PRIOR-GUIDED IMPORTANCE WEIGHTING FOR SUPERVISED FINE-TUNING OF MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Supervised fine-tuning (SFT) is critical for adapting pretrained multimodal large language models (MLLMs) to downstream tasks, yet it is often performed with limited target-domain data, making the training outcome highly sensitive to how supervision is utilized. Recent data-centric fine-tuning methods address this challenge by selecting or reweighting target samples, but they often rely on restrictive assumptions or additional components, and their emphasis on better fitting the target distribution may still lead to over-reliance on scarce target supervision. In this work, we view limited-data MLLM fine-tuning as a trade-off between target adaptation and pretrained prior preservation, and propose Prior-Guided Importance Weighting (PGIW), an importance-sampling-based framework that incorporates pretrained model guidance without accessing the original pretraining data. Our theoretical analysis derives an optimal balancing coefficient under an expected Kullback-Leibler divergence risk, while importance weighting enables PGIW to approximate the unavailable pretraining-distribution term using target-domain samples. We further instantiate PGIW as an efficient fine-tuning algorithm for MLLMs. Experiments on multimodal benchmarks demonstrate that PGIW improves target-task performance and cross-domain generalization over standard fine-tuning baselines. The source code is available at this URL.

1 INTRODUCTION

Multimodal large language models (MLLMs) have demonstrated remarkable capabilities in integrating visual and linguistic information, enabling a wide range of applications such as visual question answering (Li et al., 2023), image captioning (Alayrac et al., 2022), and multimodal reasoning (Liu et al., 2023b). In practice, these capabilities are typically acquired through a two-stage pipeline: large-scale multimodal pretraining, followed by supervised fine-tuning (SFT) on task-specific training datasets. Despite the strong representational power obtained during pretraining, the performance of the fine-tuned model critically depends on the quality of the SFT data (Zhou et al., 2023). However, in contrast to the massive data used for pretraining or the broad knowledge distribution implicitly represented by downstream test data, the available fine-tuning datasets are often severely limited. This data scarcity issue is further exacerbated in the multimodal setting, where the increased modality complexity substantially raises the data requirement for effective adaptation (Alayrac et al., 2022; Huang et al., 2023). These challenges have motivated a growing body of data-centric fine-tuning methods, which aim to more efficiently leverage limited SFT data through techniques such as data selection, reweighting, and importance sampling.

Existing data-centric fine-tuning methods for LLMs and MLLMs typically aim to improve supervised fine-tuning by reweighting or selecting training samples so that the effective training distribution better approximates the real-world knowledge implicitly reflected by downstream test data. While effective, this target-focused perspective may still lead to over-reliance on scarce target supervision in limited-data regimes, especially when the fine-tuning set is narrow or domain-biased. Moreover, many such approaches rely on auxiliary components, such as scoring or reference models, which introduce additional computational overhead and system complexity (Wang et al., 2025).

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.

Besides, several LLM-oriented importance sampling methods assume that the distribution changes before and after SFT between the small and large model are proportional (Liu et al., 2025; 2024). This assumption becomes considerably less applicable in the multimodal setting, where learning dynamics can vary substantially across scales due to heterogeneous modality interactions. As a result, existing importance weighting strategies do not readily extend to MLLM fine-tuning, leaving effective and principled importance weighting methods for MLLMs largely unexplored.

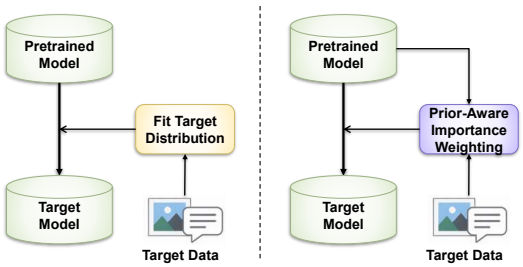


Figure 1: **Comparison between conventional target-focused fine-tuning and our prior-guided approach**. **(Left)** Most existing data-centric fine-tuning methods implicitly aim to approximate the underlying target data distribution by selecting or reweighting target samples. **(Right)** We formulate fine-tuning as a prior-guided risk minimization problem that balances target-domain supervision with pretrained-model guidance, encouraging the model to adapt to the target task while mitigating over-reliance on limited target data.

In this work, we propose **P**rior-**G**uided **I**mportance-**W**eighting (PGIW) for SFT of MLLMs. As illustrated in Fig. 1, PGIW does not rely solely on limited target-domain supervision; instead, it incorporates guidance from the pretrained model during fine-tuning. Specifically, we formulate limited-data MLLM fine-tuning as a trade-off between target adaptation and pretrained prior preservation under an expected Kullback–Leibler (K-L) divergence risk. By leveraging importance weighting, PGIW combines target-domain supervision with prior guidance induced by model-based density-ratio estimates, without requiring access to the original pretraining data. Our theoretical analysis derives an optimal balancing weight that controls this trade-off, and we further instantiate it as a practical fine-tuning algorithm for MLLMs. Experiments evaluate PGIW on visual recognition, cross-domain generalization, and multimodal reasoning, including comparisons with prior-preserving fine-tuning baselines.

Our main contributions are summarized as follows:

- **Theory.** We formulate SFT of MLLMs as a prior-target balancing problem under an expected KL-divergence risk. Leveraging the principle of importance sampling, we establish PGIW, a prior-guided importance-weighting framework that combines target-domain supervision with pretrained-model guidance without access to pretraining data, and derive an optimal balancing weight that controls this trade-off.
- **Algorithm.** Based on the theoretical analysis, we propose the practical algorithm of PGIW for MLLM training. Specifically, the proposed algorithm is architecture-agnostic and can be readily applied to different MLLM architectures.
- **Experiments.** We evaluate PGIW across in-domain visual recognition, cross-domain generalization, and multimodal reasoning tasks using LLaVA-1.5-7B and Qwen2-VL-7B. PGIW improves visual recognition and cross-domain generalization and outperforms the compared regularization, distillation, and fixed importance-weighting baselines on both reasoning benchmarks with both backbones.

2 Preliminaries

2.1 Importance Sampling

Importance sampling is a standard technique for estimating expectations under a target distribution using samples drawn from a different distribution (Liu et al., 2025). Let $p(x)$ denote a target distribution and $q(x)$ a sampling distribution. For any measurable function f, the expectation under p can be expressed as

$$\mathbb{E}_p[f(x)] = \sum_{x \in \mathcal{X}} q(x) \frac{p(x)}{q(x)} f(x) = \mathbb{E}_q[w(x)\, f(x)]\,, \tag{1}$$

where $\mathcal{X}$ denotes the sample space and $w(x) = \frac{p(x)}{q(x)}$ is the importance weight. Given samples $\{x_i\}_{i=1}^N \sim q$, the expectation can be approximated by the weighted empirical average

$$\mathbb{E}_p[f(x)] \approx \frac{1}{N} \sum_{i=1}^{N} w(x_i)\, f(x_i). \tag{2}$$

For $w(x) = \frac{p(x)}{q(x)}$ to be well defined, we assume $p \ll q$, meaning that $q(x) > 0$ whenever $p(x) > 0$. This condition concerns the underlying distributions rather than the finite samples: samples from q estimate the weighted expectation in equation 2, while $w(x)$ corrects probability-scale differences wherever $q(x) > 0$. We also assume that $w(x)f(x)$ has a finite second moment under q, which keeps the empirical estimate stable.

2.2 Asymptotic Normality and Fisher Information

When estimating an unknown parameter vector $\underline{\theta}^* \in \mathbb{R}^d$ from independent and identically distributed (i.i.d.) observations generated according to the distribution $P_{X;\underline{\theta}^*}$, let $\mathcal{D} = \{x_1, \ldots, x_N\}$ denote a collection of N i.i.d. samples. The maximum likelihood estimator (MLE) is then defined as the solution to the following optimization problem:

$$\hat{\underline{\theta}}_{\mathrm{MLE}} = \arg\max_{\underline{\theta}} \sum_{x \in \mathcal{D}} \log P_{X;\underline{\theta}}(x). \tag{3}$$

Under standard regularity assumptions on the parametric family $\{P_{X;\underline{\theta}}\}$, the MLE enjoys the classical asymptotic normality property (Van der Vaart, 2000). Specifically, as the sample size N tends to infinity, we have

$$\sqrt{N}\big(\hat{\underline{\theta}}_{\mathrm{MLE}} - \underline{\theta}^*\big) \xrightarrow{d} \mathcal{N}\big(\underline{0},\, J(\underline{\theta}^*)^{-1}\big)\,, \tag{4}$$

where $J(\underline{\theta}) \in \mathbb{R}^{d\times d}$ denotes the Fisher information matrix, defined by

$$J(\underline{\theta}) = \mathbb{E}\left[\left(\frac{\partial}{\partial \underline{\theta}} \log P_{X;\underline{\theta}}\right)\left(\frac{\partial}{\partial \underline{\theta}} \log P_{X;\underline{\theta}}\right)^{\top}\right]. \tag{5}$$

2.3 Problem Formulation

Let $P_{X,\underline{\theta}_{\mathrm{pre}}}$ denote the pretraining data distribution model. Here, $\underline{\theta} \in \mathbb{R}^d$ is the parameter vector of the model, and X denotes the joint random variable of inputs and outputs, *i.e.*, $X = (Z, Y)$. Specifically, we formulate the pretraining process as a maximum likelihood estimation process. Suppose that the pretrained model $P_{X,\hat{\underline{\theta}}_{\mathrm{pre}}}$ is obtained by pretraining on N_{pre} inaccessible samples i.i.d. drawn from $P_{X,\underline{\theta}_{\mathrm{pre}}}$, denoted by $X^{N_{\mathrm{pre}}} = \{x_1, \ldots, x_{N_{\mathrm{pre}}}\}$, i.e.,

$$\hat{\theta}_{\mathrm{pre}} = \arg\max_{\theta} \sum_{x \in X^{N_{\mathrm{pre}}}} \log P_{X;\theta}(x). \tag{6}$$

Suppose that $P_{X,\underline{\theta}_{\mathrm{tgt}}}$ denotes the distribution of the target data. Then, in fine-tuning, we are given a collection of N_{tgt} i.i.d. samples $X^{N_{\mathrm{tgt}}} = \{x_1, \ldots, x_{N_{\mathrm{tgt}}}\}$ drawn from the target distribution $P_{X,\underline{\theta}_{\mathrm{tgt}}}$. If we directly perform training to obtain $P_{X,\hat{\underline{\theta}}_{\mathrm{tgt}}}$, then the resulting estimator $\hat{\underline{\theta}}_{\mathrm{tgt}}$ can be expressed as

$$\hat{\theta}_{\mathrm{tgt}} = \arg\max_{\theta} \sum_{x \in X^{N_{\mathrm{tgt}}}} \log P_{X;\theta}(x). \tag{7}$$

In this sense, fine-tuning large models can be understood as a process of knowledge transfer from the pretrained model to the target task. In prior transfer learning approaches (Shui et al., 2021; Zhang et al., 2026a), the model is fine-tuned on a mixture of target samples and weighted pretraining

samples. Let $\hat{\underline{\theta}}_{\text{mix}}$ denote the maximum likelihood estimator obtained from N_{tgt} samples drawn from $P_{X;\underline{\theta}_{\text{tgt}}}$ and N_{pre} samples drawn from $P_{X;\underline{\theta}_{\text{pre}}}$, where the pretraining samples are assigned a transfer weight $w \in [0, +\infty)$, i.e.,

$$\hat{\underline{\theta}}_{\text{mix}} = \arg\max_{\underline{\theta}} \left(\sum_{x \in X^{N_{\text{tgt}}}} \log P_{X;\underline{\theta}}(x) \; + \sum_{x \in X^{N_{\text{pre}}}} w \log P_{X;\underline{\theta}}(x) \right) \tag{8}$$

$$= \arg\max_{\underline{\theta}} \left(N_{\text{tgt}} \mathbb{E}_{\hat{P}_{X^{N_{\text{tgt}}}}} [\log P_{X;\theta}(x)] + w N_{\text{pre}} \mathbb{E}_{\hat{P}_{X^{N_{\text{pre}}}}} [\log P_{X;\theta}(x)] \right). \tag{9}$$

