RECOVERING SILENCED GRADIENTS: NOISE-ROBUST PROMPT TUNING FOR VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Prompt tuning efficiently adapts vision-language models by learning a small set of parameters while keeping the backbone frozen. However, existing noise-robust methods suppress useful learning signals along with corrupted supervision. Through theoretical and experimental analyses, we identify a failure mode termed gradient silencing. Correctly labeled samples with low prediction confidence, which we call hard positives, can be misclassified as noisy. Mean absolute error (MAE) then attenuates their corrective gradients in proportion to that confidence. To address this limitation, we propose Geometric Gradient Rectification Prompting (GGRP), which couples geometric target rectification with silenced gradient recovery. The key idea is to use image–text similarities to determine both training targets and sample contributions, rather than only separating clean and noisy samples. Specifically, temporal momentum guidance stabilizes textual class prototypes. We then use geometric target rectification to construct training targets via optimal transport (OT) alignment. Finally, silenced gradient recovery uses the OT assignment margin to balance cross-entropy on rectified targets with MAE on observed labels. This allows hard positives to receive corrective gradients even at low prediction confidence when OT sufficiently supports their correct labels, while retaining MAE-based attenuation for ambiguous samples. Extensive experiments on seven datasets with synthetic label noise and Food101N with naturally noisy labels show that GGRP surpasses state-of-the-art methods across a range of noise settings.

1 INTRODUCTION

Vision-Language Models (VLMs) such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) align images and text in a shared embedding space through large-scale contrastive pre-training, and they show strong zero-shot recognition across many downstream datasets. Adapting these models through full fine-tuning is computationally expensive and may degrade the generalization ability of the pre-trained backbone. Prompt tuning (Zhou et al., 2022b; Khattak et al., 2023; Yao et al., 2023; Li et al., 2025; Guo et al., 2025) has emerged as a compelling alternative. Instead of updating the entire model, it learns a small number of prompt tokens for the frozen encoders, preserving transferability while reducing storage overhead.

Despite this progress, noisy labels in real-world data can bias the prompts and weaken image–text alignment (Hua & Qian, 2025; Niu & Xue, 2026). With adaptation restricted to a small set of parameters, gradients from mislabeled samples can interfere with those from correctly labeled samples. Preserving informative learning signals under noisy supervision remains a major challenge for adapting pre-trained VLMs.

To improve robustness to label noise, recent methods combine sample selection with noise-tolerant losses (Pan et al., 2025; Niu & Xue, 2026; Guo & Gu, 2024). A representative strategy partitions samples into clean and noisy subsets according to whether optimal transport assignments agree with the observed labels. It then applies mean absolute error (MAE) to the noisy subset. However, this hard partition and the associated loss jointly induce a failure mode we term *gradient silencing*. Hard positives—correctly labeled samples with low prediction confidence—are systematically misclassified as noisy early in training. Once these samples are assigned to the noisy subset, MAE scales their cross-entropy gradients by a factor proportional to confidence in the observed label, causing their

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.

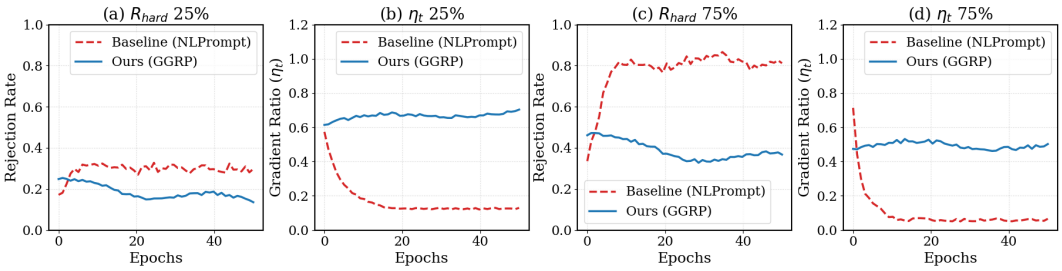


Figure 1: Gradient contribution tracking on Caltech101 with 25% and 75% symmetric label noise (curves lightly smoothed for readability). The hard rejection rate is denoted by R_{hard}, and the gradient norm ratio is represented by η_t.

corrective gradients to vanish as that confidence approaches zero. Consequently, easy samples dominate the prompt updates, leaving hard positives poorly learned. We examine this behavior in Fig. 1 by tracking hard-positive rejection and gradient contributions during training. We provide a detailed theoretical and experimental analysis in Section 3. Overall, existing prompt tuning methods still struggle to suppress label noise while preserving informative learning signals from hard positives.

Motivated by these limitations, we propose Geometric Gradient Rectification Prompting (GGRP), a framework that restores learning signals for hard positives while retaining robustness to label noise. Our key idea is to use similarities between image features and textual class prototypes to determine both the training targets and how strongly each sample contributes to learning, rather than only separating clean and noisy samples. Specifically, we first use temporal momentum guidance to stabilize the textual class prototypes used for target construction. Second, we introduce geometric target rectification via optimal transport alignment to jointly construct training targets within each batch from image–text similarities, retaining the observed-label classes as candidates. Finally, we implement silenced gradient recovery by using the gap between the two leading OT assignment scores to balance cross-entropy on rectified targets and MAE on observed labels. When the assignments provide sufficient support for their correct labels, hard positives can receive corrective gradients even when their prediction confidence is low. Overall, GGRP couples geometric target rectification with silenced gradient recovery while retaining MAE-based attenuation for ambiguous samples.

We evaluate GGRP on seven image-classification datasets under symmetric and asymmetric label noise, with corruption rates ranging from 12.5% to 75%, and on Food101N with naturally noisy labels. GGRP achieves competitive performance across these benchmarks, with gains over NLPrompt under severe label noise. Under 75% symmetric and asymmetric noise, GGRP improves the average accuracy over NLPrompt by 3.24 and 6.14 percentage points, respectively. We further conduct component ablations and sensitivity analyses to examine the roles of geometric target rectification, silenced gradient recovery, and their associated design choices.

Our contributions are summarized as follows:

- We identify and analyze gradient silencing, showing how hard sample partitioning and confidence-dependent gradient attenuation can suppress learning from hard positives.
- We propose GGRP, which combines temporal momentum guidance, geometric target rectification, and silenced gradient recovery to selectively recover corrective gradients under noisy supervision.
- We conduct extensive experiments on synthetic and real-world noisy-label benchmarks, demonstrating competitive performance and improvements over NLPrompt under severe label noise.

2 Related Work

Prompt tuning for VLMs. Parameter-efficient fine-tuning has become a standard strategy for adapting foundation models. In VLMs, prompt tuning methods such as CoOp (Zhou et al., 2022b) and MaPLe (Khattak et al., 2023) optimize compact prompt vectors while freezing the pre-trained

encoders, enabling efficient transfer to few-shot recognition (Gao et al., 2021; Zhang et al., 2022; Zhou et al., 2022a; Khattak et al., 2023), detection (Rasheed et al., 2022; Maaz et al., 2022), and related tasks. However, restricting adaptation to a small set of parameters limits the capacity to compensate for corrupted supervision, making prompt tuning brittle under label noise.

Robust learning with noisy labels. Learning with noisy labels (LNL) methods often combine sample selection with robust objectives. Co-teaching (Han et al., 2018) selects small-loss samples, whereas NLPrompt (Pan et al., 2025) partitions training data according to agreement between optimal transport assignments and observed labels, applying CE to the clean subset and MAE to the noisy subset. Such hard partitioning can route correctly labeled hard samples to the noisy subset, where MAE attenuates their gradients in proportion to prediction confidence. GGRP addresses this limitation through temporal momentum guidance, geometric target rectification, and silenced gradient recovery, selectively restoring CE gradients for geometrically supported samples while retaining MAE on ambiguous samples.

3 Analysis of Gradient Silencing

Background. Prompt tuning optimizes a continuous soft prompt $\boldsymbol{\theta}$ while keeping the vision-language model (VLM) frozen. Recent robust methods, such as NLPrompt, mitigate label noise by explicitly partitioning the training data into clean and noisy subsets. Specifically, NLPrompt uses agreement between optimal transport assignments and observed labels to partition the training data $\mathcal{D}$ into two mutually exclusive subsets: a set $\mathcal{D}_{clean}$ optimized with cross-entropy (CE) loss, and a set $\mathcal{D}_{noisy}$ constrained by robust losses such as mean absolute error (MAE). However, agreement-based partitioning can mistake correctly labeled hard samples for noise early in training; we analyze the gradient flow to demonstrate that this explicit dichotomy, when combined with the limited parameter space of prompt tuning, can suppress the optimization path for hard positive samples.

