.; He, T.; Xie, X.; Li, Y.-F.; and Qin, K. 2024a. Towards effective data-free knowledge distillation via diverse diffusion augmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 4416–4425.
- Li, Y.; Kim, Y.; Lee, D.; Kundu, S.; and Panda, P. 2024b. GenQ: Quantization in low data regimes with generative synthetic data. In *European Conference on Computer Vision*, 216–235.
- Li, Z.; Ma, L.; Chen, M.; Xiao, J.; and Gu, Q. 2022. Patch similarity aware data-free quantization for vision transformers. In *European Conference on Computer Vision*, 154–170.

- Liu, H.; Wang, Y.; Liu, H.; Sun, F.; and Yao, A. 2024. Small-scale data-free knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6008–6016.
- Liu, Y.; Zhang, W.; Wang, J.; and Wang, J. 2021a. Data-free knowledge transfer: A survey. *arXiv preprint arXiv:2112.15278*.
- Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.
- Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11976–11986.
- Lopes, R. G.; Fenu, S.; and Starner, T. 2017. Data-free knowledge distillation for deep neural networks. *arXiv preprint arXiv:1710.07535*.
- Mahendran, A.; and Vedaldi, A. 2015. Understanding deep image representations by inverting them. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5188–5196.
- Mansourian, A. M.; Ahmadi, R.; Ghafouri, M.; Babaei, A. M.; Golezani, E. B.; Ghamchi, Z. Y.; Ramezani, V.; Taherian, A.; Dinashi, K.; Miri, A.; and Kasaei, S. 2025. A Comprehensive Survey on Knowledge Distillation. *Transactions on Machine Learning Research*.
- Micaelli, P.; and Storkey, A. J. 2019. Zero-shot knowledge transfer via adversarial belief matching. *Advances in Neural Information Processing Systems*, 32.
- Naeem, M. F.; Oh, S. J.; Uh, Y.; Choi, Y.; and Yoo, J. 2020. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, 7176–7185. PMLR.
- Nayak, G. K.; Mopuri, K. R.; Shaj, V.; Radhakrishnan, V. B.; and Chakraborty, A. 2019. Zero-shot knowledge distillation in deep networks. In *International Conference on Machine Learning*, 4743–4751. PMLR.
- Pratt, S.; Covert, I.; Liu, R.; and Farhadi, A. 2023. What does a platypus look like? Generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15691–15701.
- Qi, X.; Li, R.; Peng, L.; Ling, Q.; Yu, J.; Chen, Z.; Chang, P.; Han, M.; and Xiao, J. 2025. Data-free knowledge distillation with diffusion models. In *IEEE International Conference on Multimedia and Expo*, 1–6.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.
- Ramachandran, A.; Kundu, S.; and Krishna, T. 2024. CLAMP-ViT: Contrastive data-free learning for adaptive post-training quantization of ViTs. In *European Conference on Computer Vision*.
- Schauerte, B.; and Stiefelhofen, R. 2015. On the distribution of salient objects in web images and its influence on salient object detection. *PLOS ONE*, 10(7): e0130316.
- Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.
- Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive Multi-view Coding. In *European Conference on Computer Vision*, 776–794.
- Tran, M.-T.; Le, T.; Le, X.-M.; Cai, J.; Harandi, M.; and Phung, D. 2024a. Large-scale data-free knowledge distillation for ImageNet via multi-resolution data generation. *arXiv preprint arXiv:2411.17046*.
- Tran, M.-T.; Le, T.; Le, X.-M.; Harandi, M.; Tran, Q. H.; and Phung, D. 2024b. NAYER: Noisy layer data generation for efficient and effective data-free knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23860–23869.
- Yang, A.; et al. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.
- Yin, H.; Molchanov, P.; Alvarez, J. M.; Li, Z.; Mallya, A.; Hoiem, D.; Jha, N. K.; and Kautz, J. 2020. Dreaming to distill: Data-free knowledge transfer via DeepInversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8715–8724.
- Yu, S.; Chen, J.; Han, H.; and Jiang, S. 2023. Data-free knowledge distillation via feature exchange and activation region constraint. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24266–24275.
- Zhong, Y.; Zhou, Y.; Zhang, Y.; Sui, W.; Li, S.; Li, Y.; Chao, F.; and Ji, R. 2025. Semantic alignment and reinforcement for data-free quantization of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12479–12490.
- Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2921–2929.

Supplementary Material

Overview

This supplementary material complements the main paper with further methodological details, extensive additional quantitative results, deeper ablation and analytical studies, qualitative visualizations, the complete implementation setup and extended discussions.

The contents are organized as follows:

- **Sec. (Further Methodological Details):** Completes the description of the class-discriminative spatial attribution (CDSA) interface by detailing how the spatial response is recovered for every teacher family considered in this work, i.e., feature-map-based backbones (ResNet, ConvNeXt, Swin) and CLS-token-based backbones (plain ViT). This instantiation is what makes UniDFKD applicable to arbitrary architectures.
- **Sec. (Extensive Quantitative Results):** Reports supplementary quantitative evaluations, including a comparison with state-of-the-art DFKD methods on conventional CNNs, an analysis of computational efficiency (speedup), and results on the large-scale ImageNet-1K benchmark under tight synthesis budgets.
- **Sec. (Additional Ablation and Analytical Studies):** Extends the layer-wise diagnosis to additional architectures and datasets, verifies the robustness of Categorical Semantic Conditioning (CSC) to vision-agnostic text encoders, and ablates the design and the parameter sensitivity of the spatial prior used by Spatial Semantic Anchoring (SSA).
- **Sec. (Qualitative Visualizations):** Showcases visual comparisons of the synthesized images to demonstrate the superior synthesis quality of our approach.
- **Sec. (Extended Implementation Details):** Provides comprehensive experimental settings, including hyperparameter configurations, optimization strategies, and data augmentation recipes, together with the full protocol of the layer-wise diagnosis reported in Fig.3 of the main manuscript.
- **Sec. (Extended Discussions and Limitations):** Clarifies the relationship between our framework and alternative paradigms (e.g., BN-free model inversion, data-free quantization, and diffusion-based synthesis), and frankly analyzes the potential limitations of the center-biased spatial prior.

