PREVENTING COLLAPSE IN LONG-HORIZON CONTINUAL TEST-TIME ADAPTATION FOR OPEN-VOCABULARY SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

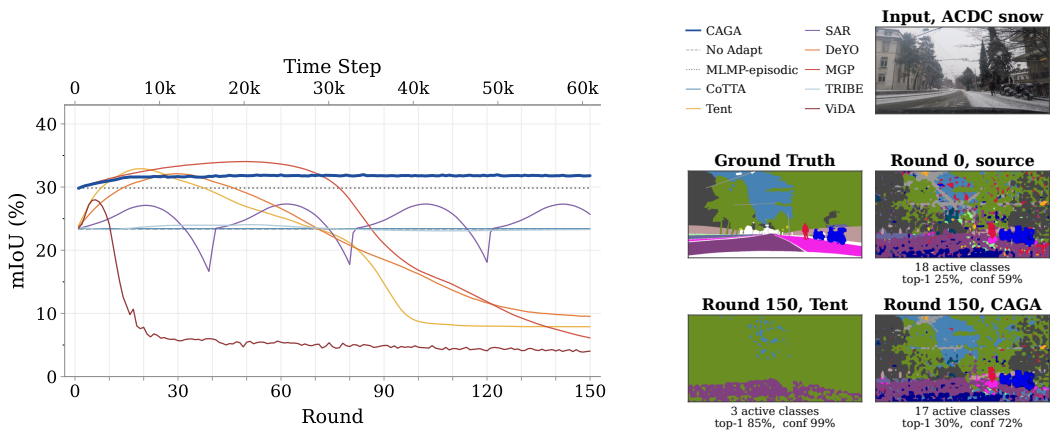
Open-vocabulary semantic segmentation (OVSS) assigns pixel-level labels from free-form text classes but, like other dense prediction models, degrades under distribution shift. Continual test-time adaptation (CTTA) addresses non-stationary shifts by continuously adapting models on unlabeled test streams. However, most existing CTTA methods are designed for closed-set classifiers with fixed prediction heads, and thus do not well match to the open-vocabulary prediction structure. We stress-test CTTA for OVSS over 150 adaptation rounds, an order of magnitude longer than common evaluation protocols. At this horizon, we uncover an adaptation–stability trade-off across existing methods: those that achieve substantial adaptation gains eventually collapse, whereas those that remain stable provide limited improvement. Although episodic resetting prevents collapse, it sacrifices the ability to accumulate useful knowledge over time. To address this, we identify two label-free signals that track collapse progression from the perspectives of optimization dynamics and prediction distribution: gradient-norm trend and class-marginal entropy. Building on these signals, we propose CAGA (Collapse-Aware Gated Anchoring), which preserves beneficial adaptation while selectively restoring the model to a previously reliable state as collapse emerges. CAGA mitigates the adaptation–stability trade-off, consistently preventing long-horizon collapse while sustaining adaptation gains across 8 benchmarks and 14 OVSS backbones. The code will be released upon acceptance.

1 INTRODUCTION

Vision-language models (VLMs), such as CLIP (Radford et al., 2021), have enabled open-vocabulary recognition by aligning visual representations with natural language. Open-vocabulary semantic segmentation (OVSS) extends this capability to dense prediction, allowing categories to be specified at inference time rather than restricted to a fixed label set (Li et al., 2022; Ghiasi et al., 2022; Rao et al., 2022; Cho et al., 2024). Despite this flexibility, OVSS remains vulnerable to distribution shifts such as adverse weather and image corruptions (Kamann & Rother, 2020; Sakaridis et al., 2021; Choi et al., 2021). Test-time adaptation (TTA) mitigates such shifts by adapting pretrained models online using unlabeled target data (Wang et al., 2021; Lee et al., 2024), and recent work has extended this paradigm to OVSS (Noori et al., 2025).

