Bridging Domain Expertise and Generalization for Performance Estimation

Supplementary Material

A. Derivation of Proposition 1

For a finite target sample set $\{(x_i, y_i)\}_{i=1}^N$, the standard empirical accuracy (*i.e.*, Hard Accuracy) of the model f_θ is defined as:

$$\mathrm{ACC}_N = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\{\widehat{y}(x_i) = y_i\}, \tag{A.1}$$

where $\widehat{y}(x) = \arg\max_j \widehat{P}_\theta(j \mid x)$ is the hard label prediction. Since the hard indicator function is non-differentiable and difficult to analyze directly, we consider the expected soft accuracy as a surrogate for theoretical analysis. The soft accuracy replaces the hard match with the probability of the true class:

$$\mathrm{ACC}_N^{\text{soft}} = \frac{1}{N}\sum_{i=1}^{N} \widehat{P}_\theta(y_i \mid x_i). \tag{A.2}$$

By the Law of Large Numbers (LLN), as $N \to \infty$, the empirical soft accuracy converges to its population expectation over the target distribution $\mathcal{D}_t$:

$$\lim_{N\to\infty} \mathrm{ACC}_N^{\text{soft}} = \mathbb{E}_{(X,Y)\sim\mathcal{D}_t}\big[\widehat{P}_\theta(Y \mid X)\big] \triangleq \mathbb{E}[\mathrm{ACC}_{\text{soft}}]. \tag{A.3}$$

Using the law of iterated expectations (conditioning on X), we expand this term as:

$$\mathbb{E}[\mathrm{ACC}_{\text{soft}}] = \int_{\mathcal{X}} \mathbb{E}_{Y|x}\big[\widehat{P}_\theta(Y \mid x)\big]\, p_t(x)\, \mathrm{d}x. \tag{A.4}$$

Inside the integral, the term $\mathbb{E}_{Y|x}[\widehat{P}_\theta(Y \mid x)]$ represents the expected confidence score of the model with respect to the label variable Y given $X = x$. By the definition of expectation for a discrete random variable, this expectation expands to the sum of the function values weighted by their probabilities:

$$\mathbb{E}_{Y|x}\big[\widehat{P}_\theta(Y \mid x)\big] = \sum_{j=1}^{K} \underbrace{P^*(j \mid x)}_{\text{True Probability}} \cdot \underbrace{\widehat{P}_\theta(j \mid x)}_{\text{Model Prediction}}, \tag{A.5}$$

where $P^*(j \mid x) \triangleq \mathbb{P}(Y = j \mid X = x)$ denotes the ground-truth class posterior, and $\widehat{P}_\theta(j \mid x)$ denotes the predictive probability of the model for class j.

Substituting this expansion back into Eq. (A.4), we obtain:

$$\mathbb{E}[\mathrm{ACC}_{\text{soft}}] = \int_{\mathcal{X}} \left(\sum_{j=1}^{K} \widehat{P}_\theta(j \mid x)\, P^*(j \mid x)\right) p_t(x)\, \mathrm{d}x. \tag{A.6}$$

Finally, approximating the outer integral with the finite target samples $\{x_i\}_{i=1}^N$ via Monte Carlo estimation yields:

$$\mathbb{E}[\mathrm{ACC}_{\text{soft}}] \;\approx\; \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} \widehat{P}_\theta(j \mid x_i)\, P^*(j \mid x_i). \tag{A.7}$$

B. Contrastive representation learning

Many contrastive embedding models are trained using the InfoNCE (Information Noise-Contrastive Estimation) objective [4, 38, 41, 53, 55]. Fundamentally, this loss shapes the embedding space by pulling representations of positive (matched) pairs closer while pushing negative (unmatched) pairs apart.

