DIMENSION-WISE DISTRIBUTIONALLY ROBUST OPTIMIZATION FOR MULTIMODAL LEARNING VIA ADAPTIVE AGGREGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal data inherently carries multiple semantic dimensions, such as imaging modality, anatomical region, and task type in medical data. Each dimension induces subpopulation shifts that may vary independently. Group Distributionally Robust Optimization (Group DRO) provides a principled framework for addressing such shifts by minimizing worst-case losses across predefined groups. However, when multimodal data carries multiple grouping dimensions, the standard practice of combining them via their Cartesian product leads to an explosive growth in the number of groups (for instance, combining 8 modalities, 5 anatomical regions, and 5 task types yields 200 groups), leaving each group with extremely sparse samples and making robust optimization unstable. To address this, we propose Dimension-wise DRO, which formulates robustness along each semantic dimension independently and adaptively aggregates dimension-wise risks into a unified objective. Our method consists of three components: (i) selecting non-redundant dimensions by pruning highly correlated axes; (ii) defining a smooth robust risk for each selected dimension via KL regularization; and (iii) aggregating dimension-wise risks with a Hedge-style adaptive weighting scheme that prioritizes harder dimensions. We theoretically prove that the proposed aggregated objective bounds the worst-case group risk across all dimensions, with a convergence rate of $O(\log T/\sqrt{T})$, matching standard SGD up to a logarithmic factor. Experiments on three multimodal benchmarks—OmniMedVQA, VQAv2, and BDD100K—demonstrate that our method consistently improves worst-group performance under both shifted and unshifted settings without sacrificing average accuracy, outperforming both annotation-dependent and annotation-free baselines.

1 INTRODUCTION

Multimodal learning has become increasingly common in real-world applications such as medical diagnosis (Cai et al., 2026), autonomous driving (Wei et al., 2025), and robotics (Han et al., 2025), where models need to integrate and reason over heterogeneous data sources. Compared with unimodal settings, multimodal learning faces a more complex form of distribution shift (Chen et al., 2025; Verma et al., 2024). In multimodal data, each modality carries its own form of shift: an imaging model may degrade on MRI while remaining accurate on CT, or generalize across brain images but fail on lung. These modality-specific shifts coexist within the same dataset, making the overall distribution shift inherently multi-dimensional. Figure 1 illustrates this phenomenon on a popular medical VQA benchmark OmniMedVQA (Hu et al., 2024): model performance varies substantially across imaging modalities (e.g., 71.5% on X-ray vs. 53.1% on CT), across anatomical regions (e.g., 72.7% on chest vs. 20.1% on heart), and across clinical tasks (e.g., 94.6% on modality recognition vs. 38.8% on lesion grading).

Group Distributionally Robust Optimization (Group DRO) (Sagawa et al., 2020) provides a principled framework for addressing such group-level distribution shifts by minimizing the worst-case loss across predefined groups. This framework has inspired a wide range of subsequent methods, which can be broadly divided into two categories according to whether they require group annotations during training. "Annotation-dependent" methods assume access to group labels and optimize worst-group performance explicitly. The original Group DRO (Sagawa et al., 2020) falls under the umbrella of

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

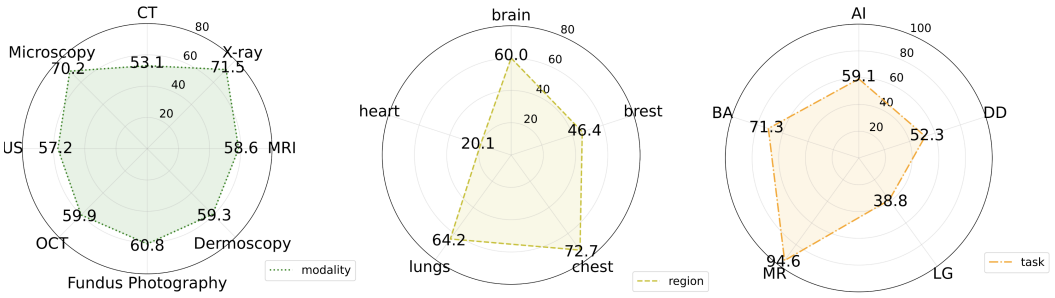

Figure 1: Multi-dimensional performance variation of Qwen2.5-VL (Bai et al., 2025) on OmniMed-VQA (Hu et al., 2024). Each radar plot corresponds to one grouping dimension: modality (left), anatomical region (middle), and clinical task (right). Each axis within a plot represents one group along that dimension, and the radial distance indicates accuracy. The plots are far from circular: performance varies substantially across groups within each dimension, and underperforming groups appear across all three dimensions rather than being concentrated in a single one dimension. This illustrates that distribution shift in multimodal medical data is inherently multi-dimensional.

this category, which minimizes the maximum loss across predefined groups. DFR (Kirichenko et al., 2023) observes that ERM already learns good feature representations, and only retrains the last layer on a small group-balanced dataset. H-DRO (Jo et al., 2026) further models intra-group uncertainty via hierarchical ambiguity sets. "Annotation-free" methods implicitly infer group structures from training signals, without access to explicit group labels. JTT (Liu et al., 2021) identifies minority samples through training dynamics and upweights them in a second training stage. SSA (Nam et al., 2022) trains a spurious attribute predictor using a small number of group-labeled samples and applies pseudo-labels for worst-group loss minimization. GIC (Han & Zou, 2024a) infers precise group labels and integrates them with invariant learning algorithms.

When applied to multimodal data with multiple grouping dimensions, both those two categories of methods face limitations. For annotation-dependent methods, the standard practice of forming groups as the Cartesian product of all dimensions leads to a challenge we refer to as **group explosion**. When a dataset carries K independent grouping dimensions, the Cartesian product yields $\prod_{k=1}^{K} G_k$ groups. As the number of groups grows, the number of samples per group shrinks correspondingly, compounding the data scarcity that multimodal scenarios commonly face. The medical VQA benchmark in Figure 1 provides a typical example: it contains 8 imaging modalities, 5 anatomical regions, and 5 clinical tasks, yielding 8×5×5=200 combinatorial groups. For annotation-free methods, these methods infer group information from a training signal that is not explicitly tied to a specific dimension. As a result, they may not explicitly target robustness along a particular dimension when robustness requirements differ across dimensions.

To address these limitations, we propose **Dimension-wise DRO**, a framework that addresses multi-dimensional robustness by enforcing it along each semantic dimension separately and adaptively aggregating the dimension-wise risks, rather than forming combinatorial groups. The framework consists of three components. **First**, to avoid redundant constraints from correlated dimensions, we measure pairwise dependence with Cramér's V and select a maximally non-redundant subset via maximum independent set. **Second**, we compute a robust risk along each selected dimension by reweighting its groups based on their empirical losses. To reduce the influence of noisy loss estimates from small groups, we add a KL regularization term that keeps the weights close to a reference distribution and admits a closed-form solution. **Third**, we aggregate the dimension-wise risks into a unified objective via a weighted sum. Since different dimensions exhibit different levels of robust loss, uniform weighting may leave harder dimensions under-optimized; we therefore adopt a Hedge-style adaptive weighting scheme that adjusts the relative importance of each dimension based on its current robust loss. Our contributions are summarized as follows:

- We propose Dimension-wise DRO, a robust training framework that enforces robustness along each semantic dimension independently and adaptively aggregates the dimension-wise risks. And we provide theoretical guarantees that minimizing the proposed aggregated

objective bounds the worst-case group risk across all dimensions, with a convergence rate of $O(\log T/\sqrt{T})$, matching standard SGD up to a logarithmic factor.

- We validate our approach on three multimodal benchmarks spanning medical VQA, natural-image VQA, and autonomous driving, showing consistent improvements in worst-group and out-of-distribution performance over both annotation-dependent and annotation-free baselines.

2 PRELIMINARIES

In this section, we define the multi-dimensional group structure and briefly review Group Distributionally Robust Optimization (Sagawa et al., 2020) and Cramér's V (Bergsma, 2013), which serve as the basis of our method.

2.1 GROUP DISTRIBUTIONALLY ROBUST OPTIMIZATION

Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ denote the sample space and $\ell(\theta; z)$ the loss function of model parameters θ on sample z. Group DRO (Sagawa et al., 2020) considers scenarios where samples belong to predefined groups. Let G denote the set of groups and P_g the conditional distribution of group g. The group loss is defined as $L_g(\theta) = \mathbb{E}_{z\sim P_g}[\ell(\theta; z)]$. Group DRO minimizes the worst-case group loss: $\min_\theta \max_{g\in G} L_g(\theta)$. This objective encourages the model to perform well on the worst-performing group. Instead of the hard maximum over groups, a common relaxation introduces a distribution over groups $\pi \in \Delta^{|G|}$ and optimizes the soft worst-case loss:

$$L_{DRO}(\theta) = \max_{\pi\in\Delta^{|G|}} \sum_{g=1}^{|G|} \pi_g L_g(\theta). \tag{1}$$

Here $\Delta^{|G|}$ denotes the probability simplex over $|G|$ groups (i.e., $\pi_g \geq 0$ and $\sum_{g=1}^{|G|} \pi_g = 1$). This formulation allows groups to be reweighted toward the worst-performing ones and provides a smooth approximation to the worst-case group loss.

Unlike conventional group robustness settings, multimodal datasets are annotated along multiple semantic attributes. In medical imaging, for example, these attributes include imaging modality, anatomical region, and task type. We call each attribute a *grouping dimension*. Each sample z is associated with K grouping dimensions. Dimension k contains G_k groups:

$$\mathcal{M}^{(k)} = \{g_1^{(k)}, g_2^{(k)}, \ldots, g_{G_k}^{(k)}\}, \tag{2}$$

where $g_i^{(k)}$ denotes the i-th group. A straightforward way to obtain a single grouping structure is to take the Cartesian product of all dimensions:

$$G = \mathcal{M}^{(1)} \times \cdots \times \mathcal{M}^{(K)}. \tag{3}$$

This produces $\prod_{k=1}^{K} G_k$ groups, which grows exponentially with K. As a result, most groups contain very few samples, leading to severe data sparsity and unstable optimization. We address this issue in Section 3 by operating on individual dimensions instead of constructing the full Cartesian product.