In realistic deployments, however, test distributions often evolve continuously instead of remaining fixed. Continual test-time adaptation (CTTA) addresses this setting by continually updating the model on unlabeled test streams (Wang et al., 2022). Most CTTA methods are developed for closed-set recognition with fixed prediction heads (Yuan et al., 2023; Liu et al., 2024), whereas OVSS relies on compatibility between adapted visual features and language-derived semantic representations (Rao et al., 2022; Cho et al., 2024). Continual visual updates may therefore perturb the pretrained cross-modal structure required for open-vocabulary recognition. Very recent attempts have explored CTTA for OVSS, but they either evaluate over relatively short adaptation horizons (Doloriel et al., 2026) or rely on additional pretrained models for stabilization (Luo et al., 2026), leaving the long-horizon stability of self-contained adaptation insufficiently understood.

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.



(a) Long-horizon performance on the ACDC dataset.

(b) Illustration of prediction collapse.

Figure 1: **Long-horizon collapse in CTTA for OVSS.** Across a broad range of methods, we observe an adaptation–stability trade-off: methods that achieve substantial adaptation gains eventually collapse, degenerating into single-class predictions, whereas those that remain stable provide only limited improvement. In contrast, the proposed CAGA (the thick blue curve) preserves adaptation gains and semantic diversity over long horizons.

We stress-test CTTA for OVSS over a longer horizon, extending evaluation to 150 rounds. As illustrated in Figure 1, improvements observed early in adaptation often fail to persist and can eventually reverse into severe performance collapse. Across a broad range of methods, we observe an adaptation–stability trade-off: methods that achieve substantial adaptation gains eventually suffer severe performance collapse in our long-horizon evaluation (Wang et al., 2021; Liu et al., 2024; Zhang et al., 2025; Doloriel et al., 2026), whereas more stable methods provide only limited adaptation improvement (Wang et al., 2022; Niu et al., 2023; Su et al., 2024; Niu et al., 2025; Ma et al., 2025). Episodic adaptation avoids cumulative collapse by resetting after each sample, but consequently discards previously acquired updates and cannot accumulate knowledge over time (Noori et al., 2025). This exposes a central challenge for continual OVSS adaptation: preserving useful adaptation without sacrificing long-horizon stability.

We analyze this challenge from two complementary perspectives: optimization dynamics and prediction distribution. An upward gradient-norm trend reflects emerging optimization instability, while decreasing class-marginal entropy indicates increasing concentration of predictions on a small subset of classes. Both signals are directly observable during unlabeled adaptation and together capture different stages of collapse progression.

Building on this analysis, we propose CAGA (Collapse-Aware Gated Anchoring), a collapse-aware restoration strategy that monitors these two label-free signals and selectively anchors the model toward reliable states encountered earlier in the stream. CAGA adaptively adjusts both the restoration target and strength according to the observed drift, preserving beneficial updates while suppressing accumulated instability. Extensive experiments across 8 benchmarks and 14 OVSS backbones show that CAGA mitigates the adaptation–stability trade-off, consistently preventing long-horizon collapse while retaining substantial adaptation gains. Our contributions are summarized as follows:

- We identify long-horizon collapse in CTTA for OVSS, revealing an adaptation–stability trade-off.
- We identify two complementary label-free signals that track collapse progression: gradient-norm trend and class-marginal entropy.
- We propose CAGA, which prevents long-horizon collapse while preserving adaptation gains across 8 benchmarks and 14 OVSS backbones.

2 Related Work

Open-vocabulary semantic segmentation. OVSS performs dense prediction while allowing categories to be specified at test time. Most existing OVSS methods build on pretrained CLIP (Radford

et al., 2021). MaskCLIP (Zhou et al., 2022), SCLIP (Wang et al., 2024), and ClearCLIP (Lan et al., 2024) expose or refine spatial representations without segmentation-specific training. NACLIP (Hajimiri et al., 2025) further improves localization through neighborhood-aware attention. Like other vision models, OVSS methods remain vulnerable to distribution shifts.