Formally, given a mini-batch of N samples, for each query sample q_i, the InfoNCE loss is defined as:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\mathrm{sim}(\mathbf{q}_i, \mathbf{k}_i^+)/\tau)}{\sum_{j=1}^{N} \exp(\mathrm{sim}(\mathbf{q}_i, \mathbf{k}_j)/\tau)}, \tag{B.1}$$

where $\mathbf{q}_i, \mathbf{k}_i^+ \in \mathbb{R}^d$ denote the embeddings of the query and its corresponding positive key, and $\{\mathbf{k}_j\}_{j=1}^N$ includes the positive key $\mathbf{k}_i^+$ and $N-1$ negative keys. The function $\mathrm{sim}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top\mathbf{v}/(\|\mathbf{u}\|\|\mathbf{v}\|)$ computes cosine similarity, and $\tau \in \mathbb{R}^+$ is a temperature parameter controlling the concentration of the distribution.

B.1. CLIP Training

CLIP applies InfoNCE to learn aligned image-text representations. Given a batch of N image-text pairs $\{(I_i, T_i)\}_{i=1}^N$, CLIP encodes images and texts into ℓ_2-normalized embeddings $\mathbf{z}_i^v, \mathbf{z}_i^t \in \mathbb{R}^d$ using vision encoder f_v and text encoder f_t respectively:

$$\mathbf{z}_i^v = \frac{f_v(I_i)}{\|f_v(I_i)\|_2}, \quad \mathbf{z}_i^t = \frac{f_t(T_i)}{\|f_t(T_i)\|_2}. \tag{B.2}$$

The symmetric contrastive loss is calculated as the average of the image-to-text ($\mathcal{L}^{I\to T}$) and text-to-image ($\mathcal{L}^{T\to I}$) losses. For the i-th sample, the image-to-text loss is:

$$\mathcal{L}_i^{I\to T} = -\log \frac{\exp(\mathrm{sim}(\mathbf{z}_i^v, \mathbf{z}_i^t)/\tau)}{\sum_{j=1}^{N} \exp(\mathrm{sim}(\mathbf{z}_i^v, \mathbf{z}_j^t)/\tau)}, \tag{B.3}$$

where $\mathrm{sim}(\mathbf{z}_i^v, \mathbf{z}_j^t) = (\mathbf{z}_i^v)^\top \mathbf{z}_j^t$ is the cosine similarity. Symmetrically, the text-to-image loss is:

$$\mathcal{L}_i^{T\to I} = -\log \frac{\exp(\mathrm{sim}(\mathbf{z}_i^t, \mathbf{z}_i^v)/\tau)}{\sum_{j=1}^{N} \exp(\mathrm{sim}(\mathbf{z}_i^t, \mathbf{z}_j^v)/\tau)}. \tag{B.4}$$

The final CLIP loss is the average of both directions:

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2N}\sum_{i=1}^{N}\left(\mathcal{L}_i^{I\to T} + \mathcal{L}_i^{T\to I}\right), \tag{B.5}$$

B.2. Mechanism of Uniformity

To understand why CLIP yields predictions with high uniformity (*i.e*., over-smoothed distributions), we analyze the InfoNCE loss through two complementary lenses: *Alignment-Uniformity decomposition* and *Gradient dynamics*.

Alignment and Uniformity. Previous work [55] decomposes the quality of learned representations into two key properties:

- **Alignment** measures the closeness of positive pairs.
- **Uniformity** quantifies how uniformly features are distributed on the hyper-sphere.

InfoNCE implicitly optimizes both objectives, but the balance between them evolves during training. In the early stage (*i.e*. Alignment phase), the model rapidly pulls positive pairs closer and the numerator term in InfoNCE dominates optimization, as positive similarities are initially low. In the late stage (*i.e*. Uniformity phase), once positive pairs are sufficiently aligned, the numerator approaches its maximum and the gradient signal from the numerator nearly diminishes. Thus the optimization shifts to minimizing the denominator of the InfoNCE loss, which encourages spreading out negative pairs uniformly across the hypersphere [46]. Consequently, the resulting prediction probabilities tend to spread out over all classes rather than concentrating on a single peak, especially when semantic ambiguity exists.

Implicit Hard Negative Mining. Another explanation takes a different but consistent perspective that the InfoNCE performs implicit hard negative mining via its exponential weighting scheme [7]:

$$\frac{\partial \mathcal{L}_{\text{InfoNCE}}}{\partial \text{sim}(\mathbf{q}_i, \mathbf{k}_j)} \propto \exp(\text{sim}(\mathbf{q}_i, \mathbf{k}_j)/\tau), \quad j \neq i, \tag{B.6}$$

which means that hard negatives receive exponentially larger gradients, causing the model to focus more on distinguishing confusing negative samples. Consequently, InfoNCE automatically emphasizes learning from the most challenging examples without explicit hard negative mining.

