TRAINING-FREE FEW-SHOT ADAPTATION VIA EXPONENTIAL TILTING OF LATENT REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Adapting frozen foundation models to new few-shot tasks is often infeasible when parameters cannot be modified e.g., under shared checkpoints or deployment policies barring mutation. We study few-shot classification under this fully frozen-model regime: no gradients, no parameter updates, no upstream data. We show that under these constraints, adaptation reduces to choosing a weighting over the support embeddings used to form class prototypes, and that exponential (Gibbs) tilting is the unique such weighting minimizing KL divergence from the empirical support distribution subject to a task-relevance constraint. We instantiate this constraint with simple support-set scores label consistency, geometric alignment, or classifier confidence yielding closed-form tilted prototypes with no backward pass or iterative solving. Across three frozen encoders (DINOv3, I-JEPA, CLIP) and several few-shot and cross-domain benchmarks, tilted prototypes improve over uniform-average baselines in most settings and are competitive with training-free adaptation methods that use strictly more information (query streams, upstream statistics), while gradient-based fine-tuning remains an upper bound under weaker constraints. This positions principled latent-distribution reweighting as a lightweight, gradient-free alternative to parameter-based few-shot adaptation.

1 INTRODUCTION

The standard paradigm for adapting pretrained models to downstream tasks assumes that some component of the model remains modifiable at prediction time, e.g. full fine-tuning, adapter, or possibly learnable prompting. This assumption is less convenient in practice. For example, shared serving infrastructures expose a single frozen checkpoint to heterogeneous downstream distributions simultaneously. Or at deployment time, regulation may warrant model immutability. These, combined with source-free adaptation, expose the statistical problem of distribution shift between pretraining and prediction/deployment while the standard solution space of gradient-based parameter updates is unavailable. We posit that adaptation is achievable by correcting for this distribution shift at the level of the induced latent distribution, without any parameter update.

We formalize this view as a change-of-measure problem. Given a frozen encoder f_θ and a small labeled support set $\mathcal{SS}$, we seek a reweighted distribution over latent representations under which the encoder's expected behavior aligns with the downstream task. Among all measures consistent with task-relevance statistics derived from $\mathcal{SS}$, exponential (Gibbs) tilting offers the KL-optimal solution. The resulting adapted prediction is then the standard decision rule evaluated under this tilted measure, circumventing optimization altogether as shown in figure 1.

While exponential tilting and KL minimization are classical tools from information geometry Csiszár (1975), their application to frozen-model test-time adaptation has not, to our knowledge, been previously proposed. Existing importance-weighting methods address covariate shift during training Shimodaira (2000), and existing TTA methods either update parameters or operate at the logit level. The adaptation problem under frozen-model constraints is a KL-minimization problem, and exponential tilting is its unique analytic solution. To see why, consider what adaptation can mean when θ is frozen. The only degrees of freedom available are the weights assigned to the n support embeddings $\{z_i\}_{i=1}^n$ when forming the prediction $p_\lambda(y) = \sum_i w_i\, p_0(y \mid z_i)$. Any adaptation method in this regime, including k-NN, prototype refinement, label propagation, and temperature scaling,

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.

implicitly defines such a weighting. Among all distributions P over $\{z_i\}$ that enforce consistency with the downstream task (=a moment constraint on the score function), the one that deviates least from the pretrained reference measure P_0 is a KL projection, and its unique solution is the Gibbs measure (Theorem $A.1$). Existing methods either (a) solve an approximation to this problem without recognizing its exact form, or (b) operate outside the constraint set by updating parameters.In summary, we formulate *frozen-model test-time adaptation as a change-of-measure problem over the induced latent distribution*, and proposing exponential tilting as the principled solution.

Parameter updates retain an empirical advantage where applicable Table 2. Our goal for comparison is against the null adaptation baseline under identical operational constraints. Within that constrained setting, exponential tilting constitutes a principled and, as we demonstrate, empirically non-trivial inference-time solution while being independent of architecture, training objective, and optimization details.

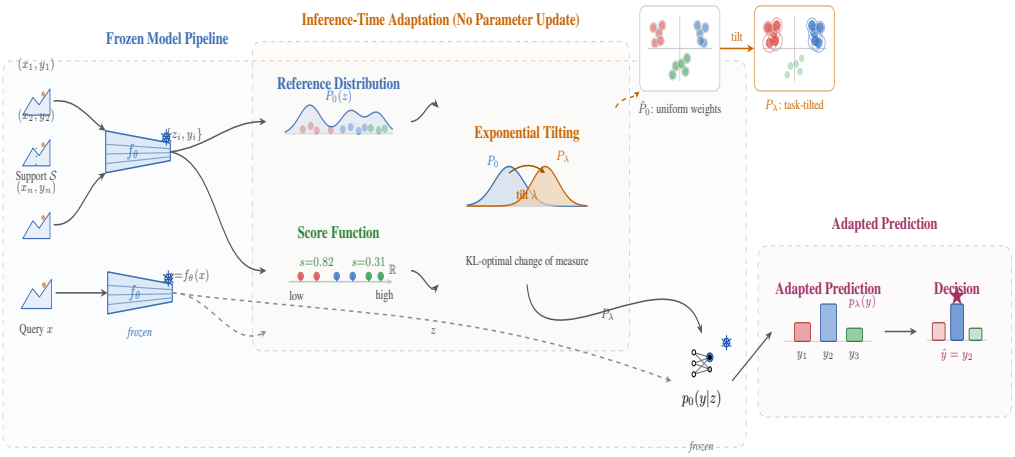


Figure 1: Inference-time adaptation as a change of measure. With a frozen encoder f_θ and classifier $p_0(y \mid z)$, the support set $\mathcal{SS}$ induces a reference latent distribution $\hat{P}_0$. Exponential tilting guided by a task-relevant score function produces a reweighted distribution P_λ the KL-optimal measure that aligns the model's behaviour with the downstream task. The final prediction $\hat{y}$ is the $\arg\max$ of the expected posterior under P_λ. The entire pipeline operates `without gradient updates` or access to source data.

Contributions

1) We identify frozen-model few-shot adaptation as a change-of-measure problem and show that exponential tilting is its KL-optimal solution.
2) We derive task-similarity score functions from a labeled support set and show how they parametrize the tilted measure without any access to upstream data.

2 Related Work

Parameter-updating test-time adaptation. Most TTA methods adapt at inference by updating parameters: self-supervised auxiliary losses Sun et al. (2020), entropy minimization on normalization/affine parameters Wang et al. (2020); Niu et al. (2022), augmentation-consistency objectives Zhang et al. (2022), continual self-training Wang et al. (2024b), test-time-tuned prompts Zhou et al. (2022); Khattak et al. (2023); Shu et al. (2022), or low-rank/adapter updates Hu et al. (2022). All require gradients or mutable state, and entropy-based variants are known to accumulate errors under small batches or strong shift Niu et al. (2022). These fall outside the frozen-model regime studied here and appear only as reference points (Appendix A).

Training-free and backpropagation-free adaptation. A smaller line adapts without gradients: cache/retrieval methods that combine similarity-weighted stored features with the frozen model's output (Tip-Adapter Zhang et al. (2021), TDA Karmanov et al. (2024), retrieval with upstream data Han et al. (2025)); classifier-adjustment methods that update prototypes from a test stream (T3A Iwasawa & Matsuo (2021), NEO Murphy et al. (2025), the optimization-free component of OFTTA Wang et al. (2024a)); online distribution-statistics estimation (DOTA Han et al. (2024)); and Gaussian class-conditional alignment relying on CLIP priors and a cross-sample knowledge bank (ADAPT Zhang et al. (2025)). Each requires a test stream, upstream data, or a text encoder, or uses a heuristic update rule (Table 2). We use only a labeled support set and a vision encoder, and derive the update as the solution of a KL-constrained problem (Eq. 3).

Few-shot learning with frozen encoders. Prototype/metric methods classify by distance to support-set class representatives Snell et al. (2017), and simple baselines on strong pretrained embeddings rival episodic meta-learners Chen et al. (2019); Tian et al. (2020). Transductive extensions exploit unlabeled queries via prototype refinement, label propagation, information maximization, or fine-tuning Ren et al. (2018); Zhu & Koniusz (2023); Boudiaf et al. (2020); Dhillon et al. (2020) typically modifying the decision rule or running an iterative/optimization procedure per episode. We keep the classifier fixed and apply a single closed-form reweighting of the latent support distribution; our transductive variant only augments the reference measure with pseudo-labeled queries (Section 4.2), with no iteration.

Post-hoc calibration and prior correction. Temperature scaling rescales confidence without task-specific structure Guo et al. (2017) and degrades under dataset shift Ovadia et al. (2019); label-shift correction reweights posteriors by estimated class priors Saerens et al. (2002); Lipton et al. (2018). Both act on logits/probabilities. We instead reweight the latent-embedding distribution over which predictions are aggregated, and recover label-shift correction as a special case (Corollary .5).

Importance weighting and exponential tilting. Importance weighting corrects covariate shift during training Shimodaira (2000); exponential tilting is the KL/I-projection solution under moment constraints Csiszár (1975); Cover (1999) and the maximum-entropy distribution consistent with them Jaynes (1957). In learning it has modified training objectives (tilted ERM Li et al. (2020)) and defined uncertainty sets in DRO Duchi et al. (2021) both acting on losses or worst-case distributions during training. To our knowledge it has not been used to reweight a frozen encoder's latent distribution at inference. On a finite support set the tilted weights reduce to a softmax over scores, structurally related to cache models' exponentiated-affinity weights Zhang et al. (2021) and attention-style aggregation Snell et al. (2017). Our contribution is the change-of-measure derivation of these weights from a KL-projection objective, the support-set score functions, and evaluation under strict frozen-model constraints.