2.2 CATEGORICAL ASSOCIATION MEASURE: CRAMÉR'S V

Let A and B be two categorical grouping dimensions with finite category sets $\mathcal{A} = \{1, \ldots, r\}$ and $\mathcal{B} = \{1, \ldots, c\}$. Given n samples $\{(a_\ell, b_\ell)\}_{\ell=1}^n$, we form the contingency table $\mathbf{N} \in \mathbb{N}^{r\times c}$ where $\mathbf{N}_{ij}$ counts the number of samples with $(A = i, B = j)$. Let $\mathbf{N}_{i\cdot} = \sum_{j=1}^{c} \mathbf{N}_{ij}$, $\mathbf{N}_{\cdot j} = \sum_{i=1}^{r} \mathbf{N}_{ij}$, and $n = \sum_{i=1}^{r}\sum_{j=1}^{c} \mathbf{N}_{ij}$. Under the independence model, the expected count at cell (i, j) is $\mathbf{E}_{ij} = \frac{\mathbf{N}_{i\cdot}\mathbf{N}_{\cdot j}}{n}$. The Pearson chi-square statistic is

$$\chi^2(A, B) = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(\mathbf{N}_{ij} - \mathbf{E}_{ij})^2}{\mathbf{E}_{ij}}. \tag{4}$$

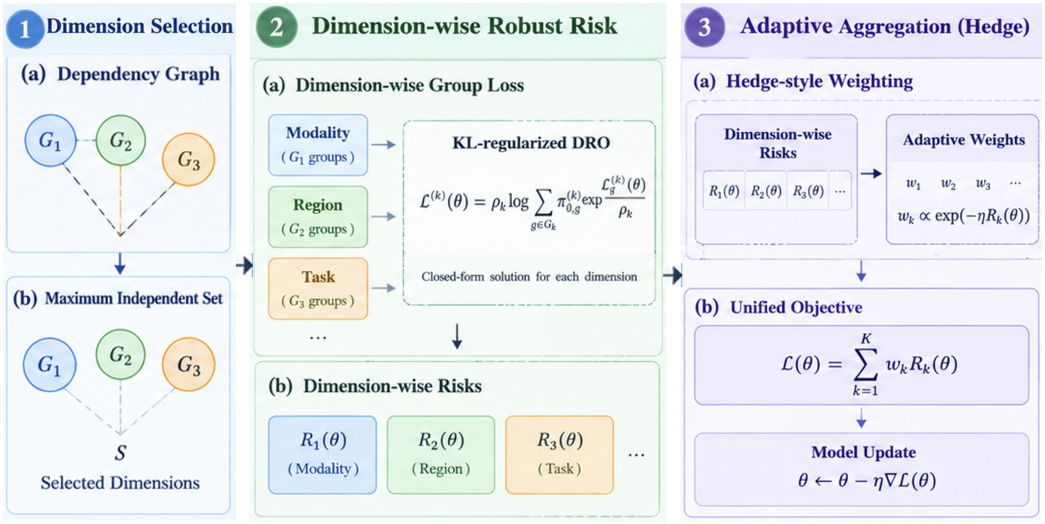

Figure 2: Overview of Dimension-wise DRO. The method proceeds in three stages. (1) **Dimension selection**: pairwise dependence between grouping dimensions is measured by Cramér's V, and a maximally non-redundant subset is selected via maximum independent set. (2) **Dimension-wise robust risk**: for each selected dimension, a KL-regularized Group DRO objective computes a robust risk over its groups, which admits a closed-form solution. (3) **Adaptive aggregation**: the dimension-wise risks are combined with a Hedge-style weighting scheme that dynamically adjusts the weight of each dimension based on its current robust loss.

Cramér's V (Bergsma, 2013) is a normalized association strength measure defined as

$$V(A,B) = \sqrt{\frac{\chi^2(A,B)}{n \cdot \min(r-1,\ c-1)}}, \qquad V(A,B) \in [0,1]. \tag{5}$$

Cramér's V Bergsma (2013) measures the strength of association between two categorical variables. This makes it well suited for our setting, where the grouping dimensions are categorical. We use $V(A,B)$ to quantify pairwise dependence between grouping dimensions throughout the paper, specifically in the dimension selection step of our method.

3 Method

In this section, we introduce the proposed Dimension-wise Group Distributionally Robust Optimization **(Dimension-wise DRO)** framework. Our method consists of three stages. First, we select a set of non-redundant grouping dimensions by measuring pairwise dependence with Cramér's V and solving a maximum independent set problem (Section 3.1). Second, for each selected dimension, we define a dimension-wise robust risk using a KL-regularized form of Group DRO (Section 3.1). Third, we aggregate the dimension-wise risks into a unified objective with a Hedge-style adaptive weighting scheme (Section 3.2). This design avoids constructing combinatorial groups while preserving robustness along each semantic dimension. An overview of the framework is shown in Figure 2.

3.1 Dimension-wise Group Risk

The K grouping dimensions defined in Section 2.1 are not necessarily independent. For example, in driving data, time of day largely determines illumination. When two dimensions are highly correlated, enforcing robustness along both of them imposes largely redundant constraints without adding new robustness guarantees. Before performing robust training, we therefore prune redundant dimensions and retain a subset that is as independent as possible.

Dimension selection. To this end, we measure pairwise dependence between dimensions using Cramér's V (Eq. equation 5), which yields a symmetric dependence matrix $\mathbf{V} \in [0,1]^{K \times K}$. A

large V_{ij} indicates that dimensions i and j carry similar semantic information and should not both be retained. We then construct a dependency graph over dimensions, where an edge connects two dimensions whose dependence exceeds a threshold t, and select a non-redundant subset by finding a maximum independent set—the largest set of dimensions with no edges among them. Solving this problem exactly is NP-hard, so we use a standard greedy heuristic in practice. Full details of the graph construction and the selection procedure are provided in Appendix A.

After selecting a set of non-redundant dimensions, we now define the robust risk along each dimension. Let $\ell(\theta; z)$ denote the loss of model parameters θ on sample z. For dimension k and group g, let $P_g^{(k)}$ denote the distribution of samples in group g along dimension k. The group loss is $L_g^{(k)}(\theta) = \mathbb{E}_{z\sim P_g^{(k)}}[\ell(\theta; z)]$. Following the soft Group DRO formulation in Eq. 1, we define the soft worst-case group risk for dimension k as

$$\max_{\pi^{(k)}\in\Delta^{G_k}} \sum_{g=1}^{G_k} \pi_g^{(k)} L_g^{(k)}(\theta), \tag{6}$$

where $\pi^{(k)}$ is a distribution over the G_k groups of dimension k and $\pi_g^{(k)}$ denotes the weight assigned to group g.

However, $\pi^{(k)}$ is unconstrained beyond the simplex, so when some groups contain very few samples, their empirical losses are noisy. To control this, we add a KL regularization term that keeps $\pi^{(k)}$ close to a reference distribution $\pi_0^{(k)}$:

$$L^{(k)}(\theta) = \max_{\pi^{(k)}\in\Delta^{G_k}} \sum_{g=1}^{G_k} \pi_g^{(k)} L_g^{(k)}(\theta) - \rho_k \, D_{KL}(\pi^{(k)} \| \pi_0^{(k)}), \tag{7}$$

where $\rho_k > 0$ is a regularization parameter and $\pi_0^{(k)}$ is typically chosen as the empirical group distribution. This design serves two purposes. First, the KL term prevents $\pi^{(k)}$ from deviating too far from $\pi_0^{(k)}$, avoiding over-concentration on a single group. Second, it admits a unique closed-form solution (details in Appendix B.3):

$$L^{(k)}(\theta) = \rho_k \log\left(\sum_{g=1}^{G_k} \pi_{0,g}^{(k)} \exp\left(\frac{L_g^{(k)}(\theta)}{\rho_k}\right)\right). \tag{8}$$

This closed form allows us to compute the dimension-wise robust risk efficiently without solving an inner optimization problem at each step.

3.2 Multi-Dimensional Robust Objective

Given the dimension-wise robust risks, we now aggregate them into a single training objective. One option is to take the maximum over dimensions, but this would only propagate gradients from the single hardest dimension and leave the remaining dimensions unoptimized. We therefore use a weighted sum:

$$L_{\text{multi}}(\theta) = \sum_{k=1}^{K} w_k L^{(k)}(\theta), \tag{9}$$

where w_k controls the relative importance of dimension k. This form allows all dimensions to contribute to the gradient.

The remaining question is how to set the weights w_k. Fixed weights do not adapt to the current robust loss of each dimension, so a dimension that is already well optimized continues to receive the same weight as one that is still far from robust. We therefore update the weights adaptively using a Hedge-style rule, which increases the weight of dimensions with higher robust loss. Let $L^{(k)}(\theta_t)$ denote the robust loss of dimension k at iteration t. The update is

$$\tilde{w}_k^{(t+1)} = w_k^{(t)} \cdot \exp\left(\eta_t \cdot L^{(k)}(\theta_{t+1})\right), \tag{10}$$

where $\eta_t = \eta_0/t$ is the weight learning rate at iteration t. This rule increases the weight of dimensions with higher robust loss, directing the model's attention toward dimensions that remain difficult. We

Algorithm 1 Dimension-wise DRO with Hedge-Style Weight Update

Input: Model parameters θ_0, dimension weights $w_0 = [1/K, \ldots, 1/K]$, weight learning rate η_0, model learning rate α, lower bound β
for $t = 0$ to $T - 1$ **do**
  Sample a mini-batch $\mathcal{B}$
  for each dimension $k = 1$ to K **do**
    for each group $g = 1$ to G_k **do**
      Compute group loss: $L_g^{(k)}(\theta_t) = \frac{1}{|\mathcal{B}_g^{(k)}|} \sum_{z \in \mathcal{B}_g^{(k)}} \ell(\theta_t; z)$
    end for
    Compute robust weights $\pi^{(k)}$ via Eq. 8
    Compute dimension loss: $L^{(k)}(\theta_t) = \sum_g \pi_g^{(k)} L_g^{(k)}(\theta_t)$
  end for
  Compute total loss: $L_{\text{total}} = \sum_{k=1}^{K} w_k^{(t)} L^{(k)}(\theta_t)$
  Update model: $\theta_{t+1} = \theta_t - \alpha_t \nabla_\theta L_{\text{total}}$
  Weight update (using new model):
  for each dimension $k = 1$ to K **do**
    $\tilde{w}_k = w_k^{(t)} \cdot \exp\left(\eta_t \cdot L^{(k)}(\theta_{t+1})\right)$
  end for
  $w_k^{(t+1)} = \tilde{w}_k / \sum_j \tilde{w}_j$
  $w_k^{(t+1)} = \max(w_k^{(t+1)}, \beta)$, then renormalize
end for
Return: Trained model parameters θ_T

then normalize the weights:

$$\bar{w}_k^{(t+1)} = \frac{\tilde{w}_k^{(t+1)}}{\sum_{j=1}^{K} \tilde{w}_j^{(t+1)}}. \tag{11}$$

To prevent any dimension from being entirely suppressed, we clip the normalized weights at a lower bound β and renormalize:

$$w_k^{(t+1)} = \frac{\max\left(\bar{w}_k^{(t+1)}, \beta\right)}{\sum_{j=1}^{K} \max\left(\bar{w}_j^{(t+1)}, \beta\right)}, \tag{12}$$

where $\beta > 0$ is a small constant. This ensures that every dimension remains active, so that the model continues to capture shifts along all selected dimensions.

Overall algorithm. The complete training procedure is summarized in Algorithm 1. At each iteration, we first update the model parameters using the current weights, then update the dimension weights using the losses computed from the new model.

3.3 Theoretical Guarantees

Dimension-wise DRO optimizes each dimension separately and then aggregates the results. This design raises two questions. First, does minimizing the aggregated objective actually control the worst-case risk across all dimensions? Second, since the dimension weights w_k and the model parameters θ are updated in a coupled manner, does the overall procedure still converge? We address both questions below.

Does the aggregated objective control the worst-case risk? Dimension-wise DRO optimizes a weighted sum of dimension-wise risks, but the bound concerns the worst-case risk across dimensions and groups. These two quantities can differ substantially: if one dimension has a much larger loss than the others, the weighted sum can be much smaller than the worst-case risk, so minimizing the former does not automatically control the latter.

The gap is controlled by our choice of the Hedge update. We adopt Hedge precisely because its update rule produces weights of a softmax form over the per-dimension losses, which provides a

bound on the gap between the maximum and the weighted sum. Expanding the Hedge recursion (Appendix B.16, Step 2) shows that the final weights approximate this softmax form. By the softmax approximation property (Lemma B.14), the gap is at most $\log K/\tau$, where $\tau = \eta_0(T-1)$. This gap grows only logarithmically in K and can be made arbitrarily small by increasing τ. Theorem 3.1 formalizes this argument.