Let us derive the gradient of InfoNCE loss, which is an equivalent form of the formulation Eq. (B.1):

$$\mathcal{L}_i = -\log \frac{\exp(s_i^+/\tau)}{\exp(s_i^+/\tau) + \sum_{j\neq i}\exp(s_j^-/\tau)}, \tag{B.7}$$

where s_i^+ denotes the positive pair similarity and s_j^- denotes the negative pair similarity. For a negative pair similarity s_j^-, the gradient is:

$$\begin{aligned}
\frac{\partial \mathcal{L}_i}{\partial s_j^-} &= \frac{\partial}{\partial s_j^-}\left[-\log \frac{\exp(s_i^+/\tau)}{\exp(s_i^+/\tau) + \sum_{k\neq i}\exp(s_k^-/\tau)}\right] \\
&= \frac{\partial}{\partial s_j^-}\left[\log\left(\exp(s_i^+/\tau) + \sum_{k\neq i}\exp(s_k^-/\tau)\right) - \frac{s_i^+}{\tau}.\right] \\
&= \frac{1}{\tau}\frac{\exp(s_j^-/\tau)}{\exp(s_i^+/\tau) + \sum_{k\neq i}\exp(s_k^-/\tau)}.
\end{aligned} \tag{B.8}$$

Therefore, the gradient magnitude is proportional to the exponential term:

$$\frac{\partial \mathcal{L}_i}{\partial s_j^-} \propto \exp(s_j^-/\tau), \tag{B.9}$$

which shows that negative samples with higher similarity (hard negatives) receive larger gradients. This implicit weighting mechanism naturally emphasizes hard negatives during optimization, leading to more uniform feature distribution to separate these challenging cases on the hypersphere, as discussed in [7, 46]. This training dynamic is fundamentally intertwined with the temperature parameter τ. Since CLIP trains a learnable temperature which typically converges to a very low value ($\tau \approx 0.01$), this low τ imposes severe uniformity pressure on the feature space by aggressively magnifying the gradient signal from hard negatives. Though low τ mathematically causes the softmax output to be sharper (high confidence) by magnifying logit differences, the resultant high feature uniformity is the root cause of the observed calibration issue: it leads to overly uniform prediction distributions for inputs lacking strong semantic alignment, as we observe in CLIP.

C. Datasets

We focus on natural image classification and evaluate our method on 10 benchmark datasets encompassing both natural and synthetic distribution shifts. Specifically, we use MNIST [24], CIFAR-10 [22], CIFAR-100 [22], ImageNet [44], Tiny-ImageNet, FMoW [5], and four datasets from the BREEDS benchmark [45]: Living-17, Nonliving-26, Entity-13, and Entity-30. Tiny-ImageNet is a compact subset of ImageNet comprising 200 classes. BREEDS (Benchmark for Robustness under Evolving Distribution Shifts) constructs subpopulation shifts by partitioning ImageNet classes hierarchically, where training and test sets contain different fine-grained subclasses within the same superclass, (*e.g*. different dog breeds), simulating realistic deployment scenarios. For each source dataset, we evaluate performance on corresponding shifted test sets under distribution shift.

Table 5. Summary of datasets and their corresponding distribution shifts. For CIFAR, ImageNet, Tiny-ImageNet and MNIST, the suffixes '-N' and '-S' denote *Natural* and *Synthetic* distribution shifts, respectively. For BREEDS, '-S' and '-N' denote *Same* and *Novel* subpopulations. BREEDS-S consists of L17-S, NL26-S, E13-S, and E30-S; BREEDS-N consists of L17-N, NL26-N, E13-N, and E30-N.