Further Methodological Details

Architecture-specific Instantiation of CDSA

The interface in Eq.6 requires only the spatial response $\mathbf{R}^f \in \mathbb{R}^{H_p \times W_p \times D_f}$; backbones differ solely in how this response is recovered, depending on whether an explicit spatial representation survives until classification. We describe the two cases below and omit the superscript f for brevity.

Feature-map-based architectures. For ResNet, ConvNeXt, and hierarchical transformers such as Swin, the response is the final classifier-compatible feature map taken

before global pooling,

$$\mathbf{R}(\hat{x}) = \Phi(\mathbf{F}_{\text{final}}(\hat{x})) \in \mathbb{R}^{H_p \times W_p \times D_f}, \quad (\text{S1})$$

where $\Phi(\cdot)$ absorbs the architecture-specific terminal normalization when present. Only the source of $\mathbf{F}_{\text{final}}$ differs within this family: ResNet uses its final convolutional feature map, for which Eq.6 reduces to standard class-activation mapping; ConvNeXt uses its normalized final convolutional representation; and Swin restores its final normalized tokens to the spatial grid of the last stage.

CLS-token-based architectures. Plain ViTs retain no such feature map, and their raw attention indicates spatial saliency without category-specific evidence. We therefore intercept the attention-weighted values before token aggregation. Let index 0 denote the CLS token and indices $1:N$ the $N = H_p W_p$ patch tokens. In the final multi-head attention block, the CLS-to-patch attention of head $h \in \{1, \dots, H\}$ is

$$\mathbf{a}^{(h)} = \text{Softmax} \left(\frac{\mathbf{q}_0^{(h)} (\mathbf{K}_{1:N}^{(h)})^\top}{\sqrt{d_h}} \right) \in \mathbb{R}^N, \quad (\text{S2})$$

with d_h the per-head dimension. Rather than summing over patches, we keep the contribution of each patch n across all heads and merge them through the output projection $\mathbf{W}^O$:

$$\mathbf{r}_n = \text{Concat} \left(a_n^{(1)} \mathbf{V}_n^{(1)}, \dots, a_n^{(H)} \mathbf{V}_n^{(H)} \right) \mathbf{W}^O \in \mathbb{R}^{D_f}, \quad (\text{S3})$$

and rearranging $\{\mathbf{r}_n\}_{n=1}^N$ onto the patch grid gives $\mathbf{R}(\hat{x}) \in \mathbb{R}^{H_p \times W_p \times D_f}$.

Extensive Quantitative Results

Comparison on General CNN Architectures

Table S1 compares existing data-free knowledge distillation (DFKD) methods with our UniDFKD on conventional CNN architectures. As observed, previous approaches have already achieved remarkable success on these standard benchmarks, with their distillation performance closely approaching the accuracy of student models trained on real data. The performance ceiling on these conventional CNNs is therefore highly saturated, making it exceedingly difficult for new methods to obtain large absolute accuracy gains.

Despite this saturation, our proposed UniDFKD remains highly competitive. Most notably, compared with CPNet, the latest state-of-the-art method that heavily relies on architecture-specific BN priors, UniDFKD delivers entirely on-par performance. Under several configurations, it even surpasses all existing baselines and establishes new state-of-the-art records. For instance, UniDFKD achieves the highest accuracy in the W40-2 $\rightarrow$ W16-2 (94.10%) and V11 $\rightarrow$ R18 (92.44%) settings on CIFAR-10, as well as in the W40-2 $\rightarrow$ W40-1 (71.99%) and V11 $\rightarrow$ R18 (71.85%) settings on CIFAR-100.

These quantitative results indicate that the advantages of UniDFKD are not restricted to BN-free or heterogeneous scenarios. While it fundamentally resolves the catastrophic degradation of existing methods on modern architectures, it simultaneously retains state-of-the-art efficacy in traditional BN-based settings, validating the universality and robustness of our explicit semantic prior framework.

Method	CIFAR10					CIFAR100					TinyImageNet
	R34 R18	W40-2 W16-2	W40-2 W16-1	W40-2 W40-1	V11 R18	R34 R18	W40-2 W16-2	W40-2 W16-1	W40-2 W40-1	V11 R18	R34 R18
Teacher	95.7	94.87	94.87	94.87	92.25	78.05	75.83	75.83	75.83	71.32	66.44
Student	95.2	93.95	91.12	93.94	95.2	77.1	73.56	65.31	72.19	77.1	64.87
DeepInversion (Yin et al. 2020)	93.26	89.72	83.04	86.85	90.36	61.32	61.34	53.77	68.58	54.13	-
DAFL (Chen et al. 2019)	92.22	81.55	65.71	81.33	81.1	74.47	43.7	20.88	42.83	54.16	-
DFQ (Choi et al. 2020)	94.61	92.01	86.14	91.69	90.84	77.01	64.79	51.27	54.43	66.21	-
ZSKT (Micaelli and Storkey 2019)	93.32	89.66	83.74	86.07	89.46	67.74	54.59	36.6	53.6	54.31	-
CMI (Fang et al. 2021)	94.84	92.52	90.01	92.78	91.13	77.04	68.75	57.91	68.88	70.56	64.01
Fast (Fang et al. 2022)	94.05	92.45	89.29	92.51	90.53	74.34	65.12	54.02	63.91	67.44	-
MAD (Do et al. 2022)	94.9	92.64	-	-	-	77.31	64.05	-	-	-	62.32
Spaceship (Yu et al. 2023)	95.39	93.25	90.38	93.56	92.27	77.41	69.95	58.06	68.78	71.41	64.04
SSD-KD (Liu et al. 2024)	94.26	93.11	89.96	93.23	90.67	75.16	65.28	55.61	64.57	68.77	-
NAYER (Tran et al. 2024b)	95.21	94.07	91.94	94.15	<u>92.37</u>	<u>77.54</u>	71.72	62.23	71.80	71.75	64.17
CPNet (Chen et al. 2025)	95.26	<u>94.09</u>	92.00	94.22	92.18	77.68	71.90	63.23	<u>71.98</u>	<u>71.80</u>	64.91
UniDFKD (Ours)	<u>95.33</u>	94.10	<u>91.98</u>	<u>94.19</u>	92.44	77.37	<u>71.77</u>	<u>62.65</u>	71.99	71.85	<u>64.71</u>