However, in MLLM fine-tuning, $X^{N_{\text{pre}}}$ and its empirical distribution $\hat{P}_{X^{N_{\text{pre}}}}$ are generally inaccessible. We therefore use $\hat{P}_{X^{N_{\text{tgt}}}} \frac{P_{X;\hat{\underline{\theta}}_{\text{pre}}}}{P_{X;\hat{\underline{\theta}}_{\text{tgt}}}}$ as a model-based surrogate for the empirical pretraining term in equation 9. This surrogate does not reconstruct the original pretraining samples. Instead, the N_{tgt} target samples approximate the pretraining term's contribution to the likelihood objective wherever the target distribution is nonzero, while the density ratio corrects probability-scale differences in that region. Pretraining mass outside this region cannot be recovered from target data and is assumed to have limited effect on the target objective. The proposed estimator can be expressed as

$$\hat{\underline{\theta}}_{\text{iw}} = \arg\max_{\underline{\theta}} \left(N_{\text{tgt}} \mathbb{E}_{\hat{P}_{X^{N_{\text{tgt}}}}} [\log P_{X;\theta}(x)] + w N_{\text{pre}} \mathbb{E}_{\hat{P}_{X^{N_{\text{tgt}}}} \frac{P_{X;\hat{\underline{\theta}}_{\text{pre}}}}{P_{X;\hat{\underline{\theta}}_{\text{tgt}}}}} [\log P_{X;\theta}(x)] \right), \tag{10}$$

$$= \arg\max_{\underline{\theta}} \left(\sum_{x \in X^{N_{\text{tgt}}}} \log P_{X;\underline{\theta}}(x) + \frac{w N_{\text{pre}}}{N_{\text{tgt}}} \sum_{x \in X^{N_{\text{tgt}}}} \frac{P_{X;\hat{\underline{\theta}}_{\text{pre}}}(x)}{P_{X;\hat{\underline{\theta}}_{\text{tgt}}}(x)} \log P_{X;\underline{\theta}}(x) \right), \tag{11}$$

From equation 11, we observe that the resulting training objective can be decomposed into two terms. The first term corresponds to the conventional loss for fine-tuning, while the second term is a prior-guided term that aims to preserve knowledge encoded in the pretrained model. In the second term, the effective **importance weight** assigned to each sample is given by the product of two factors. The first factor is $\frac{P_{X;\hat{\underline{\theta}}_{\text{pre}}}(x)}{P_{X;\hat{\underline{\theta}}_{\text{tgt}}}(x)}$, which is the **density ratio** between the pretrained model and the target model evaluated at the sample, and the second factor is a global **balancing weight** between the two terms, which can be expressed as

$$\alpha \; = \; \frac{w \, N_{\text{pre}}}{N_{\text{tgt}}}. \tag{12}$$

In this work, we first derive the optimal **balancing weight** α^* from a theoretical perspective, and then discuss how to implement the entire importance weighting framework in a practical training algorithm. We adopt the expected K-L divergence (Cover, 1999) between the true target distribution $P_{X;\underline{\theta}_{\text{tgt}}}$ and the fine-tuned model $P_{X;\hat{\underline{\theta}}_{\text{iw}}}$ as the risk measure, and determine the optimal balancing weight α by minimizing this measure. Formally, the optimization problem is given by

$$\alpha^* = \arg\min_{\alpha} \mathbb{E}\Big[D\Big(P_{X;\underline{\theta}_{\text{tgt}}} \,\|\, P_{X;\hat{\underline{\theta}}_{\text{iw}}} \Big) \Big], \tag{13}$$

where D denotes the K-L divergence.

3 Prior-Guided Importance Weighting

3.1 Theory of the Optimal Balancing Weight

In analyzing the target parameter estimation problem, directly evaluating the K-L measure is generally intractable. We therefore conduct our analysis primarily in the asymptotic regime, where the behavior

of the estimator admits a more tractable characterization. Following prior works (Zhang et al., 2026a;b), we assume that N_{tgt} and N_{pre} are asymptotically comparable and that the discrepancy between the target and pretrained parameters is sufficiently small, i.e., $||\underline{\theta}_{\text{tgt}} - \underline{\theta}_{\text{pre}}|| = O(\frac{1}{\sqrt{N_{\text{tgt}}}})$. The following theorem characterizes the optimal balancing weight.

Theorem 1. *(proved in Appendix B.1) For the fine-tuning from $P_{X;\underline{\theta}_{\text{pre}}}$ to $P_{X;\underline{\theta}_{\text{tgt}}}$, we assume that $\underline{\theta}_{\text{tgt}}, \underline{\theta}_{\text{pre}} \in \mathbb{R}^d$ and $||\underline{\theta}_{\text{tgt}} - \underline{\theta}_{\text{pre}}|| = O(\frac{1}{\sqrt{N_{\text{tgt}}}})$. Then, by minimizing the K-L measure $\mathbb{E}\left[D(P_{X;\underline{\theta}_{\text{tgt}}}||P_{X;\hat{\underline{\theta}}_{\text{iw}}})\right]$, the **optimal balancing weight** α^* is*

$$\alpha^* = \frac{N_{\text{pre}}}{N_{\text{tgt}} + tN_{\text{pre}}N_{\text{tgt}}}, \tag{14}$$

where

$$t \triangleq \frac{(\underline{\theta}_{\text{pre}} - \underline{\theta}_{\text{tgt}})^\top J(\underline{\theta}_{\text{pre}})(\underline{\theta}_{\text{pre}} - \underline{\theta}_{\text{tgt}})}{d}. \tag{15}$$

Here, t is a scalar ; $J(\underline{\theta}_{\text{pre}}) \in \mathbb{R}^{d\times d}$ denotes the Fisher information matrix evaluated at $\underline{\theta}_{\text{pre}}$, and $(\underline{\theta}_{\text{pre}} - \underline{\theta}_{\text{tgt}}) \in \mathbb{R}^d$ is the coordinate-wise difference between the two parameter vectors.

Remark 2. *Theorem 1 builds on the global balancing principle in (Zhang et al., 2026a). The new challenge is to realize this principle when the pretraining samples are inaccessible. PGIW addresses this challenge through a two-level weighting scheme that combines a global balancing weight with sample-wise model-based density ratios, and instantiates it as a practical self-consistent fine-tuning algorithm. Thus, the main advance lies in the pretraining-data-free realization rather than in the closed-form coefficient alone.*

We analyze the optimal balancing weight α^*. Through a simple transformation, we can obtain $\alpha^* = \frac{\frac{1}{N_{\text{tgt}}}}{\frac{1}{N_{\text{pre}}}+t}$. By equation 15, we can interpret t as a measure of the distributional discrepancy between the pretraining data and the target data. Our analysis shows that when $t \to 0$, the α^* converges to $\frac{N_{\text{pre}}}{N_{\text{tgt}}}$, indicating that the optimization relies more heavily on guidance from the pretrained model. In contrast, when t becomes large, α^* approaches zero, implying that the contribution of pretrained prior information is largely diminished.

In equation 14, a key technical challenge is how to characterize the quantity $\underline{\theta}_{\text{pre}} - \underline{\theta}_{\text{tgt}}$ in equation 15. We employ a method referred to as the **natural-gradient** or **Fisher-gradient** (Amari, 1998; Zhang et al., 2026b). Specifically, we approximate the discrepancy by the gradient preconditioned with the inverse Fisher information matrix, i.e.,

$$\underline{\theta}_{\text{pre}} - \underline{\theta}_{\text{tgt}} \approx J(\underline{\theta}_{\text{pre}})^{-1}G, \tag{16}$$

where G denotes the gradient of the target samples evaluated at the pretrained parameters. In our framework, the Fisher information matrix is already computed when evaluating the $J(\underline{\theta}_{\text{pre}})$ in equation 15, and can thus be directly reused.

We observe that, compared with equation 7, the proposed estimator equation 11 based on importance weighting preserves more information from the pretrained model. Theoretical results in Theorem 1 demonstrate that retaining such information can enhance the training of the target model. Moreover, preserving more information from the pretrained model naturally contributes to improved cross-domain generalization, a phenomenon that we further validate empirically in our experiments.

3.2 Multimodal Large Model Framework

We build upon a pretrained MLLM and reorganize its internal information flow into a structured pipeline that is particularly suited for generative visual classification. Rather than treating the MLLM as a monolithic end-to-end system, we explicitly expose and fix the roles of its visual encoder, multimodal projector, and language model, enabling a clear separation between visual representation extraction, cross-modal alignment, and linguistic reasoning.

Given an input image v, a frozen vision encoder first produces a sequence of visual tokens

$$z_v = \phi(v), \tag{17}$$

where $\phi(\cdot)$ denotes a pretrained visual backbone. These visual tokens are then mapped into the language embedding space through a pretrained multimodal projection module, yielding projected visual representations that are compatible with the token embeddings of the language model. In parallel, a fixed textual prompt u is tokenized and embedded into a sequence of textual tokens z_u using the input embedding layer of the LLM. The projected visual tokens and the textual prompt tokens are subsequently concatenated at the token level to form a unified multimodal sequence:

$$z = \mathrm{Concat}(z_v, z_u). \tag{18}$$

The resulting multimodal token sequence z is fed into the language model, which performs autoregressive generation over the combined visual–textual context. Formally, the model defines a conditional distribution

$$P_{Y|Z;\underline{\theta}}(y \mid z), \tag{19}$$

where θ denotes the adaptable model parameters, with the remaining parameters held fixed. This formulation allows the language model to reason jointly over visual evidence and textual class semantics within a single autoregressive framework.

Generative Classification via Token-Level Fusion. For downstream classification, we formulate prediction as constrained text generation. A fixed prompt enumerating all candidate classes is appended to the visual token sequence, and the model generates the class name corresponding to the input image. By performing multimodal fusion explicitly at the token level, this framework avoids task-specific classifier heads and preserves the representational structure of the pretrained MLLM.

Unless otherwise specified, the vision encoder, multimodal projector, and the backbone language model are all kept frozen. Adaptation uses LoRA: in this implementation, the theoretical parameter θ collects the trainable LoRA parameters, with the pretrained reference corresponding to their initialization. Accordingly, d, G, and J refer to the dimension, gradient statistic, and Fisher information in these coordinates.

3.3 Practical Algorithm

We consider supervised fine-tuning of a pretrained MLLM for image classification. Each target-domain sample is denoted by $x = (z, y)$, where z is the input which has been illustrated in equation 18, and $y \in \{0, \ldots, C-1\}$ is a class label. Following standard practice, we adopt a fixed textual prompt that enumerates all class indices and formulate classification as next-token prediction. The procedures of our method are summarized in Algorithm 1. In the following paragraphs, we provide detailed explanations of it.

Prior-Guided Importance Weighting. As analyzed in Section 3.1, we can use importance weighting to preserve prior information and facilitate the fine-tuning process. For each sample (z, y), we define the density ratio

$$r(z,y) = \frac{P_{X;\underline{\hat{\theta}}_{\mathrm{pre}}}(x)}{P_{X;\underline{\hat{\theta}}_{\mathrm{tgt}}}(x)} = \frac{P_{Z,Y;\underline{\hat{\theta}}_{\mathrm{pre}}}(z,y)}{P_{Z,Y;\underline{\hat{\theta}}_{\mathrm{tgt}}}(z,y)} = \frac{P_{Y|Z;\underline{\hat{\theta}}_{\mathrm{pre}}}(y \mid z) P_Z(z)}{P_{Y|Z;\underline{\hat{\theta}}_{\mathrm{tgt}}}(y \mid z) P_Z(z)} = \frac{P_{Y|Z;\underline{\hat{\theta}}_{\mathrm{pre}}}(y \mid z)}{P_{Y|Z;\underline{\hat{\theta}}_{\mathrm{tgt}}}(y \mid z)}, \tag{20}$$

where the numerator is evaluated using the frozen pretrained model and the denominator uses the current fine-tuned model. Since MLLMs model the conditional label distribution given the multimodal input, we treat the input marginal $P_Z(z)$ as a fixed reference measure shared by the pretrained and fine-tuned models. Under this conditional modeling view, the joint ratio reduces to the conditional likelihood ratio in equation 20.

Following the formulation in equation 11, our theoretical framework is instantiated during training primarily through a modification of the loss function. Specifically, the training objective $\mathcal{L}_{\mathrm{Train}}$ consists of two components: a conventional loss term $\mathcal{L}_{\mathrm{Base}}$ and an additional importance-weighted

loss term $\mathcal{L}_{\mathrm{IW}}$ that incorporates prior information, and the weighted coefficient α between them is analytically determined by Theorem 1. The expression of $\mathcal{L}_{\mathrm{Train}}$ is

$$\mathcal{L}_{\mathrm{Train}} = \mathcal{L}_{\mathrm{Base}} + \alpha \mathcal{L}_{\mathrm{IW}}, \tag{21}$$

where the two terms are

$$\mathcal{L}_{\mathrm{Base}} = \mathbb{E}_{(z,y)\sim\mathcal{D}_{\mathrm{tgt}}}\big[\,\ell(\hat{\theta}_{\mathrm{iw}}; z, y)\big], \qquad \mathcal{L}_{\mathrm{IW}} = \mathbb{E}_{(z,y)\sim\mathcal{D}_{\mathrm{tgt}}}\big[r(z,y)\,\ell(\hat{\theta}_{\mathrm{iw}}; z, y)\big], \tag{22}$$

where ℓ denotes the cross-entropy loss, and $\mathcal{D}_{\mathrm{tgt}}$ denotes the training dataset used for fine-tuning. Therefore, the final expression of $\mathcal{L}_{\mathrm{Train}}$ is

$$\mathcal{L}_{\mathrm{Train}} = \mathbb{E}_{(z,y)\sim\mathcal{D}_{\mathrm{tgt}}}\big[\,(1 + r(z,y)\alpha)\;\ell(\hat{\theta}_{\mathrm{iw}}; z, y)\big]. \tag{23}$$

In addition, the computation of α^* in equation 12 requires an estimate of N_{pre}. In practice, the literal pretraining sample size of an MLLM is often only specified at a coarse token-count scale. Accordingly, we consider the following two practical strategies for specifying α^*.