Dataset	**Shift Type**	**Shift Dataset(s)**
MNIST	–	QMNIST, USPS, SVHN
CIFAR-10	C10-N	CIFAR-10v2 (re-sampled natural test set)
	C10-S	CIFAR-10-C (19 corruption types × 5 severity levels = 95 corruptions)
CIFAR-100	C100-N	CIFAR-100 test set (standard evaluation)
	C100-S	CIFAR-100-C (19 corruption types × 5 severity levels = 95 corruptions)
ImageNet	IN-N	ImageNet-V2 (3 variants: Matched-Frequency, Threshold-0.7, Top-Images), ImageNet-Sketch
	IN-S	ImageNet-C (19 corruption types × 5 severity levels = 95 corruptions)
Tiny-ImageNet	IN200-N	Imagenet200-V2 (3 variants), Imagenet200-Sketch, ImageNet-Reality
	IN200-S	ImageNet200-C (19 corruption types × 5 severity levels = 95 corruptions)
FMoW	–	FMoW (OOD-val and OOD-test)
BREEDS	BREEDS-S	IN-N and IN-S with same subpopulation hierarchies
	BREEDS-N	IN-N and IN-S with novel subpopulation hierarchies

For MNIST, we consider QMNIST [59], USPS [18], and SVHN [34] as shifted variants. For CIFAR-10, we use CIFAR-10v2 [30] as a natural shift (C10-N) and CIFAR-10-C [15] for synthetic corruptions (C10-S). For CIFAR-100, we use the standard test set as the natural shift benchmark (C100-N) and CIFAR-100-C [15] for synthetic corruptions (C100-S). For ImageNet, we assess natural shifts (IN-N) by ImageNet-V2 [39] and ImageNet-Sketch [54], and synthetic corruption (IN-S) by ImageNet-C [15]. For Tiny-ImageNet, the natural shifts (IN200-N) include ImageNet-Reality [16] and the corresponding 200 matching classes from ImageNet-V2 and ImageNet-Sketch. The synthetic shift (IN200-S) utilizes the 200 matching classes extracted from ImageNet-C. Note that the corruption benchmarks (CIFAR-10/100-C, ImageNet-C, ImageNet200-C) each contain 19 corruption types (*e.g.*, Gaussian noise, motion blur, frost) with 5 severity levels per type, resulting in 95 distinct test conditions per dataset. FMoW naturally contains temporal and geographical distribution shifts and the out-of-distribution test data contain images from time periods and geographic regions unseen during training. For BREEDS datasets, we employ identical natural and synthetic shift protocols as ImageNet (V2 and C variants), models are evaluated on test sets containing either the training-time fine-grained categories (BREEDS-S) or novel categories from the same coarse-grained class (BREEDS-N), which isolates the subpopulation shift effects.

D. Baselines

We provide detailed formulations and implementation descriptions of all baselines compared in our experiments.

Importance Re-weighting (IM). The IM method estimates the target error as a re-weighted source error, where the weights are obtained as the ratio between the densities of target and source data across confidence bins. Following [3] this effectively corresponds to using a single slice in the classifier confidence space.

Average Confidence (AC). THe AC baseline directly estimates the target error by computing the average of one minus the maximum softmax confidence over the unlabeled target samples.

Difference of Confidence (DoC). It is also known as DOC-Feat, which models the error as the difference between the source and target confidence distributions[11]. The formulation is $\hat{\varepsilon}_{\text{DoC}} = \mathbb{E}_{x\sim D_S}\left[\mathbb{I}\left[\arg\max_{j\in\mathcal{Y}} f_\theta(j \mid x) \neq y\right]\right] + \mathbb{E}_{x\sim D_T}\left[1-\max_{j\in\mathcal{Y}} f_\theta(y \mid x)\right] - \mathbb{E}_{x\sim D_S}\left[1-\max_{j\in\mathcal{Y}} f_\theta(j \mid x)\right]$

Generalized Disagreement Equality (GDE). It estimates the target error as the disagreement ratio between the predictions of two independently trained models $f_\theta(x)$ and $f_{\theta'}(x)$ on the target data[19], which can be formulated as $\hat{\varepsilon}_{\text{GDE}} = \mathbb{E}_{x\sim D_T}\left[\mathbb{I}(\arg\max_{j\in\mathcal{Y}} f_\theta(j\,|\,x) \neq \arg\max_{j\in\mathcal{Y}} f_{\theta'}(j\,|\,x))\right]$.