Test-time adaptation. TTA commonly adapts models through entropy minimization (Wang et al., 2021), with subsequent work improving sample reliability and controlling overconfidence (Niu et al., 2022; Lee et al., 2024; Ma et al., 2025; Zhang et al., 2025), or using self-distillation (Niu et al., 2025). CLIP-based methods explore prompt optimization, weight averaging, and feature caches (Shu et al., 2022; Osowiechi et al., 2024; Karmanov et al., 2024). Recently, MLMP (Noori et al., 2025) explores TTA for OVSS by combining multi-level features with multi-prompt entropy minimization. However, its episodic protocol prevents knowledge accumulation over time.

Continual test-time adaptation. Retaining updates across a stream introduces cumulative drift. Existing CTTA methods stabilize adaptation through teacher supervision, memory, or adapters (Yuan et al., 2023; Su et al., 2024; Liu et al., 2024), recovery toward pretrained parameters (Niu et al., 2023; Wang et al., 2022), or gradient-subspace constraints evaluated under closed-set prediction (Duan et al., 2025; Xiong et al., 2026). Very recent efforts have extended CTTA to OVSS, but existing approaches either evaluate diversity regularization and source anchoring over only three to five passes (Doloriel et al., 2026) or rely on an auxiliary pretrained segmentation model for pseudo-label refinement (Luo et al., 2026). In contrast, the proposed CAGA maintains stability over prolonged horizons using only label-free signals available during adaptation.

3 Analysis

We observe that entropy-minimization-based TTA can initially improve performance yet eventually collapse over a long horizon (see Figure 1). We analyze this behavior from two complementary perspectives: optimization dynamics and prediction distribution, which motivate the two label-free signals used by CAGA.

3.1 Why Does entropy Minimization Collapse?

Let $\mathbf{p}_m = (p_{m,1}, \ldots, p_{m,C}) \in \mathbb{R}^C$ denote the class-probability vector of the m-th prediction among M predictions, and let $\bar{\mathbf{p}} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{p}_m$ denote their class marginal. The entropy of prediction m is $H(\mathbf{p}_m) = -\sum_{c=1}^{C} p_{m,c} \log p_{m,c}$. Let S be uniformly distributed over the M predictions, and let $Y \in \{1, \ldots, C\}$ be a class-valued random variable such that $\Pr(Y = c \mid S = m) = p_{m,c}$. Their mutual information decomposes as:

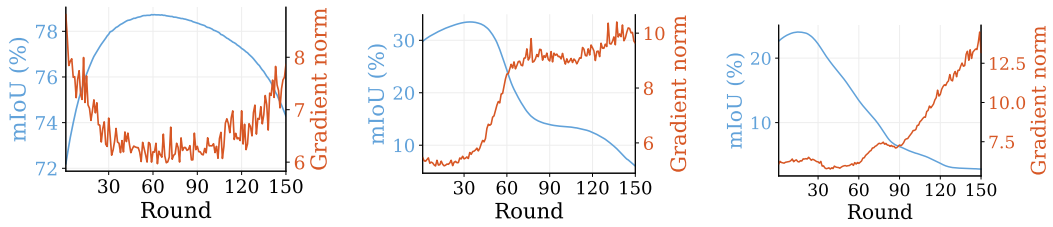
$$I(S;Y) = H(\bar{\mathbf{p}}) - \frac{1}{M}\sum_{m=1}^{M} H(\mathbf{p}_m). \tag{1}$$

Eq. (1) separates marginal class diversity from per-prediction uncertainty. Standard entropy minimization reduces only the latter, allowing deterministic solutions with low marginal diversity, including single-class collapse. In OVSS, this tendency can be amplified because all spatial predictions are coupled through the shared adapted visual representation. Correlated preferences toward the same classes can therefore reinforce a common update bias, progressively concentrating predictions onto a small subset of categories. As shown in Figure 1(b), Tent (Wang et al., 2021) predicts most pixels as belonging to the vegetation class at round 150.

Recent efforts mitigate this degeneracy with marginal-diversity regularization (Doloriel et al., 2026). While this can delay collapse, we show in Appendix A1 that, for any finite regularization weight, its corrective gradients vanish as predictions approach a common one-hot solution. This motivates CAGA to intervene directly in parameter space rather than relying solely on objective regularization.