Theorem 3.1 (Generalization Bound). *Let $\ell(\theta; z) \in [0, M]$ be a bounded loss function. For each dimension $k \in [K]$, let $n_{\min}^{(k)} = \min_g n_g^{(k)}$ be the minimum group sample size. Let θ_T be the final model returned by the algorithm after T iterations, where T is the total number of training iterations and $\eta > 0$ is the learning rate of the Hedge update. Let $w_{T,k}$ denote the Hedge weight for dimension k at iteration T, and let $\hat{F}(\theta_T) = \sum_{k=1}^{K} w_{T,k}\hat{L}^{(k)}(\theta_T)$ be the empirical weighted objective. Define $\tau = \eta_0 \log T$. Then, with probability at least $1-\delta$ over the training data,*

$$\max_{k\in[K]} \max_{g\in[G_k]} \mathbb{E}_{P_g^{\text{test}}}[\ell(\theta_T)] \;\le\; \hat{F}(\theta_T) + \frac{\log K}{\tau} + M \max_{k\in[K]} \sqrt{\frac{\log(2KG_k/\delta)}{2n_{\min}^{(k)}}},$$

The bound decomposes into three terms: the empirical weighted objective $\hat{F}(\theta_T)$, the approximation gap $\log K/\tau$, and a statistical term that depends only on the minimum group size across dimensions. Theorem 3.1 shows that minimizing the proposed objective controls the worst-case group risk across all dimensions.

Does the coupled update converge? The algorithm alternates between updating θ and w. Standard convergence analysis assumes that F decreases at every iteration, so that the total decrease accumulates and the algorithm converges. This assumption does not hold here: after θ is updated, the subsequent update of w can make F increase, partially offsetting the decrease from the gradient step. As a result, F is no longer guaranteed to decrease monotonically, and the standard telescoping argument does not directly apply.

We handle this by bounding the additional change in F caused by the Hedge update. Since Hedge changes w by $O(1/t)$ at step t (Lemma B.8), this additional change accumulates to $O(\log T)$ over T iterations, which is the source of the extra logarithmic factor in the final rate. Theorem B.11 formalizes this argument.

Theorem 3.2 (Convergence Rate). *Let $\{(\theta_t, w_t)\}_{t=0}^{T-1}$ be the iterates generated by the algorithm, where $\theta_t \in \Theta$ denotes the model parameters and $w_t \in \Delta_K$ denotes the dimension weights at iteration t. Let $F(\theta, w) = \sum_{k=1}^{K} w_k L^{(k)}(\theta)$ be the objective, and let $F^* = \inf_{\theta\in\Theta, w\in\Delta_K} F(\theta, w)$. Set the model step size $\alpha_t = \alpha/\sqrt{t+1}$ with $\alpha > 0$, and the Hedge step size $\eta_t = \eta_0/t$ with $\eta_0 M \le 1/2$, where M is the loss upper bound from Assumption B.2. Under Assumptions B.1, B.2, and B.3, the iterates satisfy:*

$$\min_{0\le t<T} \mathbb{E}\big[\|\nabla_\theta F(\theta_t, w_t)\|^2\big] \le \frac{C\log T}{\sqrt{T}},$$

where $C > 0$ is a constant depending on L_F (the smoothness constant from Lemma B.6), $\mathcal{G}$ (the gradient bound), M, α, η_0, and $\Delta_F = F(\theta_0, w_0) - F^*$.

Theorem B.11 establishes that Dimension-wise DRO converges to a stationary point at a rate of $O(\log T/\sqrt{T})$, matching that of standard stochastic optimization up to this logarithmic factor. Proofs of both theorems are provided in Appendix B.

4 Experiments

4.1 Experimental Setup

We evaluate on three multimodal benchmarks, each carrying multiple grouping dimensions. OmniMedVQA (Hu et al., 2024) is a medical VQA benchmark annotated along imaging modality, anatomical region, and clinical task. BDD100K (Yu et al., 2020) is an autonomous driving detection benchmark annotated along weather, time of day, and scene type. For both datasets, the grouping dimensions are chosen so that they capture distinct semantic axes. VQAv2 (Goyal et al., 2017) is a natural-image VQA benchmark with several original annotations, including question type, answer

Table 1: Worst-group and average accuracy on OmniMedVQA, VQAv2, and BDD100K under their original (unshifted) distributions. BDD100K is evaluated with mAP. All results are averaged over three runs with different random seeds. Boldface indicates the best performance.

		Unshifted OmniMedVQA		**Unshifted VQAv2**		**Unshifted BDD100K**	
Method	**Group label**	Worst Acc	Average Acc	Worst Acc	Average Acc	Worst mAP	Average mAP
ERM	×	22.1	84.5	32.3	59.0	0.250	**0.430**
JTT	×	24.6	84.0	33.5	57.3	0.251	0.414
CnC	×	23.1	83.8	34.2	57.5	0.253	0.419
GIC	×	24.4	83.5	32.7	56.4	0.258	0.403
SSA	×	22.5	83.2	35.1	57.8	0.260	0.421
Group DRO	✓	24.9	84.2	37.9	60.1	0.257	0.425
LISA	✓	22.7	84.1	35.9	58.6	0.251	0.413
DFR	✓	22.3	83.0	33.1	57.2	0.257	0.418
H-DRO	✓	25.1	81.3	37.4	56.9	0.259	0.415
Ours	✓	**28.4**	**85.3**	**38.5**	**61.0**	**0.263**	**0.430**

Table 2: Worst-group and average accuracy on OmniMedVQA, VQAv2, and BDD100K under shifted distributions. BDD100K is evaluated with mAP. All results are averaged over three runs with different random seeds. Boldface indicates the best performance.

		Shifted OmniMedVQA		**Shifted VQAv2**		**Shifted BDD100K**	
Method	**Group label**	Worst Acc	Average Acc	Worst Acc	Average Acc	Worst mAP	Average mAP
Group DRO	✓	16.9	78.2	21.5	51.2	0.208	0.412
LISA	✓	14.8	77.9	20.6	49.6	0.201	0.410
DFR	✓	14.5	77.6	19.8	47.7	0.204	0.409
H-DRO	✓	17.1	76.3	20.3	47.9	0.209	0.404
Ours	✓	**20.4**	**79.7**	**23.6**	**52.8**	**0.211**	**0.418**

type, and visual class. We apply the dimension selection procedure in Section A to these annotations and retain two dimensions, question type and visual class. These three datasets span distinct domains and task types, allowing us to test whether Dimension-wise DRO generalizes beyond a single application. Full dataset statistics and the specific groups within each dimension are provided in Appendix C.

We evaluate each method in two settings. In the *unshift* setting, training and test samples are drawn from the same distribution, and we report worst-group accuracy on the test set. In the *shift* setting, we leave out the smallest group from training and report worst-group accuracy on the full test set. This directly tests whether a method remains robust when a subgroup is absent from training.

We compare against two families of baselines that cover the main approaches to group robustness. Annotation-free methods (ERM, JTT (Liu et al., 2021), CnC (Zhang et al., 2022), GIC (Han & Zou, 2024b), SSA (Nam et al., 2022)) do not require group labels, while annotation-dependent methods (Group DRO (Sagawa et al., 2020), LISA (Yao et al., 2022), DFR (Kirichenko et al., 2023), H-DRO (Jo et al., 2026)) assume access to group annotations. For Group DRO and other annotation-dependent baselines that require a single grouping structure, we select the dimension that yields the best validation worst-group accuracy and use it as the grouping structure for training. We do not form groups as the Cartesian product of all dimensions, as this consistently degrades worst-group performance due to data sparsity (Appendix D.2).

Implementation details. For OmniMedVQA and VQAv2, we fine-tune Qwen2.5-VL-2B(Bai et al., 2025) for 2 epochs with batch size 32 using AdamW (Loshchilov & Hutter, 2019) and a learning rate of 1×10^{-5}. For BDD100K, we train YOLOv5-s (Jocher, 2021) for 50 epochs with batch size 32 using AdamW and a learning rate of 1×10^{-5}. All experiments are run on 8 RTX 4090 GPUs. The Hedge learning rate is $\eta_0 = 0.1$, the KL regularization parameter is $\rho_k = 0.05$, and the dependency threshold is $t = 0.7$. Results are averaged over three random seeds. Full hyperparameter settings are provided in Appendix C. We report worst-group and average accuracy for OmniMedVQA and VQAv2, and worst-group and average mAP for BDD100K. The worst-group defined as the minimum accuracy over all groups formed by the Cartesian product of the selected dimensions. This provides a stricter evaluation than taking the minimum over single-dimension groups, since combinatorial groups are often much smaller and harder to optimize.

Table 3: Ablation studies on VQAv2 under unshift and shift settings. We report worst-group and average accuracy (%). Boldface indicates the best performance

Configuration	Unshift		Shift	
	Worst Acc	Average Acc	Worst Acc	Average Acc
Dimension-wise DRO (full)	**38.5**	**61.0**	**23.6**	**52.8**
w/o dimension selection	37.6	59.4	22.4	51.7
w/o KL regularization	38.0	60.7	22.7	52.5
w/o Hedge (uniform weights)	38.2	60.5	23.1	52.5

4.2 Main Results

Tables 1 and 2 report worst-group and average performance on the three benchmarks under unshift and shift settings. Across all datasets and both settings, Dimension-wise DRO achieves the best worst-group performance, indicating that it improves worst-group robustness across dimensions. The gains are larger under shift than under unshift: on OmniMedVQA, worst-group accuracy improves over the strongest baseline by 3.5% under shift vs. 3.3% under unshift; on VQAv2, 2.1% vs. 0.6%; on BDD100K, 0.02 vs. 0.03. Dimension-wise DRO also outperforms annotation-free methods such as JTT, CnC, and GIC, even though these methods do not require group labels. Annotation-free methods improve over ERM but remain below Dimension-wise DRO, suggesting that inferring a single proxy signal does not fully capture dimension-specific robustness. Finally, the improvement in worst-group performance does not come at the cost of average accuracy: on all three datasets, Dimension-wise DRO's average accuracy is on par with or better than the best baseline.

4.3 Ablation Studies

We ablate the three components of Dimension-wise DRO (dimension selection, KL regularization, and Hedge-style weight update) on VQAv2. Table 3 reports worst-group and average accuracy under both unshift and shift settings. **Dimension selection.** Replacing the MIS-selected dimensions with all annotated dimensions reduces worst-group accuracy from 38.5% to 37.6% under unshift and from 23.6% to 22.4% under shift. This is consistent with our motivation that correlated dimensions impose redundant constraints, which become harmful when the model must allocate capacity across multiple axes. **KL regularization.** Replacing the KL-regularized risk with the hard maximum reduces worst-group accuracy from 38.5% to 38.0% under unshift and from 23.6% to 22.7% under shift. This suggests that the smoothing effect of KL regularization helps stabilize training when some groups contain very few samples. **Hedge weight update.** Replacing the adaptive weights with uniform weights reduces worst-group accuracy from 38.5% to 38.2% under unshift and from 23.6% to 23.1% under shift. This suggests that adaptively reweighting dimensions based on their current robust loss contributes to the final performance.

Additional results. Due to space constraints, we provide additional experimental results in the appendix D. Appendix D.1 reports full ablation results on OmniMedVQA and BDD100K. Appendix D.3 provides the sensitivity analysis for the KL regularization parameter ρ_k and the Hedge learning rate η_0, as well as the Cramér's V heatmap for VQAv2.