Average Thresholded Confidence (ATC). ATC estimates target error by identifying a threshold t such that the fraction of source data points with scores below t matches the validation error on source data. The target error is then estimated as the proportion of target examples falling below this threshold: $\hat{\varepsilon}_{\text{ATC}} = \mathbb{E}_{x\sim D_T}\left[\mathbb{I}(s(f_\theta(x)) < t)\right]$, where $s(\cdot)$ denotes a scalar score function relating positively with

the performance of the model. As proposed in [10], there are two variants considered: (1) **ATC-MC**, which uses the maximum softmax confidence $s_{\text{MC}} = \max_{j \in \mathcal{Y}} f_\theta(j \mid x)$, and (2) **ATC-NE**, which uses the negative entropy score $s_{\text{NE}} = -\sum_{j \in \mathcal{Y}} f_\theta(j \mid x) \log f_\theta(j \mid x)$.

Projection Normalization (ProjNorm). ProjNorm [61] originally proposed a parameter-space metric that quantifies the distributional shift between source and target domains. The original method does not directly estimate the target accuracy but instead demonstrates that the projection norm strongly correlates with the true target error. In our evaluation, since our goal is to directly predict model performance, we follow a practical implementation adapted from [31], which converts the original ProjNorm metric into an approximate accuracy estimator by comparing the output distributions between source and target data. This enables a fair, quantitative comparison under the same Mean Absolute Estimation Error (MAE) metric.

Confidence Optimal Transport (COT). In [31], COT introduces an optimal-transport-based (OT) estimator that measures the Wasserstein distance between the empirical distribution of model confidence vectors on the unlabeled target set and the empirical source label distribution. Unlike confidence-based estimators (*e.g.*, Average Confidence) which may underestimate error by selecting the pseudo-label distribution as reference, COT uses the source label distribution under the assumption that $P_T(\vec{y}) \approx P_S(\vec{y})$. The paper further proposes **COTT** (COT with Thresholding), which learns a threshold on validation data and estimates error as the fraction of target samples whose per-sample transport costs exceed this threshold.

E. CWF on Additional Datasets

To further validate the effectiveness of the confidence-weighted fusion (CWF), we conduct supplementary experiments on additional datasets including Tiny-ImageNet, Living-17, and Nonliving-26. We adopt the same evaluation protocol as described in Sec. 5.3 to assess how well the fused predictions align semantically with the ground truth using Semantic Alignment Score (SAS). Note that this hierarchy-based evaluation is only applicable to datasets whose class labels conform to the WordNet taxonomy, which restricts our evaluation to the aforementioned datasets that satisfy this prerequisite. As shown in Figs. 5 to 7, the fused predictions on Living-17 and Nonliving-26 consistently exhibit markedly higher semantic consistency with the true labels than those of the base model. However, on Tiny-ImageNet, while the fused distribution substantially surpasses predictions of CLIP, it exhibits slightly degraded performance relative to the base model.

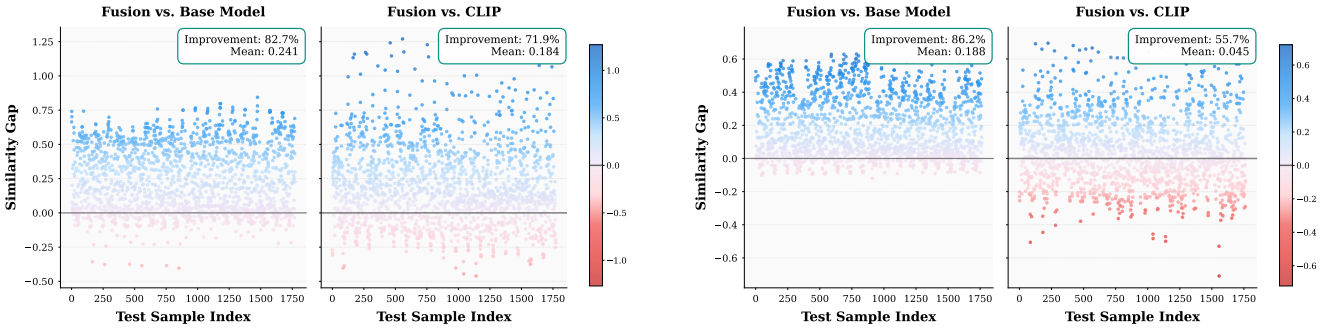
This phenomenon suggests that poor performance of CLIP on this particular dataset adversely affects the fusion outcome, where the predictions of base model are compromised by weaker contributions of CLIP. This observation highlights a critical caveat: while CLIP exhibits remarkable zero-shot generalization capabilities, its knowledge coverage remains bounded. In certain specialized domains, its predictions can fall considerably short of those from task-specific base models. Nevertheless, this finding underscores the merit of our fusion-based approach over directly adopting CLIP predictions as an approximation of the ideal distribution. The integration of CLIP represents a calculated trade-off: while it may introduce adverse effects in limited scenarios, the confidence-weighted fusion mechanism predominantly yields improved predictive distributions across the majority of settings by effectively leveraging the complementary strengths of both models.