Table S1: Performance comparison of existing DFKD methods and UniDFKD. The best results are highlighted in bold and the runner-up is underlined. The results of previous methods are taken from their original papers. “R” indicates ResNet, “W” corresponds to WideResNet, and “V” stands for VGG in this paper. “-” denotes the result is not reported.

	DeepInv	CMI	Fast	NAYER	Ours
CIFAR-10	53.34	82.19	83.71	85.86	93.91
Time (hours)	36.87	33.43	5.97	5.69	3.36
CIFAR-100	14.51	42.42	23.65	57.51	70.41
Time (hours)	37.22	35.70	6.99	6.86	3.65
Average Speed Up	1.00×	1.07×	5.75×	5.95×	10.59×

Table S2: Comparison of distillation accuracy (%) and training time (hours) on CIFAR-10 and CIFAR-100 under the ViT-S $\rightarrow$ ViT-T setting. Our method not only achieves the highest distillation performance but also significantly improves synthesis efficiency. The “Average Speed Up” is calculated relative to the DeepInv baseline.

Computational Efficiency Comparison

Data-free knowledge distillation traditionally suffers from substantial computational overhead, primarily because the image synthesis process is iterative and time-consuming. In Table S2, we therefore evaluate the computational efficiency of our method against existing baselines under the ViT-S $\rightarrow$ ViT-T setting on the CIFAR-10 and CIFAR-100 datasets.

As shown in Table S2, traditional optimization-based inversion methods such as DeepInversion and CMI are highly inefficient, requiring between 33 and 38 hours to complete the distillation process. Recent methods such as Fast and NAYER improve efficiency by accelerating the synthesis phase, reducing the time to approximately 5.6 to 7 hours, yet UniDFKD pushes this boundary significantly further. Our method requires only 3.36 hours on CIFAR-10 and 3.65 hours on CIFAR-100, achieving a remarkable average speedup of $10.59\times$ relative to the DeepInversion baseline.

Crucially, this exceptional efficiency does not come at the expense of distillation quality. On the contrary, UniDFKD simultaneously achieves the highest distillation accuracy, outperforming the second-fastest method (NAYER) by a substantial absolute margin of 8.05% (93.91% vs.

Data Ratio	1%		5%	
Accuracy	Top 1	Top 5	Top 1	Top 5
CMI	3.61	12.44	16.22	42.71
NAYER	2.68	9.45	18.27	46.55
UniDFKD	19.18	41.72	36.32	66.24

Table S3: Accuracy (%) on ImageNet-1K under the DeiT-S $\rightarrow$ DeiT-T setting with limited synthesis budgets. The “Data Ratio” (1% and 5%) indicates the number of synthesized images relative to the full scale of the original ImageNet-1K training set. UniDFKD significantly outperforms both CMI and NAYER across all extremely low data regimes.

85.86%) on CIFAR-10 and 12.9% (70.41% vs. 57.51%) on CIFAR-100. This demonstrates that, by leveraging explicit architecture-agnostic semantic priors, UniDFKD not only provides stronger and more accurate guidance for high-fidelity data generation, but also facilitates much faster convergence. Overall, UniDFKD yields a markedly superior trade-off between training efficiency and distillation performance compared with existing approaches (Tran et al. 2024b).

Evaluation on Large-Scale Datasets

Extending DFKD to large-scale datasets such as ImageNet-1K (IN-1K) inherently poses severe computational and storage challenges. Synthesizing the full scale of the IN-1K training dataset (over 1.2 million images at a high resolution) consumes prohibitive amounts of GPU memory and disk storage, rendering it practically infeasible in standard resource-constrained environments. Consequently, to demonstrate the superiority and practicality of our method on large-scale datasets, we evaluate its robustness under extremely limited synthesis budgets, specifically synthesizing only 1% and 5% of the original IN-1K dataset scale.

Table S3 presents the top-1 and top-5 distillation accuracy on IN-1K under the DeiT-S $\rightarrow$ DeiT-T setting. In the extremely stringent 1% data regime, existing baseline methods suffer from catastrophic performance degradation: NAYER and CMI achieve a mere 2.68% and 3.61% top-1 accuracy, respectively, indicating their inability to extract effective task-relevant distributions with so few synthesis iterations. In stark contrast, UniDFKD exhibits exceptional robustness, securing 19.18% top-1 and 41.72% top-5 accuracy under the same 1% budget.

As the synthesis budget increases to 5%, the performance gap widens further. UniDFKD achieves 36.32% top-1 and 66.24% top-5 accuracy, which substantially outperforms NAYER (18.27% top-1) and CMI (16.22% top-1) by over 18% in absolute top-1 accuracy. These results demonstrate that our unified semantic prior framework effectively maximizes the information density of each generated sample. By enforcing strictly aligned semantic structures and class-discriminative spatial localization, UniDFKD achieves highly efficient knowledge transfer even with extremely sparse synthetic data, highlighting its scalability and robustness for large-scale applications.

Additional Ablation and Analytical Studies

Extended Layer-wise Analysis

To further validate our preliminary observations, Fig. S1 and Fig. S2 extend the layer-wise analysis to additional modern architectures (ConvNeXt and ViT) across CIFAR-10 and the more challenging ImageNet-100 dataset. The diagnostic protocol strictly follows the one described in Sec. .