3.1 Theoretical Analysis

Gradient Derivation. Let $s_{\tilde{y}}$ denote the prediction probability for the target class $\tilde{y}$. For samples classified as noisy by the selection mechanism, the model minimizes the MAE loss $\mathcal{L}_{\text{MAE}} = 2(1-s_{\tilde{y}})$. We derive the gradient with respect to the learnable prompt $\boldsymbol{\theta}$:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{MAE}} &= -2\sum_{k=1}^{K}\frac{\partial s_{\tilde{y}}}{\partial z_k}\nabla_{\boldsymbol{\theta}} z_k \\ &= -2s_{\tilde{y}}\left(\nabla_{\boldsymbol{\theta}} z_{\tilde{y}} - \sum_{k=1}^{K} s_k \nabla_{\boldsymbol{\theta}} z_k\right).\end{aligned} \tag{1}$$

The prompt affects all logits, so Eq. (1) includes terms from both the target class and the other classes. All of these terms share the factor $s_{\tilde{y}}$, which scales the entire gradient. Assuming the logit Jacobians are bounded, i.e., $\|\nabla_{\boldsymbol{\theta}} z_k\|_2 \le M$ for all classes, which is natural for normalized CLIP features and finite temperature, Eq. (1) gives

$$\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{MAE}}\|_2 \le 2s_{\tilde{y}}\left(\|\nabla_{\boldsymbol{\theta}} z_{\tilde{y}}\|_2 + \sum_{k=1}^{K} s_k\|\nabla_{\boldsymbol{\theta}} z_k\|_2\right) \le 4Ms_{\tilde{y}}. \tag{2}$$

Extension to GCE-family Robust Losses. The same low-confidence silencing effect is not unique to MAE. For generalized cross-entropy (GCE) (Zhang & Sabuncu, 2018), defined as $\mathcal{L}_{\text{GCE}} = (1-s_{\tilde{y}}^q)/q$ with $q \in (0,1]$, the gradient with respect to logits is

$$\nabla_{\mathbf{z}}\mathcal{L}_{\text{GCE}} = s_{\tilde{y}}^q(\mathbf{s} - \mathbf{e}_{\tilde{y}}). \tag{3}$$

Thus, whenever $q > 0$, the gradient magnitude is still modulated by $s_{\tilde{y}}^q$ and vanishes as $s_{\tilde{y}} \to 0$. MAE corresponds to the $q = 1$ case up to a constant factor, whereas CE appears as the $q \to 0$ limiting case and preserves a non-vanishing corrective gradient. Thus, gradient silencing also affects other bounded robust losses applied to low-confidence samples (Appendix B.1).

Gradient Silencing of Hard Positives. The fundamental failure stems from the misalignment between rigid binary partitioning and the robust loss landscape. Specifically, hard positives (e.g., atypical views), which inherently exhibit low prediction confidence, can be misclassified into the $\mathcal{D}_{noisy}$ subset by the selection mechanism. Once relegated to this subset, these samples encounter gradient silencing under the MAE loss. Combining Eq. (2) with the low-confidence condition yields

$$\lim_{s_{\tilde{y}} \to 0} \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{MAE}}\|_2 = 0. \tag{4}$$

Unlike the cross-entropy loss, which produces a large corrective signal for low-confidence targets, bounded robust losses such as MAE and GCE attenuate the gradients from these data points in the low-confidence regime. This gradient suppression is amplified by the intrinsic nature of prompt tuning, which optimizes a compact prompt $\boldsymbol{\theta}$ within a frozen feature space. When the pulling force from hard positives is strongly attenuated, the optimization becomes dominated by easy samples in $\mathcal{D}_{clean}$ and the learned prompt is biased toward easy-sample prototypes. As this bias grows, hard positives remain far from the learned decision region and continue to be selected as noisy in subsequent epochs. It is worth noting that MAE also attenuates gradients when $s_{\tilde{y}} \to 1$, which corresponds to already well-fitted samples and is therefore expected rather than harmful. The failure mode we focus on is specifically the low-confidence regime: hard positives mistakenly routed to $\mathcal{D}_{noisy}$ require strong corrective updates, yet MAE suppresses precisely this restorative signal when $s_{\tilde{y}} \to 0$.

3.2 Experimental Analysis

Analysis Method. To empirically verify this theoretical deduction, we design a Gradient Contribution Tracking experiment on the Caltech101 dataset. We conduct experiments under 25% and 75% symmetric noise rates. The clean ground-truth labels are used only for this diagnostic analysis, not for training. At epoch t, each clean example is ranked by its clean-label confidence $c_{i,t} = p_t(y_i \mid \mathbf{x}_i)$ under the online prompt. The bottom 20% form the Hard Clean set $\mathcal{H}_t$, and the remaining clean examples form the Non-hard Clean set $\mathcal{E}_t$. During the training of the baseline algorithm, we monitor two metrics. The first, the Hard Selection Rejection Rate (R_{hard}), measures the proportion of Hard Clean samples routed into $\mathcal{D}_{noisy}$:

$$R_{hard,t} = \frac{|\{i \in \mathcal{H}_t : i \in \mathcal{D}_{noisy,t}\}|}{|\mathcal{H}_t|}. \tag{5}$$

Second, the Effective Gradient Norm Ratio (η_t) quantifies the relative per-sample contribution of hard samples compared with non-hard samples:

$$\eta_t = \frac{\frac{1}{|\mathcal{H}_t|} \sum_{i \in \mathcal{H}_t} \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_i\|_2}{\frac{1}{|\mathcal{E}_t|} \sum_{i \in \mathcal{E}_t} \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_i\|_2 + \varepsilon_g}, \tag{6}$$

where ε_g is a small constant used only for numerical stability.

Analysis results. As shown in Fig. 1, under 75% symmetric noise, the baseline rejection rate (R_{hard}) rises to approximately 80%, while the gradient norm ratio (η_t) falls below 0.1. Under 25% noise, rejection is less frequent and η_t remains above 0.1, although hard clean samples still contribute smaller per-sample gradient norms than non-hard clean samples. These results indicate stronger gradient attenuation for hard clean samples under the higher noise rate.

4 Methodology

We propose Geometric Gradient Rectification Prompting (GGRP), a framework designed to mitigate the gradient silencing of hard positives inherent in noisy-label prompt tuning. Fig. 2 summarizes the framework. We first formulate the problem setting, then introduce temporal momentum guidance, geometric target rectification, and silenced gradient recovery. Algorithm 1 gives the full training procedure (Appendix A).

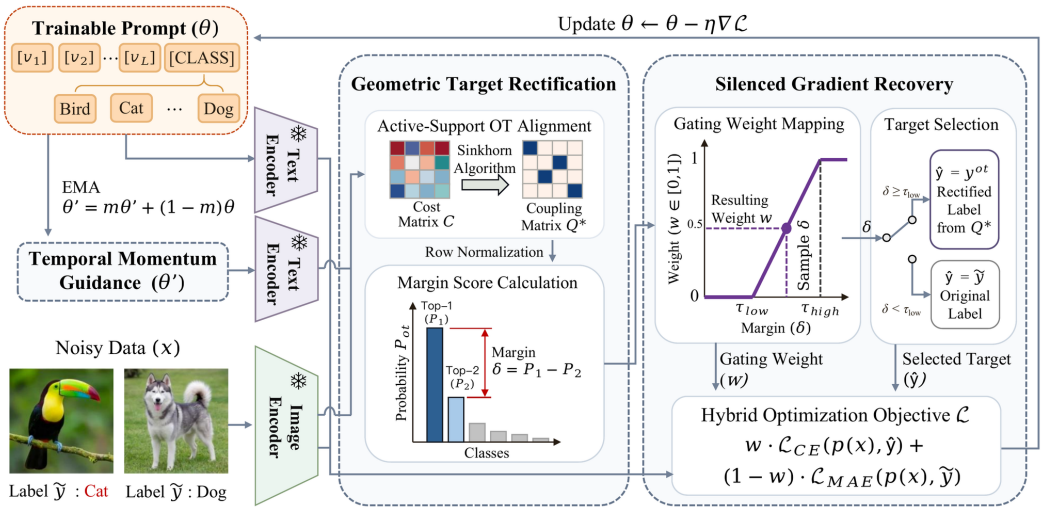


Figure 2: Overview of our proposed Geometric Gradient Rectification Prompting (GGRP).

4.1 Preliminaries

We consider vision-language prompt tuning under noisy supervision. Let $\mathcal{D} = \{(\mathbf{x}_i, \tilde{y}_i)\}_{i=1}^N$ be a training dataset where $\tilde{y}_i \in \{1, \ldots, K\}$ denotes the potentially corrupted label. We utilize a vision-language model (e.g., CLIP) with a frozen image encoder Φ_I and text encoder Φ_T. We optimize a set of continuous learnable vectors $\boldsymbol{\theta} \in \mathbb{R}^{L \times D}$ (the soft prompt) while keeping the backbone frozen. For clarity, we write the normalized image and text features as

$$\mathbf{f}_I(\mathbf{x}) = \frac{\Phi_I(\mathbf{x})}{\|\Phi_I(\mathbf{x})\|_2}, \qquad \mathbf{f}_T(\boldsymbol{\theta}; c) = \frac{\Phi_T(\boldsymbol{\theta}; c)}{\|\Phi_T(\boldsymbol{\theta}; c)\|_2}. \tag{7}$$

The class logit is defined as $z_c = \langle \mathbf{f}_I(\mathbf{x}), \mathbf{f}_T(\boldsymbol{\theta}; c) \rangle / \tau$, and the prediction probability is $p(c|\mathbf{x}) = \mathrm{softmax}(\mathbf{z})_c$, where τ is the temperature parameter.

4.2 Geometric Gradient Rectification Prompting

temporal momentum guidance. Reliable geometric target rectification requires stable textual class prototypes. Using the online prompt to construct these prototypes makes them fluctuate as stochastic gradient descent updates the prompt on noisy data. To stabilize the prototypes used for OT alignment, we employ temporal momentum guidance. Let $\boldsymbol{\theta}_t$ denote the online prompt parameters at step t. We maintain a parallel guide $\boldsymbol{\theta}'$ evolving via an exponential moving average (EMA):

$$\boldsymbol{\theta}'_t = m \cdot \boldsymbol{\theta}'_{t-1} + (1-m) \cdot \boldsymbol{\theta}_t \tag{8}$$

where m is the momentum coefficient. Since the image encoder is frozen, the image features are fixed normalized CLIP embeddings. The EMA guide stabilizes the text prototypes $\mathbf{F}'_T \in \mathbb{R}^{K \times D}$ and therefore stabilizes the image–text geometric landmarks used by geometric target rectification.

geometric target rectification. A direct mini-batch OT formulation with a hard all-class marginal can be problematic when $B \ll K$: many classes are absent from the batch, yet an all-class column marginal still forces them to receive transport mass. This issue is especially relevant for fine-grained datasets such as StanfordCars. We therefore solve OT on a batch-adaptive active class support rather than over all K classes.