(a) Proxy-based finite-N_{pre} approximation. As a practical proxy, we estimate the effective pretraining sample size N_{pre} by multiplying the effective target sample size N_{tgt} with an estimated ratio of pretraining to target tokens. We then compute α^* using equation 12.

(b) Infinite-pretraining approximation. If such a proxy is unavailable in practice, one may instead adopt the limiting form of α^* as $N_{\mathrm{pre}} \to \infty$, namely,

$$\alpha^* = \lim_{N_{\mathrm{pre}}\to\infty} \frac{N_{\mathrm{pre}}}{N_{\mathrm{tgt}} + tN_{\mathrm{pre}}N_{\mathrm{tgt}}} = \frac{1}{t\,N_{\mathrm{tgt}}}. \tag{24}$$

In this way, α^* becomes independent of N_{pre}. From a practical perspective, this fallback is well suited to modern large-model scenarios, in which the effective pretraining size is typically far larger than the target fine-tuning size. In all our experiments, we adopt strategy (b).

Diagonal Empirical Fisher. For the efficient computation of the Fisher information matrix in equation 14 for MLLMs, we adopt the diagonal empirical Fisher approximation (He et al., 2024; Amari, 1998), which computes a diagonal matrix through the accumulation of outer products of gradients. Moreover, to estimate the parameter discrepancy via the Fisher-gradient via equation 16, we compute the gradient statistic G, which is the average gradient of the target loss evaluated at the pretrained parameters. Both the diagonal empirical Fisher matrix J and the gradient statistic G require only a small subset of target-domain samples $\mathcal{D}_{\mathrm{tgt}}^{\mathrm{sub}} \subset \mathcal{D}_{\mathrm{tgt}}$ to compute, rather than the full target dataset. This design makes the proposed procedure lightweight and computationally efficient, while providing sufficiently accurate statistics for initialization and distance estimation.

Self-Consistent Importance Reweighting. In equation 20, the explicit introduction of $\hat{\underline{\theta}}_{\mathrm{tgt}}$ can complicate the practical implementation, as it requires a separate conventional fine-tuning procedure on the training data (cf. equation 7). To address this issue, we adopt a *self-consistent importance reweighting* strategy. Specifically, we replace $\hat{\underline{\theta}}_{\mathrm{tgt}}$ with the proposed estimator $\hat{\underline{\theta}}_{\mathrm{iw}}$ in equation 11, which is initialized as $\hat{\underline{\theta}}_{\mathrm{iw}} = \hat{\underline{\theta}}_{\mathrm{pre}}$. During fine-tuning, the importance ratios are computed using the current estimate $\hat{\underline{\theta}}_{\mathrm{iw}}$ and used to construct the training loss for gradient-based optimization. As the model parameters evolve, the induced importance weights are updated accordingly, yielding a self-consistent refinement of the weighting scheme.

4 Experiments

4.1 Experimental Setup

Datasets and models. We evaluate PGIW on ImageNet-100 for in-domain recognition, DomainNet for cross-domain generalization, and ScienceQA and M3CoT for multimodal reasoning (Appendix D). DomainNet has six domains: Real (R), Clipart (C), Painting (P), Sketch (S), Infograph (I), and Quickdraw (Q). Arrows indicate training-to-test domain transfer, e.g., R→C means training on Real and testing on Clipart. We use LLaVA-1.5-7B and Qwen2-VL-7B, with the vision encoder, multimodal projector, and backbone language model frozen. The DomainNet experiments use LLaVA-1.5-7B. All fine-tuning methods use LoRA modules in the transformer attention layers.

Algorithm 1 PGIW Algorithm

Input: Model type $P_{Y|Z;\underline{\hat{\theta}}}$ and pretrained parameter $\underline{\hat{\theta}}_{\text{pre}}$, parameter dimension d, the pretraining data size N_{pre}, target dataset $D_{\text{tgt}} = \{(z^j_{\text{tgt}}, y^j_{\text{tgt}})\}_{j=1}^{N_{\text{tgt}}}$, sampled data $\mathcal{D}^{\text{sub}}_{\text{tgt}} = \{(z^j, y^j)\}_{j=1}^{|\mathcal{D}^{\text{sub}}_{\text{tgt}}|}$, loss function ℓ, // $|\mathcal{D}^{\text{sub}}_{\text{tgt}}| \ll N_{\text{tgt}}$

Parameter: Learning rate η.

Initialize: initialize the target model parameter $\underline{\hat{\theta}}_{\text{iw}} = \underline{\hat{\theta}}_{\text{pre}}$, initialize weights $\alpha = 0$

Output: a well-trained $\underline{\hat{\theta}}_{\text{iw}}$ for target model $P_{Y|Z;\underline{\hat{\theta}}_{\text{iw}}}$

$$G^{d\times 1} \leftarrow \frac{1}{|\mathcal{D}^{\text{sub}}_{\text{tgt}}|} \sum_{(z^j,y^j)\in\mathcal{D}^{\text{sub}}_{\text{tgt}}} \nabla_{\underline{\theta}} \ell(\hat{\theta}_{\text{pre}}; z^j, y^j)$$

$$\boldsymbol{J}^{d\times d} \leftarrow \text{Diag}\left(\frac{1}{|\mathcal{D}^{\text{sub}}_{\text{tgt}}|} \sum_{(z^j,y^j)\in\mathcal{D}^{\text{sub}}_{\text{tgt}}} \nabla_{\underline{\theta}} \ell(\hat{\theta}_{\text{pre}}; z^j, y^j) \odot \nabla_{\underline{\theta}} \ell(\hat{\theta}_{\text{pre}}; z^j, y^j) \right)$$

$$\underline{\theta}_{\text{pre}} - \underline{\theta}_{\text{tgt}} \leftarrow J^{-1}G$$

$$t \leftarrow \frac{(\underline{\theta}_{\text{pre}} - \underline{\theta}_{\text{tgt}})^\top J (\underline{\theta}_{\text{pre}} - \underline{\theta}_{\text{tgt}})}{d}$$

$$\alpha \leftarrow \frac{N_{\text{pre}}}{N_{\text{tgt}} + t N_{\text{pre}} N_{\text{tgt}}}$$

repeat // Self-Consistent Importance Reweighting

$$\mathcal{L}_{\text{Train}} \leftarrow \frac{1}{|D_{\text{tgt}}|} \sum_{(z^j,y^j)\in D_{\text{tgt}}} \ell(\hat{\theta}_{\text{iw}}; z, y) \left(1 + \frac{P_{Y|Z;\underline{\hat{\theta}}_{\text{pre}}}(y|z)}{P_{Y|Z;\underline{\hat{\theta}}_{\text{iw}}}(y|z)} \alpha \right)$$

$$\underline{\hat{\theta}}_{\text{iw}} \leftarrow \underline{\hat{\theta}}_{\text{iw}} - \eta \nabla_{\underline{\theta}} \mathcal{L}_{\text{Train}}$$

until $\underline{\hat{\theta}}_{\text{iw}}$ converges;

Table 1: Top-1 accuracy (%) on ImageNet-100.

Method	LLaVA	Qwen	Avg.
Zero-shot	48.3	55.2	51.8
LoRA	59.7	66.5	63.1
PGIW (Ours)	**61.1**	**67.8**	**64.5**

Table 2: Cross-domain accuracy (%) on DomainNet with Real as the source domain.

Method	R→C	R→P	R→S	R→I	R→Q	Avg
Zero-shot	30.2	28.7	22.5	31.1	25.3	27.6
LoRA	42.8	40.3	35.7	44.1	38.6	40.3
PGIW (Ours)	**47.1**	**45.6**	**41.2**	**48.3**	**43.7**	**45.2**

Baselines. The baseline labeled *LoRA* uses only the standard cross-entropy objective. On ScienceQA and M3CoT, we additionally compare L2-SP (LI et al., 2018), which penalizes parameter deviations from initialization; EWC (Kirkpatrick et al., 2017), which weights these deviations by Fisher information; KD (Hinton et al., 2015), which regularizes predictions toward the frozen pretrained model via KL divergence; and Fixed-IW, which uses a fixed global coefficient instead of PGIW's analytically determined coefficient. Their objectives and coefficients are given in Appendix E. Zero-shot results are also reported for the visual recognition benchmarks.

Training and evaluation. We optimize the LoRA parameters with AdamW. PGIW follows Section 3.3, using the infinite-pretraining approximation and a diagonal empirical Fisher estimate computed once before fine-tuning. Predictions are formulated as class- or answer-index prediction. We report top-1 accuracy for ImageNet-100 and DomainNet and answer accuracy for ScienceQA and M3CoT, averaged over multiple runs. In Tables 1, 3, and 4, Avg. is the mean across the two backbones.

4.2 Results and Analysis

Visual recognition and cross-domain generalization. On ImageNet-100 (Table 1), PGIW improves over LoRA by 1.4 and 1.3 percentage points on LLaVA and Qwen, respectively. These gains support its effectiveness for in-domain adaptation. On DomainNet (Table 2), PGIW improves accuracy in all five target domains, raising the average from 40.3% to 45.2%. The largest gain is on Sketch (+5.5 points), supporting the benefit of prior-guided adaptation under domain shift.

Table 3: Answer accuracy (%) on ScienceQA.

Method	LLaVA	Qwen	Avg.
LoRA	91.42	96.35	93.89
L2-SP	91.82	95.97	93.90
EWC	91.30	94.84	93.07
KD	91.87	96.39	94.13
Fixed-IW	91.42	96.44	93.93
PGIW (Ours)	**91.96**	**96.65**	**94.31**

Table 4: Answer accuracy (%) on M3CoT.

Method	LLaVA	Qwen	Avg.
LoRA	54.83	63.20	59.02
L2-SP	55.05	63.68	59.37
EWC	54.87	63.98	59.43
KD	55.00	63.89	59.45
Fixed-IW	55.26	63.76	59.51
PGIW (Ours)	**56.86**	**64.50**	**60.68**

Comparison with prior-preserving baselines. On M3CoT (Table 4), PGIW achieves 56.86% on LLaVA and 64.50% on Qwen, exceeding the strongest alternative in each column by 1.60 and 0.52 points, respectively. Relative to Fixed-IW, the gains are 1.60 and 0.74 points, supporting the practical value of the analytically determined coefficient. On ScienceQA (Table 3), PGIW achieves 91.96% and 96.65% on LLaVA and Qwen, respectively, improving over the strongest alternative for each backbone by 0.09 and 0.21 points. These margins are smaller than on M3CoT, indicating that the benefit varies across tasks and backbones.

Computational overhead and training behavior. PGIW requires an additional frozen-model forward pass. The ScienceQA timing measurements in Appendix H show per-step overheads of 27.4% for LLaVA and 39.6% for Qwen, alongside a one-time preparation cost. Training curves are reported in Appendix I; iteration counts alone do not determine wall-clock efficiency.

5 Related Work

Our work is related to data-centric fine-tuning, supervised fine-tuning of large multimodal models, and prior-preserving adaptation. Importance weighting and data selection have been widely studied for improving the efficiency and robustness of large-model fine-tuning, including layer-wise or module-wise sampling strategies and instruction-data selection methods (Pan et al., 2024; Liu et al., b; Zhang et al., 2025a; Yang et al., 2025b). In parallel, supervised fine-tuning has become a standard paradigm for adapting LLMs and MLLMs to downstream instruction-following and vision–language tasks (Brown et al., 2020; Liu et al., 2022; Ouyang et al., 2022; Liu et al., 2023a; Zhu et al., 2023). Another line of work emphasizes preserving pretrained knowledge during adaptation to achieve robust fine-tuning, through parameter-space regularization, Fisher-based constraints, feature preservation, causal regularization, or distillation-style regularization (Aljundi et al., 2018; Kirkpatrick et al., 2017; He et al., 2023; Zheng et al., 2023; Yang et al., 2024). Different from these studies, we formulate MLLM supervised fine-tuning as a prior-guided importance-weighting problem and derive an analytically motivated balancing weight that incorporates pretrained prior information without requiring access to pretraining data. We provide a more detailed version of the related work in the Appendix C.