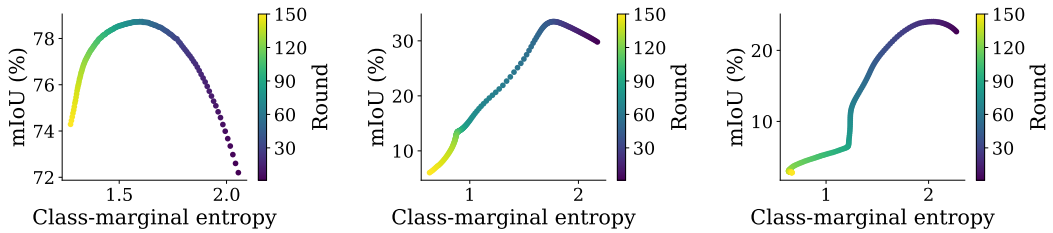
3.2 Optimization Signal: Gradient-Norm Trend

Recent analysis shows that gradients induced by confident mispredictions can exhibit persistent directional leakage into a knowledge-preserving subspace, whose accumulation can progressively



(a) VOC20-C ($r = -0.92$, $\rho = -0.84$) (b) ACDC ($r = -0.95$, $\rho = -0.90$) (c) Cityscapes-C ($r = -0.78$, $\rho = -0.88$)

Figure 2: **Gradient-norm trend tracks long-horizon performance degradation.** Across datasets, gradient norm decreases during early adaptation and rises as performance deteriorates, with strong negative correlation to mIoU. r and ρ denote Pearson and Spearman correlations, respectively.



(a) VOC20-C (b) ACDC (c) Cityscapes-C

Figure 3: **Class-marginal entropy tracks prediction collapse.** Across datasets, peak segmentation performance occurs at different absolute entropy levels, whereas post-peak degradation is consistently accompanied by decreasing class-marginal entropy. Colors indicate adaptation rounds.

erode pretrained representations (Xiong et al., 2026). Motivated by this observation, we model the adaptation gradient within a local monitoring window $\mathcal{W}_k$ as:

$$g_t = \mu_k + \xi_t, \qquad t \in \mathcal{W}_k, \tag{2}$$

where μ_k denotes a coherent component induced by persistent gradient leakage, and ξ_t captures zero-mean residual variation with $\mathbb{E}[\xi_t] = 0$. This decomposition gives:

$$\mathbb{E}\big[\|g_t\|_2^2\big] = \mathbb{E}\big[\|\xi_t\|_2^2\big] + \|\mu_k\|_2^2. \tag{3}$$

Hence, for comparable residual variation, stronger persistent leakage increases the expected squared gradient norm. Moreover, Jensen's inequality gives:

$$\mathbb{E}[\|g_t\|_2] \geq \|\mathbb{E}[g_t]\|_2 = \|\mu_k\|_2, \tag{4}$$

showing that a growing coherent component also raises a lower bound on the observable gradient magnitude.

Individual gradient norms fluctuate across samples, so CAGA instead monitors the temporal slope of window-averaged gradient norms. As shown in Figure 2, this slope decreases during the early stage of adaptation, when performance improves, and rises as performance begins to deteriorate, providing a label-free indicator of emerging optimization instability.

However, as collapse approaches saturation, entropy gradients vanish, as we further demonstrate in Appendix A1, making the gradient-norm signal less sensitive to further deterioration. We therefore complement it with the prediction-distribution signal introduced next.

3.3 Prediction Signal: Class-Marginal Entropy

Let $u = (1/C, \ldots, 1/C)$ denote the uniform distribution over classes, the class-marginal entropy can be written as:

$$H(\bar{\mathbf{p}}) = \log C - D_{\mathrm{KL}}(\bar{\mathbf{p}} \| u), \tag{5}$$