Table 6. **Ablation of thresholding strategy.** The performance difference between the standard FRAP and FRAP w/o Thresholding denoted as FRAP $_{\text{(w/o thd)}}$

Dataset Family	Shift Type	FRAP	FRAP$_{\text{(w/o thd)}}$
MNIST	–	**12.42**±0.10	16.78±0.04
CIFAR-10	C10-N	**2.14**±0.02	6.13±0.04
	C10-S	**3.24**±0.05	7.26±0.05
CIFAR-100	C100-N	**4.00**±0.03	7.97±0.05
	C100-S	11.42±0.10	**8.53**±0.06
ImageNet	IN-N	**2.16**±0.03	12.31±0.07
	IN-S	**4.76**±0.01	9.16±0.06
Tiny-ImageNet	IN200-N	**7.46**±0.04	14.7±0.05
	IN200-S	**9.98**±0.04	12.37±0.06
FMoW	–	**3.42**±0.02	15.74±0.05
BREEDS-S	L17-S	**8.55**±0.05	10.98±0.06
	NL26-S	**9.30**±0.07	12.63±0.07
	E13-S	**8.29**±0.05	11.69±0.06
	E30-S	**6.94**±0.06	11.73±0.07
BREEDS-N	L17-N	5.37±0.05	**4.09**±0.04
	NL26-N	6.66±0.09	**5.63**±0.06
	E13-N	**5.90**±0.06	6.66±0.05
	E30-N	**5.54**±0.07	6.4±0.05
Average		**6.53**	10.04

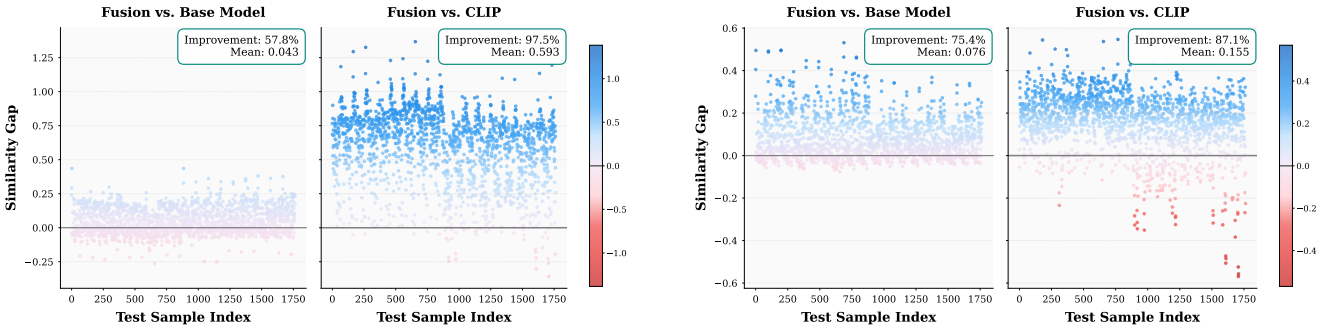
F. Thresholding Strategy Ablation

As detailed in Sec. 5.4, the thresholding strategy is introduced to address the inherent error of soft accuracy and the gap between the fused predictions and the true labels. This strategy transforms the continuous estimation problem into a more robust binary decision. By calibrating the threshold δ on the source validation set $\mathcal{D}_s$, we effectively align the fraction of samples predicted as incorrect (*i.e.*, $\text{Est}(x) < \delta$)