For a mini-batch $\mathcal{B}$, let $\mathcal{Y}_\mathcal{B} = \{\tilde{y}_i : i \in \mathcal{B}\}$ be the observed noisy-label classes in the batch. Using the detached EMA prototypes, we also obtain the top-r guide-predicted classes for each sample and denote their union across the mini-batch by $\mathcal{T}_\mathcal{B}$. The active support is

$$\mathcal{A}_\mathcal{B} = \mathcal{Y}_\mathcal{B} \cup \mathcal{T}_\mathcal{B}. \tag{9}$$

By construction, the active support retains the true class of every correctly labeled sample in the mini-batch, regardless of its prediction confidence. This ensures that low-confidence hard positives retain their correct class as a candidate for geometric assignment. The top-r guide predictions further expand the support with alternative classes, allowing OT to consider corrections to corrupted labels. This support keeps the OT problem small, $B \times |\mathcal{A}_\mathcal{B}|$, and avoids assigning mass to classes that are neither observed in the batch nor geometrically plausible under the guide.

Let n_k denote the number of training samples with noisy label $\tilde{y}_i = k$, and define the empirical noisy-label prior $\pi_k = n_k/N$. To avoid the uniform-prior bias while also avoiding a hard all-class batch constraint, we restrict and renormalize the empirical prior on the active support:

$$b_a^\mathcal{B} = \frac{\max(\pi_a, 1/N)}{\sum_{\ell \in \mathcal{A}_\mathcal{B}} \max(\pi_\ell, 1/N)}, \qquad a \in \mathcal{A}_\mathcal{B}. \tag{10}$$

The floor $1/N$ only prevents zero-mass candidate classes and does not introduce a tunable hyperparameter. The cost matrix is defined by the negative cosine similarity between fixed image features and EMA text prototypes:

$$\mathbf{C}_{ia}^\mathcal{B} = -\langle \mathbf{f}_I(\mathbf{x}_i), \mathbf{f}'_{T,a} \rangle, \qquad a \in \mathcal{A}_\mathcal{B}. \tag{11}$$

The optimal coupling matrix $\mathbf{Q}^* \in \mathbb{R}_+^{B \times |\mathcal{A}_\mathcal{B}|}$ is obtained by solving the entropic OT problem

$$\begin{aligned} \mathbf{Q}^* = \arg\min_{\mathbf{Q} \geq 0} \langle \mathbf{C}^\mathcal{B}, \mathbf{Q} \rangle + \epsilon \sum_{i,a} \mathbf{Q}_{ia} \log \mathbf{Q}_{ia} \\ \text{s.t.} \quad \mathbf{Q}\mathbf{1}_{|\mathcal{A}_\mathcal{B}|} = \frac{1}{B}\mathbf{1}_B, \quad \mathbf{Q}^\top \mathbf{1}_B = \mathbf{b}^\mathcal{B}. \end{aligned} \tag{12}$$

where $0 \log 0$ is defined as zero and ϵ controls the smoothness. This formulation uses empirical class-frequency information without enforcing equal class mass and without forcing every class to appear in every mini-batch. Sinkhorn-Knopp solves this problem; Appendix F details the support and prior design.

We derive the rectified conditional assignment score via row-wise renormalization:

$$P_{\text{ot}}(y_a|\mathbf{x}_i) = \frac{\mathbf{Q}^*_{ia}}{\sum_{\ell \in \mathcal{A}_\mathcal{B}} \mathbf{Q}^*_{i\ell}}, \qquad a \in \mathcal{A}_\mathcal{B}, \tag{13}$$

and set $P_{\text{ot}}(y_k|\mathbf{x}_i) = 0$ for $k \notin \mathcal{A}_\mathcal{B}$. These scores should be interpreted as detached prior-constrained assignment scores rather than calibrated posteriors.

While OT reflects batch-level geometric evidence, boundary samples may still exhibit ambiguity. To quantify separation among the active candidates, we define the top-1/top-2 OT assignment margin δ_i:

$$\delta_i = P_{\text{ot}}(y_{k_1}|\mathbf{x}_i) - P_{\text{ot}}(y_{k_2}|\mathbf{x}_i), \tag{14}$$

where k_1 and k_2 are the highest- and second-highest-scoring classes in $\mathcal{A}_\mathcal{B}$; if the active support contains only one class, the second score is set to zero. The OT margin measures separation between the leading candidates within the active support and controls the CE contribution in the hybrid objective. Larger margins favor CE on the rectified target, while margins at or below τ_{low} retain MAE on the observed label.

silenced gradient recovery. We implement silenced gradient recovery by incorporating δ_i into the objective through assignment-margin gating. The score is mapped to a gating weight $w_i \in [0, 1]$ via a linear ramp function:

$$w_i = \text{clip}\left(\frac{\delta_i - \tau_{\text{low}}}{\tau_{\text{high}} - \tau_{\text{low}}}, 0, 1\right) \tag{15}$$

This weight controls the relative contributions of CE on the rectified target and MAE on the observed label. Let $y_i^{ot} = \arg\max_k P_{\text{ot}}(y_k|\mathbf{x}_i)$ denote the rectified label derived from the optimal transport coupling. The hybrid target $\hat{y}_i$ is determined as:

$$\hat{y}_i = \begin{cases} y_i^{ot} & \text{if } \delta_i \geq \tau_{\text{low}} \quad \text{(Rectified Label from } \mathbf{Q}^*\text{)} \\ \tilde{y}_i & \text{otherwise} \quad \text{(Original Noisy Label)} \end{cases} \tag{16}$$

For mislabeled samples, recovery of the true class depends on its inclusion in the active support; a large margin alone does not establish target correctness (Appendix G).

Table 1: Top-1 test accuracy (%) under varying levels of **Symmetric Noise**. Subtable (a) reports the arithmetic average across seven datasets; subtables (b)–(h) report per-dataset accuracy. Bold indicates the highest displayed accuracy in each column. Results with standard deviations are provided in Appendix H.

(a) Average (All Datasets)

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	74.23	67.59	61.66	52.91	42.09	28.83
JoAPR (CVPR'24)	76.36	73.03	71.16	68.94	62.58	54.22
NLPrompt (CVPR'25)	80.21	78.95	78.02	74.65	71.46	62.39
NA-MVP (CVPR'26)	80.07	78.80	77.21	74.19	70.31	64.81
GGRP (Ours)	**80.32**	**79.29**	**78.49**	**75.41**	**72.28**	**65.63**

(b) Flowers102

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	88.93	83.50	77.93	70.10	55.60	37.17
JoAPR (CVPR'24)	85.57	81.23	74.60	70.23	67.90	66.93
NLPrompt (CVPR'25)	93.87	92.57	92.73	89.90	84.77	76.80
NA-MVP (CVPR'26)	92.13	91.02	90.44	87.96	84.12	78.38
GGRP (Ours)	**94.27**	**93.44**	**93.42**	**90.48**	**85.00**	**78.73**

(c) DTD

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	56.00	49.57	43.30	34.37	27.83	17.27
JoAPR (CVPR'24)	58.07	57.70	56.33	53.03	48.05	29.90
NLPrompt (CVPR'25)	62.97	61.23	59.17	55.17	49.03	39.80
NA-MVP (CVPR'26)	**63.27**	**61.41**	59.36	**55.72**	**50.83**	44.21
GGRP (Ours)	62.10	60.57	**59.48**	55.48	49.46	**44.45**

(d) EuroSAT

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	76.50	69.23	61.67	52.33	37.63	26.70
JoAPR (CVPR'24)	75.13	61.10	60.90	63.63	38.97	27.33
NLPrompt (CVPR'25)	82.53	79.53	78.13	66.70	63.53	43.80
NA-MVP (CVPR'26)	81.77	77.43	72.08	64.61	55.27	42.84
GGRP (Ours)	**83.06**	**80.35**	**78.42**	**67.27**	**64.17**	**43.95**

(e) OxfordPets

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	76.50	66.73	60.33	47.03	35.77	24.60
JoAPR (CVPR'24)	84.00	83.26	83.20	83.10	82.40	74.40
NLPrompt (CVPR'25)	86.17	86.00	85.33	84.87	83.63	70.77
NA-MVP (CVPR'26)	**88.21**	**88.47**	**87.93**	86.02	85.11	82.64
GGRP (Ours)	87.70	87.46	87.20	**86.46**	**85.20**	**82.90**

(f) StanfordCars

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	66.20	59.70	53.40	45.90	35.67	22.90
JoAPR (CVPR'24)	68.60	66.30	62.83	56.67	48.50	39.40
NLPrompt (CVPR'25)	**69.37**	**68.80**	**67.20**	65.63	62.83	58.30
NA-MVP (CVPR'26)	67.68	66.92	65.87	63.74	60.31	56.03
GGRP (Ours)	69.01	68.23	66.85	**66.16**	**63.46**	**58.90**

(g) UCF101

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	69.03	63.40	58.23	49.73	40.83	26.30
JoAPR (CVPR'24)	72.83	71.17	70.37	67.63	65.30	57.67
NLPrompt (CVPR'25)	74.83	73.40	72.83	70.33	68.10	60.53
NA-MVP (CVPR'26)	**75.31**	**74.05**	72.88	70.34	66.72	62.14
GGRP (Ours)	74.27	73.10	**72.91**	**71.20**	**68.46**	**62.60**

(h) Caltech101

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	86.43	81.03	76.73	70.90	61.33	46.90
JoAPR (CVPR'24)	90.30	90.45	89.90	88.27	86.93	83.93
NLPrompt (CVPR'25)	91.73	91.13	90.77	89.93	88.30	86.70
NA-MVP (CVPR'26)	**92.14**	**92.31**	**91.88**	**90.96**	89.84	87.42
GGRP (Ours)	91.82	91.86	91.12	90.80	**90.20**	**87.86**

Hybrid Optimization Objective. The final objective harmonizes geometric consistency with robust learning. We combine a cross-entropy (CE) loss for high-margin rectified targets and a mean absolute error (MAE) loss for noise robustness:

$$\mathcal{L} = \frac{1}{B}\sum_{i=1}^{B}\left[w_i \mathcal{L}_{\text{CE}}(p(\mathbf{x}_i), \hat{y}_i) + (1 - w_i)\mathcal{L}_{\text{MAE}}(p(\mathbf{x}_i), \tilde{y}_i)\right] \tag{17}$$

where $\mathcal{L}_{\text{MAE}}(p, y) = \|\mathbf{e}_y - p\|_1 = 2(1 - p_y)$. Gradients do not propagate through the active support, Sinkhorn solver, OT margin, or gate. Strong geometric evidence enables recovery; weak evidence retains robust attenuation (Appendix B.2).