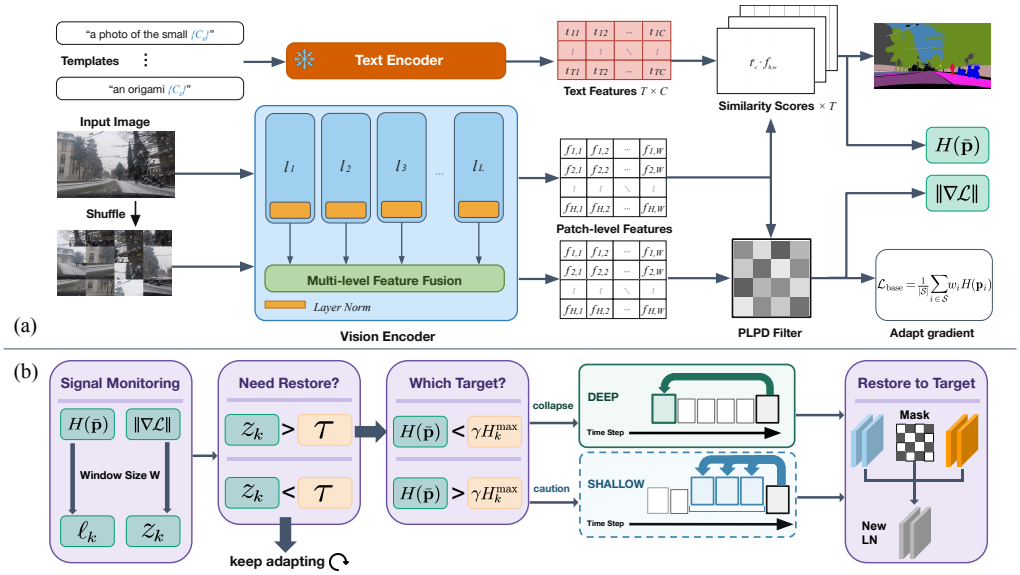


Figure 4: **Overview of the proposed CAGA.** (a) Base adaptation pipeline: multiple text prompts are encoded and combined with multi-level visual features to produce segmentation predictions, while reliability-filtered entropy minimization updates only the vision-side LayerNorm parameters. (b) Collapse-aware restoration: CAGA monitors the gradient-norm trend and class-marginal entropy over time, triggering restoration when the gradient-norm trend exceeds a threshold. The entropy state then determines whether to perform deep restoration toward the permanent anchor or shallow restoration toward a recent snapshot. The selected target is applied through stochastic parameter restoration of the adapted LayerNorm parameters.

where D_{KL} denotes the KL divergence. Hence, $H(\bar{\mathbf{p}})$ decreases as the predicted class marginal becomes more concentrated, i.e., as $D_{\mathrm{KL}}(\bar{\mathbf{p}}|u)$ increases. In the extreme case where nearly all predictions collapse to a single class, $H(\bar{\mathbf{p}})$ approaches 0. As shown in Figure 3, post-peak mIoU degradation is consistently accompanied by decreasing $H(\bar{\mathbf{p}})$. Together with the gradient trend, this provides complementary signals of optimization instability and class concentration.

4 Method

Figure 4 provides an overview of the proposed CAGA. The base adaptation continuously extracts useful information from the unlabeled stream, while a collapse-aware restoration mechanism intervenes only when the monitored dynamics indicate instability. CAGA operates entirely without target labels or auxiliary pretrained models.

4.1 Base Adaptation

Following recent TTA for OVSS designs (Noori et al., 2025) and reliability-aware entropy minimization (Lee et al., 2024), we freeze the text encoder and adapt only the LayerNorm affine parameters of the visual encoder. Multiple text templates and multi-level visual features are used to obtain dense open-vocabulary predictions (Noori et al., 2025). The overall test-time optimization pipeline is illustrated in Figure 4(a).

For each test sample, we minimize prediction entropy over reliable predictions (Lee et al., 2024). Reliability is determined using both prediction entropy and prediction-level perturbation difference (PLPD), measured from a patch-shuffled view of the input. Predictions with sufficiently low entropy and high PLPD are retained and reweighted according to their confidence and perturbation sensitivity. The resulting objective is:

$$\mathcal{L}_{\mathrm{base}} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} w_i H(\mathbf{p}_i), \tag{6}$$

where $\mathcal{S}$ denotes the retained predictions and w_i their reliability weights. If no prediction passes the filter, no adaptation update is performed.