5 Conclusion

We introduced Dimension-wise DRO, a robust training framework for multimodal data with multiple grouping dimensions that enforces robustness along each semantic dimension independently and aggregates the results with a Hedge-style adaptive weighting scheme, avoiding combinatorial group construction while requiring only single-dimension annotations. Experiments on medical VQA, natural-image VQA, and autonomous driving detection show that Dimension-wise DRO consistently improves worst-group performance under both shift and unshift settings without sacrificing average accuracy, and ablations confirm that dimension selection, KL regularization, and Hedge weighting each contribute to the final performance. An interesting direction for future work is to investigate multi-dimensional robust optimization in reinforcement learning settings, where the data distribution evolves together with the policy, creating a form of shift that static grouping structures cannot capture.

AI USE STATEMENT

A large language model (LLM) was employed for language polishing and grammar correction. All scientific ideas, experimental design, analysis, and conclusions were generated solely by the authors. The LLM did not contribute to any research ideation.

ETHICS STATEMENT

Our work uses publicly available benchmarks and open model checkpoints and involves no humar-subiects or private data. We do not foresee direct negative societal impacts beyond the general risksof improving multimodal language models.

REPRODUCIBILITY STATEMENT

All interventions, thresholds, and evaluation settings are specified in Appendices, with complete per-model-task results tabulated there. Code for reproducing the interventions and figures wilbe released upon acceptance.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Wicher Bergsma. A bias-correction for cramér's v and tschuprow's t. *Journal of the Korean Statistical Society*, 42(3):323–328, 2013.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and trends in Machine Learning*, 8(3-4):231–357, 2015.

Qing Cai, Shaozhe Wang, Fan Zhang, Cheng Zhang, Yu Liu, Huafeng Li, Junyu Dong, and David Zhang. Multimodal artificial intelligence for disease diagnosis: Advances, applications, and challenges. *Pattern Recognition*, pp. 113456, 2026.

MingCai Chen, Baoming Zhang, Zongbo Han, Wenyu Jiang, Yanmeng Wang, Shuai Feng, Yuntao Du, and Bingkun Bao. Test-time selective adaptation for uni-modal distribution shift in multi-modal data. In *Forty-second International Conference on Machine Learning*, 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Xiaofeng Han, Shunpeng Chen, Zenghuang Fu, Zhe Feng, Lue Fan, Dong An, Changwei Wang, Li Guo, Weiliang Meng, Xiaopeng Zhang, et al. Multimodal fusion and vision-language models: A survey for robot vision. *Information Fusion*, pp. 103652, 2025.

Yujin Han and Difan Zou. Improving group robustness on spurious correlation requires preciser group inference. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17480–17504. PMLR / OpenReview.net, 2024a. URL https://proceedings.mlr.press/v235/han24g.html.

Yujin Han and Difan Zou. Improving group robustness on spurious correlation requires preciser group inference. In *International Conference on Machine Learning*, pp. 17480–17504. PMLR, 2024b.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22170–22183, 2024.

Sung Ho Jo, Seonghwi Kim, and Minwoo Chae. Mitigating spurious correlation via distributionally robust learning with hierarchical ambiguity sets. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

Glenn Jocher. YOLOv5: A state-of-the-art real-time object detection system. https://github.com/ultralytics/yolov5, 2021.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=Zb6c8A-Fghk.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pp. 6781–6792. PMLR, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Jun Hyun Nam, Jaehyung Kim, Jaeho Lee, and Jinwoo Shin. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=_F9xpOrqyX9.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Aayush Atul Verma, Amir Saeidi, Shamanthak Hegde, Ajay Therala, Fenil Denish Bardoliya, Nagaraju Machavarapu, Shri Ajay Kumar Ravindhiran, Srija Malyala, Agneet Chatterjee, Yezhou Yang, et al. Evaluating multimodal large language models across distribution shifts and augmentations. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 5314–5324. IEEE, 2024.

Chuheng Wei, Ziye Qin, Ziyan Zhang, Guoyuan Wu, and Matthew J Barth. Integrating multi-modal sensors: A review of fusion techniques for intelligent vehicles. In *2025 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1817–1824. IEEE, 2025.

Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, pp. 25407–25437. PMLR, 2022.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 2633–2642. IEEE, 2020.

Michael Zhang, Nimit Sharad Sohoni, Hongyang R. Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: a contrastive approach for improving robustness to spurious correlations. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26484–26516. PMLR, 2022. URL https://proceedings.mlr.press/v162/zhang22z.html.

A SELECTING NON-REDUNDANT GROUPING DIMENSIONS

The K grouping dimensions defined in Section 2.1 are not necessarily independent. For example, in driving data, time of day largely determines illumination. When two dimensions are highly correlated, enforcing robustness along both of them imposes largely redundant constraints without adding new robustness guarantees. Before performing robust training, we therefore prune redundant grouping dimensions and retain a set of dimensions that are as independent as possible.

To identify redundant dimensions, we need to quantify how strongly two dimensions are associated. We use Cramér's V (Eq. equation 5) as the association measure: for any two dimensions (i, j),

$$V_{ij} = V(i, j), \tag{13}$$

and computing V_{ij} over the dataset yields a symmetric matrix $\mathbf{V} \in [0, 1]^{K \times K}$ with $V_{ii} = 1$. A large V_{ij} indicates that dimensions i and j carry similar semantic information and should not both be retained.

To select a non-redundant subset, we interpret $\mathbf{V}$ as a graph over dimensions.We build a *grouping-dimension dependency graph* $G_t = (\mathcal{V}, \mathcal{E}_t)$ with $\mathcal{V} = \{1, \ldots, K\}$. Two dimensions i and j are connected by an edge if their dependence exceeds the threshold t, i.e.,

$$\mathcal{E}_t = \{(i, j) : V_{ij} \geq t\}. \tag{14}$$

Here $t \in [0, 1]$ is the threshold above which two dimensions are considered redundant.

Given G_t, selecting a non-redundant subset is equivalent to finding a maximum independent set: we want the largest set of dimensions with no edges among them,

$$S^\star = \arg\max_{S \subseteq \mathcal{V}} |S| \quad \text{s.t.} \quad \forall i \neq j \in S,\ (i, j) \notin \mathcal{E}_t. \tag{15}$$

Solving MIS exactly is NP-hard, so we use a standard greedy heuristic in practice.

B THEORETICAL ANALYSIS

This section establishes two main results for our multi-dimensional Group DRO framework:

1. **Convergence** (Theorem B.11): The algorithm achieves a stationary-point convergence rate of $O(\log T/\sqrt{T})$ for the objective $F(\theta, w)$.
2. **Generalization** (Theorem B.16): The final model's worst-group test loss is controlled by the empirical objective plus statistical terms scaling as $O(\sqrt{\log(KG/\delta)/n_{\min}})$.

The analysis proceeds as follows. Section B.1 introduces the setup and assumptions. Section B.3 derives the closed-form solution of the inner KL-regularized DRO problem. Section B.4 establishes convergence, building on smoothness (Lemma B.6) and Hedge weight stability (Lemma B.8). Section B.5 proves the generalization bound, using the log-sum-exp Lipschitz property (Lemma B.12) and a softmax approximation lemma (Lemma B.14).

B.1 PROBLEM SETUP

Let $\ell(\theta; z) \in [0, M]$ be the base loss function, where θ are the model parameters and $z \in \mathcal{Z}$ is a data sample.

Multi-dimensional grouping. We consider K pre-defined grouping dimensions. For each dimension $k \in [K]$, the data is partitioned into G_k mutually exclusive groups. For a group $g \in [G_k]$, let $B_g^{(k)}$ denote the set of training samples belonging to group g in dimension k, and let $n_g^{(k)} = |B_g^{(k)}|$ be its size. Define the minimum group size for dimension k as $n_{\min}^{(k)} = \min_{g \in [G_k]} n_g^{(k)}$.

Group and dimension losses. For each group g in dimension k, the population and empirical group losses are

$$L_g^{(k)}(\theta) = \mathbb{E}_{z \sim P_g^{(k)}}[\ell(\theta; z)], \qquad \hat{L}_g^{(k)}(\theta) = \frac{1}{n_g^{(k)}} \sum_{i \in B_g^{(k)}} \ell(\theta; z_i).$$

For each dimension k, the KL-regularized Group DRO loss is defined as

$$L^{(k)}(\theta) = \max_{\pi^{(k)} \in \Delta_{G_k}} \left\{ \sum_{g=1}^{G_k} \pi_g^{(k)} L_g^{(k)}(\theta) - \rho_k D_{\mathrm{KL}}(\pi^{(k)} \| \pi_0^{(k)}) \right\},$$

where $\rho_k > 0$ is the regularization parameter and $\pi_0^{(k)} \in \Delta_{G_k}$ is the reference distribution (typically uniform).

Multi-dimensional objective. Given dimension weights $w = (w_1, \ldots, w_K) \in \Delta_K$, the multi-dimensional objective is

$$F(\theta, w) = \sum_{k=1}^{K} w_k L^{(k)}(\theta), \qquad \hat{F}(\theta, w) = \sum_{k=1}^{K} w_k \hat{L}^{(k)}(\theta).$$

B.2 Assumptions

Assumption B.1 (Smoothness of Base Loss)**.** *The loss function $\ell(\theta; z)$ is L-smooth in θ uniformly over $z \in \mathcal{Z}$:*

$$\|\nabla_\theta \ell(\theta_1; z) - \nabla_\theta \ell(\theta_2; z)\| \le L\|\theta_1 - \theta_2\|, \qquad \forall \theta_1, \theta_2 \in \Theta,\ z \in \mathcal{Z}.$$

Assumption B.2 (Boundedness of Base Loss)**.** *The loss function is uniformly bounded:*

$$0 \le \ell(\theta; z) \le M, \qquad \forall \theta \in \Theta,\ z \in \mathcal{Z}.$$

Assumption B.3 (Bounded Gradient of Base Loss)**.** *The gradient of the loss function is uniformly bounded:*

$$\|\nabla_\theta \ell(\theta; z)\| \le \mathcal{G}, \qquad \forall \theta \in \Theta,\ z \in \mathcal{Z}.$$

Assumption B.4 (Loss Stability in Late Training)**.** *There exists a threshold $T_0 < T$ such that for all $t \ge T_0$ and all dimensions $k \in [K]$, the per-dimension empirical losses satisfy*

$$\left| \hat{L}^{(k)}(\theta_{t+1}) - \hat{L}^{(k)}(\theta_T) \right| \le \epsilon$$

for some small constant $\epsilon > 0$.

B.3 Closed-Form Solution of KL-Regularized Group DRO

The inner maximization in $L^{(k)}(\theta)$ is a concave problem over the simplex with a KL regularizer. The following lemma shows it admits a closed-form solution, which is essential for both computing the algorithm's updates and analyzing their smoothness (Lemma B.6).