5 Experiments

5.1 Setup

We evaluate on seven datasets spanning diverse visual recognition tasks: Flowers102 (Nilsback & Zisserman, 2008), OxfordPets (Parkhi et al., 2012), and StanfordCars (Krause et al., 2013) for fine-grained classification; Caltech101 (Fei-Fei et al., 2004) for general objects; DTD (Cimpoi et al., 2014) for textures; EuroSAT (Helber et al., 2019) for satellite imagery; and UCF101 (Soomro et al., 2012) for actions. Food101N (Lee et al., 2018) additionally provides a benchmark with naturally occurring web-label noise. The synthetic-noise comparisons include CoOp (Zhou et al., 2022b), which learns continuous text prompts with frozen encoders; JoAPR (Guo & Gu, 2024), which combines adaptive data partitioning with label refurbishment; NLPrompt (Pan et al., 2025), which uses optimal transport to partition samples for CE/MAE training; and NA-MVP (Niu & Xue, 2026), which uses bi-directional multi-view prompts and region-aware optimal transport alignment. Table 3 also includes generalized cross-entropy (GCE) (Zhang & Sabuncu, 2018) as a loss-based reference.

Table 2: Top-1 test accuracy (%) under varying levels of **Asymmetric Noise**. Subtable (a) reports the arithmetic average across seven datasets; subtables (b)–(h) report per-dataset accuracy. Values are accuracy point estimates. Bold indicates the highest displayed accuracy in each column. Results with standard deviations are provided in Appendix H.

(a) Average (All Datasets)

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	73.23	63.62	51.49	38.53	25.90	14.70
JoAPR (CVPR'24)	74.10	72.54	67.06	62.30	50.42	36.42
NLPrompt (CVPR'25)	79.79	78.33	75.36	68.29	63.23	47.84
NA-MVP (CVPR'26)	80.13	78.30	74.93	70.07	63.74	**54.34**
GGRP (Ours)	**80.80**	**79.46**	**75.91**	**71.07**	**63.84**	53.98

(b) Flowers102

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	86.97	74.70	60.43	42.60	26.53	12.60
JoAPR (CVPR'24)	85.17	79.63	73.97	73.83	53.37	13.27
NLPrompt (CVPR'25)	93.80	93.40	91.77	81.10	73.63	55.33
NA-MVP (CVPR'26)	92.32	90.69	87.90	84.67	74.43	56.45
GGRP (Ours)	**93.93**	**94.15**	**92.32**	**84.73**	**74.74**	**56.86**

(c) DTD

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	55.60	47.75	38.07	29.63	20.53	11.70
JoAPR (CVPR'24)	52.40	56.63	53.10	48.93	40.20	28.26
NLPrompt (CVPR'25)	62.30	60.60	56.47	50.80	40.27	28.37
NA-MVP (CVPR'26)	64.36	62.48	57.32	50.97	45.48	35.39
GGRP (Ours)	**64.88**	**63.03**	**57.44**	**51.08**	**46.40**	**36.88**

(d) EuroSAT

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	76.00	66.27	53.83	41.17	28.00	17.43
JoAPR (CVPR'24)	69.37	67.30	59.40	47.60	33.93	17.50
NLPrompt (CVPR'25)	80.13	77.13	71.43	54.30	**66.33**	32.73
NA-MVP (CVPR'26)	**81.38**	77.07	70.07	60.91	48.11	34.55
GGRP (Ours)	80.63	**77.34**	**71.56**	**65.05**	49.09	**34.68**

(e) OxfordPets

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	76.10	66.20	52.53	38.73	26.63	14.90
JoAPR (CVPR'24)	82.90	83.40	79.07	75.84	52.74	43.57
NLPrompt (CVPR'25)	86.00	84.97	82.40	77.53	66.33	48.60
NA-MVP (CVPR'26)	**88.90**	87.00	**83.87**	78.76	72.13	63.40
GGRP (Ours)	**88.90**	**87.40**	83.73	**80.22**	**72.97**	**64.80**

(f) StanfordCars

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	65.77	57.13	46.23	33.73	22.37	12.80
JoAPR (CVPR'24)	66.47	61.70	51.50	42.03	30.80	22.97
NLPrompt (CVPR'25)	69.77	67.53	64.23	59.03	50.90	39.50
NA-MVP (CVPR'26)	66.62	65.61	63.37	59.64	**57.45**	**53.37**
GGRP (Ours)	**70.52**	**68.60**	**64.45**	**59.86**	56.07	49.96

(g) UCF101

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	67.23	58.07	46.47	34.43	23.67	13.17
JoAPR (CVPR'24)	72.07	69.80	64.10	59.17	56.07	47.46
NLPrompt (CVPR'25)	74.90	73.53	71.03	65.97	58.97	49.27
NA-MVP (CVPR'26)	75.37	73.75	**72.15**	**67.85**	**63.41**	**55.96**
GGRP (Ours)	**75.50**	**74.28**	71.19	66.87	60.83	52.27

(h) Caltech101

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	84.93	75.23	62.87	49.43	33.57	20.33
JoAPR (CVPR'24)	90.30	89.30	88.30	88.73	85.80	81.90
NLPrompt (CVPR'25)	91.60	91.17	90.20	89.27	86.17	81.07
NA-MVP (CVPR'26)	**91.96**	**91.50**	89.81	87.69	85.19	81.24
GGRP (Ours)	91.24	91.40	**90.65**	**89.70**	**86.80**	**82.40**

Table 3: Top-1 test accuracy (%) on Food101N with naturally occurring label noise. NA-MVP (Niu & Xue, 2026) uses RN50, 16-shot native noisy labels, 50 epochs, and seed 2, evaluated on 30,300 clean test images. Bold indicates the highest displayed accuracy.

Metric	**CoOp**	**GCE**	**JoAPR**	**NLPrompt**	**NA-MVP**	**GGRP**
Accuracy (%)	69.50	71.32	72.57	76.46	76.59	**76.93**

For the synthetic benchmarks, we follow the CoOp/NLPrompt few-shot setup with 16 training examples per class and CLIP ResNet-50 as the default backbone. GGRP optimizes 16 shared context tokens using SGD with an initial learning rate of 0.002, cosine annealing, and 100 training epochs. We inject noise only into training labels and leave test labels unchanged. Symmetric noise replaces a selected label uniformly with an alternative class, whereas asymmetric noise uses a fixed successor-class mapping. Both settings use six corruption rates: 12.5%, 25.0%, 37.5%, 50.0%, 62.5%, and 75.0%. We report top-1 test accuracy (%); the summary in each synthetic-noise table is the arithmetic average across the seven datasets. Detailed protocols and GGRP hyperparameters are provided in Appendix C.

5.2 Performance Comparison

Tables 1 and 2 report synthetic-noise evaluations on seven datasets under symmetric and asymmetric corruptions from 12.5% to 75.0%, while Table 3 evaluates naturally noisy supervision on Food101N. Tables 1 and 2 report mean accuracies for readability. The corresponding standard deviations across runs are reported alongside the means in Appendix H: Table 9 for symmetric noise and Table 10 for asymmetric noise. In subtable (a) of each appendix table, we separately average the dataset-level mean accuracies and standard deviations across the seven datasets. Across these settings, GGRP remains competitive with NLPrompt at low noise and shows clearer gains over NLPrompt as corruption

Table 4: Component ablation on Caltech101 under symmetric noise. Values are test accuracies (%).

Method	**25.0%**	**50.0%**	**75.0%**
EMA guide only (w/o OT, w/o dynamic gating)	88.13	71.60	53.87
EMA + OT (w/o dynamic gating; MAE loss only)	87.63	88.83	84.93
Generic OT (all classes, uniform prior; no gate)	90.18	87.76	83.62
Confidence pseudo-labels (same gate)	89.54	85.21	77.86
Active OT + MAE (same gate; CE replaced)	91.37	90.46	86.19
Active OT + CE (w/o gating)	89.43	81.68	69.57
Full GGRP	91.86	90.80	87.86

becomes severe. At 75.0% symmetric noise, GGRP improves the average accuracy over NLPrompt from 62.39% to 65.63%, with a notable gain on OxfordPets (82.90% vs. 70.77%). Under the harder asymmetric setting, GGRP also improves the 75.0% average over NLPrompt from 47.84% to 53.98%, while CoOp drops to 14.70%. Against NA-MVP, GGRP achieves higher average accuracy at the five asymmetric-noise rates from 12.5% to 62.5%; at 75.0%, NA-MVP reaches 54.34% versus 53.98% for GGRP. The gains over NLPrompt suggest that combining geometric target rectification with silenced gradient recovery can help when hard partitioning and MAE attenuation suppress hard positive samples, while the comparison with NA-MVP shows that these gains do not extend to every baseline. Food101N shows a consistent real-noise gain over NLPrompt (76.93% vs. 76.46%). Further analysis of these comparisons is provided in Appendix D.