6 Conclusion

In this paper, we presented PGIW, a prior-guided importance-weighting framework for SFT of MLLMs. We formulated fine-tuning as a trade-off between target adaptation and pretrained prior preservation, and derived an optimal balancing weight under an expected K-L divergence risk. Since the original pretraining data are inaccessible in practice, PGIW approximates the pretraining-distribution term by an importance-weighted expectation over target samples using model-based density-ratio estimates. This yields an efficient fine-tuning algorithm with diagonal empirical Fisher estimation and self-consistent reweighting. Experiments demonstrate improvements in visual recognition and cross-domain generalization, with additional comparisons on multimodal reasoning benchmarks supporting the benefits of prior-guided weighting over alternative adaptation objectives.

AI USE STATEMENT

In this work, we used OpenAI Codex to assist with interpreting author-provided experimental results and improve the language and readability of the manuscript. We have not used generative AI tools to generate synthetic datasets, develop theoretical models, formulate mathematical claims, write proofs, propose hypotheses, implement the proposed method, translate the manuscript, or clean and reformat datasets. Qualitative and thematic data analysis are not applicable to this work. We have reviewed all AI-assisted work, checked numerical comparisons against our experimental results, and verified that the revised text accurately reflects our methods, findings, and limitations. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies prior-guided importance weighting for fine-tuning multimodal large language models using existing benchmarks: ImageNet-100, DomainNet, ScienceQA, and M3CoT. It does not introduce new data collection or human-subject experiments.

REPRODUCIBILITY STATEMENT

The manuscript describes the PGIW objective, the derivation of the balancing coefficient, and the practical training procedure, with pseudocode provided in Algorithm 1. The appendix presents the proof of the main theoretical result. The methodology specifies the infinite-pretraining approximation, diagonal empirical Fisher estimation, and self-consistent importance reweighting used in our implementation. The experimental section identifies the pretrained backbones, LoRA adaptation strategy, baselines, and evaluation metrics, while the appendix describes the datasets, baseline objectives and coefficients, and additional timing and training-trajectory results. An anonymous source-code link is provided in the abstract to support reproduction of the experiments.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 139–154, 2018.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

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

Guande He, Jianfei Chen, and Jun Zhu. Preserving pre-trained features helps calibrate fine-tuned language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Yifei He, Shiji Zhou, Guojun Zhang, Hyokun Yun, Yi Xu, Belinda Zeng, Trishul Chilimbi, and Han Zhao. Robust multi-task learning with excess risks. In *International Conference on Machine Learning*, pp. 18094–18114. PMLR, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36: 72096–72109, 2023.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Meng Li, Guangda Huzhang, Haibo Zhang, Xiting Wang, and Anxiang Zeng. Optimal transport-based token weighting scheme for enhanced preference optimization. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21311–21334, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1035. URL https://aclanthology.org/2025.acl-long.1035/.

Xuhong LI, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2825–2834. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/li18a.html.

Aiwei Liu, Haoping Bai, Zhiyun Lu, Yanchao Sun, Xiang Kong, Xiaoming Simon Wang, Jiulong Shan, Albin Madappally Jose, Xiaojiang Liu, Lijie Wen, et al. Tis-dpo: Token-level importance sampling for direct preference optimization with estimated weights. In *The Thirteenth International Conference on Learning Representations*, a.

Alisa Liu, Xiaochuang Han, Yizhong Wang, Yulia Tsvetkov, Yejin Choi, and Noah A. Smith. Tuning language models by proxy. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=dribhnhm1i.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023b.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 61–68, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.8. URL https://aclanthology.org/2022.acl-short.8/.

Yi Liu, Guoyin Wang, Shicheng Li, Feifan Song, and Xu Sun. ATLANTIS: Weak-to-strong learning via importance sampling. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1042–1052, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.52. URL https://aclanthology.org/2025.acl-long.52/.

Yuxi Liu, Renjia Deng, Yutong He, Xue Wang, Tao Yao, and Kun Yuan. Misa: Memory-efficient llms optimization with module-wise importance sampling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, b.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Rui Pan, Xiang Liu, Shizhe Diao, Renjie Pi, Jipeng Zhang, Chi Han, and Tong Zhang. Lisa: Layerwise importance sampling for memory-efficient large language model fine-tuning. *Advances in Neural Information Processing Systems*, 37:57018–57049, 2024.

Changjian Shui, Zijian Li, Jiaqi Li, Christian Gagné, Charles X Ling, and Boyu Wang. Aggregating from multiple target-shifted sources. In *International conference on machine learning*, pp. 9638–9648. PMLR, 2021.

Satoshi Suzuki, Shin'ya Yamaguchi, Shoichiro Takeda, Taiga Yamane, Naoki Makishima, Naotaka Kawata, Mana Ihori, Tomohiro Tanaka, Shota Orihashi, and Ryo Masumura. Difference vector equalization for robust fine-tuning of vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9278–9286, 2026.

Junjiao Tian, Chengyue Huang, and Zsolt Kira. Rethinking weight decay for robust fine-tuning of foundation models. *Advances in Neural Information Processing Systems*, 37:22418–22440, 2024.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Shaobo Wang, Xiangqi Jin, Ziming Wang, Jize Wang, Jiajun Zhang, Kaixin Li, Zichen Wen, Zhong Li, Conghui He, Xuming Hu, and Linfeng Zhang. Data whisperer: Efficient data selection for task-specific LLM fine-tuning via few-shot in-context learning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23287–23305, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1135. URL https://aclanthology.org/2025.acl-long.1135/.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7949–7961. IEEE, 2022.

Ning Yang, Hai Lin, Yibo Liu, Baoliang Tian, Guoqing Liu, and Haijun Zhang. Token-importance guided direct preference optimization. *arXiv preprint arXiv:2505.19653*, 2025a.

Xu Yang, Chen Liu, and Ying Wei. Data selection matters: Towards robust instruction tuning of large multimodal models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=oLJMsGMfqr.

Zitong Yang, Aonan Zhang, Sam Wiseman, Xiang Kong, Ke Ye, and Dong Yin. Memory retaining finetuning via distillation. In *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*, 2024.

Bolin Zhang, Jiahao Wang, Qianlong Du, Jiajun Zhang, Zhiying Tu, and Dianhui Chu. A survey on data selection for llm instruction tuning. *Journal of Artificial Intelligence Research*, 83, 2025a.

Qingyue Zhang, Haohao Fu, Guanbo Huang, Yaoyuan Liang, Chang Chu, Tianren Peng, Yanru Wu, Qi Li, Yang Li, and Shao-Lun Huang. A high-dimensional statistical method for optimizing transfer quantities in multi-source transfer learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Qingyue Zhang, Chang Chu, Haohao Fu, Tianren Peng, Yanru Wu, Guanbo Huang, Yang Li, and Shao-Lun Huang. Unified optimization of source weights and transfer quantities in multi-source transfer learning: An asymptotic framework. *arXiv preprint arXiv:2601.10779*, 2026a.

Qingyue Zhang, Chang Chu, Tianren Peng, Qi Li, Xiangyang Luo, Zhihao Jiang, and Shao-Lun Huang. Lora-da: Data-aware initialization for low-rank adaptation via asymptotic analysis. *ICML*, 2026b.

Junhao Zheng, Qianli Ma, Shengjie Qiu, Yue Wu, Peitian Ma, Junlong Liu, Huawen Feng, Xichen Shang, and Haibin Chen. Preserving commonsense knowledge from pre-trained language models via causal inference. *arXiv preprint arXiv:2306.10790*, 2023.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.

Wenxuan Zhou, Shujian Zhang, Lingxiao Zhao, and Tao Meng. T-reg: Preference optimization with token-level reward regularization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27876–27889, 2025.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

APPENDIX

A Notation **15**

B Proof **16**

B.1 Proof of Theorem 1 16

C Extended Related Work **23**

C.1 Importance Sampling and Data Selection 23

C.2 Supervised fine-tuning (SFT) 23

C.3 Robust Fine-Tuning and Prior-Guided Fine-Tuning 23

D Dataset Description **24**

E Additional Baseline Details **24**

F Broader Impacts **24**

G Limitations **25**

H Time Complexity **25**

I Training Loss on ScienceQA (Qwen) **26**

A NOTATION

Table 5: Summary of key notation.

Notation	Description
$X = (Z, Y)$	Joint variable of multimodal input Z and output label Y.
$x = (z, y)$	A target-domain training sample.
$\mathcal{D}_{\mathrm{tgt}}$	Target-domain fine-tuning dataset.
$\mathcal{D}_{\mathrm{tgt}}^{\mathrm{sub}}$	A small subset of target samples used to estimate gradient and Fisher statistics.
$P_{X;\theta}$	Parametric data distribution induced by model parameter θ.
$P_{Y\|Z;\theta}(y\|z)$	Conditional output distribution modeled by the MLLM.
$P_{X;\theta_{\mathrm{pre}}}$, $P_{X;\theta_{\mathrm{tgt}}}$	Pretraining and target data distributions.
$\hat{\theta}_{\mathrm{pre}}$	Parameter of the pretrained model.
$\hat{\theta}_{\mathrm{tgt}}$	Estimator obtained by standard fine-tuning on target data.
$\hat{\theta}_{\mathrm{iw}}$	Estimator obtained by the proposed importance-weighted objective.
N_{pre}, N_{tgt}	Effective pretraining sample size and target fine-tuning sample size.
w	Transfer weight assigned to the pretraining-distribution term in the ideal objective.
α, α^*	Balancing coefficient and its theoretically optimal value.
$J(\theta)$	Fisher information matrix evaluated at θ.
G	Gradient statistic computed from target samples at $\hat{\theta}_{\mathrm{pre}}$.
t	Fisher-weighted discrepancy between pretraining and target parameters.
$r(z, y)$	Model-based density-ratio estimate used for importance weighting.
$D(P\|\|Q)$	Kullback–Leibler divergence between distributions P and Q.
$\ell(\theta; z, y)$	Cross-entropy loss for sample (z, y).
$\mathcal{L}_{\mathrm{Base}}$, $\mathcal{L}_{\mathrm{IW}}$	Standard target loss and importance-weighted prior-guided loss.
$\mathcal{L}_{\mathrm{train}}$	Final training objective combining $\mathcal{L}_{\mathrm{Base}}$ and $\alpha\mathcal{L}_{\mathrm{IW}}$.

B PROOF

B.1 PROOF OF THEOREM 1

Proof. We first consider the scenario in which the model parameter is one-dimensional. Based on this setting, we establish the following lemma. After proving this lemma, we extend the analysis to the high-dimensional case and complete the proof of the theorem.

Lemma 3. *For the fine-tuning from $P_{X;\theta_{\text{pre}}}$ to $P_{X;\theta_{\text{tgt}}}$, we assume that $\theta_{\text{tgt}}, \theta_{\text{pre}} \in \mathbb{R}$ and $|\theta_{\text{tgt}} - \theta_{\text{pre}}| = O(\frac{1}{\sqrt{N_{\text{tgt}}}})$. Then, by minimizing the K-L measure $\mathbb{E}\left[D(P_{X;\theta_{\text{tgt}}}||P_{X;\hat{\theta}_{\text{iw}}})\right]$, the **optimal balancing weight** α^* is*

$$\alpha^* = \frac{N_{\text{pre}}}{N_{\text{tgt}} + tN_{\text{pre}}N_{\text{tgt}}}, \tag{25}$$

where

$$t \triangleq J(\theta_{\text{pre}})(\theta_{\text{pre}} - \theta_{\text{tgt}})^2. \tag{26}$$

Proof. According to equation 12, it then suffices to determine this optimal w in order to obtain the optimal balancing weight α. Therefore, in the proof, we first derive an explicit expression of the K-L measure as a function of w. We then obtain the optimal value of w, from which the optimal balancing weight α follows directly.