Lemma B.5 (Closed-Form Solution of KL-Regularized Group DRO)**.** *Consider a single dimension k with G_k groups and group losses $L_g^{(k)}(\theta)$. Let $\pi_{0,g}^{(k)} > 0$ be a reference distribution and $\rho_k > 0$ a fixed regularization parameter. Consider the KL-regularized Group DRO problem*

$$L^{(k)}(\theta) = \max_{\pi^{(k)} \in \Delta^{G_k}} \left\{ \sum_{g=1}^{G_k} \pi_g^{(k)} L_g^{(k)}(\theta) - \rho_k D_{\mathrm{KL}}(\pi^{(k)} \| \pi_0^{(k)}) \right\}.$$

Then the problem admits a unique optimal solution of the closed form

$$\pi_g^{(k)*}(\theta) = \frac{\pi_{0,g}^{(k)} \exp\left(\frac{L_g^{(k)}(\theta)}{\rho_k}\right)}{\sum_{j=1}^{G_k} \pi_{0,j}^{(k)} \exp\left(\frac{L_j^{(k)}(\theta)}{\rho_k}\right)}, \quad g \in [G_k].$$

And $L^{(k)}(\theta)$ can be expressed in closed form:

$$L^{(k)}(\theta) = \rho_k \log\left(\sum_{g=1}^{G_k} \pi_{0,g}^{(k)} \exp\left(\frac{L_g^{(k)}(\theta)}{\rho_k}\right) \right).$$

Proof. The objective function is:

$$f(\pi) = \sum_{g=1}^{G_k} \pi_g L_g^{(k)}(\theta) - \rho_k \sum_{g=1}^{G_k} \pi_g \log\left(\frac{\pi_g}{\pi_{0,g}^{(k)}}\right).$$

Since the negative entropy term $-\rho_k \sum_g \pi_g \log \pi_g$ is strictly concave in π (for $\rho_k > 0$) and the linear term is concave, the overall objective $f(\pi)$ is strictly concave on the simplex Δ^{G_k}. Hence a unique maximizer exists.

To find it, we form the Lagrangian with multiplier $\lambda \in \mathbb{R}$ for the simplex constraint $\sum_g \pi_g = 1$:

$$\mathcal{L}(\pi, \lambda) = \sum_{g=1}^{G_k} \pi_g L_g^{(k)}(\theta) - \rho_k \sum_{g=1}^{G_k} \pi_g \log\left(\frac{\pi_g}{\pi_{0,g}^{(k)}}\right) + \lambda\left(1 - \sum_{g=1}^{G_k} \pi_g\right).$$

Taking the derivative with respect to π_g and setting to zero:

$$\frac{\partial \mathcal{L}}{\partial \pi_g} = L_g^{(k)}(\theta) - \rho_k\left(\log\left(\frac{\pi_g}{\pi_{0,g}^{(k)}}\right) + 1\right) - \lambda = 0.$$

Rearranging:$\log\left(\frac{\pi_g}{\pi_{0,g}^{(k)}}\right) = \frac{L_g^{(k)}(\theta) - \lambda - \rho_k}{\rho_k}$. Then, exponentiating both sides:

$$\pi_g = \pi_{0,g}^{(k)} \exp\left(\frac{L_g^{(k)}(\theta) - \lambda - \rho_k}{\rho_k}\right) = \pi_{0,g}^{(k)} \exp\left(\frac{L_g^{(k)}(\theta)}{\rho_k}\right) \cdot \exp\left(-\frac{\lambda + \rho_k}{\rho_k}\right).$$

Let $C = \exp(-(\lambda + \rho_k)/\rho_k)$ be a positive constant independent of g. Then:

$$\pi_g = C \cdot \pi_{0,g}^{(k)} \exp\left(\frac{L_g^{(k)}(\theta)}{\rho_k}\right).$$

The constant C is determined by the simplex constraint $\sum_g \pi_g = 1$:

$$\sum_{g=1}^{G_k} \pi_g = C \sum_{j=1}^{G_k} \pi_{0,j}^{(k)} \exp\left(\frac{L_j^{(k)}(\theta)}{\rho_k}\right) = 1 \quad \Longrightarrow \quad C = \frac{1}{\sum_{j=1}^{G_k} \pi_{0,j}^{(k)} \exp\left(\frac{L_j^{(k)}(\theta)}{\rho_k}\right)}.$$

Substituting C back gives the closed form:

$$\pi_g^{(k)*}(\theta) = \frac{\pi_{0,g}^{(k)} \exp\left(\frac{L_g^{(k)}(\theta)}{\rho_k}\right)}{\sum_{j=1}^{G_k} \pi_{0,j}^{(k)} \exp\left(\frac{L_j^{(k)}(\theta)}{\rho_k}\right)}.$$

Substituting $\pi_g^{(k)*}$ back into $f(\pi)$ and simplifying yields $L^{(k)}(\theta) = \rho_k \log \sum_g \pi_{0,g} e^{L_g^{(k)}/\rho_k}$. Strict concavity guarantees that this stationary point is the unique global maximizer. □

B.4 Convergence Rate

To analyze convergence, we first establish that $F(\theta, w)$ is smooth in θ. The main challenge is that the optimal group weights $\pi^{(k)*}(\theta)$ depend on θ through a softmax, introducing additional curvature proportional to $\mathcal{G}^2/\rho_k$.

Lemma B.6 (Smoothness of F w.r.t. θ)**.** *Let $F(\theta, w) = \sum_{k=1}^{K} w_k L^{(k)}(\theta)$ with $\sum_{k=1}^{K} w_k = 1$. Under Assumptions B.1, B.2, and B.3, F is L_F-smooth in θ:*

$$\|\nabla_\theta F(\theta_1, w) - \nabla_\theta F(\theta_2, w)\|_2 \le L_F \|\theta_1 - \theta_2\|_2,$$

where $L_F = \max_{k \in [K]} \left(L_\ell + \frac{2\mathcal{G}^2}{\rho_k}\right)$.

Proof. The proof proceeds in four steps.

Step 1: Properties of group losses. For each group loss $L_g^{(k)}(\theta) = \frac{1}{n_g^{(k)}} \sum_{i \in B_g^{(k)}} \ell(\theta; z_i)$, the gradient is the average of sample gradients: $\nabla L_g^{(k)}(\theta) = \frac{1}{n_g^{(k)}} \sum_{i \in B_g^{(k)}} \nabla \ell(\theta; z_i)$. We will use the following three properties:

- *Smoothness*: By Assumption B.1,
$$\|\nabla L_g^{(k)}(\theta_1) - \nabla L_g^{(k)}(\theta_2)\| \le \frac{1}{n_g^{(k)}} \sum_i \|\nabla \ell(\theta_1; z_i) - \nabla \ell(\theta_2; z_i)\| \le L_\ell \|\theta_1 - \theta_2\|.$$
- *Bounded gradient*: By Assumption B.3, $\|\nabla L_g^{(k)}(\theta)\| \le \frac{1}{n_g^{(k)}} \sum_i \|\nabla \ell(\theta; z_i)\| \le \mathcal{G}$.
- *Boundedness*: By Assumption B.2, $0 \le L_g^{(k)}(\theta) \le M$.

Step 2: Lipschitz continuity of $\pi^{(k)*}(\theta)$. By Lemma B.5, the optimal group weight is
$$\pi_g^{(k)*}(\theta) = \frac{\pi_{0,g}^{(k)} \exp\Big(L_g^{(k)}(\theta)/\rho_k\Big)}{\sum_{j=1}^{G_k} \pi_{0,j}^{(k)} \exp\Big(L_j^{(k)}(\theta)/\rho_k\Big)}.$$
Define $v_g(\theta) = \log \pi_{0,g}^{(k)} + L_g^{(k)}(\theta)/\rho_k$, so that $\pi^{(k)*}(\theta) = \text{softmax}(v(\theta))$.
The gradient of softmax is
$$\nabla \pi_g^{(k)*}(\theta) = \pi_g^{(k)*}(\theta) \left(\nabla v_g(\theta) - \sum_{j=1}^{G_k} \pi_j^{(k)*}(\theta) \nabla v_j(\theta) \right).$$
By Step 1, $\|\nabla v_g(\theta)\| = \|\nabla L_g^{(k)}(\theta)\|/\rho_k \le \mathcal{G}/\rho_k$. Therefore,
$$\|\nabla \pi_g^{(k)*}(\theta)\| \le \pi_g^{(k)*}(\theta) \left(\frac{\mathcal{G}}{\rho_k} + \sum_j \pi_j^{(k)*}(\theta) \frac{\mathcal{G}}{\rho_k} \right) = \frac{2\mathcal{G}}{\rho_k} \pi_g^{(k)*}(\theta).$$
Summing over g:
$$\|\nabla \pi^{(k)*}(\theta)\|_1 = \sum_{g=1}^{G_k} \|\nabla \pi_g^{(k)*}(\theta)\| \le \frac{2\mathcal{G}}{\rho_k} \sum_{g=1}^{G_k} \pi_g^{(k)*}(\theta) = \frac{2\mathcal{G}}{\rho_k}.$$
By the mean value theorem,
$$\|\pi^{(k)*}(\theta_1) - \pi^{(k)*}(\theta_2)\|_1 \le \frac{2\mathcal{G}}{\rho_k} \|\theta_1 - \theta_2\|. \tag{*}$$

Step 3: Smoothness of $L^{(k)}$. The gradient of $L^{(k)}$ is $\nabla L^{(k)}(\theta) = \sum_{g=1}^{G_k} \pi_g^{(k)*}(\theta) \nabla L_g^{(k)}(\theta)$. Then,
$$\begin{aligned}
&\|\nabla L^{(k)}(\theta_1) - \nabla L^{(k)}(\theta_2)\| \\
&= \left\| \sum_g \pi_g^{(k)*}(\theta_1) \nabla L_g^{(k)}(\theta_1) - \sum_g \pi_g^{(k)*}(\theta_2) \nabla L_g^{(k)}(\theta_2) \right\| \\
&\le \underbrace{\left\| \sum_g \pi_g^{(k)*}(\theta_1) \Big(\nabla L_g^{(k)}(\theta_1) - \nabla L_g^{(k)}(\theta_2) \Big) \right\|}_{A} \\
&\quad + \underbrace{\left\| \sum_g \Big(\pi_g^{(k)*}(\theta_1) - \pi_g^{(k)*}(\theta_2) \Big) \nabla L_g^{(k)}(\theta_2) \right\|}_{B}.
\end{aligned}$$

For term A, by Step 1 and $\sum_g \pi_g^{(k)*}(\theta_1) = 1$: $A \le \sum_g \pi_g^{(k)*}(\theta_1)\|\nabla L_g^{(k)}(\theta_1) - \nabla L_g^{(k)}(\theta_2)\| \le L_\ell\|\theta_1 - \theta_2\|$.

For term B: $B \le \sum_g |\pi_g^{(k)*}(\theta_1) - \pi_g^{(k)*}(\theta_2)| \cdot \|\nabla L_g^{(k)}(\theta_2)\| \le \frac{2\mathcal{G}}{\rho_k}\|\theta_1 - \theta_2\| \cdot \mathcal{G} = \frac{2\mathcal{G}^2}{\rho_k}\|\theta_1 - \theta_2\|$.

Combining A and B: $\|\nabla L^{(k)}(\theta_1) - \nabla L^{(k)}(\theta_2)\| \le \left(L_\ell + \frac{2\mathcal{G}^2}{\rho_k}\right)\|\theta_1 - \theta_2\|$.

Thus $L^{(k)}$ is L_k-smooth with $L_k = L_\ell + 2\mathcal{G}^2/\rho_k$.