5.3 Ablative Analysis

Component Contribution: Table 4 organizes the component analysis on Caltech101 around target construction, loss choice, and dynamic gating. *EMA guide only* retains temporal momentum guidance but removes geometric target rectification and silenced gradient recovery, serving as a reduced baseline. The *EMA + OT (MAE only)* variant retains temporal momentum guidance and geometric target rectification, but replaces silenced gradient recovery with MAE-only training. *Generic OT* uses all-class support and a uniform prior, with ungated CE on OT targets; it provides a joint reference for the support, prior, and gating design. To examine the source of target guidance, the *Confidence pseudo-labels* variant replaces OT assignments with argmax targets from detached EMA class probabilities. It retains the margin-gating rule and the MAE fallback on observed labels.

Active OT + MAE retains active-support OT, the empirical prior, and the gating rule, but replaces rectified-target CE with MAE to examine the role of the loss on corrected targets. *Active OT + CE (w/o gating)* retains the same OT formulation but applies CE to every assigned target, testing the role of selective weighting. *Full GGRP* combines temporal momentum guidance, geometric target rectification with the empirical prior, and silenced gradient recovery through the hybrid objective that weights rectified-target CE against MAE on the observed label.

Additional Experiments in the Appendix. We examine sampler/prior/support choices on StanfordCars (Appendix E.1, Table 5); training efficiency on Caltech101 (Appendix E.2, Table 6); sensitivity to the number of guide candidates on StanfordCars (Appendix E.3, Table 7); and gating-threshold sensitivity on Caltech101 (Appendix E.4, Table 8). These studies assess design choices, computational cost, and parameter sensitivity.

6 Conclusion

In this work, we conduct a systematic analysis of noisy supervision in prompt tuning to reveal the failure mode of gradient silencing. We demonstrate that hard sample partitioning routes hard positives to the noisy subset, where confidence-dependent MAE attenuation suppresses their corrective gradients alongside those of mislabeled samples. Building on these insights, we introduce Geometric Gradient Rectification Prompting (GGRP), a framework that couples geometric target rectification with silenced gradient recovery. GGRP combines temporal momentum guidance, geometric target rectification, and silenced gradient recovery to selectively recover corrective gradients for geometrically supported hard positives while retaining robust attenuation for ambiguous samples. Extensive experiments show that GGRP matches or surpasses state-of-the-art methods across a range of noise settings, supporting the effectiveness of geometric target rectification and silenced gradient recovery.

AI USE STATEMENT

Generative AI tools were used to assist with implementation of the method, translation and language editing of manuscript text, literature search and discovery, and LaTeX formatting. During revision, these tools also assisted in checking consistency between reported results and supporting experimental records and in refining the wording and scope of the claims made in the paper. The authors take responsibility for the final content of this paper, including its claims, references, and reported results.

REFERENCES

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPRW*, 2004.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-Adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021.

Boyang Guo, Liang Li, Jiehua Zhang, Yaoqi Sun, Chenggang Yan, and Xichun Sheng. Prompt learning with knowledge regularization for pre-trained vision-language models. *IEEE Transactions on Multimedia*, pp. 1–12, 2025. doi: 10.1109/TMM.2025.3639886.

Yuncheng Guo and Xiaodong Gu. JoAPR: Cleaning the lens of prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28695–28705, June 2024.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pp. 8527–8537, 2018.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.

Li Hua and Jin Qian. AnomalyNLP: Noisy-label prompt learning for few-shot industrial anomaly detection. *Electronics*, 14(20):4016, 2025. doi: 10.3390/electronics14204016.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pp. 4904–4916. PMLR, 2021.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal prompt learning. In *CVPR*, pp. 19113–19122, 2023.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *ICCVW*, 2013.

Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. CleanNet: Transfer learning for scalable image classifier training with label noise. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Yilun Li, Miaomiao Cheng, Xu Han, and Wei Song. Divergence-enhanced knowledge-guided context optimization for visual-language prompt tuning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. Class-agnostic object detection with multi-modal transformer. In *ECCV*. Springer, 2022.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, 2008.

Lu Niu and Cheng Xue. Noise-aware few-shot learning through bi-directional multi-view prompt alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41646–41656, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Niu_Noise-Aware_Few-Shot_Learning_through_Bi-directional_Multi-View_Prompt_Alignment_CVPR_2026_paper.html.

Bikang Pan, Qun Li, Xiaoying Tang, Wei Huang, Zhen Fang, Feng Liu, Jingya Wang, Jingyi Yu, and Ye Shi. NLPrompt: Noise-label prompt learning for vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19963–19973, 2025.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. In *NeurIPS*, 2022.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6757–6767, June 2023. doi: 10.1109/CVPR52729.2023.00653.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-Adapter: Training-free adaption of CLIP for few-shot classification. In *Computer Vision – ECCV 2022*, pp. 493–510, 2022. doi: 10.1007/978-3-031-19833-5_29.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pp. 16816–16825, 2022a.

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

A Full Training Procedure

Algorithm 1 Geometric Gradient Rectification Prompting

Input: Noisy training dataset $\mathcal{D}$, empirical class prior $\boldsymbol{\pi}$, batch size B, and number r of top-ranked guide classes per sample.
Initialize: Guide prompt $\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta}$.
while training **do**
  Draw a mini-batch $\mathcal{B}$ of B examples from $\mathcal{D}$ using the class-aware sampler.
  Extract fixed image features and EMA text prototypes without gradient.
  Build active class support $\mathcal{A}_{\mathcal{B}}$ from noisy labels and top-r guide classes.
  Restrict and renormalize the empirical prior to $\mathbf{b}^{\mathcal{B}}$.
  Compute cost $\mathbf{C}^{\mathcal{B}}_{ia} = -\langle \mathbf{f}_I(\mathbf{x}_i), \mathbf{f}'_{T,a} \rangle$.
  $\mathbf{Q}^* \leftarrow \text{Sinkhorn}(\mathbf{1}/B, \mathbf{b}^{\mathcal{B}}, \mathbf{C}^{\mathcal{B}}, \epsilon)$.
  Normalize rows to obtain P_{ot}, rectified labels y^{ot}, and margins δ.
  Compute online probabilities $p(\mathbf{x})$ using $\boldsymbol{\theta}$.
  Compute gate w_i via Eq. (15) and target $\hat{y}_i$.
  Optimize Eq. (17) with gradients stopped through OT and the gate.
  Update $\boldsymbol{\theta}' \leftarrow m\boldsymbol{\theta}' + (1-m)\boldsymbol{\theta}$.
end while

B Additional Gradient Analysis

B.1 Extension to GCE-family Robust Losses

The low-confidence silencing effect is not unique to MAE. For generalized cross-entropy (GCE) (Zhang & Sabuncu, 2018), defined as $\mathcal{L}_{\text{GCE}} = (1 - s_{\tilde{y}}^q)/q$ with $q \in (0, 1]$, the gradient with respect to logits is

$$\nabla_{\mathbf{z}} \mathcal{L}_{\text{GCE}} = s_{\tilde{y}}^q (\mathbf{s} - \mathbf{e}_{\tilde{y}}). \tag{18}$$

Thus, for any $q > 0$, the gradient magnitude is modulated by $s_{\tilde{y}}^q$ and vanishes as $s_{\tilde{y}} \to 0$. CE corresponds to the $q \to 0$ limiting case and preserves a non-vanishing corrective gradient.

B.2 Gradient Analysis of GGRP

To formalize how GGRP restores the learning signal, we examine the gradient flow with respect to the online logits $\mathbf{z}$. Let $\mathbf{z} \in \mathbb{R}^K$ denote the logits for sample $\mathbf{x}_i$, where $z_k = \langle \mathbf{f}_I(\mathbf{x}_i), \mathbf{f}_T(\boldsymbol{\theta}; k) \rangle / \tau$ and $\mathbf{p} = \text{softmax}(\mathbf{z})$. During backpropagation, no gradient is propagated through the active support, the EMA guide, the Sinkhorn solver, the OT margin, or the discrete target selection. Thus the gating weight w_i and the hybrid target $\hat{y}_i$ are treated as constants. For the per-sample objective in Eq. (17), the logit gradient is

$$\nabla_{\mathbf{z}} \mathcal{L}_i = \underbrace{w_i (\mathbf{p} - \mathbf{e}_{\hat{y}_i})}_{\text{CE injection}} + \underbrace{(1 - w_i) 2 p_{\tilde{y}_i} (\mathbf{p} - \mathbf{e}_{\tilde{y}_i})}_{\text{MAE suppression}}. \tag{19}$$

Eq. (19) makes the role of GGRP explicit: the CE branch supplies a non-vanishing corrective signal for geometrically supported samples, while the MAE branch retains the low-confidence suppression needed for ambiguous noisy labels.

Consider a hard positive sample whose noisy label is in fact correct, i.e., $\tilde{y}_i = k$, but whose online confidence is low ($p_k \to 0$). Under MAE alone, its target-logit gradient is

$$\frac{\partial \mathcal{L}_{\text{MAE}}}{\partial z_k} = 2 p_k (p_k - 1) \xrightarrow[p_k \to 0]{} 0, \tag{20}$$

which recovers the gradient silencing effect identified in Section 3. GGRP changes this limit only when the detached momentum guide provides sufficient geometric evidence. If the OT assignment matches the semantic class, $\hat{y}_i = k$, the target-logit component of Eq. (19) becomes

$$\frac{\partial \mathcal{L}_i}{\partial z_k} = w_i (p_k - 1) + (1 - w_i) 2 p_k (p_k - 1). \tag{21}$$

Taking the hard-positive limit gives

$$\lim_{p_k \to 0} \frac{\partial \mathcal{L}_i}{\partial z_k} = -w_i. \tag{22}$$

Therefore, for a hard positive correctly assigned by OT with $\delta_i \geq \tau_{\text{high}}$, we have $w_i = 1$ and the CE branch restores the full non-vanishing logit-level gradient $\partial \mathcal{L}_i / \partial z_k \to -1$. For intermediate margins, $\tau_{\text{low}} \leq \delta_i < \tau_{\text{high}}$, the restorative force is injected proportionally to w_i rather than activated abruptly. For ambiguous samples with $\delta_i < \tau_{\text{low}}$, $w_i = 0$ and Eq. (19) reduces to the MAE gradient on the original noisy label, suppressing CE-driven updates when the geometric evidence is insufficient. At the prompt level, this non-vanishing logit signal produces a corrective prompt update as long as the corresponding logit Jacobian is non-degenerate. The conclusion is therefore conditional on reliable detached OT assignment; if the geometric assignment is wrong, the CE branch may reinforce an error, which motivates the active support and margin gate. Thus, GGRP conditionally revives hard-positive gradients through OT-guided CE injection while preserving robust gradient attenuation for low-margin ambiguous samples.