3 METHODOLOGY

We present a training-free test-time adaptation method that operates by reweighting latent representations induced by a frozen encoder. We first formalize the problem setting and constraints, then cast adaptation as a change of measure in latent space, followed by score construction and prediction under the resulting tilted distribution. Throughout, we emphasize that adaptation occurs solely through distributional reweighting and does not modify model parameters, logits, or decision rules.

Problem Setting and Constraints. We consider a frozen encoder f and a frozen classifier head $p_0(y \mid z)$. Given an input image $x \in \mathcal{X}$, the encoder produces a latent representation $z = f(x), f$ fixed. The classifier head defines a conditional distribution over labels given the latent representation, $p_0(y \mid z), p_0$ fixed.

At test time, we are given a small labeled support set $\mathcal{D}_{\text{few}} = \{(x_i, y_i)\}_{i=1}^{n}, \qquad n \ll 100$, drawn from a downstream task distribution that may differ from the pretraining distribution.

The study operates under several constraints. The encoder parameters and the classifier head are both frozen, and no gradient updates or optimization occur at test time. Only a small number of test examples are available, corresponding to a low- or few-shot setting. Additionally, there is no access

to upstream data or any of their statistics. The goal is to adapt predictions on downstream query examples using only $\mathcal{D}_{\text{few}}$, while respecting these constraints.

Latent Distributions Induced by Frozen Encoders. Although the encoder f is deterministic, applying it to a dataset induces a distribution over latent representations. Given a data distribution $P(x)$, the encoder induces a pushforward measure $P(z)$ defined by $z = f(x)$. For a finite dataset, this corresponds to the empirical distribution

$$\hat{P}(z) = \frac{1}{n} \sum_{i=1}^{n} \delta(z - f(x_i)), \tag{1}$$

where $\delta(\cdot)$ denotes the Dirac delta function.

Model predictions depend on expectations taken with respect to such induced latent distributions. When the downstream task differs from pretraining, the induced distribution over latent space changes even though the model parameters remain fixed. This motivates adapting predictions by modifying the distribution over latent representations, rather than the model itself.

Gibbs Tilting as Domain Adaptation. Exponential tilting arises from an operational constraint, not convenience. In the label-aware case, let $s(z_i) = \log p_0(y_i \mid z_i)$. We impose

$$\mathbb{E}_{z \sim P}[s(z)] = \frac{1}{n} \sum_{i=1}^{n} \log p_0(y_i \mid z_i), \tag{2}$$

i.e., P matches the empirical average log-likelihood on the support set, downweighting label-inconsistent regions and upweighting task-consistent ones. By exponential family theory Wainwright & Jordan (2008), this moment constraint uniquely yields the maximum-entropy (Gibbs) distribution. Thus, exponential tilting gives the minimum-deviation adaptation of P_0 consistent with supervision Csiszár (1975); Cover (1999); Duchi et al. (2021).

Adaptation as a KL-Optimal Change of Measure. When both the encoder and classifier are frozen, test-time adaptation can only operate by modifying how latent representations are aggregated at inference. This corresponds to selecting a new probability measure $P(z)$ over latent space, while preserving absolute continuity with respect to a reference measure $P_0(z)$ induced by the frozen encoder. We formalize adaptation as a constrained inference problem. Given a task-relevant score function $s(z)$ constructed from the few-shot support set, we seek a distribution $P(z)$ that enforces consistency with downstream supervision while remaining maximally faithful to frozen inference. This leads to the following variational formulation:

$$P^\star = \arg\min_{P} \ \mathrm{KL}(P \,\|\, P_0) \quad \text{s.t.} \quad \mathbb{E}_{z \sim P}[s(z)] = c, \tag{3}$$

where c is a task-dependent constraint determined by the support set.

The solution to equation 3 is given by an exponentially tilted distribution

$$P_\lambda(z) = \frac{P_0(z) \exp(\lambda s(z))}{Z(\lambda)}, \qquad Z(\lambda) = \int P_0(z) \exp(\lambda s(z))\, dz, \tag{4}$$

where $\lambda \geq 0$ is the Lagrange multiplier associated with the expectation constraint. This distribution is the *unique* measure that satisfies the task constraint while inducing the smallest possible deviation from P_0 in Kullback-Leibler divergence Csiszár (1975); Cover (1999).

This KL-optimality endows exponential tilting with a clear operational meaning in our setting. Rather than heuristically reweighting embeddings, tilting implements the least-invasive inference-time adaptation compatible with downstream supervision. Any alternative reweighting that enforces the same task constraint would necessarily induce a larger divergence from frozen inference. In practice, we take P_0 to be the empirical distribution induced by the few-shot support embeddings, ensuring that no assumptions are made about upstream data or latent geometry beyond what is observed at test time. Exponential tilting then reallocates probability mass toward task-consistent regions of latent space, without modifying encoder parameters, classifier weights, or decision rules. This perspective provides a unifying interpretation of test-time adaptation under frozen-model constraints, adaptation is achieved not by learning new representations, but by performing KL-optimal inference under a task-modified latent measure.

Score Construction from Few-Shot Data. Exponential tilting provides the mechanism for reweighting, but it requires a scalar score function defined over latent representations. This score determines which embeddings should receive higher or lower probability mass under the adapted measure.

Few-shot data provide sparse constraints on which latent regions are associated with correct predictions. The role of the score is to encode these constraints in a form suitable for exponential tilting. The score function $s(z)$ injects task information from the few-shot support set into the latent distribution. We consider two simple and complementary instantiations.

Label-aware score. For a labeled support example (x_i, y_i) with embedding $z_i = f(x_i)$, we define

$$s_{\text{label}}(z_i) = \log p_0(y_i \mid z_i). \tag{5}$$

This score increases the weight of embeddings that are already consistent with the observed labels under the frozen classifier. It can be viewed as a likelihood-based reweighting that emphasizes label-consistent regions of latent space.

Geometry-aware score. We also extend the idea to regimes where labels may not be available at all, or may be available too sparsely at inference time to estimate a class-conditioned statistic. Pretrained foundation models are known to organize data into semantically meaningful geometries, where task-relevant information is reflected in directions of variation, local neighborhoods, and class-aligned submanifolds Assran et al. (2023). The score estimates which regions of the representation space are over- or under-represented under the downstream distribution relative to the pretrained reference measure. That is, it is a likelihood surrogate derived from latent geometry instead of label conditioning.

Let $\hat{\mu}$ and $\hat{\Sigma}$ denote the empirical mean and covariance of the support embeddings $\{z_i\}_{i=1}^n$. We define

$$s_{\text{geom}}(z) = \log \frac{\mathcal{N}(z; \hat{\mu}, \hat{\Sigma})}{\mathcal{N}(z; 0, I)}. \tag{6}$$

This score emphasizes embeddings that align with the geometry of the support set relative to a simple reference distribution. To ensure stability in low-shot regimes, the covariance estimate may be regularized; we avoid complex density estimation or high-capacity modeling.

Confidence-aware score. When neither reliable labels nor a well-conditioned support covariance are available (e.g., very low shot counts where $\hat{\Sigma}$ is poorly estimated), we consider a simpler score requiring no label information beyond what the frozen classifier already outputs:

$$s_{\text{conf}}(z) = \max_y p_0(y \mid z). \tag{7}$$

This score upweights embeddings the frozen classifier is already confident about, independent of whether that confidence agrees with the true label. It is a strictly weaker signal than s_{label} (which uses the ground-truth label y_i), but requires no covariance estimation, making it the most robust choice in extreme low-shot regimes.

The label-aware and geometry-aware scores can be used individually or combined additively to form a composite task-relevant score.

Prediction Under the Tilted Distribution. We now formalize how predictions are computed once the latent measure has been tilted, and show that standard prediction rules admit a natural and deterministic extension under the adapted distribution.

Given the tilted distribution $P_\lambda(z)$, predictions are formed by taking expectations of the frozen classifier,

$$p_\lambda(y) = \mathbb{E}_{z \sim P_\lambda(z)} \left[p_0(y \mid z) \right]. \tag{8}$$

The final prediction for a query example is obtained via the standard decision rule

$$\hat{y} = \arg\max_y p_\lambda(y). \tag{9}$$

Because $p_\lambda(y)$ above is defined over the support set alone, it does not by itself distinguish between queries. In practice, this marginal is realized as a set of *tilted class prototypes* $\tilde{\mu}_c = \sum_i \tilde{w}_{i,c}\, z_i$, where $\tilde{w}_{i,c} \propto w_i\, p_0(c \mid z_i)$ and $w_i \propto \exp(\lambda s(z_i))$ are the global tilting weights of Corollary .2. A query embedding $z = f(x)$ is then classified against $\{\tilde{\mu}_c\}$ using the same fixed cosine-softmax rule $p_0(\cdot \mid z)$ as the frozen baseline. This is consistent with the "Tilted (Inductive)" description in Section 4.2 and is the exact rule implemented in the accompanying notebook.

The classifier head and decision rule remain unchanged. The same tilted distribution P_λ, constructed from the support set, is used for all queries within an episode. Adaptation therefore operates at the level of the latent distribution, rather than on a per-query parameter or score update.

Concretely, tilting weights $w_i \propto \exp(\lambda s(z_i))$ (Corollary .2) are computed once over the *entire* support set $\{z_i\}_{i=1}^n$, not per class. The adapted class marginal is then, exactly as in Theorem .3,

$$p_\lambda(y = c) = \sum_{i=1}^{n} w_i\, p_0(c \mid z_i), \qquad \hat{y} = \arg\max_c\, p_\lambda(y = c). \tag{10}$$

Because w_i is shared across classes and each term is weighted by the frozen classifier's soft posterior $p_0(c \mid z_i)$, every support embedding contributes to every class's score in proportion to how consistent it is with that class under p_0. This is what allows adaptation to be non-trivial even at $K{=}1$: although each class has only one labeled example, its predicted score is a weighted combination over all $n = NK$ support embeddings (across all classes), not merely a within-class average, so cross-class embeddings that are highly consistent with class c under the frozen classifier still shift $p_\lambda(y = c)$. Tilting reduces to a no-op only in the degenerate case $n = 1$, not $K = 1$.