Step 4: Smoothness of F. Since $\nabla_\theta F(\theta, w) = \sum_{k=1}^K w_k \nabla L^{(k)}(\theta)$,

$$\begin{aligned}\|\nabla_\theta F(\theta_1, w) - \nabla_\theta F(\theta_2, w)\| &\le \sum_{k=1}^K w_k \|\nabla L^{(k)}(\theta_1) - \nabla L^{(k)}(\theta_2)\| \\ &\le \sum_{k=1}^K w_k L_k \|\theta_1 - \theta_2\| \\ &\le \left(\max_{k\in[K]} L_k\right)\|\theta_1 - \theta_2\|,\end{aligned}$$

Therefore, F is L_F-smooth with $L_F = \max_{k\in[K]} L_k = \max_{k\in[K]}\left(L_\ell + \frac{2\mathcal{G}^2}{\rho_k}\right)$. □

Lemma B.7 (Lipschitz Continuity of F w.r.t. w)**.** *Under Assumption B.2, for any fixed θ and any $w_1, w_2 \in \Delta_K$:*

$$|F(\theta, w_1) - F(\theta, w_2)| \le M\|w_1 - w_2\|_1.$$

Proof. By the definition of F:

$$|F(\theta, w_1) - F(\theta, w_2)| = \left|\sum_{k=1}^K (w_{1,k} - w_{2,k})L^{(k)}(\theta)\right| \le M\sum_{k=1}^K |w_{1,k} - w_{2,k}| = M\|w_1 - w_2\|_1.$$

□

Lemma B.8 (Cumulative Change of Hedge Weights)**.** *Consider the Hedge update with decreasing step size $\eta_t = \eta_0/t$:*

$$w_{t+1,k} = \frac{w_{t,k}\exp(\eta_t \ell_{t,k})}{\sum_{j=1}^K w_{t,j}\exp(\eta_t \ell_{t,j})},$$

where $\ell_{t,k} = \hat{L}^{(k)}(\theta_{t+1}) \in [0, M]$ and $w_{1,k} = 1/K$. Assume $\eta_t \le 1/2$ for all t (which holds for $\eta_0 M \le 1/2$). Then the cumulative change of weights satisfies:

$$\sum_{t=1}^{T-1}\|w_{t+1} - w_t\|_1 \sum_{t=1}^{T-1}\|w_{t+1} - w_t\|_1 \le (2 + C_0)M\eta_0 \log T,$$

where C_0 is a constant.

Proof. The proof proceeds in four steps.

Step 1: Rewrite the Hedge update. Define the normalization factor $Z_t = \sum_{j=1}^K w_{t,j}\exp(\eta_t \ell_{t,j})$. Then $w_{t+1,k} = \frac{w_{t,k}\exp(\eta_t \ell_{t,k})}{Z_t}$. Hence

$$w_{t+1,k} - w_{t,k} = w_{t,k}\left(\frac{\exp(\eta_t \ell_{t,k})}{Z_t} - 1\right). \tag{16}$$

Step 2: Linearize the exponential. Since $\eta_t \ell_{t,k} \in [0, \eta_0 M]$ and η_0 is chosen such that $\eta_0 M \le 1/2$, we use the Taylor expansion:

$$\exp(\eta_t \ell_{t,k}) = 1 + \eta_t \ell_{t,k} + R_{t,k}, \qquad |R_{t,k}| \le \frac{1}{2}(\eta_t \ell_{t,k})^2 e^{\eta_t \ell_{t,k}} \le (\eta_t M)^2.$$

Similarly,

$$Z_t = 1 + \eta_t \bar{\ell}_t + \bar{R}_t, \qquad \bar{\ell}_t = \sum_{j=1}^{K} w_{t,j} \ell_{t,j}, \qquad |\bar{R}_t| \le (\eta_t M)^2.$$

Then

$$\frac{\exp(\eta_t \ell_{t,k})}{Z_t} = \frac{1 + \eta_t \ell_{t,k} + R_{t,k}}{1 + \eta_t \bar{\ell}_t + \bar{R}_t}.$$

Using $\frac{1+a}{1+b} = 1 + (a-b) + O(b^2)$ for small a, b:

$$\frac{\exp(\eta_t \ell_{t,k})}{Z_t} - 1 = \eta_t(\ell_{t,k} - \bar{\ell}_t) + O\big((\eta_t M)^2\big).$$

More precisely, since $|\ell_{t,k} - \bar{\ell}_t| \le 2M$,

$$\left| \frac{\exp(\eta_t \ell_{t,k})}{Z_t} - 1 \right| \le 2\eta_t M + 2C_0(\eta_t M)^2. \tag{17}$$

where C_0 is a constant.

Step 3: Bound the ℓ_1 distance. Using the above 16 and 17,

$$\begin{aligned} \|w_{t+1} - w_t\|_1 &= \sum_{k=1}^{K} w_{t,k} \left| \frac{\exp(\eta_t \ell_{t,k})}{Z_t} - 1 \right| \\ &\le \sum_{k=1}^{K} w_{t,k} (2\eta_t M + 2C_0(\eta_t M)^2) \\ &= 2\eta_t M + 2C_0(\eta_t M)^2. \\ &\le \eta_t M(2 + C_0) \end{aligned}$$

Step 4: Sum over t. Summing over $t = 1, \dots, T-1$, we obtain

$$\sum_{t=1}^{T-1} \|w_{t+1} - w_t\|_1 \le (2 + C_0) M \eta_0 \sum_{t=1}^{T-1} \frac{1}{t}.$$

Using $\sum_{t=1}^{T-1} \frac{1}{t} = O(\log T)$, it follows that

$$\sum_{t=1}^{T-1} \|w_{t+1} - w_t\|_1 \le (2 + C_0) M \eta_0 \log T.$$

This completes the proof. □

Lemma B.9 (Single-Step Descent)**.** *Under Lemma B.6 and $\mathbb{E}[\|\tilde{g}_t\|^2] \le \mathcal{G}^2$, the iterates satisfy:*

$$\mathbb{E}[F(\theta_{t+1}, w_t)] \le \mathbb{E}[F(\theta_t, w_t)] - \frac{\alpha_t}{2} \mathbb{E}[\|\nabla_\theta F(\theta_t, w_t)\|^2] + \frac{L_F \alpha_t^2}{2} \mathcal{G}^2.$$

Proof. By L_F-smoothness (Lemma B.6):

$$F(\theta_{t+1}, w_t) \le F(\theta_t, w_t) + \langle \nabla_\theta F(\theta_t, w_t), \theta_{t+1} - \theta_t \rangle + \frac{L_F}{2} \|\theta_{t+1} - \theta_t\|^2.$$

Substituting $\theta_{t+1} - \theta_t = -\alpha_t \tilde{g}_t$:

$$F(\theta_{t+1}, w_t) \le F(\theta_t, w_t) - \alpha_t \langle \nabla_\theta F(\theta_t, w_t), \tilde{g}_t \rangle + \frac{L_F \alpha_t^2}{2} \|\tilde{g}_t\|^2.$$

Taking conditional expectation and using $\mathbb{E}[\tilde{g}_t \mid \theta_t, w_t] = \nabla_\theta F(\theta_t, w_t)$:

$$\mathbb{E}[F(\theta_{t+1}, w_t) \mid \theta_t, w_t] \le F(\theta_t, w_t) - \alpha_t \|\nabla_\theta F(\theta_t, w_t)\|^2 + \frac{L_F \alpha_t^2}{2} \mathbb{E}[\|\tilde{g}_t\|^2 \mid \theta_t, w_t].$$

Using $\mathbb{E}[\|\tilde{g}_t\|^2] \leq \mathcal{G}^2$ and taking full expectation:

$$\begin{aligned}\mathbb{E}[F(\theta_{t+1}, w_t)] &\leq \mathbb{E}[F(\theta_t, w_t)] - \alpha_t \mathbb{E}[\|\nabla_\theta F(\theta_t, w_t)\|^2] + \frac{L_F \alpha_t^2}{2}\mathcal{G}^2 \\ &\leq \mathbb{E}[F(\theta_t, w_t)] - \frac{\alpha_t}{2}\mathbb{E}[\|\nabla_\theta F(\theta_t, w_t)\|^2] + \frac{L_F \alpha_t^2}{2}\mathcal{G}^2.\end{aligned}$$

□

Lemma B.10 (Per-Iteration Descent with Weight Drift)**.** *Combining Lemmas B.9, B.7:*

$$\mathbb{E}[F(\theta_{t+1}, w_{t+1})] \leq \mathbb{E}[F(\theta_t, w_t)] - \frac{\alpha_t}{2}\mathbb{E}[\|\nabla_\theta F(\theta_t, w_t)\|^2] + \frac{L_F \alpha_t^2}{2}\mathcal{G}^2 + M\mathbb{E}[\|w_{t+1} - w_t\|_1].$$

Proof. By Lemma B.7:

$$|F(\theta_{t+1}, w_{t+1}) - F(\theta_{t+1}, w_t)| \leq M\|w_{t+1} - w_t\|_1.$$

Taking expectation and substituting Lemma B.9 for $\mathbb{E}[F(\theta_{t+1}, w_t)]$ yields the result. □

We now state and prove the main convergence guarantee. The rate $O(\log T/\sqrt{T})$ differs from the standard $O(1/\sqrt{T})$ subgradient rate by an extra $\log T$ factor, arising from two independent sources: the harmonic series $\sum \alpha_t^2$ (gradient noise) and the cumulative Hedge weight drift (Lemma B.8).

Theorem B.11 (Convergence Rate)**.** *Let $\{(\theta_t, w_t)\}_{t=0}^{T-1}$ be the iterates generated by the algorithm, where $\theta_t \in \Theta$ denotes the model parameters and $w_t \in \Delta_K$ denotes the dimension weights at iteration t. Let $F(\theta, w) = \sum_{k=1}^K w_k L^{(k)}(\theta)$ be the objective, and let $F^* = \inf_{\theta\in\Theta, w\in\Delta_K} F(\theta, w)$. Set the model step size $\alpha_t = \alpha/\sqrt{t+1}$ with $\alpha > 0$, and the Hedge step size $\eta_t = \eta_0/t$ with $\eta_0 M \leq 1/2$, where M is the loss upper bound from Assumption B.2. Under Assumptions B.1, B.2, and B.3, the iterates satisfy:*

$$\min_{0\leq t<T} \mathbb{E}\big[\|\nabla_\theta F(\theta_t, w_t)\|^2\big] \leq \frac{C\log T}{\sqrt{T}},$$

where $C > 0$ is a constant depending on L_F (the smoothness constant from Lemma B.6), $\mathcal{G}$ (the gradient bound), M, α, η_0, and $\Delta_F = F(\theta_0, w_0) - F^*$.

Proof. We write $F_t = F(\theta_t, w_t)$ and $\nabla F_t = \nabla_\theta F(\theta_t, w_t)$ for brevity. The proof proceeds in three steps:

Step 1: Telescoping the per-iteration inequality. By Lemma B.10, each iteration satisfies

$$\frac{\alpha_t}{2}\,\mathbb{E}[\|\nabla F_t\|^2] \leq \underbrace{\mathbb{E}[F_t] - \mathbb{E}[F_{t+1}]}_{\text{descent (telescopes)}} + \underbrace{\frac{L_F\alpha_t^2}{2}\,\mathcal{G}^2}_{\text{gradient noise}} + \underbrace{M\,\mathbb{E}[\|w_{t+1} - w_t\|_1]}_{\text{weight movement}}.$$

Summing over $t = 0, \ldots, T-1$, the descent terms telescope ($\sum_t(\mathbb{E}[F_t] - \mathbb{E}[F_{t+1}]) = \mathbb{E}[F_0] - \mathbb{E}[F_T] \leq F(\theta_0, w_0) - F^*$), yielding

$$\sum_{t=0}^{T-1}\frac{\alpha_t}{2}\,\mathbb{E}[\|\nabla F_t\|^2] \leq \Delta_F + \frac{L_F\mathcal{G}^2}{2}\sum_{t=0}^{T-1}\alpha_t^2 + M\sum_{t=0}^{T-1}\mathbb{E}[\|w_{t+1} - w_t\|_1],$$

where $\Delta_F = F(\theta_0, w_0) - F^*$.