C Experimental Protocols

Datasets and baselines. The seven synthetic-noise datasets cover fine-grained classification (Flowers102 (Nilsback & Zisserman, 2008), OxfordPets (Parkhi et al., 2012), StanfordCars (Krause et al., 2013)), general object recognition (Caltech101 (Fei-Fei et al., 2004)), texture recognition (DTD (Cimpoi et al., 2014)), satellite imagery (EuroSAT (Helber et al., 2019)), and action recognition (UCF101 (Soomro et al., 2012)). Food101N (Lee et al., 2018) is used for naturally occurring web-label noise. We compare with CoOp (Zhou et al., 2022b), JoAPR (Guo & Gu, 2024), NLPrompt (Pan et al., 2025), and NA-MVP (Niu & Xue, 2026).

Noise settings. For the seven datasets under synthetic-noise protocols, we evaluate two corruption patterns: symmetric noise and asymmetric noise. Noise injection is restricted strictly to the training set while the test set remains untouched to ensure an unbiased evaluation protocol. Under the symmetric setting, clean labels are uniformly flipped to any alternative category with equal probability, representing a purely stochastic label corruption process. In contrast, asymmetric noise deterministically maps each selected label to a fixed successor class. This asymmetric setting aims to mimic realistic annotation errors where human labelers might confuse fine-grained categories with high visual similarity. To systematically evaluate robustness, we control the severity of corruption by varying the noise rate across different levels to observe performance trends. Due to its inherent structural correlation, asymmetric noise typically poses a more severe challenge to model convergence and generalization compared to the uniform distribution of symmetric noise. It serves as a rigorous benchmark for evaluating the ability of the model to rectify misleading gradients. For Food101N, we do not inject artificial corruption; instead, we train on the provided web-collected noisy labels and evaluate under the same real-world noisy-label protocol as the compared methods.

Implementation details. We follow the setup of CoOp and NLPrompt for fair comparison. We use SGD with an initial learning rate of 0.002 and cosine annealing. Unless otherwise specified, CLIP with ResNet-50 is the default backbone, the number of training epochs is 100, and 16 shared context tokens are used with the class token placed at the end. For synthetic-noise benchmarks, we sample a 16-shot training set per dataset. For GGRP, $\tau_{\text{low}} = 0.05$, $\tau_{\text{high}} = 0.12$, and $r = 3$ by default. The active support contains noisy-label classes in the current batch and the top-r guide-predicted classes per sample. We use a class-aware mini-batch sampler that increases class coverage and prioritizes classes not sampled in recent batches. Experiments were conducted in PyTorch on NVIDIA RTX 4090 GPUs.

NA-MVP results. We report NA-MVP (Niu & Xue, 2026) accuracies on all seven synthetic-noise datasets at all six noise rates, with arithmetic averages across datasets, and on Food101N. Standard deviations for the synthetic-noise results are reported in Appendix H. The Food101N result uses our independent RN50, 16-shot reproduction with 50 training epochs and seed 2, reporting final-epoch test accuracy.

C.1 GGRP HYPERPARAMETERS

Unless otherwise specified, GGRP uses a mini-batch size of $B = 32$ and an EMA momentum coefficient of $m = 0.99$. The active support includes the observed-label classes in the mini-batch and the top-r guide-predicted classes per sample, with $r = 3$.

For the entropic OT problem in Eq. (12), we set the regularization coefficient to $\epsilon = 0.05$. The Sinkhorn-Knopp solver uses at most 100 iterations with a convergence tolerance of 10^{-6}. Image and text features are ℓ_2-normalized. The classification logit scale is $1/\tau = 100$, corresponding to a temperature of $\tau = 0.01$, whereas the OT cost is the unscaled negative cosine similarity.

The margin gate uses $\tau_{\text{low}} = 0.05$ and $\tau_{\text{high}} = 0.12$. As defined in Eq. (15), the CE weight is zero for margins at or below τ_{low}, increases linearly between the two thresholds, and reaches one at or above τ_{high}.

D DETAILED PERFORMANCE ANALYSIS

Robustness against symmetric noise. As shown in Table 1, GGRP is competitive with NLPrompt in low-noise regimes (e.g., 12.5% and 25.0%), where differences between GGRP and NLPrompt are often small (e.g., 80.32% for GGRP vs. 80.21% for NLPrompt on average at 12.5% noise). As the noise intensity escalates, the gain over NLPrompt from geometric target rectification becomes clearer. In scenarios with extreme noise (75%), GGRP achieves an average accuracy of 65.63% compared to 62.39% for NLPrompt. Notably, on fine-grained datasets like OxfordPets, GGRP surpasses NLPrompt by 12.13 percentage points (82.90% vs. 70.77% at 75% noise), suggesting that geometric alignment helps preserve hard-sample learning signals under severe corruption.

Robustness against asymmetric noise. The results for asymmetric noise, which mimics realistic structural confusion, are presented in Table 2. GGRP achieves higher average accuracy than NLPrompt at all six noise rates. At 75% noise, it reaches 53.98%, compared with 47.84% for NLPrompt and 14.70% for CoOp. The comparison with NA-MVP is mixed: GGRP leads by 0.67, 1.16, 0.98, 1.00, and 0.10 percentage points at noise rates of 12.5%, 25.0%, 37.5%, 50.0%, and 62.5%, respectively. At 75.0% noise, NA-MVP reaches 54.34%, exceeding GGRP by 0.36 percentage points. Thus, the improvements over NLPrompt do not establish a general advantage over NA-MVP under asymmetric noise.

Robustness on real-world noisy labels. We further evaluate GGRP on Food101N, a real-world noisy-label benchmark without synthetic label injection. As shown in Table 3, GGRP achieves 76.93% accuracy, a clear improvement over NLPrompt (76.46%) under the same setting. This suggests that geometric target rectification can effectively transfer to naturally noisy supervision and deliver consistent gains in real-world noisy-label scenarios.

E Additional Ablations

E.1 Sampler, prior, and active support

GGRP uses three design choices to make mini-batch OT reliable when the number of classes is larger than the batch size: a class-aware sampler to improve batch coverage, an empirical noisy-label prior to avoid the equal-class-mass assumption, and an active support to avoid forcing absent classes to receive mass in every batch. To isolate their effects, Table 5 compares all combinations of random versus class-aware sampling, uniform versus empirical class priors, and all-class versus active-support OT on StanfordCars, a high-class-count benchmark where $B \ll K$ is most pronounced. This ablation directly tests whether the proposed active-support empirical-prior design improves robustness beyond the base OT rectifier.

The full active-support empirical-prior design performs best, indicating that GGRP benefits from both sampler-driven coverage and a transport marginal restricted to observed or geometrically plausible classes.

Table 5: Sampler/prior/support ablation on StanfordCars.

Support	**Sampler**	**Prior**	**Sym. 50%**	**Sym. 75%**	**Asym. 50%**	**Asym. 75%**
All	Random	Uniform	63.02	54.10	55.84	35.72
All	Random	Emp.	63.68	55.36	56.61	36.84
All	Cls.	Uniform	64.91	56.94	57.83	38.02
All	Cls.	Emp.	64.76	57.61	58.31	38.56
Active	Random	Uniform	65.03	57.52	58.06	38.49
Active	Random	Emp.	65.72	58.31	58.82	39.22
Active	Cls.	Uniform	65.89	58.47	59.02	39.51
Active	Cls.	Emp.	**66.16**	**58.90**	**59.86**	**49.96**

E.2 Training efficiency analysis

In addition to standard performance comparisons, we analyze the training efficiency of GGRP on Caltech101. Table 6 reports training time per epoch, convergence epochs, and accuracy under high symmetric noise.

Table 6: Training efficiency analysis on Caltech101.

Method	**Time/Ep. (s)**	**Epochs**	**Accuracy under Noise (%)**		
			50%	**62.5%**	**75%**
JoAPR	14	200	88.27	86.93	83.93
NLPrompt	25	200	89.93	88.30	86.70
GGRP (Ours)	**6**	**100**	**90.80**	**90.20**	**87.86**

E.3 SENSITIVITY TO THE NUMBER OF GUIDE CANDIDATES r

The active support augments noisy-label classes in the current batch with the top-r guide-predicted classes for each sample. A small r keeps the OT problem compact and conservative, whereas a large r increases the chance that the correct class is included but may also introduce less plausible classes into the transport problem. Table 7 evaluates this tradeoff on StanfordCars under both symmetric and asymmetric noise. The default setting $r = 3$ corresponds to the main GGRP results in Tables 1 and 2 and consistently achieves the best or near-best accuracy. This supports a middle ground: small support is conservative, while overly large support admits weaker candidates.

Table 7: Sensitivity to the number of guide candidates r.

Setting	**Number of guide candidates r**				
	1	**2**	**3**	**5**	**10**
Sym. 50%	65.43	65.98	**66.16**	66.05	65.74
Sym. 75%	57.84	58.48	**58.90**	58.55	58.63
Asym. 50%	58.61	59.08	**59.86**	59.01	59.12
Asym. 75%	38.92	39.51	**49.96**	39.81	39.64

E.4 SENSITIVITY TO GATING THRESHOLDS

We investigate the impact of the gating thresholds on Caltech101 under 50% symmetric noise. As shown in Table 8, GGRP remains stable across a range of τ_{low} and τ_{high} values, with most settings falling between 89.4% and 90.8% accuracy. The default setting reaches the best observed accuracy, while nearby settings perform similarly.