From (Zhang et al., 2025b), we know that

$$D\left(P_{X;\theta_{\text{tgt}}}\middle\|P_{X;\hat{\theta}_{\text{iw}}}\right) = \frac{1}{2}\sum_{x\in X}\frac{\left(P_{X;\hat{\theta}_{\text{iw}}}(x) - P_{X;\theta_{\text{tgt}}}(x)\right)^2}{P_{X;\theta_{\text{tgt}}}(x)} + o(|\hat{\theta}_{\text{iw}} - \theta_{\text{tgt}}|^2), \tag{27}$$

We denote δ as a small constant, and we can rewrite equation 27 as

$$\begin{aligned}
&\mathbb{E}\left[D\left(P_{X;\theta_{\text{tgt}}}\middle\|P_{X;\hat{\theta}_{\text{iw}}}\right)\right] \\
&= \sum_{X^{N_{\text{tgt}}}} P_{X^{N_{\text{tgt}}};\theta_{\text{tgt}}}(X^{N_{\text{tgt}}}) \sum_{X^{N_{\text{pre}}}} P_{X^{N_{\text{pre}}};\theta_{\text{pre}}}(X^{N_{\text{pre}}}) D\left(P_{X;\theta_{\text{tgt}}}\middle\|P_{X;\hat{\theta}_{\text{iw}}}\right) \\
&= \sum_{\substack{X^{N_{\text{tgt}}}:|\hat{\theta}_{\text{tgt}}-\theta_{\text{tgt}}|^2<\delta \\ X^{N_{\text{pre}}}:|\hat{\theta}_{\text{pre}}-\theta_{\text{pre}}|^2<\delta}} P_{X^{N_{\text{tgt}}};\theta_{\text{tgt}}}(X^{N_{\text{tgt}}}) P_{X^{N_{\text{pre}}};\theta_{\text{pre}}}(X^{N_{\text{pre}}}) \cdot \left(\frac{1}{2}\sum_{x\in X}\frac{\left(\frac{\partial P_{X;\theta_{\text{tgt}}}(x)}{\partial\theta_{\text{tgt}}}(\hat{\theta}_{\text{iw}} - \theta_{\text{tgt}})\right)^2}{P_{X;\theta_{\text{tgt}}}(x)} + o(|\hat{\theta}_{\text{iw}} - \theta_{\text{tgt}}|^2)\right) \\
&+ \sum_{\substack{X^{N_{\text{tgt}}}, X^{N_{\text{pre}}}: \\ |\hat{\theta}_{\text{tgt}}-\theta_{\text{tgt}}|^2\geq\delta \text{ or } |\hat{\theta}_{\text{pre}}-\theta_{\text{pre}}|^2\geq\delta}} P_{X^{N_{\text{tgt}}};\theta_{\text{tgt}}}(X^{N_{\text{tgt}}}) P_{X^{N_{\text{pre}}};\theta_{\text{pre}}}(X^{N_{\text{pre}}}) \cdot D\left(P_{X;\theta_{\text{tgt}}}\middle\|P_{X;\hat{\theta}_{\text{iw}}}\right)
\end{aligned} \tag{28}$$

For the negligible term in the first term of equation 28, i.e., $o(|\hat{\theta}_{\text{iw}} - \theta_{\text{tgt}}|^2)$, it is small enough for us to omit it. We next analyze the second term. From (Zhang et al., 2025b), we know that for the sample set $X^{N_{\text{tgt}}}$ and its corresponding MLE $\hat{\theta}_{\text{tgt}}$, we have

$$P_{X^{N_{\text{tgt}}};\theta_{\text{tgt}}}(X^{N_{\text{tgt}}}) \doteq e^{-N_{\text{tgt}}D\left(\hat{P}_{X^{N_{\text{tgt}}}}\middle\|P_{X;\theta_{\text{tgt}}}\right)} \leq e^{\frac{-N_{\text{tgt}}J(\theta_{\text{tgt}})|\hat{\theta}_{\text{tgt}}-\theta_{\text{tgt}}|^2}{2}}. \tag{29}$$

We denote δ as a small constant. Therefore, for the scenario $|\hat{\theta}_{\text{tgt}} - \theta_{\text{tgt}}|^2 \geq \delta$, the probability of $P_{X^{N_{\text{tgt}}};\theta_{\text{tgt}}}(X^{N_{\text{tgt}}})$ is $O(e^{\frac{-N_{\text{tgt}}J(\theta_{\text{tgt}})\delta}{2}})$, which is exponentially decaying with N_{tgt} so that its order

is $o(\frac{1}{N_{\text{tgt}}})$. The same argument applies to $P_{X^{N_{\text{pre}}};\theta_{\text{pre}}}(X^{N_{\text{pre}}})$. Therefore, the contribution of the second term in equation 28 is of order $o\Big(\frac{1}{N_{\text{tgt}}}\Big)$. Therefore, we have

$$\begin{aligned}
&\mathbb{E}\left[D\left(P_{X;\theta_{\text{tgt}}}\middle\|P_{X;\hat{\theta}_{\text{iw}}}\right)\right] \\
=&\frac{1}{2}\mathbb{E}\left[\sum_{x\in X}\frac{\left(\frac{\partial P_{X;\theta_{\text{tgt}}}(x)}{\partial\theta_{\text{tgt}}}(\hat{\theta}_{\text{iw}}-\theta_{\text{tgt}})\right)^2}{P_{X;\theta_{\text{tgt}}}(x)}\right]+o\left(\frac{1}{N_{\text{tgt}}}\right) \\
=&\frac{1}{2}\mathbb{E}\left[\left(\hat{\theta}_{\text{iw}}-\theta_{\text{tgt}}\right)^2\right]\sum_x\frac{(\frac{\partial P_{X;\theta_{\text{tgt}}}(x)}{\partial\theta_{\text{tgt}}})^2}{P_{X;\theta_{\text{tgt}}}(x)}+o\left(\frac{1}{N_{\text{tgt}}}\right) \\
=&\frac{1}{2}\mathbb{E}\left[\left(\hat{\theta}_{\text{iw}}-\theta_{\text{tgt}}\right)^2\right]J(\theta_{\text{tgt}})+o\left(\frac{1}{N_{\text{tgt}}}\right)
\end{aligned}\tag{30}$$

We make a decomposition for the K-L measure

$$\begin{aligned}
&\mathbb{E}\left[D\left(P_{X;\theta_{\text{tgt}}}\middle\|P_{X;\hat{\theta}_{\text{iw}}}\right)\right] \\
=&\frac{1}{2}\mathbb{E}\left[\left(\hat{\theta}_{\text{iw}}-\theta_{\text{tgt}}\right)^2\right]J(\theta_{\text{tgt}})+o(\frac{1}{N_{\text{tgt}}}) \\
=&\frac{1}{2}J(\theta_{\text{tgt}})\left(\mathbb{E}\left[\left(\hat{\theta}_{\text{iw}}-E_{\hat{\theta}_{\text{iw}}}\right)^2\right]+\mathbb{E}\left[\left(E_{\hat{\theta}_{\text{iw}}}-\theta_{\text{tgt}}\right)^2\right]\right)+o(\frac{1}{N_{\text{tgt}}}).
\end{aligned}\tag{31}$$

Then, we transform the intractable term in the definition of $\hat{\theta}_{\text{iw}}$ in equation 10. Specifically, for the distribution $\hat{P}_{N_{\text{tgt}}}\frac{P_{X;\hat{\theta}_{\text{pre}}}(x)}{P_{X;\hat{\theta}_{\text{tgt}}}(x)}$, we perform the following transformation.

$$\begin{aligned}
&\hat{P}_{N_{\text{tgt}}}(x)\frac{P_{X;\hat{\theta}_{\text{pre}}}(x)}{P_{X;\hat{\theta}_{\text{tgt}}}(x)} \\
&=\frac{\hat{P}_{N_{\text{tgt}}}(x)}{P_{X;\hat{\theta}_{\text{tgt}}}(x)}\left(P_{X;\theta_{\text{pre}}}(x)+\frac{\partial P_{X;\theta_{\text{pre}}}(x)}{\partial\theta}(\hat{\theta}_{\text{pre}}-\theta_{\text{pre}})+O_p(\frac{1}{N_{\text{tgt}}})\right) \\
&=\frac{P_{X;\theta_{\text{tgt}}}(x)+\hat{P}_{N_{\text{tgt}}}(x)-P_{X;\theta_{\text{tgt}}}(x)}{P_{X;\theta_{\text{tgt}}}(x)+\frac{\partial P_{X;\theta_{\text{tgt}}}(x)}{\partial\theta}(\hat{\theta}_{\text{tgt}}-\theta_{\text{tgt}})+O_p(\frac{1}{N_{\text{tgt}}})}\left(P_{X;\theta_{\text{pre}}}(x)+\frac{\partial P_{X;\theta_{\text{pre}}}(x)}{\partial\theta}(\hat{\theta}_{\text{pre}}-\theta_{\text{pre}})\right)+O_p(\frac{1}{N_{\text{tgt}}}) \\
&=\frac{\left(P_{X;\theta_{\text{tgt}}}(x)\right)^2+\left(\hat{P}_{N_{\text{tgt}}}(x)-P_{X;\theta_{\text{tgt}}}(x)-\frac{\partial P_{X;\theta_{\text{tgt}}}(x)}{\partial\theta}(\hat{\theta}_{\text{tgt}}-\theta_{\text{tgt}})\right)P_{X;\theta_{\text{tgt}}}(x)}{\left(P_{X;\theta_{\text{tgt}}}(x)\right)^2}\Bigg(P_{X;\theta_{\text{pre}}}(x) \\
&+\frac{\partial P_{X;\theta_{\text{pre}}}(x)}{\partial\theta}(\hat{\theta}_{\text{pre}}-\theta_{\text{pre}})\Bigg)+O_p(\frac{1}{N_{\text{tgt}}}) \\
&=P_{X;\theta_{\text{pre}}}(x)+\frac{\partial P_{X;\theta_{\text{pre}}}(x)}{\partial\theta}(\hat{\theta}_{\text{pre}}-\theta_{\text{pre}})+ \\
&\left(\hat{P}_{N_{\text{tgt}}}(x)-P_{X;\theta_{\text{tgt}}}(x)-\frac{\partial P_{X;\theta_{\text{tgt}}}(x)}{\partial\theta}(\hat{\theta}_{\text{tgt}}-\theta_{\text{tgt}})\right)\frac{P_{X;\theta_{\text{pre}}}(x)}{P_{X;\theta_{\text{tgt}}}(x)}+O_p(\frac{1}{N_{\text{tgt}}})
\end{aligned}\tag{32}$$

For notational convenience in the proof, we adopt an equivalent definition of $\hat{\theta}_{\text{iw}}$ to that in equation 11, which differs only by an additional normalization factor in the denominator and does not affect the resulting optimizer, i.e.,

$$\hat{\theta}_{\text{iw}}\triangleq\arg\max_{\theta}L_n(\theta),\tag{33}$$

where

$$L_n(\theta) = \frac{N_{\text{tgt}}\mathbb{E}_{\hat{P}_{X^{N_{\text{tgt}}}}}[\log P_{X;\theta}(x)] + wN_{\text{pre}}\mathbb{E}_{\hat{P}_{X^{N_{\text{tgt}}}}\frac{P_{X;\hat{\theta}_{\text{pre}}}}{P_{X;\hat{\theta}_{\text{tgt}}}}}[\log P_{X;\theta}(x)]}{N_{\text{tgt}} + wN_{\text{pre}}}. \tag{34}$$

By combining equation 32, we derive its corresponding expected estimator as follows.

$$E_{\hat{\theta}_{\text{iw}}} \triangleq \arg\max_{\theta} L(\theta), \tag{35}$$

where

$$L(\theta) = \frac{N_{\text{tgt}}\mathbb{E}_{P_{X;\theta_{\text{tgt}}}}[\log P_{X;\theta}(x)] + wN_{\text{pre}}\mathbb{E}_{P_{X;\theta_{\text{pre}}}}[\log P_{X;\theta}(x)]}{N_{\text{tgt}} + wN_{\text{pre}}}. \tag{36}$$

Combining equation 35 with the result in (Zhang et al., 2025b), we have the following lemma to establish the relation of $E_{\hat{\theta}_{\text{iw}}}$ to the θ_{tgt} and θ_{pre}.