(a) Same shift (b) Novel shift

Figure 5. **Per-experiment Semantic Alignment Score (SAS) differences on Living-17**. We compare the confidence-weighted fusion against the base model (left column) and CLIP (right column) under (a) natural shifts and (b) synthetic corruptions. The y-axis shows $\Delta\text{SAS} = \text{SAS}_{\text{fusion}} - \text{SAS}_{\text{baseline}}$. Positive values indicate improved semantic alignment ($\Delta\text{SAS} > 0$), while negative values indicate degradation. The fusion largely enhances alignment under both conditions.



(a) Same shift (b) Novel shift

Figure 6. **Per-experiment Semantic Alignment Score (SAS) differences on Nonliving-26.** (a) same shifts and (b) novel shifts is the same with the Living-17. The fusion consistently achieves better semantic alignment than both the base model and CLIP.

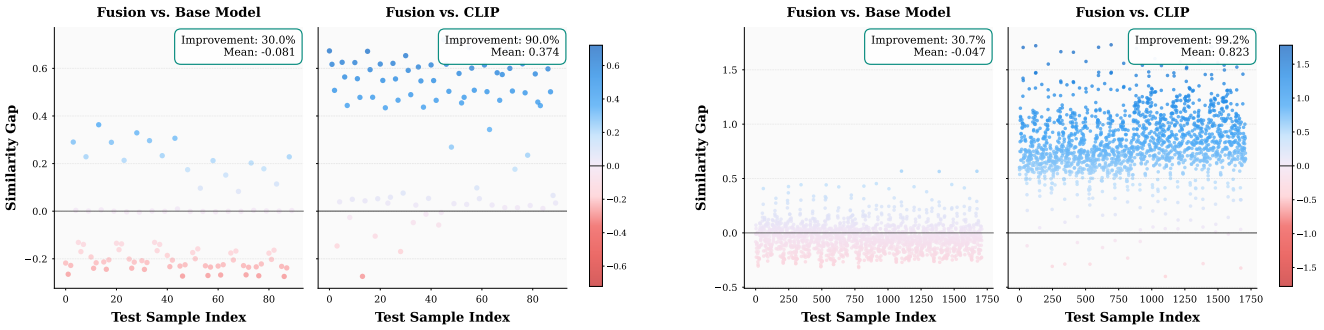


Figure 7. **Per-experiment Semantic Alignment Score (SAS) differences on Tiny-ImageNet.** (a) natural shifts and (b) synthetic corruptions is similar to the ImageNet. While the fused distribution on this dataset provides limited gains over the base model.

with the observed error rate of the base model on $\mathcal{D}_s$. This process stabilizes the estimation by correcting the scale and offset between the raw prediction score and the actual error magnitude, relying on the relative ranking of scores rather than their absolute magnitudes.

To quantify the effect of the thresholding strategy, we perform an ablation study, which serves the formulation result in Eq. (A.7) directly as the performance estimation, where the label distribution is replaced by the fused prediction produced through test-time calibration (TTC) and confidence-weighted fusion (CWF). This ablation allows us to isolate the performance gain specifically attributable to the thresholding step.

As illustrated in Tab. 6, ablation of the thresholding strategy results in a marked deterioration in estimation performance. Without thresholding, 15 out of 18 benchmark set-

Table 7. **Comparison with extensive baselines.** This table extends Tab. 4 with full results across all datasets. Background colors denote reference sources: CLIP, SigLIP, Random. Columns represent: (1) **Base**: reference pseudo-labels; (2) **Fix**: FRAP with fixed reference scaling; (3) **Dyna**: our full framework with Test-Time Calibration (TTC). **Random** reports performance using a pathological reference via Dirichlet distribution.