Step 2: Bound each term using the step sizes.

(a) Gradient noise. Since $\alpha_t = \alpha/\sqrt{t+1}$,

$$\sum_{t=0}^{T-1}\alpha_t^2 = \alpha^2\sum_{t=0}^{T-1}\frac{1}{t+1} = \alpha^2 H_T = O(\alpha^2\log T),$$

where H_T is the T-th harmonic number.

(b) Weight movement. By Lemma B.8 (which requires $\eta_0 M \le 1/2$, as assumed in the theorem):

$$\sum_{t=0}^{T-1} \mathbb{E}[\|w_{t+1} - w_t\|_1] \le (2 + C_0)\, M\eta_0 \log T.$$

(c) Gradient accumulation factor. By integral approximation ($\sum_{s=1}^{T} s^{-1/2} \ge \int_1^T x^{-1/2}\,dx = 2\sqrt{T} - 2$) (Bubeck, 2015):

$$\sum_{t=0}^{T-1} \alpha_t = \alpha \sum_{s=1}^{T} \frac{1}{\sqrt{s}} \ge 2\alpha\sqrt{T} - O(1).$$

Combining (a) and (b), the RHS of the Step 1 inequality is

$$\Delta_F + \underbrace{\frac{L_F \mathcal{G}^2 \alpha^2}{2} \log T}_{\text{from (a)}} + \underbrace{(2 + C_0) M^2 \eta_0 \log T}_{\text{from (b)}} = \Delta_F + C_1 \log T,$$

where $C_1 = \frac{L_F \mathcal{G}^2 \alpha^2}{2} + (2+C_0) M^2 \eta_0$. Note that both $\log T$ factors arise independently: the first from the harmonic series $\sum \alpha_t^2$ (gradient noise), the second from the Hedge bound (weight movement).

Step 3: Extract the minimum and rearrange. Since $\alpha_t > 0$, the weighted sum lower-bounds the minimum:

$$\sum_{t=0}^{T-1} \frac{\alpha_t}{2}\, \mathbb{E}[\|\nabla F_t\|^2] \;\ge\; \frac{1}{2} \left(\min_{0 \le t < T} \mathbb{E}[\|\nabla F_t\|^2] \right) \sum_{t=0}^{T-1} \alpha_t.$$

Combining with Step 1–2:

$$\min_{0 \le t < T} \mathbb{E}[\|\nabla F_t\|^2] \;\le\; \frac{2\big(\Delta_F + C_1 \log T\big)}{\sum_{t=0}^{T-1} \alpha_t} \;\le\; \frac{2\big(\Delta_F + C_1 \log T\big)}{2\alpha\sqrt{T} - O(1)}.$$

For T sufficiently large, $2\alpha\sqrt{T} - O(1) \ge \alpha\sqrt{T}$, so

$$\min_{0 \le t < T} \mathbb{E}\big[\|\nabla_\theta F(\theta_t, w_t)\|^2\big] \;\le\; \frac{2\,\Delta_F}{\alpha\sqrt{T}} + \frac{2\,C_1 \log T}{\alpha\sqrt{T}} \;=\; O\!\left(\frac{\log T}{\sqrt{T}}\right).$$

The first term $O(1/\sqrt{T})$ comes from the initial gap Δ_F; the second term $O(\log T/\sqrt{T})$ comes from the two accumulated error sources in Step 2 and dominates the rate. □

B.5 Multi-Dimensional Generalization Bound

To establish generalization, we need to relate the population DRO loss $L^{(k)}(\theta)$ to its empirical counterpart $\hat{L}^{(k)}(\theta)$. The key tool is that the log-sum-exp function is 1-Lipschitz in the ℓ_∞ norm, which transfers per-group concentration bounds to the aggregated loss.

Lemma B.12 (Lipschitz Property of Log-Sum-Exp)**.** *Define $\phi(x) = \rho \log \sum_{g=1}^{G} \pi_{0,g} e^{x_g/\rho}$ for $x \in \mathbb{R}^G$. Then ϕ is 1-Lipschitz with respect to the ℓ_∞ norm:*

$$|\phi(x) - \phi(y)| \le \|x - y\|_\infty = \max_g |x_g - y_g|.$$

Proof. The partial derivative is

$$\frac{\partial \phi}{\partial x_g} = \frac{\pi_{0,g} e^{x_g/\rho}}{\sum_{j=1}^{G} \pi_{0,j} e^{x_j/\rho}} = \pi_g^*(x),$$

which is a probability distribution over g. Hence $\|\nabla\phi\|_1 = \sum_g \pi_g^* = 1$. By the mean value theorem and Hölder's inequality (ℓ_1 and ℓ_∞ are dual norms),

$$|\phi(x) - \phi(y)| \le \sup_\xi \|\nabla\phi(\xi)\|_1 \cdot \|x - y\|_\infty = \|x - y\|_\infty.$$

□

Corollary B.13. *For any θ,*

$$|L^{(k)}(\theta) - \hat{L}^{(k)}(\theta)| \leq \max_{g \in [G_k]} |L_g^{(k)}(\theta) - \hat{L}_g^{(k)}(\theta)|.$$

The Hedge algorithm produces dimension weights w_T that approximate a softmax of the per-dimension losses. The following standard lemma bounds the gap between the maximum and the softmax-weighted average, which we use to relate $\max_k \hat{L}^{(k)}(\theta_T)$ to $\hat{F}(\theta_T) = \sum_k w_{T,k} \hat{L}^{(k)}(\theta_T)$.

Lemma B.14 (Softmax Approximation of Maximum)**.** *For any $a = (a_1, \ldots, a_K) \in \mathbb{R}^K$ and $\tau > 0$, define the softmax weights: $w_k = \frac{e^{\tau a_k}}{\sum_{j=1}^K e^{\tau a_j}}, \quad k = 1, \ldots, K$. Then the following inequality holds:*

$$\max_{k \in [K]} a_k \;\leq\; \sum_{k=1}^K w_k a_k \;+\; \frac{\log K}{\tau}.$$

Proof. For each k, we have $e^{\tau a_k} \leq \sum_{j=1}^K e^{\tau a_j}$. Taking logarithms yields: $\tau a_k \leq \log\left(\sum_{j=1}^K e^{\tau a_j}\right)$. Since this holds for all k, we can take the maximum over k:

$$\tau \max_k a_k \leq \log\left(\sum_{j=1}^K e^{\tau a_j}\right) \quad \Longrightarrow \quad \max_k a_k \leq \frac{1}{\tau} \log\left(\sum_{j=1}^K e^{\tau a_j}\right). \tag{1}$$

Now consider the right-hand side. Using the definition of w_k, we rewrite:

$$\log\left(\sum_{j=1}^K e^{\tau a_j}\right) = \log\left(\sum_{j=1}^K \frac{w_j}{w_j} e^{\tau a_j}\right) = \log\left(\sum_{j=1}^K w_j \cdot \frac{e^{\tau a_j}}{w_j}\right).$$

By Jensen's inequality (Boyd & Vandenberghe, 2004) (since $\log$ is concave):

$$\log\left(\sum_{j=1}^K w_j \cdot \frac{e^{\tau a_j}}{w_j}\right) \geq \sum_{j=1}^K w_j \log\left(\frac{e^{\tau a_j}}{w_j}\right) = \tau \sum_{j=1}^K w_j a_j - \sum_{j=1}^K w_j \log w_j.$$

Thus:

$$\frac{1}{\tau} \log\left(\sum_{j=1}^K e^{\tau a_j}\right) \geq \sum_{j=1}^K w_j a_j - \frac{1}{\tau} \sum_{j=1}^K w_j \log w_j. \tag{2}$$

The negative entropy term satisfies $-\sum_{j=1}^K w_j \log w_j = H(w) \leq \log K$, where $H(w)$ is the Shannon entropy. Therefore:

$$-\frac{1}{\tau} \sum_{j=1}^K w_j \log w_j \leq \frac{\log K}{\tau}.$$

Combining (1) and (2):

$$\max_k a_k \leq \sum_{j=1}^K w_j a_j - \frac{1}{\tau} \sum_{j=1}^K w_j \log w_j \leq \sum_{j=1}^K w_j a_j + \frac{\log K}{\tau}.$$

This completes the proof. □

Lemma B.15 (Single-Dimension Generalization Bound)**.** *For a fixed dimension k, with probability at least $1 - \delta$:*

$$\max_{g \in [G_k]} \mathbb{E}_{P_g^{test}}[\ell(\theta)] \leq \hat{L}^{(k)}(\theta) + M \sqrt{\frac{\log(2G_k/\delta)}{2n_{\min}^{(k)}}}.$$

Proof. According to Lipschitz Property of Log-Sum-Exp B.13, $L^{(k)}(\theta)$ is 1-Lipschitz with respect to the ℓ_∞ norm. Then

$$|L^{(k)}(\theta) - \hat{L}^{(k)}(\theta)| \leq \max_{g \in [G_k]} |L_g^{(k)}(\theta) - \hat{L}_g^{(k)}(\theta)|.$$

For each group g, by Hoeffding's inequality:

$$\mathbb{P}\left(|L_g^{(k)}(\theta) - \hat{L}_g^{(k)}(\theta)| \geq t\right) \leq 2e^{-2n_g^{(k)}t^2/M^2}.$$

Set $t = M\sqrt{\frac{\log(2G_k/\delta)}{2n_g^{(k)}}}$. Then with probability $1 - \delta/G_k$:

$$L_g^{(k)}(\theta) \leq \hat{L}_g^{(k)}(\theta) + M\sqrt{\frac{\log(2G_k/\delta)}{2n_g^{(k)}}}.$$

Taking the union bound over all $g \in [G_k]$, with probability at least $1 - \delta$:

$$\max_g L_g^{(k)}(\theta) \leq \max_g \hat{L}_g^{(k)}(\theta) + M \max_g \sqrt{\frac{\log(2G_k/\delta)}{2n_g^{(k)}}}.$$

Since $n_g^{(k)} \geq n_{\min}^{(k)}$, we have $\frac{1}{\sqrt{n_g^{(k)}}} \leq \frac{1}{\sqrt{n_{\min}^{(k)}}}$. Moreover, by the upper bound property of log-sum-exp:

$$\hat{L}^{(k)}(\theta) = \rho_k \log \sum_g \pi_{0,g} e^{\hat{L}_g^{(k)}/\rho_k} \geq \max_g \hat{L}_g^{(k)}(\theta).$$

Therefore:

$$\max_g L_g^{(k)}(\theta) \leq \hat{L}^{(k)}(\theta) + M\sqrt{\frac{\log(2G_k/\delta)}{2n_{\min}^{(k)}}}.$$

Note that $\max_g L_g^{(k)}(\theta) = \max_g \mathbb{E}_{P_g^{\text{test}}}[\ell(\theta)]$ is exactly the worst-group test loss for dimension k. This completes the proof. □

We now combine the single-dimension concentration bound (Lemma B.15) with the Hedge weight structure to obtain the main generalization guarantee for the multi-dimensional objective.