Practical Approximation. Expectations under P_λ reduce to weighted sums over the finite support set (Corollary .2), requiring one score evaluation and one weighted prediction per example with no backward pass, adding negligible overhead relative to standard frozen-model inference.

4 Experiments

We evaluate our method in few-shot image classification setting on publicaly available benchmarking datasets. Our experiments are designed to answer the following questions:

i. Can distributional reweighting alone improve few-shot performance under strict no-adaptation constraints?

ii. How does exponential tilting compare to standard frozen inference? and

iii. What is the effect of inductive versus transductive reference measures?

4.1 Experimental Setup

Datasets. We evaluate our method on standard few-shot classification benchmarks and cross-domain benchmarks (details in Appendix A).

Encoders. We evaluate three frozen visual encoders without fine-tuning: **DINOv3 ViT-S/16** Siméoni et al. (2025), a self-supervised vision encoder pretrained on a curated web-scale corpus; **I-JEPA ViT-H/14 Assran et al. (2023)**, a joint embedding predictive vision encoder with 630M parameters; and **CLIP ViT-B/16 Radford et al. (2021)**, a vision-language encoder pretrained via contrastive image-text supervision. For I-JEPA, we use the officially released ViT-H/14 checkpoint, as no public ViT-B/16 checkpoint is available. All encoders remain frozen throughout evaluation. Inference is performed in evaluation mode with gradients disabled.

Classifier. We use a standard nearest-prototype classifier with cosine similarity and a fixed temperature. Class prototypes are the mean of support-set embeddings. No linear classifier, bias, or calibration parameters are trained. Prototype computation occurs per episode but involves no learning, and the classifier, similarity metric, and temperature remain fixed across tasks, consistent with standard few-shot evaluation protocols.

Table 1: Few-shot classification accuracy (%) on 5-way tasks. All models use frozen encoders and classifiers. For exponential tilting, we report accuracy at the fixed default $\lambda = 0.25$ recommended in Section 4 (*Practical Selection of* λ); no per-episode or per-shot tuning is performed on test data. Improvements arise solely from latent distribution reweighting.

Dataset	Method	1-shot	2-shot	4-shot	8-shot	16-shot
DINOv3						
CIFAR-FS	Baseline	82.32 ± 7.90	90.87 ± 6.57	91.03 ± 4.14	95.81 ± 2.55	96.29 ± 2.38
	Tilted (Ours)	$\mathbf{84.69 \pm 7.92}$	90.80 ± 6.64	$\mathbf{94.85 \pm 4.44}$	$\mathbf{96.55 \pm 2.87}$	$\mathbf{96.93 \pm 2.74}$
CUB	Baseline	85.17 ± 9.30	91.16 ± 6.50	94.89 ± 3.91	96.31 ± 3.05	96.81 ± 2.45
	Tilted (Ours)	$\mathbf{86.19 \pm 9.50}$	$\mathbf{91.81 \pm 7.88}$	$\mathbf{95.99 \pm 7.30}$	$\mathbf{96.79 \pm 6.08}$	$\mathbf{97.13 \pm 5.42}$
mini-ImageNet	Baseline	87.33 ± 8.23	91.99 ± 6.56	93.31 ± 6.84	95.82 ± 5.95	97.13 ± 5.25
	Tilted (Ours)	$\mathbf{88.93 \pm 8.10}$	$\mathbf{93.45 \pm 7.00}$	$\mathbf{96.10 \pm 6.44}$	$\mathbf{97.52 \pm 5.88}$	$\mathbf{98.04 \pm 5.11}$
JEPA						
CIFAR-FS	Baseline	59.17 ± 12.98	65.99 ± 10.26	70.46 ± 8.98	72.55 ± 8.62	74.23 ± 7.48
	Tilted (Ours)	$\mathbf{60.21 \pm 12.91}$	$\mathbf{67.68 \pm 10.54}$	$\mathbf{73.01 \pm 9.31}$	$\mathbf{75.12 \pm 9.02}$	$\mathbf{77.79 \pm 8.03}$
CUB	Baseline	38.14 ± 9.75	43.68 ± 9.62	48.97 ± 9.82	54.36 ± 10.26	57.26 ± 9.84
	Tilted (Ours)	$\mathbf{39.03 \pm 9.84}$	$\mathbf{45.12 \pm 9.48}$	$\mathbf{50.02 \pm 10.21}$	$\mathbf{56.53 \pm 10.04}$	$\mathbf{59.04 \pm 9.71}$
mini-ImageNet	Baseline	74.92 ± 9.93	82.61 ± 7.41	87.70 ± 5.50	89.50 ± 4.63	90.84 ± 3.86
	Tilted (Ours)	$\mathbf{76.12 \pm 9.91}$	$\mathbf{84.95 \pm 7.36}$	$\mathbf{91.04 \pm 5.72}$	$\mathbf{92.62 \pm 4.91}$	$\mathbf{94.18 \pm 3.98}$

Table 2: Comparison of training-free test-time adaptation methods on CIFAR-FS dataset under strict frozen-model constraints (DINOv3 ViT-S/16). Our method operates without parameter updates, gradients, or access to upstream data, while consistently improving over non-adaptive baselines.

Method	Params?	Grads?	Opt?	Stream?	Upstream?	Accuracy (%)		
						1-shot	4-shot	8-shot
Training-Free, Constraint-Compliant Methods								
Frozen Baseline (reproduced here)	×	×	×	×	×	82.32 ± 7.90	91.03 ± 4.14	95.81 ± 2.55
k-NN in Embeddings Avdiukhin et al. (2024)	×	×	×	×	×	79.77 ± 8.74	93.77 ± 4.31	96.32 ± 2.58
Temp. Scaling (TS) Guo et al. (2017)	×	✓	✓	×	×	78.77 ± 9.12	93.05 ± 4.67	95.21 ± 2.91
Label Propagation Zhu & Koniusz (2023)	×	×	×	✓	×	76.04 ± 10.21	90.02 ± 5.13	92.64 ± 3.47
Ranked Han et al. (2025)	×	×	×	×	✓	81.05 ± 8.63	92.95 ± 4.52	95.39 ± 2.76
DOTA (adapted) Han et al. (2024)	×	×	×	✓	×	82.63 ± 8.21	93.84 ± 4.38	95.36 ± 2.69
T3A Iwasawa & Matsuo (2021)	×	×	×	✓	×	83.01 ± 8.04	93.77 ± 4.41	95.39 ± 2.71
NEO Murphy et al. (2025)	×	×	×	✓	×	78.93 ± 9.35	92.75 ± 4.63	93.51 ± 3.18
OFTTA Wang et al. (2024a)	×	×	×	✓	×	82.97 ± 8.17	93.61 ± 4.29	96.28 ± 2.49
Ours (Inductive)	×	×	×	×	×	84.69 ± 7.92	94.85 ± 4.44	96.55 ± 2.87
Ours (Transductive)	×	×	×	✓	×	86.37 ± 7.58	97.21 ± 3.41	97.43 ± 2.96

4.2 Baselines and Adaptation Variants

Frozen Baseline. Standard frozen inference using uniformly weighted class prototypes computed from the support set. No adaptation is applied.

Tilted (Inductive). Exponential tilting is applied to the empirical latent distribution induced by the labeled support embeddings only. Class prototypes are recomputed as weighted averages under the tilted distribution.

Tilted (Transductive). The reference latent distribution is augmented with unlabeled query embeddings via pseudo-labeling. Each query embedding $z_j = f(x_j)$ is assigned a label $\tilde{y}_j = \arg\max_y p_0(y \mid z_j)$ in a single forward pass using the frozen baseline classifier. These pseudo-labeled queries are combined with support embeddings to form the augmented reference distribution $\hat{P}_0(z) = \frac{1}{n+m}\Big(\sum_{i=1}^{n} \delta(z - z_i) + \sum_{j=1}^{m} \delta(z - z_j)\Big)$, where n and m are the numbers of support and query examples. Exponential tilting is applied once with the same score function as the inductive variant. No optimization, gradient updates, or iterative refinement occurs; query embeddings remain fixed, leveraging test-time latent structure to yield a more representative reference distribution.

Methods in Comparison. We compare our method against training-free baselines that operate under frozen-model constraints, as summarized in Table 2. These include:

i. A frozen Prototypical Network baseline using cosine similarity Snell et al. (2017).

ii. k-nearest neighbor classification in the frozen embedding space Avdiukhin et al. (2024).

iii. Post-hoc temperature scaling applied to frozen classifier logits Guo et al. (2017).

iv. Label propagation and prototype refinement methods that leverage unlabeled query embeddings without gradient-based optimization Ren et al. (2018); Zhu & Koniusz (2023).

v. Distributional or retrieval-based training-free adaptation methods that assume access to upstream data or statistics Han et al. (2025; 2024).

vi. Backprop-free classifier-adjustment baselines: T3A Iwasawa & Matsuo (2021), NEO Murphy et al. (2025), and the classifier-adjustment component of OFTTA Wang et al. (2024a) (OFTTA's BatchNorm-replacement component does not apply to our LayerNorm-based ViT encoders and is omitted accordingly).