Lemma 4. *$E_{\hat{\theta}_{\text{iw}}}$ is a weighted average of θ_{tgt} and θ_{pre}, given by*

$$E_{\hat{\theta}_{\text{iw}}} = \frac{N_{\text{tgt}}\theta_{\text{tgt}} + wN_{\text{pre}}\theta_{\text{pre}}}{N_{\text{tgt}} + wN_{\text{pre}}} + O\left(\frac{1}{N_{\text{tgt}}}\right). \tag{37}$$

Proof. The proof can be adapted directly from (Zhang et al., 2025b). □

Because $\hat{\theta}_{\text{iw}}$ is a maximizer of $L_n(\theta)$, $E_{\hat{\theta}_{\text{iw}}}$ is a maximizer of $L(\theta)$, we can get

$$\begin{aligned} &L_n'(\hat{\theta}_{\text{iw}}) = 0 = L_n'(E_{\hat{\theta}_{\text{iw}}}) \\ &\quad + L_n''(E_{\hat{\theta}_{\text{iw}}})(\hat{\theta}_{\text{iw}} - E_{\hat{\theta}_{\text{iw}}}) + o_p\left(\frac{1}{\sqrt{N_{\text{tgt}}}}\right), \end{aligned} \tag{38}$$

$$L'(E_{\hat{\theta}_{\text{iw}}}) = 0 = \frac{N_{\text{tgt}}}{N_{\text{tgt}} + wN_{\text{pre}}}\mathbb{E}_{P_{X;\theta_{\text{tgt}}}}\left[\frac{\partial \log P_{X;E_{\hat{\theta}_{\text{iw}}}}(x)}{\partial\theta}\right] + \frac{wN_{\text{pre}}}{N_{\text{tgt}} + wN_{\text{pre}}}\mathbb{E}_{P_{X;\theta_{\text{pre}}}}\left[\frac{\partial \log P_{X;E_{\hat{\theta}_{\text{iw}}}}(x)}{\partial\theta}\right]. \tag{39}$$

By transforming equation 38, we have

$$\begin{aligned} &\sqrt{N_{\text{tgt}} + wN_{\text{pre}}}(\hat{\theta}_{\text{iw}} - E_{\hat{\theta}_{\text{iw}}}) \\ &\quad = -\frac{\sqrt{N_{\text{tgt}} + wN_{\text{pre}}}L_n'(E_{\hat{\theta}_{\text{iw}}})}{L_n''(E_{\hat{\theta}_{\text{iw}}})} + o_p(1). \end{aligned} \tag{40}$$

Combining with equation 39, we can get

$$\begin{aligned} &\sqrt{N_{\text{tgt}} + wN_{\text{pre}}}L_n'(E_{\hat{\theta}_{\text{iw}}}) \\ &= \sqrt{N_{\text{tgt}} + wN_{\text{pre}}}\left(L_n'(E_{\hat{\theta}_{\text{iw}}}) - L'(E_{\hat{\theta}_{\text{iw}}})\right) \\ &= \frac{N_{\text{tgt}}}{\sqrt{N_{\text{tgt}} + wN_{\text{pre}}}}\left(\mathbb{E}_{\hat{P}_{X^{N_{\text{tgt}}}}}\left[\frac{\partial \log P_{X;E_{\hat{\theta}_{\text{iw}}}}(x)}{\partial\theta}\right] - \mathbb{E}_{\theta_{\text{tgt}}}[\frac{\partial \log P_{X;E_{\hat{\theta}_{\text{iw}}}}(x)}{\partial\theta}]\right) \\ &+ \frac{wN_{\text{pre}}}{\sqrt{N_{\text{tgt}} + wN_{\text{pre}}}}\left(\mathbb{E}_{\hat{P}_{X^{N_{\text{tgt}}}}\frac{P_{X;\hat{\theta}_{\text{pre}}}}{P_{X;\hat{\theta}_{\text{tgt}}}}}[\frac{\partial \log P_{X;E_{\hat{\theta}_{\text{iw}}}}(x)}{\partial\theta}] - \mathbb{E}_{\theta_{\text{pre}}}[\frac{\partial \log P_{X;E_{\hat{\theta}_{\text{iw}}}}(x)}{\partial\theta}]\right). \end{aligned} \tag{41}$$

Similar to equation 40, we have

$$(\hat{\theta}_{\text{tgt}} - \theta_{\text{tgt}}) = -\frac{\mathbb{E}_{\hat{P}_{X^{N_{\text{tgt}}}}}\left[\frac{\partial \log P_{X;\theta_{\text{tgt}}}(x)}{\partial\theta}\right]}{\mathbb{E}_{\hat{P}_{X^{N_{\text{tgt}}}}}\left[\frac{\partial^2 \log P_{X;\theta_{\text{tgt}}}(x)}{\partial\theta^2}\right]} + o_p\left(\frac{1}{\sqrt{N_{\text{tgt}}}}\right). \tag{42}$$

Moreover, we know $\mathbb{E}_{\hat{P}_{X^{N_{\mathrm{tgt}}}}}\left[\frac{\partial^2 \log P_{X;\theta_{\mathrm{tgt}}}(x)}{\partial \theta^2}\right] \xrightarrow{p} -J(\theta_{\mathrm{tgt}})$, so we have

$$(\hat{\theta}_{\mathrm{tgt}} - \theta_{\mathrm{tgt}}) = \frac{\mathbb{E}_{\hat{P}_{X^{N_{\mathrm{tgt}}}}}\left[\frac{\partial \log P_{X;\theta_{\mathrm{tgt}}}(x)}{\partial \theta}\right]}{J(\theta_{\mathrm{tgt}})} + o_p\left(\frac{1}{\sqrt{N_{\mathrm{tgt}}}}\right). \tag{43}$$

Similarly, we have

$$(\hat{\theta}_{\mathrm{pre}} - \theta_{\mathrm{pre}}) = \frac{\mathbb{E}_{\hat{P}_{X^{N_{\mathrm{pre}}}}}\left[\frac{\partial \log P_{X;\theta_{\mathrm{pre}}}(x)}{\partial \theta}\right]}{J(\theta_{\mathrm{pre}})} + o_p\left(\frac{1}{\sqrt{N_{\mathrm{pre}}}}\right). \tag{44}$$

Therefore,

$$\begin{aligned}
&\sum_{x\in X} \frac{\partial P_{X;\theta_{\mathrm{pre}}}(x)}{\partial \theta}(\hat{\theta}_{\mathrm{pre}} - \theta_{\mathrm{pre}})\frac{\partial \log P_{X;E_{\hat{\theta}_{\mathrm{iw}}}}(x)}{\partial \theta} \\
&= \sum_{x\in X} \frac{\partial P_{X;\theta_{\mathrm{pre}}}(x)}{\partial \theta}(\hat{\theta}_{\mathrm{pre}} - \theta_{\mathrm{pre}})\frac{\partial \log P_{X;\theta_{\mathrm{pre}}}(x)}{\partial \theta} + O_p(\frac{1}{N_{\mathrm{tgt}}}) \\
&= (\hat{\theta}_{\mathrm{pre}} - \theta_{\mathrm{pre}}) \sum_{x\in X} \frac{\partial P_{X;\theta_{\mathrm{pre}}}(x)}{\partial \theta}\frac{\partial \log P_{X;\theta_{\mathrm{pre}}}(x)}{\partial \theta} + O_p(\frac{1}{N_{\mathrm{tgt}}}) \\
&= (\hat{\theta}_{\mathrm{pre}} - \theta_{\mathrm{pre}}) \sum_{x\in X} P_{X;\theta_{\mathrm{pre}}}(x)\frac{\partial \log P_{X;\theta_{\mathrm{pre}}}(x)}{\partial \theta}\frac{\partial \log P_{X;\theta_{\mathrm{pre}}}(x)}{\partial \theta} + O_p(\frac{1}{N_{\mathrm{tgt}}}) \\
&= (\hat{\theta}_{\mathrm{pre}} - \theta_{\mathrm{pre}})J(\theta_{\mathrm{pre}}) + O_p(\frac{1}{N_{\mathrm{tgt}}})
\end{aligned} \tag{45}$$

Moreover,

$$\begin{aligned}
&\sum_{x\in X}\left(\hat{P}_{N_{\mathrm{tgt}}}(x) - P_{X;\theta_{\mathrm{tgt}}}(x) - \frac{\partial P_{X;\theta_{\mathrm{tgt}}}(x)}{\partial \theta}(\hat{\theta}_{\mathrm{tgt}} - \theta_{\mathrm{tgt}})\right)\frac{P_{X;\theta_{\mathrm{pre}}}(x)}{P_{X;\theta_{\mathrm{tgt}}}(x)}\frac{\partial \log P_{X;E_{\hat{\theta}_{\mathrm{iw}}}}(x)}{\partial \theta} \\
&= \sum_{x\in X}\left(\hat{P}_{N_{\mathrm{tgt}}}(x) - P_{X;\theta_{\mathrm{tgt}}}(x) - \frac{\partial P_{X;\theta_{\mathrm{tgt}}}(x)}{\partial \theta}(\hat{\theta}_{\mathrm{tgt}} - \theta_{\mathrm{tgt}})\right)\frac{\partial \log P_{X;\theta_{\mathrm{tgt}}}(x)}{\partial \theta} + O_p(\frac{1}{N_{\mathrm{tgt}}}) \\
&= \sum_{x\in X}\left(\hat{P}_{N_{\mathrm{tgt}}}(x) - P_{X;\theta_{\mathrm{tgt}}}(x)\right)\frac{\partial \log P_{X;\theta_{\mathrm{tgt}}}(x)}{\partial \theta} - (\hat{\theta}_{\mathrm{tgt}} - \theta_{\mathrm{tgt}})\sum_{x\in X}\frac{\partial P_{X;\theta_{\mathrm{tgt}}}(x)}{\partial \theta}\frac{\partial \log P_{X;\theta_{\mathrm{tgt}}}(x)}{\partial \theta} + O_p(\frac{1}{N_{\mathrm{tgt}}}) \\
&= \sum_{x\in X}\left(\hat{P}_{N_{\mathrm{tgt}}}(x) - P_{X;\theta_{\mathrm{tgt}}}(x)\right)\frac{\partial \log P_{X;\theta_{\mathrm{tgt}}}(x)}{\partial \theta} \\
&- \frac{\mathbb{E}_{\hat{P}_{X^{N_{\mathrm{tgt}}}}}\left[\frac{\partial \log P_{X;\theta_{\mathrm{tgt}}}(x)}{\partial \theta}\right]}{J(\theta_{\mathrm{tgt}})}\sum_{x\in X}\frac{\partial P_{X;\theta_{\mathrm{tgt}}}(x)}{\partial \theta}\frac{\partial \log P_{X;\theta_{\mathrm{tgt}}}(x)}{\partial \theta} + o_p\left(\frac{1}{\sqrt{N_{\mathrm{tgt}}}}\right) \\
&= \sum_{x\in X}\left(\hat{P}_{N_{\mathrm{tgt}}}(x) - P_{X;\theta_{\mathrm{tgt}}}(x)\right)\frac{\partial \log P_{X;\theta_{\mathrm{tgt}}}(x)}{\partial \theta} - \frac{\mathbb{E}_{\hat{P}_{X^{N_{\mathrm{tgt}}}}}\left[\frac{\partial \log P_{X;\theta_{\mathrm{tgt}}}(x)}{\partial \theta}\right]}{J(\theta_{\mathrm{tgt}})}J(\theta_{\mathrm{tgt}}) + o_p\left(\frac{1}{\sqrt{N_{\mathrm{tgt}}}}\right) \\
&= -\sum_{x\in X} P_{X;\theta_{\mathrm{tgt}}}(x)\frac{\partial \log P_{X;\theta_{\mathrm{tgt}}}(x)}{\partial \theta} + o_p\left(\frac{1}{\sqrt{N_{\mathrm{tgt}}}}\right) \\
&= 0 + o_p\left(\frac{1}{\sqrt{N_{\mathrm{tgt}}}}\right)
\end{aligned} \tag{46}$$

Therefore, combining equation 32 with equation 45 and equation 46, we can transform the second term of equation 41, i.e.,

$$\frac{wN_{\text{pre}}}{\sqrt{N_{\text{tgt}}+wN_{\text{pre}}}}\left(\mathbb{E}_{\hat{P}_{X^{N_{\text{tgt}}}}\frac{P_{X;\hat{\theta}_{\text{pre}}}}{P_{X;\hat{\theta}_{\text{tgt}}}}}[\frac{\partial \log P_{X;E_{\hat{\theta}_{\text{iw}}}}(x)}{\partial\theta}]-\mathbb{E}_{\theta_{\text{pre}}}[\frac{\partial \log P_{X;E_{\hat{\theta}_{\text{iw}}}}(x)}{\partial\theta}]\right)$$
$$=\frac{wN_{\text{pre}}}{\sqrt{N_{\text{tgt}}+wN_{\text{pre}}}}(\hat{\theta}_{\text{pre}}-\theta_{\text{pre}})J(\theta_{\text{pre}})+o_p(1)\xrightarrow{d}\mathcal{N}\left(0,\frac{w^2N_{\text{pre}}}{N_{\text{tgt}}+wN_{\text{pre}}}J(\theta_{\text{pre}})\right). \tag{47}$$