Table 8: Gating threshold sensitivity.

$\tau_{\text{high}} \backslash \tau_{\text{low}}$	τ_{low}				
	0.01	0.03	0.05	0.07	0.09
0.10	89.4	90.4	89.9	90.4	90.1
0.12	89.9	89.9	**90.8**	**90.8**	90.4
0.14	90.1	90.4	90.4	90.4	90.1
0.16	90.1	89.9	89.9	89.8	89.7

F Additional Notes on the OT Design

The active-support formulation addresses the mini-batch degeneracy that occurs when $B \ll K$. A hard all-class marginal forces every class, including absent and implausible classes, to receive mass in every mini-batch. This can make the Sinkhorn solution reflect the imposed marginal rather than image–text geometry. Restricting the problem to $\mathcal{A}_B$ keeps the optimization focused on classes that are either observed through noisy labels or plausible under the detached guide.

The empirical prior is estimated from noisy labels and is not assumed to be a clean estimate of the true class distribution. Its role is to prevent class collapse and avoid the stronger assumption of equal class mass. The margin gate is deliberately conservative: when the active-support OT assignment is not clearly separated, GGRP does not inject CE on the rectified target and instead falls back to MAE on the original noisy label.

G Limitations and Broader Impacts

GGRP is developed for robust prompt tuning of frozen vision-language models, and its benefits are most relevant when the pre-trained image–text geometry provides a useful signal for separating hard positives from label noise. Our evaluation covers multiple datasets, noise types, and a naturally noisy benchmark, but it does not exhaust all real-world regimes such as open-set noise, highly long-tailed class distributions, or domains with weak CLIP alignment. The active-support OT formulation and margin gate are designed to make the method conservative in these cases by avoiding all-class batch constraints and falling back to robust attenuation when geometric evidence is weak.

Dependence on pre-trained image–text geometry. The method assumes that the frozen VLM embedding space retains enough semantic structure for the detached guide to provide informative batch-level assignments. This assumption is reasonable for many CLIP-style recognition tasks, but it can be weaker when classes are visually indistinguishable, text names are poor semantic descriptors, or the target domain is far from the pretraining distribution. GGRP is therefore best viewed as a conservative rectification strategy rather than a label-verification procedure. The combination of temporal momentum guidance, geometric target rectification, and silenced gradient recovery is designed to reduce the risk of over-trusting uncertain assignments: when the OT margin is low, the objective does not inject CE on the rectified target and instead falls back to the robust MAE branch.

Batch composition and prior estimation. GGRP uses batch-level geometry together with an empirical noisy-label prior. The active-support formulation avoids the stronger and often unrealistic assumption that every mini-batch should allocate mass to all classes, especially when $B \ll K$. However, the empirical prior can still be less informative under extreme class imbalance, open-set contamination, or nonstationary label distributions. The empirical prior imposes a marginal constraint on transport mass over the active support to discourage class collapse, but it may not accurately reflect the true class distribution. When domain knowledge or a small trusted validation split is available, replacing or calibrating the empirical prior may further improve robustness.

True-class support coverage. For mislabeled samples, the observed-label union and top-r guide predictions do not guarantee inclusion of the true class. If it is excluded, OT cannot recover it in that update, and a large margin among the remaining candidates may still cause incorrect CE injection. The MAE fallback applies only when the margin is at or below τ_{low}; it does not detect the missing class.

Broader impact and responsible use. By improving robustness to imperfect labels, GGRP can reduce the need for costly full fine-tuning and repeated manual relabeling when adapting foundation vision-language models. This benefit is most appropriate for non-sensitive recognition tasks where noisy supervision is expected and model outputs remain subject to standard validation. The method should not be used as a substitute for data-quality auditing or human review in high-stakes settings. Like other methods built on pre-trained models, it may inherit biases from pretraining data or downstream benchmarks; deployments should therefore include dataset auditing, calibration checks, and task-specific fairness, privacy, and safety evaluations.

H Detailed Results with Standard Deviations

Tables 9 and 10 provide the full versions of Tables 1 and 2, including standard deviations. The mean accuracies are identical to those in the main text. In subtable (a), we separately average the seven dataset-level means and standard deviations; the value after ± is the mean of the dataset-level standard deviations.

Table 9: Top-1 test accuracy (%) under **Symmetric Noise**. All means match Table 1; standard deviations are shown below. Subtable (a) separately averages the means and standard deviations across seven datasets; (b)–(h) report individual datasets. Bold indicates the highest mean in each column.

(a) Average (All Datasets)

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	74.23 ±1.02	67.59 ±1.48	61.66 ±1.44	52.91 ±3.08	42.09 ±1.95	28.83 ±1.24
JoAPR (CVPR'24)	76.36 ±1.36	73.03 ±1.65	71.16 ±1.63	68.94 ±1.23	62.58 ±1.35	54.22 ±1.82
NLPrompt (CVPR'25)	80.21 ±0.71	78.95 ±0.79	78.02 ±1.23	74.65 ±1.62	71.46 ±1.70	62.39 ±1.23
NA-MVP (CVPR'26)	80.07 ±1.75	78.80 ±1.17	77.21 ±1.35	74.19 ±1.65	70.31 ±1.68	64.81 ±1.20
GGRP (Ours)	**80.32** ±0.80	**79.29** ±1.33	**78.49** ±1.64	**75.41** ±2.42	**72.28** ±1.86	**65.63** ±1.54

(b) Flowers102

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	88.93 ±0.30	83.50 ±0.97	77.93 ±1.39	70.10 ±1.05	55.60 ±0.60	37.17 ±0.35
JoAPR (CVPR'24)	85.57 ±0.92	81.23 ±1.46	74.60 ±0.27	70.23 ±1.58	67.90 ±0.70	66.93 ±1.29
NLPrompt (CVPR'25)	93.87 ±0.69	92.57 ±0.53	92.73 ±1.80	89.90 ±1.59	84.77 ±1.40	76.80 ±1.68
NA-MVP (CVPR'26)	92.13 ±2.63	91.02 ±1.14	90.44 ±1.33	87.96 ±2.14	84.12 ±0.90	78.38 ±2.50
GGRP (Ours)	**94.27** ±1.14	**93.44** ±0.90	**93.42** ±1.45	**90.48** ±0.77	**85.00** ±1.18	**78.73** ±0.53

(c) DTD

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	56.00 ±1.45	49.57 ±1.85	43.30 ±0.60	34.37 ±3.43	27.83 ±1.28	17.27 ±0.44
JoAPR (CVPR'24)	58.07 ±1.71	57.70 ±1.67	56.33 ±2.69	53.03 ±0.72	48.05 ±0.58	29.90 ±2.30
NLPrompt (CVPR'25)	62.97 ±0.79	61.23 ±0.30	59.17 ±1.65	55.17 ±0.58	49.03 ±2.83	39.80 ±2.30
NA-MVP (CVPR'26)	**63.27** ±0.53	**61.41** ±1.59	59.36 ±0.30	**55.72** ±1.36	**50.83** ±0.66	44.21 ±0.40
GGRP (Ours)	62.10 ±0.18	60.57 ±2.49	**59.48** ±1.44	55.48 ±1.09	49.46 ±2.22	**44.45** ±1.12

(d) EuroSAT

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	76.50 ±0.44	69.23 ±2.20	61.67 ±2.97	52.33 ±2.68	37.63 ±1.87	26.70 ±1.63
JoAPR (CVPR'24)	75.13 ±0.96	61.10 ±1.84	60.90 ±1.07	63.63 ±1.36	38.97 ±2.07	27.33 ±2.38
NLPrompt (CVPR'25)	82.53 ±1.40	79.53 ±1.59	78.13 ±2.21	66.70 ±6.92	63.53 ±3.34	43.80 ±0.47
NA-MVP (CVPR'26)	81.77 ±0.64	77.43 ±0.42	72.08 ±1.47	64.61 ±2.67	55.27 ±1.15	42.84 ±1.85
GGRP (Ours)	**83.06** ±2.22	**80.35** ±2.30	**78.42** ±5.21	**67.27** ±11.74	**64.17** ±4.53	**43.95** ±4.15

(e) OxfordPets

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	76.50 ±2.36	66.73 ±2.45	60.33 ±2.37	47.03 ±7.29	35.77 ±6.62	24.60 ±2.65
JoAPR (CVPR'24)	84.00 ±2.31	83.26 ±2.69	83.20 ±2.27	83.10 ±0.97	82.40 ±1.10	74.40 ±1.02
NLPrompt (CVPR'25)	86.17 ±1.08	86.00 ±0.21	85.33 ±0.70	84.87 ±1.15	83.63 ±1.87	70.77 ±1.51
NA-MVP (CVPR'26)	**88.21** ±1.01	**88.47** ±2.49	**87.93** ±2.33	86.02 ±0.53	85.11 ±2.55	82.64 ±1.19
GGRP (Ours)	87.70 ±0.88	87.46 ±2.37	87.20 ±1.20	**86.46** ±0.86	**85.20** ±1.72	**82.90** ±2.03

(f) StanfordCars

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	66.20 ±0.59	59.70 ±0.53	53.40 ±1.04	45.90 ±1.85	35.67 ±0.49	22.90 ±1.54
JoAPR (CVPR'24)	68.60 ±0.41	66.30 ±0.70	62.83 ±1.93	56.67 ±0.45	48.50 ±1.46	39.40 ±0.85
NLPrompt (CVPR'25)	**69.37** ±0.26	**68.80** ±0.71	**67.20** ±0.44	65.63 ±0.16	62.83 ±0.70	58.30 ±0.80
NA-MVP (CVPR'26)	67.68 ±2.50	66.92 ±0.51	65.87 ±1.07	63.74 ±1.09	60.31 ±2.54	56.03 ±1.50
GGRP (Ours)	69.01 ±0.52	68.23 ±0.05	66.85 ±0.52	**66.16** ±1.01	**63.46** ±0.80	**58.90** ±1.16