We exclude MCP and DPE because both require a text encoder to build CLIP-style vision–language prototypes or caches, and are therefore incompatible with the vision-only DINO and JEPA backbones used for the results in Table 1 and Table 2; a CLIP-backbone comparison including MCP and DPE is left to Appendix Table 9 in future revisions. ADAPT Zhang et al. (2025) is concurrent, zero-shot work relying on CLIP priors and a cross-sample knowledge bank, which differs from our labeled few-shot, auxiliary-free setting (Section 2).

Score Functions and Tilting. Unless otherwise specified, we use the label-aware score s_{label} (Section 3, "Label-aware score"), which produced the results in Table 1 and Table 2, consistent with Table 10. The confidence-aware score s_{conf} is a fallback for regimes with unreliable labels or an ill-conditioned support covariance, and is evaluated separately in the score-function ablation (Table 10). Tilting weights are computed as $w_i \propto \exp(\lambda s(z_i))$, with λ within range (0.25 - 2.00) for all experiments. The value of λ is not tuned per episode or per dataset.

Practical Selection of λ. Performance is robust across $\lambda \in [0.25, 2.0]$, with the maximum accuracy gap across this range at $+1.37\%$ at 1-shot and $+0.68\%$ at 16-shot (Table 5). We recommend $\lambda = 0.25$ as a conservative default, requiring no per-task tuning, validation data, or entropy estimation.

Evaluation Metrics We report average classification accuracy over 100 randomly sampled episodes. Results are presented as mean accuracy $\pm$ standard deviation. Since no probabilistic calibration is learned, accuracy is the primary evaluation metric.

Training-Free Test-Time Protocol. All experiments strictly follow a training-free test-time adaptation protocol. Encoder parameters remain fixed throughout evaluation, and no gradient-based optimization is performed at any stage. Inference is carried out without gradient tracking, and no backward pass or optimizer is used. Query embeddings are treated as immutable and are never altered during inference or adaptation. These constraints ensure that all reported improvements arise solely from reweighting the empirical latent distribution, without updating model parameters or modifying the underlying inference procedure. Overall, the results demonstrate that exponential tilting provides a simple and effective mechanism for few-shot adaptation under strict constraints. By operating entirely at the level of latent distributions, the method enables adaptive behavior while preserving frozen inference.

Results Table 1 reports few-shot classification accuracy across datasets, encoders, and shot numbers. Exponential tilting consistently improves performance over the frozen baseline for both DINO and JEPA encoders. The inductive variant yields stable gains across all shot regimes, while the transductive variant achieves the largest improvements, highlighting the benefit of incorporating unlabeled test embeddings through distributional reweighting. Importantly, all gains are obtained without gradient-based optimization, parameter updates, or modification of the classifier or decision rule. Performance improvements arise solely from reweighting the empirical latent distribution induced by the frozen encoder.

4.3 Ablation on Tilting Strength

Table 4 reports accuracy across $\lambda \in [0.25, 2.0]$ for all shot regimes. Performance is robust across this range, with λ=0.25 recommended as a default. Full analysis is in Appendix B.1.

5 DISCUSSION

Effectiveness of Latent Distribution Reweighting. Table 1 shows that exponential tilting consistently improves few-shot accuracy across datasets and shot settings, with one exception: DINOv3 on CIFAR-FS at 2-shot, where tilted and baseline accuracies are statistically indistinguishable (90.80 ± 6.64 vs. 90.87 ± 6.57). For DINOv3, absolute gains range from -0.07% (CIFAR-FS, 2-shot) to $+3.82\%$ (CIFAR-FS, 4-shot); I-JEPA shows comparable gains of $+0.89\%$ to $+3.56\%$ despite a lower baseline. Since the encoder, classifier, and decision rule are held fixed, these comparisons isolate the effect of latent distribution reweighting. For DINOv3, gains generally shrink as shot count increases, consistent with the empirical latent distribution becoming more representative of the task; I-JEPA does not follow this trend, instead peaking at 16-shot on CIFAR-FS and mini-ImageNet. Thus, the attenuation of gains with shot count holds for DINOv3 but is not consistent across encoders. This also explains why gains are largest at $K=1$ rather than vanishing: the tilted marginal in Theorem .3 pools information across the full N-point support set for every class, so even single-example classes benefit from cross-class reweighting, unlike prototype schemes that average only within-class examples (Snell et al., 2017).

Comparison Under Strict Frozen-Model Constraints. Table 2 situates exponential tilting among training-free test-time adaptation methods operating under the same deployment constraints. Our inductive variant outperforms every constraint-compliant baseline at every reported shot count, including methods that exploit additional structure it does not use, such as streaming or pseudo-labeled test batches (DOTA, T3A, NEO, OFTTA) or upstream statistics (Ranked). This holds despite the inductive variant using only the labeled support set, with no access to query embeddings at inference time. The transductive variant, which additionally incorporates unlabeled query information, widens this margin further, achieving the highest accuracy among constraint-compliant methods at every reported shot count.

Comparisons under relaxed adaptation constraints. We further evaluate exponential tilting under relaxed adaptation constraints, compare inductive and transductive reference measures, and examine its combination with linear-head fine-tuning. These additional experiments and detailed comparisons are provided in Appendix A Table 3.

Cross-Domain Generalization. Appendix Tables 6, 7, and 8 report cross-domain results across CIFAR, CUB, CARS, STL-10, ImageNet variants, and DomainNet. Exponential tilting consistently improves over frozen inference across all domains and shot regimes, with gains largest under severe distribution shift and lowest at 16-shot, consistent with in-domain saturation behavior.

Mechanistic Interpretation of Performance Gains. Analysis of tilting weights confirms that high-confidence, geometrically aligned embeddings receive increased mass while ambiguous ones are downweighted. see Appendix for more detail B.1.

6 CONCLUSION

We cast few-shot adaptation under fully frozen-model constraints as a KL-constrained change of measure over the support-set latent distribution, and showed that exponential tilting is its unique closed-form solution. Instantiated with label-, geometry-, or confidence-based scores, tilting yields reweighted class prototypes at negligible computational cost, with no gradients, parameter updates, or iterative solving. Across three frozen encoders and multiple few-shot and cross-domain benchmarks, tilted prototypes improve over uniform-average baselines in most settings and are competitive with training-free methods that use strictly more information, while gradient-based fine-tuning remains a natural upper bound under comparatively weaker constraints. The method's benefit is bounded by the frozen encoder's representation quality: where support-set structure is not class-separable, reweighting cannot recover it. These results indicate that distributional reweighting at inference is a viable, lightweight complement to parameter-based adaptation in settings where model mutation is unavailable or undesirable; extending the approach to richer score functions and non-stationary test distributions is left to future work.

AI USE STATEMENT

In this work, we used generative AI tools solely for grammar checking and polishing the writing of the manuscript. We did not use generative AI tools for research ideation, experimental design, code development, data analysis, or generation of results, figures, or tables. All experiments, analyses, and findings reported in this paper were conducted and produced by the authors without AI assistance. We have reviewed all AI-assisted edits to ensure they affected only language and presentation, and did not alter the technical content, claims, or results of the work. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

See the ICLR 2027 AI Policy for Authors for more details.

ETHICS STATEMENT

This work does not involve human subjects, crowdsourcing, or new data collection. All experiments use publicly available benchmark datasets (CIFAR-FS, CUB, mini-ImageNet, CARS, STL-10, CIFAR-10, CIFAR-100, ImageNet and its distribution-shifted variants, and DomainNet), which contain no personally identifiable information. We do not foresee direct harmful applications: exponential tilting is a lightweight inference-time reweighting of a frozen encoder's latent distribution and introduces no new capabilities beyond those of the underlying pretrained encoders evaluated (DINOv3 ViT-S/16, I-JEPA ViT-H/14, CLIP ViT-B/16). We have no conflicts of interest to disclose.

REPRODUCIBILITY STATEMENT

We describe our experimental setup, including encoders, datasets, shot regimes, baselines, and hyperparameters, in Section 4.1. The full derivation of exponential tilting as a KL-optimal change of measure is given in Section 3, including the constrained optimization problem (Eq. equation 3) and the resulting closed-form tilted distribution P_λ (Eq. equation 4), with the corresponding optimality proof in Theorem .1 and its empirical, finite-support form in Corollary .2. Implementation details for all baselines are given in Section 4.2 (Methods in Comparison), covering constraint-compliant methods (k-NN, temperature scaling, label propagation, Ranked, DOTA, T3A, NEO, OFTTA) and the parameter-update upper bounds (Tent, Tip-Adapter-F, FT-Linear). Per-dataset and per-domain breakdowns of all results are given in Appendix A, including cross-domain transfer (Appendix B, Table 6), ImageNet distribution shifts (Table 7), DomainNet (Table 8), the CLIP ViT-B/16 backbone (Table 9), and combination with linear-head fine-tuning (Appendix A, Table 3). Ablation results for the tilting strength λ (Appendix B.1, Table 4, Table 5) and for the choice of score function (Table 10) are also reported in Appendix A. All reported datasets (CIFAR-FS, CUB, mini-ImageNet, CARS, STL-10, CIFAR-10, CIFAR-100, ImageNet and its distribution-shifted variants, and DomainNet) are publicly available.

AUTHOR CONTRIBUTIONS

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Dmitrii Avdiukhin, Vaggos Chatziafratis, Orr Fischer, and Grigory Yaroslavtsev. Embedding dimension of contrastive learning and k-nearest neighbors. *Advances in Neural Information Processing Systems*, 37:41359–41393, 2024.

Malik Boudiaf, Imtiaz Ziko, Jérôme Rony, José Dolz, Pablo Piantanida, and Ismail Ben Ayed. Information maximization for few-shot learning. *Advances in Neural Information Processing Systems*, 33:2445–2457, 2020.

Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pp. 146–158, 1975.

Guneet Singh Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. 2020.

John C Duchi, Peter W Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Jisu Han, Jaemin Na, and Wonjun Hwang. Ranked entropy minimization for continual test-time adaptation. *arXiv preprint arXiv:2505.16441*, 2025.