Consistent with the findings in the main paper, the results demonstrate a clear and universal trend across different architectures and datasets. Statistics from shallow blocks provide only marginal improvements over the unconstrained baseline (w/o Stat. Acc). As the block index increases, however, the single-block distillation accuracy rises progressively while the FID steadily decreases. Notably, matching the statistics of the deepest layers alone can already recover most of the performance obtained by matching all blocks combined.

This extended analysis provides strong empirical evidence that the phenomenon is not restricted to specific models or scales. It confirms that the true utility of statistical priors fundamentally lies in their implicit alignment of deep, class-discriminative semantic representations, which further justifies our motivation to construct an explicit, architecture-agnostic semantic prior.

Robustness to Vision-Agnostic Text Encoders

To construct the categorical semantic bank, our primary experiments utilize CLIP, a vision-language model natively aligned with visual representations. A compelling question therefore arises: *does the success of our explicit semantic prior rely strictly on this pre-existing cross-modal alignment?* To address it, Table S4 investigates the robustness of UniDFKD by substituting CLIP with a purely vision-agnostic language encoder (Qwen3-Embedding-0.6B).

As illustrated in the “w/o Cent.” (without mean-centering) columns, directly utilizing raw embeddings from the pure

Embedding Model	CIFAR-10		CIFAR-100	
	w/o Cent.	w/ Cent.	w/o Cent.	w/ Cent.
CLIP ViT-B/32	93.50	93.88	69.10	70.40
CLIP ViT-H/14	93.91	93.91	69.79	<u>70.41</u>
Qwen3-Emb.-0.6B	93.79	<u>93.91</u>	68.53	70.44

Table S4: Ablation on the text encoder used to build the class-semantic bank. We report student accuracy (%) on CIFAR-10 and CIFAR-100 for two vision-language encoders (CLIP ViT-H/14 and ViT-B/32) and a vision-agnostic, language-only encoder (Qwen3-Embedding-0.6B), each evaluated without (w/o Cent.) and with (w/ Cent.) mean-centered prompt embeddings.

language model yields highly competitive results on CIFAR-10 (93.79%). It nevertheless exhibits a noticeable performance drop on the more challenging CIFAR-100 dataset (68.53%) compared with the robust CLIP ViT-H/14 baseline (69.79%). This discrepancy is expected, as embeddings from pure language models typically suffer from representation anisotropy (i.e., vectors occupying a narrow cone in the latent space) and lack explicit geometric connections to visual structures.

Crucially, once a simple mean-centering operation (“w/ Cent.”) is applied to calibrate the embedding space—shifting the textual embeddings by their global mean to mitigate modality bias—the performance of the vision-agnostic encoder recovers dramatically. The centered Qwen3 embeddings achieve 93.91% on CIFAR-10 and 70.44% on CIFAR-100, which matches and even marginally surpasses the best-performing vision-aligned CLIP ViT-H/14 model (93.91% and 70.41%, respectively).

These empirical findings substantiate a core mechanism of UniDFKD: the efficacy of Categorical Semantic Conditioning (CSC) does not stem from any leakage of cross-modal visual memory. Instead, it leverages the rich and structurally consistent relational hierarchies—such as inter-class semantic distances and intra-class descriptive diversity—that are intrinsically embedded within pure language models. This structural alignment effectively guides the generative process and confirms the generalization and robustness of our framework with respect to arbitrary advanced text encoders.

Ablation on Spatial Semantic Anchoring

To validate the design rationale and the sensitivity of the Gaussian prior in our Spatial Semantic Anchoring (SSA) module, we evaluate several spatial prior variants across three representative distillation settings: BN-homogeneous (RN34 $\rightarrow$ RN18), LN-homogeneous (ViT-S $\rightarrow$ ViT-T), and heterogeneous (Swin-T $\rightarrow$ CNX-N).

As shown in Table S5, replacing the prior with a *Constant Map*, defined as $U(0.5)$, imposes a spatially uniform constraint on the class-discriminative attribution. Despite its simplicity, this variant still yields competitive performance (e.g., 95.01% on RN34 $\rightarrow$ RN18), suggesting that encouraging a globally regular attribution pattern already provides an

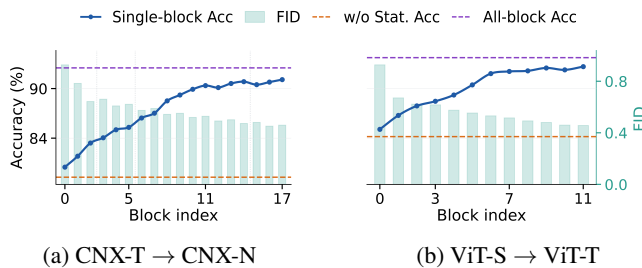


Figure S1: Layer-wise analysis on CIFAR-10. We report the single-block accuracy and FID obtained using statistics from different teacher blocks, together with the accuracy without statistical guidance and the all-block result.

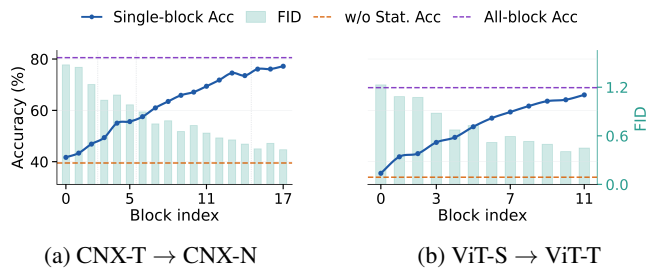


Figure S2: Layer-wise analysis on ImageNet-100. We report the single-block accuracy and FID obtained using statistics from different teacher blocks, together with the accuracy without statistical guidance and the all-block result.