(g) UCF101

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	69.03 ±1.39	63.40 ±1.18	58.23 ±1.19	49.73 ±1.73	40.83 ±0.57	26.30 ±1.24
JoAPR (CVPR'24)	72.83 ±2.13	71.17 ±1.04	70.37 ±2.65	67.63 ±1.79	65.30 ±2.08	57.67 ±2.21
NLPrompt (CVPR'25)	74.83 ±0.54	73.40 ±1.71	72.83 ±1.43	70.33 ±0.79	68.10 ±0.92	60.53 ±1.03
NA-MVP (CVPR'26)	**75.31** ±2.42	**74.05** ±1.73	72.88 ±2.24	70.34 ±1.52	66.72 ±1.99	62.14 ±0.28
GGRP (Ours)	74.27 ±0.33	73.10 ±0.68	**72.91** ±1.28	**71.20** ±0.80	**68.46** ±1.74	**62.60** ±0.76

(h) Caltech101

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	86.43 ±0.64	81.03 ±1.20	76.73 ±0.55	70.90 ±3.56	61.33 ±2.23	46.90 ±0.86
JoAPR (CVPR'24)	90.30 ±1.07	90.45 ±2.14	89.90 ±0.52	88.27 ±1.76	86.93 ±1.44	83.93 ±2.67
NLPrompt (CVPR'25)	91.73 ±0.21	91.13 ±0.50	90.77 ±0.38	89.93 ±0.15	88.30 ±0.86	86.70 ±0.85
NA-MVP (CVPR'26)	**92.14** ±2.49	**92.31** ±0.32	**91.88** ±0.68	**90.96** ±2.21	89.84 ±2.00	87.42 ±0.66
GGRP (Ours)	91.82 ±0.35	91.86 ±0.53	91.12 ±0.41	90.80 ±0.67	**90.20** ±0.80	**87.86** ±1.04

Table 10: Top-1 test accuracy (%) under **Asymmetric Noise**. All means match Table 2; standard deviations are shown below. Subtable (a) separately averages the means and standard deviations across seven datasets; (b)–(h) report individual datasets. Bold indicates the highest mean in each column.

(a) Average (All Datasets)

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	73.23 ±0.86	63.62 ±1.52	51.49 ±1.78	38.53 ±1.62	25.90 ±1.32	14.70 ±0.81
JoAPR (CVPR'24)	74.10 ±1.08	72.54 ±1.12	67.06 ±1.33	62.30 ±1.65	50.42 ±1.91	36.42 ±1.48
NLPrompt (CVPR'25)	79.79 ±0.85	78.33 ±1.14	75.36 ±1.63	68.29 ±1.58	63.23 ±1.55	47.84 ±1.96
NA-MVP (CVPR'26)	80.13 ±1.55	78.30 ±1.28	74.93 ±1.46	70.07 ±1.11	63.74 ±1.65	**54.34** ±1.50
GGRP (Ours)	**80.80** ±1.48	**79.46** ±2.61	**75.91** ±1.50	**71.07** ±1.71	**63.84** ±1.17	53.98 ±1.40

(b) Flowers102

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	86.97 ±0.93	74.70 ±1.79	60.43 ±2.77	42.60 ±2.89	26.53 ±1.69	12.60 ±0.68
JoAPR (CVPR'24)	85.17 ±1.38	79.63 ±2.11	73.97 ±0.56	73.83 ±2.74	53.37 ±1.72	13.27 ±1.74
NLPrompt (CVPR'25)	93.80 ±1.82	93.40 ±1.81	91.77 ±2.39	81.10 ±2.14	73.63 ±3.56	55.33 ±2.66
NA-MVP (CVPR'26)	92.32 ±1.55	90.69 ±0.94	87.90 ±2.32	84.67 ±1.02	74.43 ±1.54	56.45 ±2.33
GGRP (Ours)	**93.93** ±1.18	**94.15** ±0.59	**92.32** ±1.07	**84.73** ±2.12	**74.74** ±1.48	**56.86** ±1.07

(c) DTD

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	55.60 ±1.95	47.75 ±0.80	38.07 ±1.69	29.63 ±0.70	20.53 ±0.51	11.70 ±0.89
JoAPR (CVPR'24)	52.40 ±0.34	56.63 ±0.58	53.10 ±0.94	48.93 ±2.04	40.20 ±2.24	28.26 ±1.46
NLPrompt (CVPR'25)	62.30 ±0.93	60.60 ±1.97	56.47 ±0.98	50.80 ±1.93	40.27 ±0.65	28.37 ±0.47
NA-MVP (CVPR'26)	64.36 ±1.07	62.48 ±1.02	57.32 ±2.28	50.97 ±0.29	45.48 ±2.55	35.39 ±1.02
GGRP (Ours)	**64.88** ±0.27	**63.03** ±2.46	**57.44** ±0.40	**51.08** ±0.72	**46.40** ±2.35	**36.88** ±0.99

(d) EuroSAT

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	76.00 ±0.44	66.27 ±2.82	53.83 ±3.43	41.17 ±2.11	28.00 ±1.72	17.43 ±0.60
JoAPR (CVPR'24)	69.37 ±0.66	67.30 ±0.64	59.40 ±0.80	47.60 ±0.68	33.93 ±1.27	17.50 ±2.03
NLPrompt (CVPR'25)	80.13 ±1.28	77.13 ±1.48	71.43 ±2.83	54.30 ±1.51	**66.33** ±1.06	32.73 ±0.47
NA-MVP (CVPR'26)	**81.38** ±1.51	77.07 ±1.45	70.07 ±0.86	60.91 ±1.06	48.11 ±1.73	34.55 ±1.87
GGRP (Ours)	80.63 ±5.49	**77.34** ±10.00	**71.56** ±2.80	**65.05** ±2.36	49.09 ±1.50	**34.68** ±1.41

(e) OxfordPets

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	76.10 ±0.38	66.20 ±2.09	52.53 ±1.54	38.73 ±1.69	26.63 ±1.66	14.90 ±0.27
JoAPR (CVPR'24)	82.90 ±1.02	83.40 ±1.64	79.07 ±1.96	75.84 ±2.38	52.74 ±2.27	43.57 ±1.20
NLPrompt (CVPR'25)	86.00 ±0.31	84.97 ±0.49	82.40 ±2.56	77.53 ±3.64	66.33 ±2.89	48.60 ±3.40
NA-MVP (CVPR'26)	**88.90** ±1.66	87.00 ±2.36	**83.87** ±1.70	78.76 ±1.12	72.13 ±2.48	63.40 ±0.55
GGRP (Ours)	**88.90** ±0.61	**87.40** ±1.32	83.73 ±3.09	**80.22** ±2.55	**72.97** ±0.44	**64.80** ±1.55

(f) StanfordCars

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	65.77 ±0.56	57.13 ±0.92	46.23 ±1.56	33.73 ±1.13	22.37 ±1.38	12.80 ±0.57
JoAPR (CVPR'24)	66.47 ±1.59	61.70 ±0.77	51.50 ±2.45	42.03 ±1.12	30.80 ±1.14	22.97 ±1.31
NLPrompt (CVPR'25)	69.77 ±0.22	67.53 ±0.40	64.23 ±0.63	59.03 ±0.54	50.90 ±1.26	39.50 ±2.41
NA-MVP (CVPR'26)	66.62 ±1.73	65.61 ±1.18	63.37 ±2.32	59.64 ±1.11	**57.45** ±0.39	**53.37** ±0.45
GGRP (Ours)	**70.52** ±0.16	**68.60** ±1.06	**64.45** ±1.20	**59.86** ±2.52	56.07 ±0.76	49.96 ±2.02

(g) UCF101

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	67.23 ±0.94	58.07 ±1.46	46.47 ±0.77	34.43 ±1.22	23.67 ±0.98	13.17 ±0.70
JoAPR (CVPR'24)	72.07 ±1.63	69.80 ±1.04	64.10 ±2.19	59.17 ±1.07	56.07 ±2.57	47.46 ±1.60
NLPrompt (CVPR'25)	74.90 ±1.03	73.53 ±1.34	71.03 ±1.47	65.97 ±0.61	58.97 ±0.99	49.27 ±2.10
NA-MVP (CVPR'26)	75.37 ±2.03	73.75 ±1.65	**72.15** ±0.44	**67.85** ±1.63	**63.41** ±1.37	**55.96** ±2.29
GGRP (Ours)	**75.50** ±1.99	**74.28** ±2.36	71.19 ±1.24	66.87 ±0.91	60.83 ±0.93	52.27 ±1.92

(h) Caltech101

Method	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp (IJCV'22)	84.93 ±0.81	75.23 ±0.73	62.87 ±0.69	49.43 ±1.58	33.57 ±1.33	20.33 ±1.93
JoAPR (CVPR'24)	90.30 ±0.91	89.30 ±1.04	88.30 ±0.38	88.73 ±1.54	85.80 ±2.18	81.90 ±0.99
NLPrompt (CVPR'25)	91.60 ±0.35	91.17 ±0.49	90.20 ±0.57	89.27 ±0.70	86.17 ±0.43	81.07 ±2.24
NA-MVP (CVPR'26)	**91.96** ±1.27	**91.50** ±0.37	89.81 ±0.28	87.69 ±1.51	85.19 ±1.51	81.24 ±1.98
GGRP (Ours)	91.24 ±0.65	91.40 ±0.47	**90.65** ±0.70	**89.70** ±0.81	**86.80** ±0.71	**82.40** ±0.81