Zongbo Han, Jialong Yang, Guangyu Wang, Junfan Li, Qianli Xu, Mike Zheng Shou, and Changqing Zhang. Dota: Distributional test-time adaptation of vision-language models. *arXiv preprint arXiv:2409.19375*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2427–2440, 2021.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14162–14171. IEEE, 2024.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19113–19122, 2023.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. *arXiv preprint arXiv:2007.01162*, 2020.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pp. 3122–3130. PMLR, 2018.

Alexander Murphy, Michal Danilowski, Soumyajit Chatterjee, and Abhirup Ghosh. Neo: No-optimization test-time adaptation through latent re-centering, 2025. URL `https://arxiv.org/abs/2510.05635`.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pp. 16888–16905. PMLR, 2022.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018.

Marco Saerens et al. Adjusting the outputs of a classifier to new a priori probabilities. *Neural Computation*, 2002.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. In *Journal of Statistical Planning and Inference*, 2000.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.

Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *European conference on computer vision*, pp. 266–282. Springer, 2020.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

S. Wang, J. Wang, H. Xi, B. Zhang, L. Zhang, and H. Wei. Optimization-free test-time adaptation for cross-person activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(4):1–27, January 2024a.

Yanshuo Wang, Jie Hong, Ali Cheraghian, Shafin Rahman, David Ahmedt-Aristizabal, Lars Petersson, and Mehrtash Harandi. Continual test-time domain adaptation via dynamic sample selection. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1690–1699. IEEE, 2024b.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems*, 35:38629–38642, 2022.

Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

Youjia Zhang, Youngeun Kim, Young-Geun Choi, Hongyeob Kim, Huiling Liu, and Sungeun Hong. Backpropagation-free test-time adaptation via probabilistic gaussian alignment. In *Advances in Neural Information Processing Systems*, 2025.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

Hao Zhu and Piotr Koniusz. Transductive few-shot learning with prototype-based label propagation by iterative graph refinement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23996–24006, 2023.

APPENDIX

.1 THEORETICAL PROPERTIES OF EXPONENTIAL LATENT TILTING

We present theoretical results that formalize exponential tilting as a principled, training-free mechanism for test-time adaptation under frozen-model constraints.

Theorem .1 (KL-Optimal Latent Measure Adaptation)**.** *Let P_0 be a reference probability measure over latent space $\mathcal{Z}$, and let $s : \mathcal{Z} \to \mathbb{R}$ be a measurable score function. For any $c \in \mathbb{R}$, consider the constrained optimization problem*

$$\min_{P \ll P_0} \mathrm{KL}(P \,\|\, P_0) \quad \textit{subject to} \quad \mathbb{E}_{z\sim P}[s(z)] = c. \tag{11}$$

Then the unique solution is given by the exponentially tilted distribution

$$P_\lambda(z) = \frac{P_0(z)\exp(\lambda s(z))}{Z(\lambda)}, \tag{12}$$

where $\lambda \in \mathbb{R}$ is chosen such that $\mathbb{E}_{z\sim P_\lambda}[s(z)] = c$, and $Z(\lambda) = \int P_0(z)\exp(\lambda s(z))\, dz$.

Proof. The result follows from convex duality. Define the Lagrangian

$$\mathcal{L}(P, \lambda, \nu) = \mathrm{KL}(P \,\|\, P_0) + \lambda\big(\mathbb{E}_P[s(z)] - c\big) + \nu\Big(\int dP - 1\Big). \tag{13}$$

Taking the functional derivative with respect to P and setting it to zero yields

$$\log \frac{dP}{dP_0}(z) + 1 + \lambda s(z) + \nu = 0, \tag{14}$$

which implies

$$\frac{dP}{dP_0}(z) \propto \exp(\lambda s(z)). \tag{15}$$

Normalization yields the stated form. Uniqueness follows from the strict convexity of the KL divergence. □

Corollary .2 (Empirical Latent Tilting)**.** *Let the reference measure P_0 be the empirical distribution induced by a finite support set of latent representations*

$$P_0 = \frac{1}{n}\sum_{i=1}^{n} \delta_{z_i},$$

where $z_i = f(x_i)$ are embeddings produced by a frozen encoder. Let $s : \mathcal{Z} \to \mathbb{R}$ be a real-valued score function defined on latent space, and assume that $\sum_{i=1}^{n} \exp(\lambda s(z_i)) < \infty$ for the relevant range of λ.

Then the solution to the KL-constrained optimization problem in Theorem .1 is the discrete distribution

$$P_\lambda = \sum_{i=1}^{n} w_i\, \delta_{z_i}, \qquad w_i = \frac{\exp(\lambda s(z_i))}{\sum_{j=1}^{n}\exp(\lambda s(z_j))}. \tag{16}$$

Moreover, for any measurable function $g : \mathcal{Z} \to \mathbb{R}$, expectations under the tilted distribution reduce to weighted empirical averages,

$$\mathbb{E}_{z\sim P_\lambda}[g(z)] = \sum_{i=1}^{n} w_i\, g(z_i). \tag{17}$$

Theorem .3 (Bayes-Optimal Marginal Prediction under a Tilted Latent Measure)**.** *Let $p_0(y \mid z)$ be a fixed conditional classifier and let P_λ be a probability measure over latent space $\mathcal{Z}$. Consider the class of predictors that output a marginal distribution $q(y) \in \Delta(\mathcal{Y})$, independent of z.*

Then the predictive distribution

$$p_\lambda(y) = \mathbb{E}_{z\sim P_\lambda}\left[p_0(y \mid z)\right] \tag{18}$$

uniquely minimizes the expected log-loss

$$\mathbb{E}_{z\sim P_\lambda}\left[\mathbb{E}_{y\sim p_0(\cdot|z)}[-\log q(y)]\right] \tag{19}$$

over all marginal predictors $q(y)$.

Proof. The objective can be written as

$$\mathbb{E}_{y\sim p_\lambda}[-\log q(y)],$$

where $p_\lambda(y) = \mathbb{E}_{z\sim P_\lambda}[p_0(y \mid z)]$. This is the cross-entropy between p_λ and q, which is minimized uniquely when $q = p_\lambda$. □

Proposition .4 (Empirical Form of the Tilted Measure)**.** *Let*

$$P_0(z) = \frac{1}{n}\sum_{i=1}^{n} \delta(z - z_i) \tag{20}$$

be the empirical distribution induced by latent embeddings $\{z_i\}_{i=1}^n$. Then the tilted distribution assigns weights

$$w_i = \frac{\exp(\lambda s(z_i))}{\sum_{j=1}^n \exp(\lambda s(z_j))}, \tag{21}$$

and expectations under P_λ reduce to weighted sums over the support set.

Proof. Substituting the empirical form of P_0 into the definition of P_λ yields

$$P_\lambda(z) = \frac{\sum_i \delta(z - z_i)\exp(\lambda s(z_i))}{\sum_j \exp(\lambda s(z_j))}, \tag{22}$$

from which the weights follow directly. □

Corollary .5 (Label-Shift Correction as a Special Case)**.** *Assume the score function depends only on the predicted label,*

$$s(z) = \log \pi\big(y(z)\big), \tag{23}$$

where π denotes class-prior ratios and $y(z) = \arg\max_y p_0(y \mid z)$. Then exponential tilting of $P_0(z)$ induces posterior reweighting equivalent to classical label-shift correction.

Proof. All latent representations assigned to the same predicted class receive identical scores. Exponential tilting therefore redistributes probability mass uniformly within each class, resulting in posterior rescaling by class-prior ratios, as in classical label-shift adaptation. □

Proposition .6 (First-Order Stability for Small λ)**.** *For sufficiently small λ, the tilted predictive distribution admits the expansion*

$$p_\lambda(y) = p_0(y) + \lambda \operatorname{Cov}_{P_0}\big(p_0(y \mid z), s(z)\big) + O(\lambda^2). \tag{24}$$

Proof. The result follows from a first-order Taylor expansion of the exponential weights and the normalization constant $Z(\lambda)$. □

A Experiments and Results

Datasets. We evaluate our method on standard few-shot classification benchmarks, including CIFAR-FS, CUB, and mini-ImageNet. To assess robustness under distribution shift, we additionally consider cross-domain few-shot benchmarks, where models are trained on mini-ImageNet and evaluated on CUB, Cars, and Places, as well as ImageNet and its variants (ImageNet-V2, ImageNet-S, ImageNet-A, and ImageNet-R), and DomainNet. All experiments follow the standard episodic few-shot learning protocol. In each episode, N classes are sampled from the dataset, and for each class, K labeled support examples and Q query examples are drawn. Unless otherwise specified, we use a 5-way 5-shot setting with 15 query samples per class.

Data-Driven Selection of λ. While Table 5 demonstrates that performance is robust across a wide range of λ values, practical deployment may benefit from a principled heuristic. We propose selecting λ based on the entropy of the frozen classifier's predictions over the support set,

$$\lambda^* = \frac{\kappa}{H(\hat{p})}, \qquad H(\hat{p}) = -\sum_y \hat{p}(y) \log \hat{p}(y), \tag{25}$$

where $\hat{p}(y) = \frac{1}{n}\sum_{i=1}^{n} p_0(y \mid z_i)$ is the mean predictive distribution over support embeddings and $\kappa > 0$ is a fixed scalar (we use $\kappa = 0.5$ throughout), distinct from the mixing coefficient α raised in prior review, which this formulation no longer uses (Section 3, "Prediction Under the Tilted Distribution"). Intuitively, when the frozen classifier is already confident on the support set (low entropy), modest tilting suffices; when uncertainty is high, stronger reweighting is warranted. This heuristic requires no validation data, no held-out labels, and no optimization, remaining consistent with the training-free deployment constraints of the method.