Prior Design	RN34 → RN18 (BN-Homo)	ViT-S → ViT-T (LN-Homo)	Swin-T → CNX-N (Hetero)
Constant Map ($U(0.5)$)	95.01	93.20	93.15
Corner-Biased Gaussian ($\max(\pi) - \pi$)	92.20	90.60	89.20
Center Gaussian (Fixed)	95.12	93.30	93.55
Center Gaussian (Sampled)	95.33	93.91	93.64

Table S5: Ablation on the design and sensitivity of the spatial prior in SSA. We compare our sampled center-biased Gaussian prior against a constant map (no spatial preference), a corner-biased prior (reversed preference), and a fixed deterministic center prior. Performance is evaluated across BN-homogeneous, LN-homogeneous, and heterogeneous settings.

effective baseline for synthesis. Applying a *Corner-Biased Gaussian*—formulated as $\max(\pi) - \pi$, a reversed prior that encourages class-discriminative evidence toward peripheral regions—instead results in a clear performance degradation across all three settings; for instance, the accuracy drops to 89.20% in the Swin-T → CNX-N setting. This negative control indicates that the spatial orientation of the prior is important: forcing discriminative evidence toward the image boundaries introduces an unfavorable inductive bias for the object-centric benchmarks considered in this work.

Furthermore, we compare our default approach against a *Fixed Center Gaussian*, whose peak is strictly anchored to the exact image center without spatial jitter. Although the fixed-center variant remains robust and slightly outperforms the uniform prior, it is consistently surpassed by our *Sampled Center Gaussian*. By sampling the center coordinates (μ_x, μ_y) from predefined uniform intervals, the sampled prior introduces controlled spatial variation, allowing discriminative regions to appear at diverse yet plausible central locations rather than being constrained to an identical position across all samples. Overall, these results support the use of a center-oriented spatial prior and show that moderate stochasticity in its location provides a favorable balance between spatial regularity and diversity.

Qualitative Visualizations

Figure S3 compares synthetic samples generated by CMI, NAYER, and UniDFKD using ViT-S teachers on CIFAR-10, CIFAR-100, and ImageNet-100. CMI mainly produces fragmented high-frequency patterns, while NAYER improves semantic diversity but still exhibits strong texture artifacts

and irregular spatial layouts. In contrast, UniDFKD generates samples with more coherent class-related structures and better-localized discriminative regions. This advantage becomes more evident on ImageNet-100, where UniDFKD recovers richer object- and scene-level patterns rather than relying on repetitive local textures. These observations suggest that CSC provides category-consistent semantic guidance, while SSA organizes discriminative evidence into spatially meaningful regions without sacrificing diversity. Overall, the qualitative results indicate that UniDFKD synthesizes more informative and structurally coherent samples, consistent with its superior distillation performance.

Extended Implementation Details

Detailed Experimental Settings

Datasets and Architectures. We evaluate UniDFKD on three widely adopted benchmarks: CIFAR-10, CIFAR-100 (Krizhevsky and Hinton 2009), and ImageNet-100. Specifically, CIFAR-10 and CIFAR-100 consist of 50,000 training and 10,000 validation images, categorized into 10 and 100 classes, respectively. To ensure a fair comparison with prior data-free knowledge distillation methods, we maintain the native 32×32 spatial resolution exclusively for the classical ResNet-34 to ResNet-18 (RN34→RN18) setting. For all other heterogeneous distillation scenarios, the input images are upsampled to 224×224 via bilinear interpolation to strictly align with the default input configurations of modern architectures. ImageNet-100 is a 100-class subset of the large-scale ILSVRC-2012 dataset (Deng et al. 2009). We adopt the rigorously established CMC split (Tian, Krishnan, and Isola 2020), which comprises 126,689 training

images and 5,000 validation images, providing a more challenging, high-resolution testbed.

To comprehensively evaluate the architectural generalizability of our framework, our experiments span four representative architecture families that encompass distinct structural paradigms (convolution- vs. attention-based) and normalization types (BN vs. LN): (1) the classic Batch Normalization (BN)-based **ResNet** (RN) (He et al. 2016), serving as the foundational baseline for traditional CNNs; (2) the Layer Normalization (LN)-based **ConvNeXt** (CNX) (Liu et al. 2022), representing modernized CNNs that effectively bridge the performance gap with Transformers; (3) the quintessential **Vision Transformer** (ViT) (Dosovitskiy et al. 2021), which pioneers the application of patch-based global self-attention in vision tasks; and (4) the **Swin Transformer** (Swin) (Liu et al. 2021b), a state-of-the-art hierarchical Transformer distinguished by its efficient shifted window-based local attention. This diverse selection rigorously validates the robustness of UniDFKD across highly heterogeneous knowledge transfer scenarios.

Comparison Methods. We benchmark our proposed UniDFKD against a carefully selected set of representative Data-Free Knowledge Distillation (DFKD) methods that span various technical paradigms: **DeepInversion** (Yin et al. 2020), a pioneering optimization-based method that establishes the foundational paradigm of Batch Normalization (BN) statistics matching; **Fast** (Fang et al. 2022), a widely recognized generator-based framework serving as our efficiency-focused baseline; **CMI** (Fang et al. 2021), which introduces contrastive learning to explicitly mitigate mode collapse and represents diversity-driven synthesis; and **NAYER** (Tran et al. 2024b), a recent state-of-the-art approach that incorporates semantic embeddings to further enhance generative quality. This diverse baseline selection provides a comprehensive testbed to validate the superiority and generalization capability of our framework.

Teacher Preparation. All teachers are trained by us from random initialization on the real training split of each dataset, without external data or pre-trained weights. For RN34 we follow the protocol of prior DFKD works (Fang et al. 2021; Tran et al. 2024b): SGD (momentum 0.9, weight decay 5×10^{-4} , initial learning rate 0.1 with cosine annealing, batch size 128) for 200 epochs with padded random cropping and horizontal flipping. For the LN-based CNX, ViT, and Swin teachers we adopt the official `torchvision` recipes: AdamW (weight decay 0.05, peak learning rate 1×10^{-3} , cosine schedule with linear warm-up), label smoothing of 0.1, RandAugment, and Mixup/CutMix for 300 epochs. The same recipes produce the from-scratch students reported as the real-data upper bound. All teachers are frozen in `eval()` mode and shared across all competing methods.