For the first term of equation 41, applying the Central Limit Theorem, we can get

$$\frac{N_{\text{tgt}}}{\sqrt{N_{\text{tgt}}+wN_{\text{pre}}}}\left(\mathbb{E}_{\hat{P}_{X^{N_{\text{tgt}}}}}\left[\frac{\partial \log P_{X;E_{\hat{\theta}_{\text{iw}}}}(x)}{\partial\theta}\right]-\mathbb{E}_{\theta_{\text{tgt}}}[\frac{\partial \log P_{X;E_{\hat{\theta}_{\text{iw}}}}(x)}{\partial\theta}]\right)\xrightarrow{d}$$
$$\mathcal{N}\left(0,\frac{N_{\text{tgt}}}{N_{\text{tgt}}+wN_{\text{pre}}}\left(\mathbb{E}_{\theta_{\text{tgt}}}\left[\left(\frac{\partial \log P_{X;E_{\hat{\theta}_{\text{iw}}}}(x)}{\partial\theta}\right)^2\right]-\mathbb{E}_{\theta_{\text{tgt}}}\left[\left(\frac{\partial \log P_{X;E_{\hat{\theta}_{\text{iw}}}}(x)}{\partial\theta}\right)\right]^2\right)\right. \tag{48}$$

By taking Taylor expansion of $E_{\hat{\theta}_{\text{iw}}}$ at θ_{tgt}, we can get

$$\begin{aligned}
&\mathbb{E}_{\theta_{\text{tgt}}}\left[\left(\frac{\partial \log P_{X;E_{\hat{\theta}_{\text{iw}}}}(x)}{\partial\theta}\right)^2\right]\\
&=\mathbb{E}_{\theta_{\text{tgt}}}\left[\left(\frac{\partial \log P_{X;\theta_{\text{tgt}}}(x)}{\partial\theta}+\frac{\partial}{\partial\theta}\frac{\partial \log P_{X;\theta_{\text{tgt}}}(x)}{\partial\theta}(E_{\hat{\theta}_{\text{iw}}}-\theta_{\text{tgt}})+O(\frac{1}{N_{\text{tgt}}})\right)^2\right]\\
&=J(\theta_{\text{tgt}})+2(E_{\hat{\theta}_{\text{iw}}}-\theta_{\text{tgt}})\mathbb{E}_{\theta_{\text{tgt}}}\left[\frac{\partial \log P_{X;E_{\hat{\theta}_{\text{iw}}}}(x)}{\partial\theta}\frac{\partial^2 \log P_{X;E_{\hat{\theta}_{\text{iw}}}}(x)}{\partial\theta^2}\right]+O(\frac{1}{N_{\text{tgt}}})\\
&=J(\theta_{\text{tgt}})+O(\frac{1}{\sqrt{N_{\text{tgt}}}})
\end{aligned} \tag{49}$$

and

$$\begin{aligned}
&\mathbb{E}_{\theta_{\text{tgt}}}\left[\left(\frac{\partial \log P_{X;E_{\hat{\theta}_{\text{iw}}}}(x)}{\partial\theta}\right)\right]^2\\
&=\mathbb{E}_{\theta_{\text{tgt}}}\left[\left(\frac{\partial \log P_{X;\theta_{\text{tgt}}}(x)}{\partial\theta}+\frac{\partial}{\partial\theta}\frac{\partial \log P_{X;\theta_{\text{tgt}}}(x)}{\partial\theta}(E_{\hat{\theta}_{\text{iw}}}-\theta_{\text{tgt}})+O(\frac{1}{N_{\text{tgt}}})\right)\right]^2\\
&=\mathbb{E}_{\theta_{\text{tgt}}}\left[\left(\frac{\partial}{\partial\theta}\frac{\partial \log P_{X;\theta_{\text{tgt}}}(x)}{\partial\theta}(E_{\hat{\theta}_{\text{iw}}}-\theta_{\text{tgt}})+O(\frac{1}{N_{\text{tgt}}})\right)\right]^2\\
&=(E_{\hat{\theta}_{\text{iw}}}-\theta_{\text{tgt}})^2E_{\theta_{\text{tgt}}}\left[\left(\frac{\partial}{\partial\theta}\frac{\partial \log P_{X;\theta_{\text{tgt}}}(x)}{\partial\theta}\right)\right]^2+o(\frac{1}{N_{\text{tgt}}})\\
&=O(\frac{1}{N_{\text{tgt}}})
\end{aligned} \tag{50}$$

Combining equation 49 and equation 50 to equation 48, we have

$$\begin{aligned}
&\frac{N_{\text{tgt}}}{\sqrt{N_{\text{tgt}}+wN_{\text{pre}}}}\left(\mathbb{E}_{\hat{P}_{X^{N_{\text{tgt}}}}}\left[\frac{\partial \log P_{X;E_{\hat{\theta}_{\text{iw}}}}(x)}{\partial\theta}\right]-\mathbb{E}_{\theta_{\text{tgt}}}\left[\frac{\partial \log P_{X;E_{\hat{\theta}_{\text{iw}}}}(x)}{\partial\theta}\right]\right)\\
&\xrightarrow{d}\mathcal{N}\left(0,\frac{N_{\text{tgt}}}{N_{\text{tgt}}+wN_{\text{pre}}}J(\theta_{\text{tgt}})\right)
\end{aligned} \tag{51}$$

We observe from the second line of equation 47 that it depends only on $\hat{\theta}_{\text{pre}}$. Consequently, it is statistically independent of equation 51. By combining equation 41, equation 47 and equation 51, we can get

$$\sqrt{N_{\text{tgt}} + wN_{\text{pre}}} L_n'(E_{\hat{\theta}_{\text{iw}}}) \xrightarrow{d} \mathcal{N}\left(0, \frac{N_{\text{tgt}}}{N_{\text{tgt}} + wN_{\text{pre}}} J(\theta_{\text{tgt}}) + \frac{w^2 N_{\text{pre}}}{N_{\text{tgt}} + wN_{\text{pre}}} J(\theta_{\text{pre}})\right). \tag{52}$$

Given that θ_{tgt}, θ_{pre}, and $E_{\hat{\theta}_{\text{iw}}}$ are already assumed to be sufficiently close, we can use this, along with a Taylor expansion, to estimate the distance between their Fisher information matrices. We conclude that the discrepancy among $J(\theta_{\text{tgt}})$, $J(\theta_{\text{pre}})$, and $J(E_{\hat{\theta}_{\text{iw}}})$ is of the order $O\left(\frac{1}{\sqrt{N_{\text{tgt}}}}\right)$.

$$\begin{aligned}
J(\theta_{\text{pre}}) &= \mathbb{E}_{\theta_{\text{pre}}}\left[\left(\frac{\partial}{\partial\theta}\log P_{X;\theta_{\text{pre}}}\right)^2\right] \\
&= \mathbb{E}_{\theta_{\text{pre}}}\left[\left(\frac{\partial}{\partial\theta}\log P_{X;\theta_{\text{tgt}}} + \frac{\partial^2 \log P_{X;\theta_{\text{tgt}}}}{\partial\theta^2}(\theta_{\text{pre}} - \theta_{\text{tgt}}) + O(\frac{1}{N_{\text{tgt}}})\right)^2\right] \\
&= \mathbb{E}_{\theta_{\text{pre}}}\left[\left(\frac{\partial}{\partial\theta}\log P_{X;\theta_{\text{tgt}}}\right)^2\right] + O(\frac{1}{\sqrt{N_{\text{tgt}}}}) \\
&= \sum_{x\in\mathcal{X}} P_{X;\theta_{\text{pre}}}(x)\left(\frac{\partial}{\partial\theta}\log P_{X;\theta_{\text{tgt}}}(x)\right)^2 + O(\frac{1}{\sqrt{N_{\text{tgt}}}}) \\
&= \sum_{x\in\mathcal{X}} \left(P_{X;\theta_{\text{tgt}}}(x) + \frac{\partial P_{X;\theta_{\text{tgt}}}}{\partial\theta}(\theta_{\text{pre}} - \theta_{\text{tgt}}) + O(\frac{1}{N_{\text{tgt}}})\right)\cdot\left(\frac{\partial}{\partial\theta}\log P_{X;\theta_{\text{tgt}}}(x)\right)^2 + O(\frac{1}{\sqrt{N_{\text{tgt}}}}) \\
&= \sum_{x\in\mathcal{X}} P_{X;\theta_{\text{tgt}}}(x)\left(\frac{\partial}{\partial\theta}\log P_{X;\theta_{\text{tgt}}}(x)\right)^2 + O(\frac{1}{\sqrt{N_{\text{tgt}}}}) \\
&= J(\theta_{\text{tgt}}) + O(\frac{1}{\sqrt{N_{\text{tgt}}}})
\end{aligned} \tag{53}$$

Therefore, following the argument in equation 53, when the pretrained parameter and the target parameter are sufficiently close, their corresponding Fisher information matrices are also close. Combined with Lemma 4, this implies that $J\left(E_{\hat{\theta}_{\text{iw}}}\right)$ is also close to both $J(\theta_{\text{pre}})$ and $J(\theta_{\text{tgt}})$, i.e.,

$$J(\theta_{\text{pre}}) = J\left(E_{\hat{\theta}_{\text{iw}}}\right) + O(\frac{1}{\sqrt{N_{\text{tgt}}}}) \tag{54}$$

Moreover, for the denominator of equation 40, we have the following standard result (Newey & McFadden, 1994):

$$L_n''\left(E_{\hat{\theta}_{\text{iw}}}\right) \xrightarrow{p} -J\left(E_{\hat{\theta}_{\text{iw}}}\right). \tag{55}$$

Combining equation 54 and equation 55, the denominator of equation 40 can be replaced by $-J(\theta_{\text{pre}})$ without affecting the asymptotic result. Combining this with equation 40 and equation 52, we have

$$\frac{N_{\text{tgt}} + wN_{\text{pre}}}{\sqrt{N_{\text{tgt}} + w^2 N_{\text{pre}}}}(\hat{\theta}_{\text{iw}} - E_{\hat{\theta}_{\text{iw}}}) \xrightarrow{d} N\left(0, \frac{1}{J(\theta_{\text{pre}})}\right). \tag{56}$$

Substituting equation 37 and equation 56 into equation 31, we have

$$\begin{aligned}
&\mathbb{E}\left[D\left(P_{X;\theta_{\text{tgt}}}\middle\|P_{X;\hat{\theta}_{\text{iw}}}\right)\right] \\
&= \frac{1}{2}\frac{N_{\text{tgt}} + w^2 N_{\text{pre}}}{(N_{\text{tgt}} + wN_{\text{pre}})^2} + \frac{1}{2}\frac{(wN_{\text{pre}})^2 J(\theta_{\text{pre}})(\theta_{\text{tgt}} - \theta_{\text{pre}})^2}{(N_{\text{tgt}} + wN_{\text{pre}})^2} + o(\frac{1}{N_{\text{tgt}}}).
\end{aligned} \tag{57}$$

By differentiating with respect to w, we obtain the optimal value of weight is

$$w^* = \frac{1}{1 + tN_{\text{pre}}}. \tag{58}$$

Combining with equation 12, we have

$$\alpha^* = \frac{N_{\text{pre}}}{N_{\text{tgt}} + tN_{\text{pre}}N_{\text{tgt}}}. \tag{59}$$

□

We next extend Lemma 3 to the high-dimensional setting. Similar to equation 31, we can get

$$\begin{aligned}
&\mathbb{E}\left[D\left(P_{X;\underline{\theta}_{\text{tgt}}} \middle\| P_{X;\hat{\underline{\theta}}_{\text{iw}}}\right)\right] \\
&= \frac{1}{2} tr\left(J(\underline{\theta}_{\text{tgt}})\mathbb{E}\left[\left(\hat{\underline{\theta}}_{\text{iw}} - E_{\hat{\underline{\theta}}_{\text{iw}}}\right)\left(\hat{\underline{\theta}}_{\text{iw}} - E_{\hat{\underline{\theta}}_{\text{iw}}}\right)^\top\right]\right) \\
&\quad + \frac{1}{2} tr\left(J(\underline{\theta}_{\text{tgt}})\mathbb{E}\left[\left(E_{\hat{\underline{\theta}}_{\text{iw}}} - \underline{\theta}_{\text{tgt}}\right)\left(E_{\hat{\underline{\theta}}_{\text{iw}}} - \underline{\theta}_{\text{tgt}}\right)^\top\right]\right) + o\left(\frac{1}{N_{\text{tgt}}}\right).
\end{aligned} \tag{60}$$

Similar to equation 56,we can get

$$\frac{N_{\text{tgt}} + wN_{\text{pre}}}{\sqrt{N_{\text{tgt}} + w^2 N_{\text{pre}}}}\left(\hat{\underline{\theta}}_{\text{iw}} - E_{\hat{\underline{\theta}}_{\text{iw}}}\right) \xrightarrow{d} \mathcal{N}\left(0, J(\underline{\theta}_{\text{pre}})^{-1}\right). \tag{61}$$