Upper Bounds and the Benefit of Frozen Inference. Parameter-update methods in Table 2 are designated as upper bounds for a precise reason: they operate under strictly weaker constraints than the proposed method. Specifically, they require gradient computation, optimizer state, and the ability to mutate model parameters at test time capabilities that are unavailable in the deployment settings this work targets. A method with access to additional degrees of freedom solving the same classification problem constitutes a natural performance ceiling for a more constrained approach. This designation is therefore a statement about the constraint hierarchy, not a claim that frozen inference is generally competitive with fine-tuning. The concrete benefits of maintaining frozen inference in the proposed approach are fourfold. First, *determinism and reproducibility*: exponential tilting is a closed-form weighted average with no stochastic optimization dynamics, batch order sensitivity, or initialization dependence, producing identical predictions for identical inputs regardless of evaluation order. Second, *stability across test batches*: gradient-based test-time methods such as TENT Wang et al. (2020) are known to suffer from error accumulation and model collapse under distribution shift or small batch sizes Niu et al. (2022), whereas exponential tilting applies a fixed deterministic reweighting with no such failure mode. Third, *computational efficiency*: adaptation requires only a single forward pass per input followed by a weighted average over n support embeddings, with no backward pass, optimizer state, or iterative convergence, making it strictly faster than any gradient-based alternative. Fourth, *constraint compliance*: in regulated, shared, or privacy-sensitive deployment infrastructures where parameter mutation is prohibited or audited, exponential tilting is the only viable adaptation mechanism among those compared, as all parameter-update baselines violate these constraints by definition.

Performance under linear head fine-tuning. In these experiments, the frozen-head assumption is relaxed while the large encoder remains fixed. Under linear-head fine-tuning, our method achieves the best performance among lightweight adaptation approaches. As Table 3 shows, combining our approach with a linear head outperforms prior optimization-based methods (e.g., TENT, Tip-Adapter-F) across all shot regimes. Notably, the gap between exponential tilting under frozen constraints and linear-head fine-tuning is much smaller than the gap to fully frozen inference. For example, in the 5-shot setting, tilting recovers a large fraction of the fine-tuning gains without gradients or parameter updates, highlighting that improvements primarily arise from latent-space distributional correction rather than representation learning.

Table 3: Comparison of adaptation methods on CIFAR-FS under the evaluated deployment constraints using a DINOv3 ViT-S/16 encoder. The table reports whether each method updates parameters, uses gradients or optimization, requires a data stream, and accesses upstream data. Parameter-update methods are included as reference methods for comparison with the training-free setting.

Method	**Params?**	**Grads?**	**Opt?**	**Stream?**	**Upstream?**	**Accuracy (%)**		
						1-shot	**5-shot**	**10-shot**
Tent Wang et al. (2020)	✓	✓	✓	✓	×	43.52	49.04	51.35
Tip-Adapter-F Zhang et al. (2021)	✓	✓	✓	×	×	83.63	94.99	95.36
FT-Linear Radford et al. (2021)	✓	✓	✓	×	×	82.49	94.92	96.67
Ours (Linear Head FT)	✓	✓	✓	×	×	85.53	95.80	97.61

When to prefer exponential tilting over linear probing. Table 2 shows that frozen exponential tilting lags linear head fine-tuning by $\sim 3\%$ across shot regimes, and whether this gap justifies linear probing depends on context. Tilting is advantageous in parameter-immutable deployments (e.g., regulated environments or model-as-a-service), extreme low-shot settings (1–2 examples per class) where linear probing is underdetermined, and latency- or compute-constrained scenarios, as it requires only a weighted average over support embeddings. Combining tilting with linear fine-tuning consistently outperforms linear probing alone, demonstrating that distributional reweighting and parameter learning are complementary at minimal cost.

A.1 Ablation and Further Analysis

We conduct a detailed ablation study on the exponential tilting strength λ to better understand the behavior, robustness, and limitations of the proposed adaptation mechanism under the strict frozen-model, training-free setting. All experiments follow the same evaluation protocol as in the main results.

Variance in Low-Shot Regimes. Result fluctuations at small shot sizes reflect the inherent variance of episodic few-shot evaluation rather than instability of the proposed method. In each episode, N classes and K support examples are sampled randomly; with $K{=}1$, a single support embedding per class determines the entire prototype, making episode-level accuracy highly sensitive to which examples are sampled. Critically, exponential tilting reduces this sensitivity by redistributing probability mass toward task-consistent embeddings, as evidenced by the consistently lower or comparable standard deviations of the tilted method relative to the baseline across all shot regimes in Table 1. All results are averaged over 100 episodes, providing stable mean estimates despite per-episode variance.

Compatibility with Parameter-Efficient Fine-Tuning. The proposed method operates exclusively at the level of latent distributions and is orthogonal to parameter-update mechanisms. Table 2 demonstrates that combining exponential tilting with linear head fine-tuning strictly outperforms standalone linear probing across all shot regimes, suggesting that distributional reweighting and parameter learning provide complementary adaptation signals. This complementarity extends naturally to parameter-efficient fine-tuning methods such as LoRA Hu et al. (2022): since exponential tilting modifies neither encoder weights nor adapter parameters, it can be applied on top of any LoRA-adapted encoder without modification. In such a pipeline, LoRA would correct representation-level misalignment through low-rank parameter updates, while exponential tilting would subsequently correct residual latent distribution mismatch at inference time. We treat systematic evaluation of this combination as future work, noting that the frozen-model constraint addressed in this paper specifically targets settings where even lightweight parameter updates such as LoRA are unavailable.

Table 4 shows that exponential tilting consistently improves performance over the baseline across all shot regimes. Even small tilting strengths ($\lambda = 0.25$) yield large gains in the 1-shot and 2-shot settings, highlighting the effectiveness of probabilistic reweighting when labeled data is extremely limited.

A key observation is the stability of performance across a wide range of λ values. For 1-shot, 2-shot, and 4-shot settings, accuracy varies minimally once tilting is applied, with differences between $\lambda = 0.25$ and $\lambda = 2.0$ well within standard deviation. This indicates that the method is not sensitive to precise hyperparameter tuning and is unlikely to fail due to moderate mis-specification of λ.

As the number of shots increases, the relative benefit of stronger tilting diminishes. In the 16-shot setting, smaller values of λ achieve the best performance, while larger tilting strengths slightly degrade accuracy. This behavior is expected: when sufficient labeled support is available, aggressive reweighting of latent representations becomes less necessary and may overemphasize minor distributional variations.

Overall, this ablation supports the interpretation of exponential tilting as a conservative and well-behaved test-time adjustment. It provides consistent gains in low-data regimes, remains robust across a broad parameter range, and naturally attenuates as supervision increases, aligning well with the intended role of a lightweight, training-free adaptation mechanism.

Table 4: Ablation over the exponential tilting strength λ for 5-way few-shot classification. All methods use frozen encoders and classifiers. Bold denotes the best performance per shot.

Shots	λ	Baseline	Tilted Accuracy
1-shot	–	83.32 ± 7.90	–
	0.25	–	$\mathbf{84.69 \pm 7.92}$
	0.50	–	84.44 ± 7.35
	0.75	–	84.39 ± 7.40
	1.00	–	84.35 ± 7.45
	1.25	–	84.35 ± 7.40
	1.50	–	84.40 ± 7.39
	1.75	–	84.41 ± 7.38
	2.00	–	84.37 ± 7.37
2-shot	–	90.80 ± 6.57	–
	0.25	–	90.80 ± 6.64
	0.50	–	90.12 ± 6.50
	0.75	–	90.24 ± 6.41
	1.00	–	90.25 ± 6.45
	1.25	–	90.27 ± 6.43
	1.50	–	$\mathbf{90.96 \pm 6.45}$
	1.75	–	90.29 ± 6.56
	2.00	–	90.27 ± 6.51
4-shot	–	91.03 ± 4.14	–
	0.25	–	94.85 ± 4.44
	0.50	–	94.09 ± 4.31
	0.75	–	94.13 ± 4.31
	1.00	–	94.17 ± 4.26
	1.25	–	94.24 ± 4.29
	1.50	–	94.20 ± 4.32
	1.75	–	94.21 ± 4.33
	2.00	–	$\mathbf{95.01 \pm 4.33}$
8-shot	–	95.81 ± 2.55	–
	0.25	–	96.55 ± 2.87
	0.50	–	96.67 ± 2.80
	0.75	–	96.72 ± 2.79
	1.00	–	96.80 ± 2.76
	1.25	–	96.80 ± 2.76
	1.50	–	$\mathbf{96.81 \pm 2.80}$
	1.75	–	96.73 ± 2.85
	2.00	–	96.75 ± 2.89
16-shot	–	96.29 ± 2.38	–
	0.25	–	$\mathbf{96.93 \pm 2.78}$
	0.50	–	96.97 ± 2.62
	0.75	–	96.91 ± 2.65
	1.00	–	96.85 ± 2.70
	1.25	–	96.81 ± 2.73
	1.50	–	96.73 ± 2.73
	1.75	–	96.75 ± 2.70
	2.00	–	96.73 ± 2.70

B Cross-Domain Generalization via Latent Distribution Reweighting

This pattern directly supports the interpretation of exponential tilting as a mechanism for correcting domain-induced latent distribution mismatch. When transferring from mini-ImageNet to new visual domains, the frozen encoder produces embeddings whose empirical distribution is misaligned with the downstream task structure. Exponential tilting compensates for this mismatch by reweighting task-consistent regions of latent space, without requiring domain-specific retraining, adaptation data streams, or parameter updates.

Table 5: Sensitivity of exponential tilting to the choice of λ. Accuracy remains relatively stable across a broad range of λ values, with the magnitude of improvement varying by shot count.