Training Details. With the teachers fixed, we first construct the semantic bank that drives synthesis. Diverse yet category-faithful captions are collected by querying two instruction-tuned LLMs, DeepSeekV4 (DeepSeek-AI 2026) and Qwen3 (Yang et al. 2025), with the template below, yielding 200 captions per class.

```
Your task is to write me 10 different
image captions that include and
visually describe a scene around a
concept. Your concept is {object_text}.
Output one single grammatically correct
caption that is no longer than 20
words. Do not output any notes, word
counts, facts, etc. Do not use fluffy,
poetic language. Output 10 single
sentence only, beginning with 'A
photo of'. Note: Be creative and avoid
repetition.
```

Captions are encoded by a frozen OpenCLIP ViT-H/14 text encoder (Cherti et al. 2023), and the per-class embeddings are filtered by spherical k -means with $K = 7$ and $r = 15$ retained embeddings per class. For SSA, the Gaussian prior is resampled per batch in units of the attribution grid $H_p \times W_p$, whose center is $(c_x, c_y) = ((W_p - 1)/2, (H_p - 1)/2)$. The Gaussian center is jittered symmetrically around it, $\mu_x \sim \mathcal{U}[c_x - \rho W_p, c_x + \rho W_p]$ and $\mu_y \sim \mathcal{U}[c_y - \rho H_p, c_y + \rho H_p]$, while the widths are drawn proportionally to the grid, $\sigma_x \sim \mathcal{U}[\beta_{\min} W_p, \beta_{\max} W_p]$ and $\sigma_y \sim \mathcal{U}[\beta_{\min} H_p, \beta_{\max} H_p]$, with $\rho = 1/7$ and $(\beta_{\min}, \beta_{\max}) = (1/7, 3/14)$ shared by all architectures. Since the jitter is symmetric, the prior stays centered in expectation while its exact position and extent vary across batches; on the 14×14 grid of ViT this instantiates to $\mu_x, \mu_y \in [4.5, 8.5]$ and $\sigma_x, \sigma_y \in [2, 3]$, and the intervals scale accordingly for architectures with a different grid size. The amplitude is fixed to $A = 1$, since the cosine objective is scale-invariant and only the offset-to-amplitude ratio is identifiable; the offset is resampled per batch as $B \sim \mathcal{U}[0, 0.3]$, leaving a uniform floor of up to 30% of the Gaussian peak.

During the data synthesis phase, we employ the Adam optimizer with a learning rate of 0.1 to optimize each batch of 200 images. The number of synthesis iterations per batch is set to 40, 60, and 120 for CIFAR-10, CIFAR-100, and ImageNet-100, respectively. To ensure a stable and diverse initial data distribution, we conduct a synthesis-only warm-up for 20 epochs to populate the replay pool prior to student training. During synthesis, the balancing weights for the cross-entropy loss and our proposed SSA regularization are empirically set to $\lambda_{CE} = 1.0$ and $\lambda_{SSA} = 5$, respectively.

We emphasize that being “architecture-agnostic” does not imply being “architecture-exclusive.” The core objective of UniDFKD is to eliminate the reliance on BN statistics without preventing their utilization. Therefore, our proposed semantic priors are designed to be universally applicable regardless of the teacher’s architecture. For BN-free teachers, these semantic priors alone provide the requisite guidance for effective synthesis, keeping the framework fully functional. Conversely, when BN statistics are available, we retain them as an auxiliary distributional regularizer rather than discarding valuable teacher-side knowledge. In short, BN statistics and our semantic priors are highly complementary, naturally combining to benefit BN-based teachers while maintaining robust performance for BN-free ones.

After the synthesis-only warm-up, the student is trained for 100 effective epochs on the expanding replay pool, with 400 updates per epoch and a batch size of 256. Follow-

ing standard DFKD protocols (Tian, Krishnan, and Isola 2020; Tran et al. 2024b), the overall optimization objective comprises a KL divergence loss to align the student’s predictions with the teacher’s soft labels. We use AdamW ($\beta_1, \beta_2 = 0.9, 0.999$; weight decay 0.05) with a peak learning rate of 1×10^{-3} , a 20-epoch linear warm-up followed by cosine decay to 1×10^{-6} . Replay samples are augmented by random resized cropping (scale $[0.6, 1.0]$, bicubic), horizontal flipping, RandAugment (2 operations, magnitude 5), random erasing ($p = 0.25$), and Mixup or CutMix with equal probability ($\alpha_{\text{mix}} = 0.8, \alpha_{\text{cut}} = 1.0$); the teacher and the student receive the same view. We fix $T = 20$ and $\lambda_{\text{SSD}} = 3$ in all experiments, and report the best top-1 validation accuracy averaged over three random seeds (0, 1, and 2). To ensure a fair comparison, all baseline methods are trained using the identical data augmentation pipeline, optimizer settings, and total training iterations. Furthermore, we comprehensively sweep their respective hyperparameter combinations on the evaluated benchmarks, reporting the best-performing configurations for each method. All experiments were implemented in PyTorch 2.5.1 and conducted on two NVIDIA A800 GPUs (80GB).

Protocol of the Layer-wise Diagnosis in Fig.3

We select the representative method CMI (Fang et al. 2021) to conduct our baseline layer-wise experiments. Unlike NAYER (Tran et al. 2024b), which inherently introduces the confounding effects of semantic embeddings, CMI enhances the diversity of synthesized data primarily through an auxiliary generator and contrastive objectives, independent of the BN prior. By keeping these generative components strictly frozen during the layer-wise evaluation, we uniquely isolate the matched feature statistics. Consequently, any variation in student accuracy can be directly attributed to the effectiveness of the statistical priors at different network depths.