So we can know that $\frac{(N_{\text{tgt}} + wN_{\text{pre}})^2}{N_{\text{tgt}} + w^2 N_{\text{pre}}}\mathbb{E}\left[\left(\hat{\underline{\theta}}_{\text{iw}} - E_{\hat{\underline{\theta}}_{\text{iw}}}\right)\left(\hat{\underline{\theta}}_{\text{iw}} - E_{\hat{\underline{\theta}}_{\text{iw}}}\right)^\top\right]$ has the limit

$$J(\underline{\theta}_{\text{pre}})^{-1} \tag{62}$$

The same to Lemma 4, we can get

$$E_{\hat{\underline{\theta}}_{\text{iw}}} = \frac{N_{\text{tgt}}\underline{\theta}_{\text{tgt}} + wN_{\text{pre}}\underline{\theta}_{\text{pre}}}{N_{\text{tgt}} + wN_{\text{pre}}} + O\left(\frac{1}{N_{\text{tgt}}}\right), \tag{63}$$

Combining equation 60, equation 61 and equation 63, we can know that the K-L measure has the form

$$\frac{d}{2}\left(\frac{N_{\text{tgt}} + w^2 N_{\text{pre}}}{\left(N_{\text{tgt}} + wN_{\text{pre}}\right)^2} + \frac{w^2 N_{\text{pre}}^2}{\left(N_{\text{tgt}} + wN_{\text{pre}}\right)^2} t\right) + o\left(\frac{1}{N_{\text{tgt}}}\right), \tag{64}$$

where

$$t \triangleq \frac{(\underline{\theta}_{\text{pre}} - \underline{\theta}_{\text{tgt}})^\top J(\underline{\theta}_{\text{pre}})(\underline{\theta}_{\text{pre}} - \underline{\theta}_{\text{tgt}})}{d}. \tag{65}$$

By differentiating with respect to w, we obtain the optimal value of weight is

$$w^* = \frac{1}{1 + tN_{\text{pre}}} \tag{66}$$

Combining with equation 12, we have

$$\alpha^* = \frac{N_{\text{pre}}}{N_{\text{tgt}} + tN_{\text{pre}}N_{\text{tgt}}}. \tag{67}$$

□

C Extended Related Work

C.1 Importance Sampling and Data Selection

Importance sampling and importance weighting provide a classical principle for estimating target-distribution objectives using samples drawn from a different proposal distribution, by reweighting each sample via a density-ratio style factor. Beyond Monte Carlo estimation, IS-style reweighting has also been widely adopted in modern large-model training. In LLM fine-tuning, recent work applies importance sampling at different granularities for efficiency and robustness, e.g., layerwise importance sampling for memory-efficient updates (LISA) (Pan et al., 2024), and module-wise importance sampling with variance and convergence analyses (MISA) (Liu et al., b).

In a different training paradigm, preference alignment methods, such as RLHF and DPO, have explored fine-grained token-level weighting, including token-level importance sampling for DPO (TIS-DPO) (Liu et al., a) and related variants (Zhou et al., 2025; Yang et al., 2025a; Li et al., 2025). These studies focus on preference-based optimization objectives and are orthogonal to our supervised fine-tuning analysis.

Closely related to reweighting, data selection methods aim to improve fine-tuning efficiency by selecting a subset of informative training samples. Recent surveys and empirical studies have explored data selection for instruction tuning and large language model fine-tuning (Zhang et al., 2025a). Data selection has also been investigated for multimodal large language models, where selecting compact yet effective visual instruction subsets can improve robustness and data efficiency (Yang et al., 2025b). While selection can be viewed as a hard form of reweighting, our work instead studies soft importance weighting and its statistical effect on estimation error in supervised fine-tuning.

C.2 Supervised fine-tuning (SFT)

Supervised fine-tuning (SFT) has emerged as a standard paradigm for adapting pretrained large language models (LLMs) to downstream tasks, enabling specific capabilities such as instruction following and domain-specific knowledge acquisition. Representative SFT approaches include prompt tuning (Brown et al., 2020), which conditions model behavior through task-specific input prompts without updating model parameters; prefix tuning (Liu et al., 2022), which introduces a small set of learnable continuous prompts while keeping the backbone model fixed; and instruction tuning (Ouyang et al., 2022), which directly fine-tunes models on supervised instruction–response pairs to explicitly shape model behaviors. In this work, we focus on the instruction tuning setting and conduct all experiments using supervised fine-tuning with instruction-formatted data.

SFT has also become a standard practice in the development of multimodal large language models, where pretrained language models are augmented with visual encoders and adapted using multimodal instruction data. Existing methods employ supervised instruction tuning on paired visual–textual data to enable a wide range of vision–language tasks, including visual question answering, image captioning, and multimodal reasoning (Liu et al., 2023a; Zhu et al., 2023). Recent empirical studies further suggest that data-centric factors, including instruction diversity, sample selection, and implicit weighting effects, play a critical role in robust supervised fine-tuning of MLLMs (Yang et al., 2025b).

C.3 Robust Fine-Tuning and Prior-Guided Fine-Tuning

Robust fine-tuning aims to improve downstream performance while retaining the zero-shot and out-of-distribution capabilities inherited from pretraining (Wortsman et al., 2022; Tian et al., 2024; Suzuki et al., 2026). Within this broad goal, we use *prior-guided fine-tuning* to refer to methods that exploit pretrained-model information, such as parameters, Fisher information, features, or predictions, to guide downstream adaptation. Early work in continual learning investigates parameter-space regularization to preserve pretrained knowledge during adaptation. Some approaches impose uniform or heuristic constraints to retain pretrained representations (Aljundi et al., 2018). Moreover, Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) provides a more principled formulation by estimating heterogeneous parameter importance via Fisher information and selectively penalizing deviations of critical parameters.

Recent work on large language models and multimodal large language models shows that naive fine-tuning with limited target-domain data may erode pretrained representations and harm generalization; to address this issue, existing approaches include preserving pretrained features during fine-tuning (He et al., 2023), Causal Effect Tuning (CET) which frames fine-tuning via a causal graph and adds a causal objective to preserve pretrained causal effects alongside task loss (Zheng et al., 2023), and Label Annealing which regularizes the fine-tuned model's predictive distribution toward that of the pretrained model via a KL loss between the model's current predictions and the pretrained model's predictions (Yang et al., 2024).

PGIW follows this prior-guided perspective, but differs by using pretrained-model information to construct both a global balancing coefficient and sample-wise model-based importance weights, rather than only constraining parameter, feature, or prediction drift.

D Dataset Description

ImageNet-100 is a subset of ImageNet consisting of 100 object categories, following standard class-balanced sampling. It serves as a large-scale visual classification benchmark for evaluating in-domain adaptation under limited supervision.

DomainNet is a multi-domain image classification dataset containing six visual domains (*Real*, *Clipart*, *Painting*, *Sketch*, *Infograph*, and *Quickdraw*). We follow the standard cross-domain evaluation protocol, training on one domain and testing on the remaining domains, which allows us to assess robustness and cross-domain generalization.

ScienceQA is a multimodal question answering dataset that combines images, textual questions, and multiple-choice answers. We cast ScienceQA into a classification-style formulation by predicting the answer index, enabling evaluation of our method on vision–language reasoning tasks beyond pure visual recognition.

M3CoT is a multimodal chain-of-thought reasoning dataset that combines images, textual questions, multiple-choice answers, and multi-step rationales. We cast M3CoT into a classification-style formulation by predicting the answer index, enabling evaluation of our method on multi-step vision–language reasoning tasks beyond simple single-step inference.

E Additional Baseline Details

All compared fine-tuning methods adapt the pretrained backbone through LoRA. The baseline labeled LoRA minimizes cross-entropy, $\mathcal{L}_{\mathrm{CE}}$. L2-SP adds $\lambda\|\theta-\theta_0\|_2^2$, where θ_0 is the initialization; EWC adds $\lambda\sum_i F_i(\theta_i-\theta_{0,i})^2$, with F_i denoting the diagonal Fisher importance. KD adds $\lambda D_{\mathrm{KL}}(p_{\mathrm{pre}}\|p_\theta)$ to match the frozen pretrained model's predictive distribution. Fixed-IW uses $(1+\lambda r_i)\ell_i$, whereas PGIW uses $(1+\alpha^* r_i)\ell_i$ with the coefficient described in Section 3.3. Table 6 lists the coefficients associated with the reported baseline results.

Table 6: Baseline coefficients for ScienceQA and M3CoT.

	ScienceQA		M3CoT	
Method	LLaVA	Qwen	LLaVA	Qwen
L2-SP	6×10^{-5}	2×10^{-4}	3×10^{-4}	0.002
EWC	100	150	100	300
KD	0.1	0.05	0.4	1.5
Fixed-IW	0.2	0.1	0.5	0.1

F Broader Impacts

This work proposes a prior-guided importance-weighting method for supervised fine-tuning of multimodal large language models. The method may improve data efficiency and cross-domain generalization in settings where labeled multimodal data are limited. Its broader impact is mainly

methodological, as it does not target a specific sensitive application or introduce new data collection procedures. As with other fine-tuning approaches, users should ensure that the target data are appropriately curated and evaluate the resulting models for robustness, fairness, and safety before deployment.

G LIMITATIONS

This work has several limitations.

- **Asymptotic analysis.** Our theoretical analysis is developed under an asymptotic regime and assumes a small discrepancy between the pretrained and target-domain parameters. While this provides a principled characterization of the proposed weighting strategy, it may not fully capture practical fine-tuning scenarios with large distribution shifts.

- **Diagonal Fisher approximation.** For computational efficiency, the proposed method uses a diagonal empirical Fisher approximation estimated from a small subset of target-domain samples. This approximation is lightweight, but it may ignore parameter correlations and therefore provide less accurate uncertainty estimation than more expressive Fisher approximations.

- **Evaluation scope.** Although experiments on multiple multimodal benchmarks demonstrate consistent improvements, the evaluation is still limited to several backbone models and classification-style tasks. Extending the analysis and experiments to broader MLLM architectures, generation tasks, and larger distribution shifts is an important direction for future work.

H TIME COMPLEXITY

Table 7: Time Cost Difference Between PGIW and Standard LoRA Training on ScienceQA

Model	Algorithm	Average Single Step Time	PGIW Preparation Time
LLaVA	LoRA	19.02s	-
LLaVA	PGIW	24.23s	00:32:59
Qwen	LoRA	3.03s	-
Qwen	PGIW	4.23s	00:25:41

The total training duration includes additional overheads such as model checkpoint saving and process switching, while the average single-step time excludes these costs, so the product of the number of timing steps and the average single-step time is less than the total training duration.

Time complexity comparisons on the ScienceQA dataset show that the average single-step time of the PGIW algorithm increases by 5.21 seconds (+27.4%) compared with LoRA for the LLaVA model, and by 1.20 seconds (+39.6%) for the Qwen model. This is because PGIW requires an extra forward pass using the pretrained model; although it brings additional computational overhead, it yields significant performance improvements and is acceptable in practical training.

I TRAINING LOSS ON SCIENCEQA (QWEN)

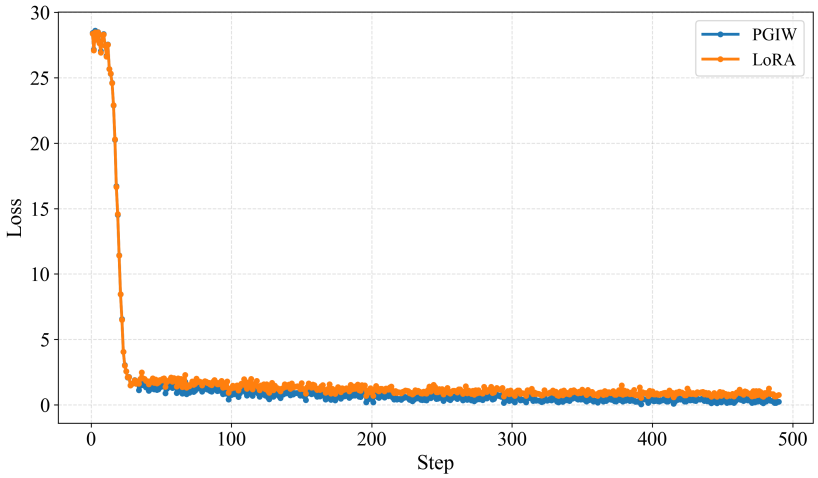


Figure 2: Loss Comparison Between PGIW and LoRA (Qwen)

Figure 2 compares the training loss of PGIW and LoRA on ScienceQA using Qwen. Both methods show a sharp initial decrease in loss, followed by a more gradual decline. After the initial decrease, PGIW maintains a lower loss than LoRA over most of the remaining training steps, indicating improved fitting under the reported loss metric.