Theorem B.16 (Main Generalization Bound)**.** *Let θ_T be the final model returned by the algorithm after T iterations, where T is the total number of training iterations and $\eta > 0$ is the learning rate of the Hedge update. Let $w_{T,k}$ denote the Hedge weight for dimension k at iteration T, and let $\hat{F}(\theta_T) = \sum_{k=1}^K w_{T,k}\hat{L}^{(k)}(\theta_T)$ be the empirical weighted objective. Define $\tau = \eta log T$. Then, with probability at least $1 - \delta$ over the training data,*

$$\max_{k\in[K]} \max_{g\in[G_k]} \mathbb{E}_{P_g^{test}}[\ell(\theta_T)] \;\leq\; \hat{F}(\theta_T) + \frac{\log K}{\tau} + M \max_{k\in[K]} \sqrt{\frac{\log(2KG_k/\delta)}{2n_{\min}^{(k)}}},$$

Proof. **Step 1: Apply the single-dimension bound to each dimension.** By Lemma B.15, for each k with probability at least $1 - \delta/K$,

$$\max_{g\in[G_k]} \mathbb{E}_{P_g^{\text{test}}}[\ell(\theta_T)] \leq \hat{L}^{(k)}(\theta_T) + M\sqrt{\frac{\log(2KG_k/\delta)}{2n_{\min}^{(k)}}}.$$

Taking the union bound over all $k \in [K]$, with probability at least $1 - \delta$,

$$\max_{k\in[K]} \max_{g\in[G_k]} \mathbb{E}_{P_g^{\text{test}}}[\ell(\theta_T)] \leq \max_{k\in[K]} \hat{L}^{(k)}(\theta_T) + M \max_{k\in[K]} \sqrt{\frac{\log(2KG_k/\delta)}{2n_{\min}^{(k)}}}. \tag{1}$$

Step 2: Express the Hedge weights in softmax form. The Hedge update is

$$w_{t+1,k} = \frac{w_{t,k} \exp\left(\eta_t \hat{L}^{(k)}(\theta_{t+1})\right)}{\sum_{j=1}^K w_{t,j} \exp\left(\eta_t \hat{L}^{(j)}(\theta_{t+1})\right)},$$

where $\eta_t = \eta_0/t$. With uniform initialization $w_{1,k} = 1/K$, expanding the recursion gives

$$w_{T,k} = \frac{\exp\Big(\sum_{t=1}^{T-1} \eta_t \hat{L}^{(k)}(\theta_{t+1})\Big)}{\sum_{j=1}^{K} \exp\Big(\sum_{t=1}^{T-1} \eta_t \hat{L}^{(j)}(\theta_{t+1})\Big)}.$$

By the convergence analysis (Theorem B.11) and the Assumption B.4, the per-dimension losses stabilize, so for sufficiently large T,

$$\sum_{t=1}^{T-1} \eta_t \hat{L}^{(k)}(\theta_{t+1}) \approx \left(\sum_{t=1}^{T-1} \eta_t\right) \hat{L}^{(k)}(\theta_T) = \eta_0 H_{T-1} \hat{L}^{(k)}(\theta_T),$$

where $H_{T-1} = \sum_{t=1}^{T-1} 1/t \approx logT$ is the $(T-1)$-th harmonic number. Substituting this into the expression for $w_{T,k}$ yields the softmax form

$$w_{T,k} \approx \frac{\exp\Big(\tau \hat{L}^{(k)}(\theta_T)\Big)}{\sum_{j=1}^{K} \exp\Big(\tau \hat{L}^{(j)}(\theta_T)\Big)}, \qquad \tau = \eta_0 logT.$$

Step 3: Relate the maximum to the weighted sum. Since $w_{T,k}$ is approximately the softmax of $\hat{L}^{(k)}(\theta_T)$ with temperature τ, Lemma B.14 applies with $a_k = \hat{L}^{(k)}(\theta_T)$:

$$\max_{k\in[K]} \hat{L}^{(k)}(\theta_T) \le \sum_{k=1}^{K} w_{T,k} \hat{L}^{(k)}(\theta_T) + \frac{\log K}{\tau} = \hat{F}(\theta_T) + \frac{\log K}{\tau}. \tag{3}$$

Step 4: Combine. Substituting (2) and (3) into (1) yields

$$\max_{k\in[K]} \max_{g\in[G_k]} \mathbb{E}_{P_g^{\text{test}}}[\ell(\theta_T)] \le \hat{F}(\theta_T) + \frac{\log K}{\tau} + M \max_{k\in[K]} \sqrt{\frac{\log(2KG_k/\delta)}{2n_{\min}^{(k)}}}.$$

This completes the proof. □

C Implementation Details

C.1 Datasets and Preprocessing

OmniMedVQA. We use the OmniMedVQA benchmark (Hu et al., 2024), which contains medical VQA samples annotated along three grouping dimensions: imaging modality (8 groups: CT, MRI, Ultrasound, X-ray, OCT, Dermoscopy, Fundus Photography, Microscopy), anatomical region (5 groups: brain, breast, chest, heart, lungs), and clinical task (5 groups: disease diagnosis, modality recognition, anatomy identification, lesion grading, biological attribute analysis).

VQAv2. We use VQAv2 (Goyal et al., 2017), which contains natural-image VQA samples with original annotations along question type, answer type, and visual content. We apply the dimension selection procedure (Section A) to these annotations and retain two dimensions: question type and visual class.

BDD100K. We use BDD100K (Yu et al., 2020) for object detection, with grouping dimensions along weather (6 groups: clear, cloudy, overcast, rainy, snowy, foggy), scene type (4 groups: highway, city street, residential, parking lot), and time of day (3 groups: daytime, dawn/dusk, night).

For OmniMedVQA and VQAv2, we fine-tune Qwen2.5-VL-2B for 2 epochs with batch size 32 using AdamW (Loshchilov & Hutter, 2019) and a learning rate of 1×10^{-5}. For BDD100K, we train YOLOv5-s (Jocher, 2021) for 50 epochs with batch size 32 using AdamW and a learning rate of 1×10^{-5}. All experiments are run on 8 RTX 4090 GPUs. Results are averaged over three random seeds. Table 4 summarizes the hyperparameter values used in all experiments. The dependency threshold t is tuned on a validation set, and the remaining hyperparameters are selected by grid search.

Table 4: Hyperparameter values used in all experiments.

Hyperparameter	**Value**
Hedge learning rate η_0	0.1
KL regularization parameter ρ_k	0.05
Weight lower bound β	0.05
Dependency threshold t	0.7
Model learning rate	1×10^{-5}
Batch size	32
Training epochs	2 / 50
Optimizer	AdamW
Random seeds	3

Table 5: Ablation results on OmniMedVQA. We report worst-group and average accuracy (%).

	Unshift		**Shift**	
Configuration	Worst	Average	Worst	Average
Dimension-wise DRO (full)	**28.4**	**85.3**	**20.4**	**79.7**
w/o dimension selection	24.1	83.6	15.8	76.2
w/o KL regularization	26.3	84.5	18.1	78.1
w/o Hedge (uniform)	27.5	84.9	19.2	78.9

D Additional Experimental Results

D.1 Full Ablation Results on OmniMedVQA and BDD100K

The main text reports ablation results on VQAv2. Here we extend the ablation to OmniMedVQA to verify that the contribution of each component is consistent across datasets. Tables 5 report worst-group and average performance under both unshift and shift settings. Removing any component degrades worst-group performance on both datasets, and the pattern is consistent with the results on VQAv2 in the main text.

D.2 Effect of Cartesian-Product Grouping

We compare using the Cartesian product of all dimensions as the grouping structure versus using the best single dimension. Table 6 reports worst-group accuracy on OmnimedVQA under the unshift setting.

Cartesian-product grouping produces many groups with very few samples each, leading to noisy group loss estimates and unstable worst-group optimization. The best single dimension avoids this issue but only uses one semantic axis. Dimension-wise DRO uses all dimensions without constructing combinatorial groups.

D.3 Sensitivity to ρ_k and η_0

We analyze the sensitivity of Dimension-wise DRO to two key hyperparameters: the KL regularization parameter ρ_k and the Hedge learning rate η_0. Figure 3 shows worst-group accuracy on VQAv2 (shift setting) as a function of each.

When ρ_k is very small (0.001), the KL constraint is weak and the result approaches the "w/o KL regularization" ablation, confirming that the e gularization term is inactive in this regime. Performance improves as ρ_k increases, peaks at $\rho_k = 0.05$, and then degrades when ρ_k becomes too large (1.0), where the group weights remain close to uniform and the robustness benefit diminishes. Performance is stable for $\rho_k \in [0.01, 0.1]$.

When η_0 is very small (0.01), the dimension weights barely change from their initial uniform values, and the result approaches the "w/o Hedge" ablation, confirming that the adaptive weighting is inactive

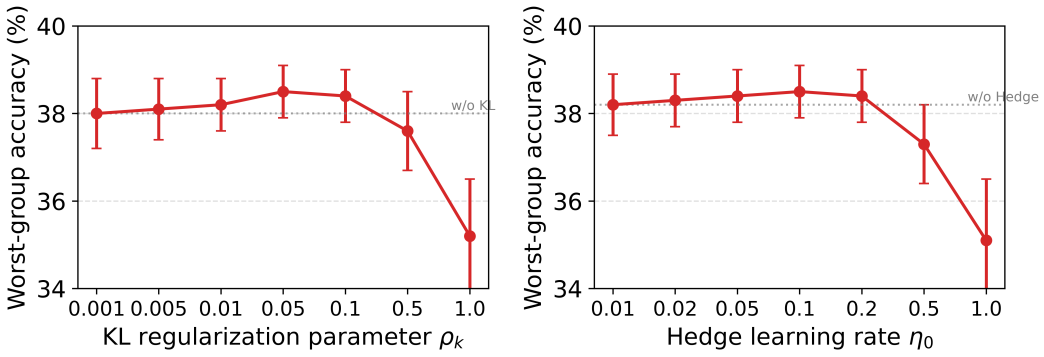

Figure 3: Sensitivity analysis on VQAv2 (shift setting). Left: KL regularization parameter ρ_k. Right: Hedge learning rate η_0. Each point is the mean of 3 runs; shaded regions show standard deviation.

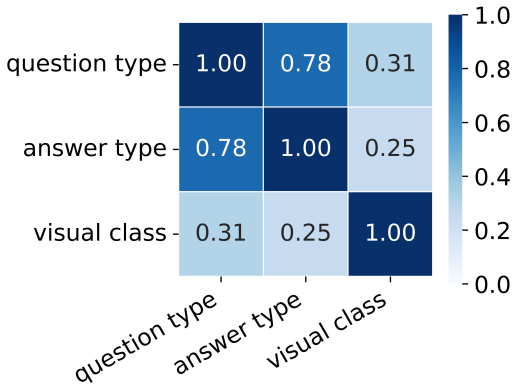

Figure 4: Cramér's V matrix for the original annotations in VQAv2. Darker cells indicate stronger dependence. The MIS procedure with threshold $t = 0.7$ removes answer type, which is highly correlated with question type.

in this regime. Performance peaks at $\eta_0 = 0.1$ and remains stable for $\eta_0 \in [0.05, 0.2]$. When η_0 is too large (1.0), the weights change abruptly between iterations, leading to unstable optimization and a clear drop in worst-group accuracy.

D.4 Cramér's V Heatmap for VQAv2

Figure 4 shows the Cramér's V matrix for the original annotations in VQAv2. Question type and answer type are highly correlated ($V = 0.78$), while visual class has low pairwise dependence with both ($V = 0.12$ and $V = 0.09$). With the dependency threshold $t = 0.7$, the MIS procedure removes answer type and retains question type and visual class.

Table 6: Worst-group accuracy (%) on OmnimedVQA (unshift) with Cartesian-product grouping versus single-dimension grouping.

Grouping structure	**Worst Acc**	**Average Acc**
Cartesian product (all dimensions)	14.1	75.3
Best single dimension	24.9	84.2