In the feature matching process, we strictly define a “stat hook” as a channel-wise mean and variance statistical pair on a normalization layer. To accommodate the intrinsic differences of normalization layers across various network architectures, we design specific feature extraction and matching strategies for ResNet-34 and ViT-S.

Statistic Matching in ResNet-34. ResNet-34 consists of 16 BasicBlocks with a hierarchical layout of 3+4+6+3. For a standard BasicBlock, the main branch contains two BN layers (bn1 and bn2), providing 2 stat hooks. For the first block of Layer 2 to Layer 4, where spatial downsampling occurs, an additional projection BN layer is appended to the shortcut branch, yielding 3 stat hooks each. Furthermore, the stem bn1 at the network entrance is treated as an independent hook in the global (all-stat) mode. Overall, ResNet-34 comprises 36 BN stat hooks, as detailed in Table S6.

To ensure a fair comparison during the preliminary block-wise diagnosis, we employ a non-cumulative constraint: when analyzing a specific block, we exclusively activate the two stat (bn1 and bn2) hooks from its main branch, regardless of whether it contains a shortcut branch, while disabling constraints from all other layers. Isolated nodes, such as the stem bn1, are also excluded from this single-block evalua-

Network Part	# Blocks	Channels (C_l)	Hooks per Block	Subtotal Hooks
Stem	-	64	1	1
Layer 1	3	64	2, 2, 2	6
Layer 2	4	128	3, 2, 2, 2	9
Layer 3	6	256	3, 2, 2, 2, 2, 2	13
Layer 4	3	512	3, 2, 2	7
Total	16	-	-	36

Table S6: Distribution of BN-stat hooks in ResNet-34.

tion. Conversely, for the global all-stat experiment, we adhere to the standard baseline protocol by utilizing all 36 hooks.

For a given BN input activation $A_l \in \mathbb{R}^{B \times C_l \times H_l \times W_l}$, we calculate the channel-wise moments across the batch and spatial dimensions:

$$\begin{aligned}\hat{\mu}_{l,c} &= \frac{1}{BH_lW_l} \sum_{b=1}^B \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} A_{l,b,c,h,w} \\ \hat{\sigma}_{l,c}^2 &= \frac{1}{BH_lW_l} \sum_{b=1}^B \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} (A_{l,b,c,h,w} - \hat{\mu}_{l,c})^2.\end{aligned}\quad (\text{S4})$$

These statistics are directly aligned with the BN running moments saved in the teacher model’s checkpoint. The matching function for the ResNet architecture is directly applied to the raw variance, formulated as follows:

$$\mathcal{L}_{\text{stat}}^{\text{BN}} = \sum_{l \in \mathcal{S}} \left(|\hat{\mu}_l - \mu_l^{\text{BN}}|_2 + \left| \hat{\sigma}_l^2 - (\sigma_l^{\text{BN}})^2 \right|_2 \right) \quad (\text{S5})$$

where $\mathcal{S}$ denotes the set of currently activated stat hooks (e.g., all 36 hooks in the global matching mode, or a specific subset during the block-wise diagnostic experiments).

Statistic Matching in ViT-S. Unlike BN, LN does not track running statistics during training. Therefore, we utilize real data to pre-compute the reference moments offline. Specifically, the pre-trained ViT-S teacher model is set to eval() mode, and forward propagation is performed on the CIFAR-10 training set (50,000 images). Given an input resolution of 224×224 and a patch size of 16, a single image yields 197 tokens (196 patch tokens and 1 CLS token). We register a hook at each LN layer to intercept its input features $A_l \in \mathbb{R}^{B \times N \times D}$ and flatten them into $A_l^{\text{flat}} \in \mathbb{R}^{(BN) \times D}$. The population mean and variance are computed along the sample and token dimensions:

$$\begin{aligned}\mu_{l,d}^{\text{real}} &= \frac{1}{BN} \sum_{b=1}^B \sum_{n=1}^N A_{l,b,n,d} \\ v_{l,d}^{\text{real}} &= \frac{1}{BN} \sum_{b=1}^B \sum_{n=1}^N (A_{l,b,n,d} - \mu_{l,d}^{\text{real}})^2.\end{aligned}\quad (\text{S6})$$

ViT-S contains 12 Transformer blocks, and each block corresponds to 2 stat hooks (norm1 and norm2), totaling 24 hooks. Including the final norm layer, the global mode comprises 25 LN hooks.

Similarly, we apply a strict non-cumulative constraint for ViT-S: when analyzing a specific Transformer block, we exclusively activate its two corresponding stat hooks (norm1

and `norm2`) while disabling constraints from all other layers. The isolated `final_norm` layer is consequently excluded from this single-block evaluation. Conversely, for the global all-stat experiment, we use all 25 LN hooks.

Consistent with the BN alignment approach in ResNet, we directly match the raw variance and formulate the statistical loss for ViT-S as follows:

$$\mathcal{L}_{\text{stat}}^{\text{LN}} = \sum_{l \in \mathcal{S}} (|\hat{\mu}_l - \mu_l^{\text{real}}|_2 + |\hat{\sigma}_l^2 - (\sigma_l^{\text{real}})^2|_2). \quad (\text{S7})$$

Extended Discussions and Limitations

On the spatial prior of SSA. The center-biased Gaussian in SSA is a simple prior and may not hold for cluttered scenes, multi-object images, or non-object-centric domains such as pathology, remote sensing, and industrial inspection. Nevertheless, the contribution of SSA is not tied to this particular form. Our analysis suggests that the effectiveness of BN priors in DFKD primarily comes from constraining semantic distributions rather than from normalization itself, while CDSA provides an architecture-agnostic interface for imposing such constraints. The Gaussian is therefore only an instantiation suited to object-centric benchmarks. Other domains may adopt a flat prior, a multi-modal prior, or a prior estimated from teacher attribution statistics. Moreover, SSA is applied as a batch-level regularizer weighted by λ_{SSA} , rather than as a hard per-sample constraint, and its performance remains stable over a broad range of prior parameters. We thus regard the semantic prior interface, rather than the center-biased Gaussian, as the transferable component.