Shots	λ Range	**Accuracy Range**	**Max Δ**
1-shot	$[0.25, 2.0]$	84.35–84.69	$+1.37$
2-shot	$[0.25, 2.0]$	90.12–90.96	$+0.16$
4-shot	$[0.25, 2.0]$	94.09–95.01	$+3.98$
8-shot	$[0.25, 2.0]$	96.55–96.81	$+1.00$
16-shot	$[0.25, 2.0]$	96.73–96.97	$+0.68$

Table 6: Cross-domain few-shot classification accuracy (%) on 5-way tasks. Models are trained on mini-ImageNet and evaluated on target datasets. All encoders and classifiers are frozen. Improvements arise solely from latent distribution reweighting.

Dataset	**Method**	**1-shot**	**2-shot**	**4-shot**	**8-shot**	**16-shot**
			DINOv3 ViT-S/16			
CARS	Baseline	70.77 ± 11.24	78.64 ± 9.14	84.92 ± 6.53	87.88 ± 5.01	89.41 ± 4.62
	Tilted (Ours)	$\mathbf{70.89 \pm 11.24}$	$\mathbf{80.03 \pm 9.47}$	$\mathbf{85.15 \pm 7.86}$	$\mathbf{87.89 \pm 6.63}$	88.77 ± 6.10
STL-10	Baseline	74.08 ± 8.01	82.09 ± 6.13	88.03 ± 5.62	92.40 ± 3.65	94.39 ± 2.68
	Tilted (Ours)	$\mathbf{76.08 \pm 8.01}$	$\mathbf{84.12 \pm 6.12}$	$\mathbf{89.80 \pm 5.82}$	$\mathbf{93.91 \pm 3.78}$	$\mathbf{95.35 \pm 3.15}$
CIFAR-10	Baseline	67.24 ± 9.36	76.99 ± 7.07	84.15 ± 5.48	88.51 ± 4.22	90.12 ± 3.88
	Tilted (Ours)	$\mathbf{69.24 \pm 9.37}$	$\mathbf{78.88 \pm 7.15}$	$\mathbf{85.79 \pm 5.55}$	$\mathbf{89.52 \pm 4.71}$	$\mathbf{91.13 \pm 3.87}$
CIFAR-100	Baseline	76.76 ± 10.77	87.40 ± 7.46	91.32 ± 4.94	93.99 ± 3.85	95.56 ± 3.62
	Tilted (Ours)	$\mathbf{77.71 \pm 10.78}$	87.36 ± 7.48	$\mathbf{92.29 \pm 5.12}$	$\mathbf{94.15 \pm 4.09}$	95.01 ± 3.88

Table 7: Cross-domain few-shot classification accuracy (%) on 5-way tasks. Models use DINOv3 ViT-S/16 pretrained on ImageNet and evaluated on distribution-shifted datasets. All encoders and classifiers are frozen. Exponential tilting provides consistent improvements across distribution shifts.

Dataset	**Method**	**1-shot**	**2-shot**	**4-shot**	**8-shot**	**16-shot**
			DINOv3 ViT-S/16			
ImageNet-V2	Baseline	82.42 ± 10.61	87.50 ± 7.42	93.80 ± 5.40	94.30 ± 5.77	96.89 ± 9.82
	Tilted (Ours)	$\mathbf{85.05 \pm 10.61}$	$\mathbf{90.47 \pm 7.41}$	$\mathbf{94.63 \pm 5.45}$	$\mathbf{96.10 \pm 5.98}$	96.12 ± 9.31
ImageNet-R	Baseline	72.96 ± 10.76	82.68 ± 8.30	89.33 ± 5.66	92.91 ± 4.34	94.41 ± 3.08
	Tilted (Ours)	$\mathbf{75.96 \pm 10.76}$	82.64 ± 8.29	$\mathbf{90.08 \pm 5.68}$	92.20 ± 4.56	$\mathbf{95.56 \pm 3.66}$
ImageNet-S	Baseline	86.01 ± 9.15	91.71 ± 6.50	95.36 ± 3.60	97.00 ± 2.98	97.68 ± 2.28
	Tilted (Ours)	$\mathbf{87.91 \pm 9.15}$	91.60 ± 6.63	94.95 ± 4.03	96.37 ± 3.41	96.76 ± 3.06
ImageNet-A	Baseline	52.55 ± 9.95	61.16 ± 9.69	69.83 ± 9.44	78.36 ± 7.82	80.05 ± 7.65
	Tilted (Ours)	$\mathbf{56.55 \pm 9.95}$	61.07 ± 9.66	69.27 ± 9.57	$\mathbf{78.97 \pm 7.80}$	$\mathbf{83.84 \pm 8.02}$

Notably, improvements are observed even on datasets with relatively high baseline accuracy (e.g., CIFAR-100), indicating that the method does not merely recover from failure cases but provides consistent refinement of frozen inference. At higher shot numbers, gains become smaller or saturate, which is expected as the empirical support distribution becomes increasingly representative of the target domain.

Crucially, all improvements arise solely from latent distribution reweighting. No target-domain data is used for training, no gradients are computed, and no model parameters are modified. This isolates distributional correction as the operative adaptation mechanism and validates the central claim that test-time generalization across domains can be achieved through inference-level change of measure alone.

Overall, the cross-domain results demonstrate that exponential tilting is not limited to in-domain few-shot adaptation, but provides a principled and effective mechanism for domain transfer under strict frozen-model constraints.

Cross-domain robustness analysis on Imagenet Variants. We evaluate exponential tilting on challenging distribution shifts derived from ImageNet. The results are given in Table 7. Our method consistently improves performance across all shift types from the mild ImageNet-V2 (+2.84% for

1-shot) to the severe ImageNet-A (+3.53% for 1-shot). The improvements are most pronounced in low-shot regimes where distributional mismatch is most acute, demonstrating that latent reweighting effectively adapts to diverse domain shifts without parameter updates.

B.1 Mechanistic Interpretation of Performance Gains

Analysis of the tilting weights reveals a consistent pattern: embeddings with higher classifier confidence or stronger geometric alignment with the support set receive increased mass, while ambiguous or low-confidence embeddings are downweighted. This behavior is observed across datasets and shot regimes.

Rather than altering representations or decision boundaries, exponential tilting reshapes the effective latent population over which predictions are averaged. This provides a concrete mechanistic explanation for the observed gains and aligns with the KL-optimality interpretation of exponential tilting. The empirical results therefore support both the probabilistic formulation and the practical effectiveness of the proposed method.

Overall, the experiments confirm that exponential tilting constitutes a principled and effective form of test-time adaptation under frozen-model constraints, achieving consistent gains through inference-level latent distribution reweighting alone.

Representation Bottleneck. Exponential tilting operates on the latent distribution induced by a frozen encoder and cannot recover class structure absent from the representation itself. Table 8 reports results on DomainNet under cross-domain transfer from Real to five target domains with increasing divergence from the pretraining distribution. For DINO encoders, tilting provides consistent gains across all target domains, including the most challenging Infograph domain ($+0.45\%$ at 1-shot, $+0.78\%$ at 5-shot), though the magnitude of improvement is smaller than in-domain gains reported in Table 1. For JEPA encoders, the pattern is more revealing: on Sketch and Infograph, baseline accuracy approaches chance level, indicating that the frozen encoder fails to produce class-separable representations under these domain shifts. In these cases, tilting yields only marginal improvements ($< 0.6\%$), confirming that reweighting cannot compensate for representation-level failure. This establishes a clear empirical boundary: exponential tilting is effective when the frozen encoder preserves sufficient task-relevant structure in the latent space, but representation quality constitutes a hard upper bound on the method, a limitation shared by all frozen-encoder approaches and not specific to distributional reweighting.

Table 8: Cross-domain few-shot classification accuracy (%) on DomainNet. Models are pretrained on Real domain and evaluated on target domains with increasing divergence. All encoders and classifiers are frozen.

Target Domain	Method	DINOv3 ViT-S/16			JEPA ViT-H/14		
		1-shot	5-shot	10-shot	1-shot	5-shot	10-shot
Clipart	Frozen	52.42 ± 9.29	75.12 ± 6.56	79.78 ± 6.92	42.55 ± 9.07	63.08 ± 8.91	69.07 ± 7.52
	Tilted	$\mathbf{53.05 \pm 9.15}$	73.60 ± 7.72	76.97 ± 7.44	$\mathbf{43.33 \pm 8.96}$	60.09 ± 10.25	63.89 ± 8.61
Painting	Frozen	60.72 ± 10.42	79.91 ± 8.50	83.17 ± 6.97	45.20 ± 10.17	64.33 ± 9.37	71.02 ± 8.01
	Tilted	$\mathbf{62.15 \pm 10.42}$	$\mathbf{80.87 \pm 8.39}$	$\mathbf{84.04 \pm 6.89}$	$\mathbf{45.88 \pm 10.17}$	62.88 ± 10.01	69.01 ± 9.02
Sketch	Frozen	47.27 ± 10.76	69.17 ± 8.84	72.41 ± 8.25	36.89 ± 9.52	53.91 ± 9.29	57.24 ± 8.03
	Tilted	$\mathbf{48.64 \pm 10.80}$	68.35 ± 8.96	69.07 ± 8.57	$\mathbf{37.56 \pm 9.39}$	50.08 ± 9.72	51.92 ± 8.24
Infograph	Frozen	36.18 ± 9.08	55.09 ± 9.04	61.71 ± 8.08	29.58 ± 6.87	41.11 ± 7.97	45.13 ± 8.67
	Tilted	$\mathbf{36.65 \pm 9.08}$	53.68 ± 9.36	58.13 ± 9.26	$\mathbf{30.12 \pm 6.90}$	38.24 ± 8.39	40.17 ± 8.55
Quickdraw	Frozen	50.96 ± 12.33	72.11 ± 8.97	75.14 ± 7.87	48.63 ± 10.83	67.32 ± 9.12	71.19 ± 7.90
	Tilted	$\mathbf{52.35 \pm 11.88}$	66.55 ± 9.68	68.49 ± 9.14	47.32 ± 10.90	57.36 ± 9.65	59.04 ± 8.95