Data	Base	Fix	Dyna	Base	Fix	Dyna	Random
MNIST	24.98$_{.21}$	**11.13**$_{.10}$	12.42$_{.10}$	32.09$_{.26}$	13.54$_{.10}$	**10.56**$_{.10}$	11.30$_{.10}$
C10-N	5.00$_{.15}$	8.48$_{.03}$	**2.14**$_{.02}$	**2.01**$_{.00}$	6.29$_{.02}$	7.54$_{.03}$	11.05$_{.04}$
C10-S	10.49$_{.13}$	8.01$_{.05}$	**3.24**$_{.05}$	9.14$_{.09}$	**5.67**$_{.04}$	7.68$_{.06}$	12.61$_{.07}$
C100-N	9.07$_{.08}$	10.90$_{.05}$	**4.00**$_{.03}$	**4.35**$_{.01}$	4.65$_{.05}$	10.22$_{.04}$	12.71$_{.06}$
C100-S	**9.51**$_{.16}$	17.07$_{.10}$	11.42$_{.10}$	**7.05**$_{.05}$	12.51$_{.09}$	17.87$_{.10}$	19.28$_{.10}$
IN-S	7.08$_{.29}$	**3.11**$_{.02}$	4.76$_{.01}$	4.02$_{.02}$	**3.57**$_{.02}$	3.74$_{.02}$	2.99$_{.02}$
IN-N	9.25$_{.24}$	**1.43**$_{.01}$	2.16$_{.03}$	3.80$_{.02}$	**0.88**$_{.01}$	1.15$_{.01}$	1.97$_{.02}$
IN200-S	12.97$_{.23}$	**4.95**$_{.04}$	9.98$_{.04}$	8.74$_{.02}$	6.33$_{.04}$	**6.14**$_{.05}$	4.86$_{.04}$
IN200-N	11.97$_{.19}$	**5.32**$_{.06}$	7.46$_{.04}$	6.53$_{.04}$	**4.31**$_{.05}$	5.43$_{.07}$	5.91$_{.07}$
FMoW	40.86$_{.10}$	**3.39**$_{.03}$	3.42$_{.02}$	41.27$_{.07}$	**2.20**$_{.02}$	2.70$_{.03}$	2.48$_{.03}$
L17-S	**4.74**$_{.23}$	6.15$_{.05}$	8.55$_{.05}$	**4.66**$_{.04}$	6.33$_{.05}$	6.81$_{.06}$	5.80$_{.06}$
NL26-S	7.79$_{.26}$	**5.64**$_{.05}$	9.30$_{.07}$	7.16$_{.04}$	**5.28**$_{.04}$	7.60$_{.09}$	7.24$_{.08}$
E13-S	10.39$_{.19}$	**4.93**$_{.05}$	8.29$_{.05}$	12.38$_{.05}$	**4.55**$_{.04}$	5.78$_{.07}$	5.75$_{.06}$
E30-S	10.53$_{.26}$	**4.07**$_{.05}$	6.94$_{.06}$	10.02$_{.05}$	**3.92**$_{.05}$	5.81$_{.08}$	5.26$_{.07}$
L17-N	1.82$_{.22}$	**5.70**$_{.05}$	5.37$_{.05}$	**1.65**$_{.02}$	5.26$_{.04}$	6.74$_{.07}$	10.20$_{.07}$
NL26-N	**0.85**$_{.18}$	5.74$_{.05}$	6.66$_{.09}$	**1.02**$_{.01}$	4.91$_{.04}$	8.48$_{.10}$	9.04$_{.09}$
E13-N	**5.19**$_{.21}$	6.98$_{.06}$	5.90$_{.06}$	**6.14**$_{.03}$	6.43$_{.05}$	8.50$_{.07}$	8.70$_{.07}$
E30-N	**3.19**$_{.22}$	6.01$_{.05}$	5.54$_{.07}$	**3.77**$_{.02}$	5.08$_{.05}$	9.05$_{.09}$	8.70$_{.07}$
Avg	10.32	6.61	**6.53**	9.21	**5.65**	7.32	8.10

tings exhibit a higher estimation error, and the average MAE rises substantially from 6.53 to 10.04. This pronounced drop underscores the effect of thresholding in our framework.

The thresholding strategy effectively alleviates the accumulated error arising from the soft-to-hard score difference and the inherent limitation of the fused predictions.