Relation to BN-free model inversion and data-free quantization. BN-free model inversion has also been explored in data-free quantization, often through transformer-specific objectives based on patch similarity or attention (Li et al. 2022; Ramachandran, Kundu, and Krishna 2024; Zhong et al. 2025). These methods differ from ours in two aspects. First, their priors depend on transformer-internal quantities and generally do not transfer to convolutional teachers, whereas CDSA abstracts a shared class-discriminative spatial response and therefore supports both BN- and LN-based architectures. Second, quantization calibration mainly requires representative activation ranges, while DFKD must recover sufficiently rich decision knowledge to train a student from random initialization. Priors effective for calibration are therefore not necessarily sufficient for distillation. Our work focuses on the stronger requirement of constructing an architecture-agnostic prior for knowledge transfer.

Relation to diffusion-based synthesis. Our language-conditioned synthesis is related to methods that generate training data with text-to-image diffusion models (Li et al. 2024a; Qi et al. 2025; He et al. 2025; Li et al. 2024b), but the two settings differ in their information assumptions. We consider an auxiliary module data-free only when it cannot independently generate target-class images. A frozen text encoder satisfies this criterion because it provides only categorical geometry, while the image evidence must still be recovered from the teacher. In contrast, a pretrained diffusion model already contains a strong natural-image prior and

can generate class-conditional samples with limited dependence on the teacher. Such approaches are valid, but effectively transfer knowledge from an external generative model under a stronger assumption than standard DFKD. We further show that replacing the vision-language encoder with a vision-agnostic language encoder retains most of the improvement (Sec.), indicating that CSC mainly contributes categorical structure rather than external visual knowledge.

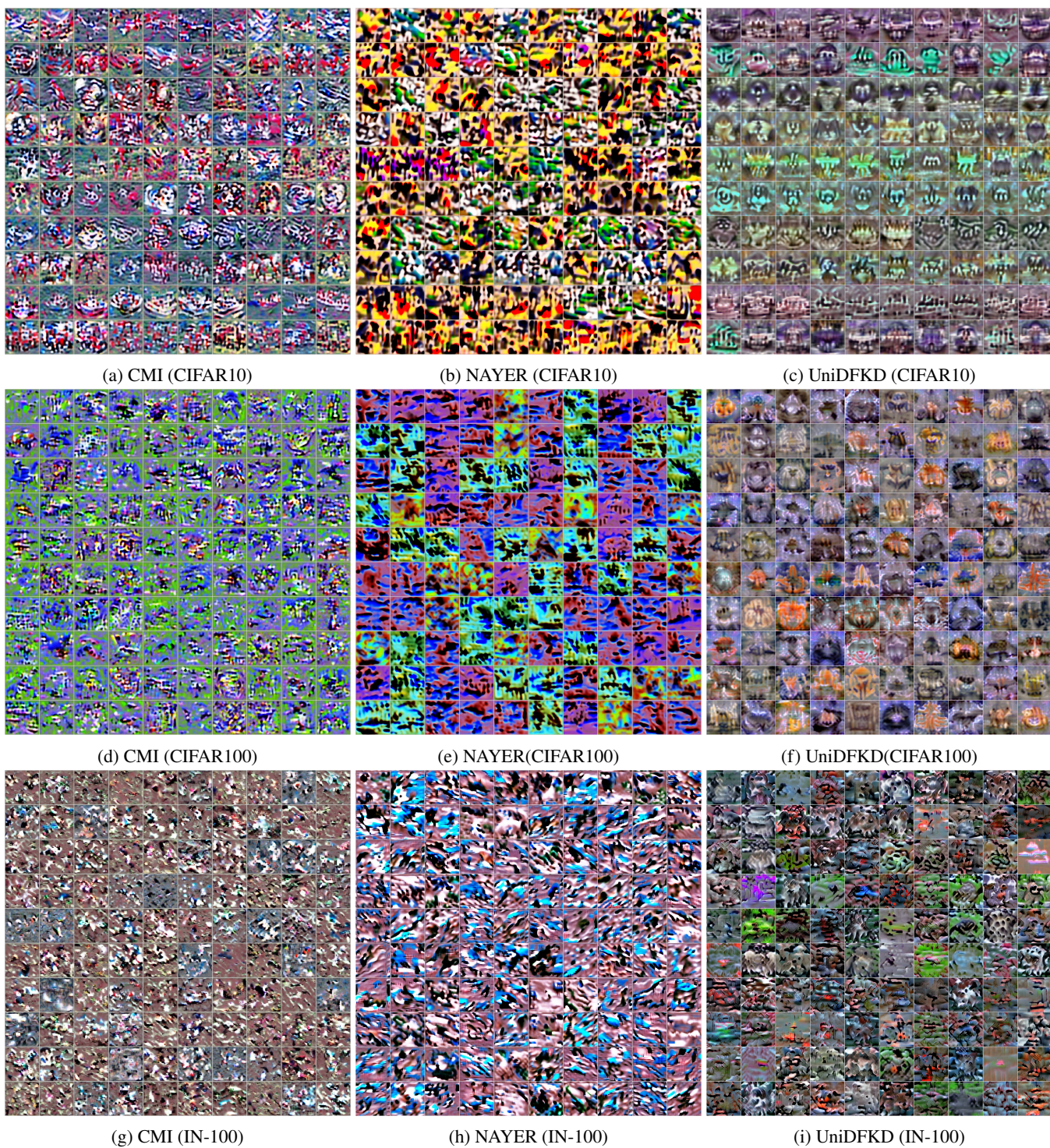


Figure S3: Representative synthetic images generated using ViT-S as the teacher on CIFAR-10, CIFAR-100, and ImageNet-100. Rows correspond to datasets, and columns compare CMI, NAYER, and UniDFKD.