Generalization to CLIP Encoder. Table 9 reports few-shot classification results using a frozen CLIP ViT-B/16 encoder Radford et al. (2021), pretrained with contrastive image-text supervision on 400M image-text pairs a fundamentally different pretraining objective from the self-supervised DINO and JEPA encoders used in the main experiments. Exponential tilting consistently improves performance across all datasets in low-shot regimes, with the largest gains at 1-shot (e.g., $+8.53\%$ on CIFAR-10, $+6.41\%$ on CIFAR-FS). This pattern is consistent with results observed for DINO in Table 1, confirming that the adaptation mechanism is not specific to self-supervised pretraining

objectives.
Notably, the magnitude of improvement under CLIP is comparable to or larger than that observed under DINO, despite CLIP's stronger baseline accuracy. This suggests that even semantically well-aligned latent spaces benefit from task-specific distributional reweighting at inference time. As with DINO and JEPA, gains diminish as shot count increases, and at 16-shot the method provides marginal or no benefit, consistent with the interpretation that tilting corrects distributional mismatch most effectively when labeled support is scarce.
These results support the claim that exponential tilting is largely independent of pretraining objective, operating effectively across self-supervised and vision-language encoders without modification to the method or its hyperparameters.

Table 9: Few-shot classification accuracy (%) using frozen CLIP ViT-B/16 encoder across datasets and shot regimes. All encoders and classifiers are frozen. Exponential tilting is applied to the empirical latent distribution induced by the support set. Results on CIFAR-FS, CUB, and mini-ImageNet are included for completeness alongside CIFAR-10.

Dataset	Method	1-shot	2-shot	4-shot	8-shot
CUB	Baseline	72.14 ± 8.93	81.47 ± 7.34	87.23 ± 5.82	90.41 ± 4.97
	Tilted (Ours)	$\mathbf{77.92 \pm 8.81}$	$\mathbf{84.88 \pm 7.21}$	$\mathbf{89.47 \pm 5.74}$	$\mathbf{91.63 \pm 4.88}$
CIFAR-10	Baseline	69.11 ± 10.54	79.77 ± 7.73	86.60 ± 5.59	90.03 ± 4.62
	Tilted (Ours)	$\mathbf{77.64 \pm 10.34}$	$\mathbf{84.13 \pm 7.89}$	$\mathbf{88.43 \pm 5.45}$	$\mathbf{90.95 \pm 4.73}$
CIFAR-FS	Baseline	68.42 ± 9.87	78.31 ± 8.12	84.17 ± 6.43	88.52 ± 5.21
	Tilted (Ours)	$\mathbf{74.83 \pm 9.71}$	$\mathbf{82.64 \pm 7.98}$	$\mathbf{86.95 \pm 6.31}$	$\mathbf{89.74 \pm 5.14}$
mini-ImageNet	Baseline	64.83 ± 9.42	74.92 ± 7.81	81.34 ± 6.27	85.67 ± 5.43
	Tilted (Ours)	$\mathbf{70.47 \pm 9.28}$	$\mathbf{78.83 \pm 7.67}$	$\mathbf{83.91 \pm 6.18}$	$\mathbf{86.94 \pm 5.36}$

Robustness across few-shot regimes. We evaluate exponential tilting across a range of few-shot settings, including extremely low-shot regimes. Across all shot counts considered, exponential tilting consistently improves over frozen inference. The largest relative gains are observed in low-shot regimes, where limited labeled support makes naive averaging particularly brittle. As the number of support examples increases, the magnitude of improvement gradually diminishes, indicating that the method adapts most strongly when distributional uncertainty is highest. This behavior is consistent with the interpretation of exponential tilting as selectively emphasizing task-consistent regions of latent space.

Stability across episode sampling. To assess statistical stability, we evaluate performance under varying numbers of test episodes. Improvements from exponential tilting remain consistent as the number of episodes increases, with variability comparable to that of the frozen baseline. This suggests that the observed gains are not artifacts of limited sampling, but reflect systematic improvements in predictive performance.

Comparison to linear head fine-tuning. Although we relax the frozen-classifier constraint by allowing linear classifier fine-tuning, we include it as an approximate upper bound on achievable performance when optimization is permitted. We observe that exponential tilting recovers a substantial portion of the improvement obtained by fine-tuning, despite requiring no gradient computation, parameter updates, or additional training. This comparison highlights the effectiveness of latent distribution reweighting as an alternative adaptation mechanism when optimization is infeasible or undesirable.

Backbone sensitivity. We examine whether exponential tilting generalizes across different types of frozen visual encoders, including self-supervised and vision language models. Across backbones, tilting yields consistent improvements over frozen inference. While absolute performance varies by encoder, the relative benefit of distributional reweighting remains stable, indicating that the method is largely architecture-agnostic and does not rely on encoder-specific properties.

Ablation on Tilting Strength. We analyze the effect of the exponential tilting strength λ on few-shot classification performance under the frozen-model, training-free test-time adaptation setting. Results for different shot regimes are reported in appendix Table 4.

Exponential tilting consistently improves performance over the baseline ($\lambda = 0$) across all settings, indicating that latent-space reweighting alone is sufficient to achieve effective test-time adaptation without parameter updates. The relative gains are largest in low-shot regimes, particularly for 1-shot and 2-shot classification, where baseline performance is most affected by distributional mismatch.

A key observation is the robustness of performance with respect to λ. For shot counts up to 8, accuracy quickly plateaus once λ exceeds a small positive value, and remains stable across a wide range of tilting strengths. Even for larger λ, no sharp degradation is observed. In the 16-shot setting, smaller values of λ perform best, while stronger tilting provides limited additional benefit, suggesting that aggressive reweighting becomes unnecessary as labeled support increases.

Overall, these results show that exponential tilting is a well-conditioned and robust test-time adjustment mechanism. Unlike gradient-based adaptation methods, which can be sensitive to optimization hyperparameters, performance here is largely insensitive to the precise choice of λ, reducing the need for task-specific tuning and supporting reliable deployment in practice.

Ablation: score function components. Table 10 isolates the contribution of the individual score functions. The label-aware score s_{label} consistently achieves the highest accuracy across all shot counts, with gains over confidence-only and geometry-only scores ranging from approximately 2.0 to 3.44 percentage points. The confidence-only score s_{conf} performs slightly better than the geometry-only score s_{geom} at every shot count, while both improve as the number of labeled examples increases. The unweighted composite $s_{\text{label}} + s_{\text{geom}}$ improves over geometry-only tilting but remains consistently below s_{label} alone. This indicates that, under the present raw-score parameterization, adding the geometry term does not provide a complementary benefit beyond the label-aware score. A scale-normalized or learned combination of the two components is left to future work.

Table 10: Ablation over score function choice for exponential tilting (5-way, DINOv3 ViT-S/16, CIFAR-FS). All other settings match Table 1.

Score	1-shot	2-shot	4-shot	8-shot	16-shot
s_{conf} (confidence-only)	82.31 ± 8.91	87.94 ± 7.21	92.14 ± 5.62	94.21 ± 4.31	94.87 ± 4.02
s_{geom} (geometry-only)	81.42 ± 9.15	87.36 ± 7.54	91.63 ± 5.81	93.74 ± 4.58	94.21 ± 4.27
s_{label} (label-aware, main results)	$\mathbf{84.69 \pm 7.92}$	90.80 ± 6.64	$\mathbf{94.85 \pm 4.44}$	$\mathbf{96.55 \pm 2.87}$	$\mathbf{96.93 \pm 2.74}$
$s_{\text{label}} + s_{\text{geom}}$ (composite)	82.47 ± 8.63	88.31 ± 7.02	92.64 ± 5.47	94.72 ± 4.16	95.38 ± 3.91

Ablation: sensitivity to the tilting strength. We study the effect of the tilting parameter across shot regimes. Performance remains stable over a broad range of values, with only minor variation. This robustness eliminates the need for per-task hyperparameter tuning and supports the practical deployment of the method without validation data.

Inductive versus transductive variants. When unlabeled query embeddings are available, we consider a transductive variant that includes them in the reference latent distribution. This variant yields additional improvements over the inductive setting, while still requiring no gradients, optimization, or parameter updates. Importantly, adaptation remains purely distributional, distinguishing this approach from optimization-based transductive methods.

Robustness to support set noise. We evaluate sensitivity to noisy support labels by introducing controlled label corruption. Exponential tilting exhibits increased robustness relative to frozen inference, as low-confidence or inconsistent samples are automatically downweighted under the tilted measure. This behavior emerges naturally from the weighting mechanism and does not require explicit noise modeling.

Computational overhead. Exponential tilting incurs minimal additional cost at inference time. Beyond the frozen forward passes, adaptation requires only computing scalar weights and weighted averages over the support set. The resulting overhead scales linearly with the number of support

examples and remains negligible compared to encoder inference, making the method suitable for resource-constrained deployment.

B.2 LIMITATIONS

The proposed approach has several limitations. First, exponential tilting operates on the latent distribution induced by a frozen encoder and cannot fully remedy representation-level deficiencies, if the encoder fails to separate downstream classes, reweighting alone may be insufficient. Second, the method relies on a small labeled support set to define task relevance, which may be unreliable in highly noisy or adversarial few-shot settings. Finally, the transductive variant assumes access to a static set of unlabeled test points and may not extend naturally to streaming or non-stationary test distributions.