4.2 Collapse-Aware Restoration

CAGA monitors the gradient-norm trend and class-marginal entropy (see Section 3) to determine when and where to restore, as illustrated in Figure 4(b). We divide the test stream into non-overlapping monitoring windows, each containing W consecutive batches, where W is the monitoring interval. At the end of each window, CAGA aggregates both signals over the window and stores the adapted LayerNorm parameters in a snapshot buffer $\mathcal{B}$.

Monitoring and triggering. Let k denote the monitoring-window index, and let g_k and $H_k = H(\bar{\mathbf{p}}_k)$ denote the mean gradient norm and class-marginal entropy in window k, respectively, and $H_k^{\max} = \max_{j \le k} H_j$. To capture the temporal trend of the gradient norm, we fit a linear regression to the latest L values of g_k, using the monitoring-window index as the time axis. Let s_k denote the resulting slope at window k. We normalize s_k using the median absolute deviation (MAD) of previous slopes $\mathcal{S}_{<k}$:

$$z_k = \frac{s_k}{1.4826\,\mathrm{MAD}(\mathcal{S}_{<k})}. \tag{7}$$

Here, z_k is the normalized gradient-trend score. Restoration is triggered only when $z_k > \tau$, where τ is the trend-trigger threshold; $H_k < \gamma H_k^{\max}$ identifies prediction collapse and determines the restoration depth, where γ is the entropy-ratio threshold.

Anchor maintenance and selection. CAGA maintains a permanent anchor θ_{A}, defined as a stored snapshot of the adapted LayerNorm parameters from a previously reliable window. We update the permanent anchor when $g_k < g_{\min}$ and $H_k \ge \gamma_a H_k^{\max}$, where $g_{\min}$ is the lowest gradient norm among previously eligible windows and γ_a is an entropy-health threshold. This excludes saturated low-gradient states from anchor selection. Upon triggering (i.e., when $z_k > \tau$), CAGA performs *deep restoration* toward this anchor if $H_k < \gamma H_k^{\max}$; otherwise, it performs *shallow restoration* toward a recent snapshot. Let $u_k \in [0, 1]$ be the percentile rank of z_k among recent triggered scores and d_k the number of windows since the permanent anchor was last updated. The shallow-restoration lag is:

$$c_k = \min(|\mathcal{B}| - 1, \lceil d_k u_k \rceil), \qquad \ell_k = \max(1, \lceil c_k u_k \rceil). \tag{8}$$

Hence, stronger trends and longer drift increase the rollback depth, subject to buffer capacity. This gated anchoring mechanism therefore selects either the permanent anchor or a recent snapshot as the restoration target.

Stochastic parameter restoration. Let θ_{tgt} denote the restoration target selected by the gated anchoring mechanism. Following CoTTA (Wang et al., 2022), we restore each adapted parameter independently with probability ρ toward θ_{tgt}:

$$\theta \leftarrow (1 - m) \odot \theta + m \odot \theta_{\mathrm{tgt}}, \qquad m_i \sim \mathrm{Bernoulli}(\rho). \tag{9}$$

Once triggered, restoration follows each accepted adaptation update until the next monitoring decision. Temporal aggregation supports a batch size of 1, and we use the same restoration configuration across all datasets. More details appear in Appendix A2.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate long-horizon CTTA on 8 OVSS benchmarks. The main evaluation includes ACDC (Sakaridis et al., 2021), Cityscapes-C (Cordts et al., 2016; Hendrycks & Dietterich, 2019), and VOC20-C (Everingham et al., 2010; Hendrycks & Dietterich, 2019). ACDC contains real adverse conditions including fog, night, rain, and snow, while Cityscapes-C and VOC20-C are constructed by applying the 15 ImageNet-C corruptions at severity level 5. For the corruption benchmarks, we use a fixed subset of 100 validation images and continuously revisit the corrupted stream for 150 adaptation rounds. Additional results on VOC21-C (Everingham et al., 2010; Hendrycks &

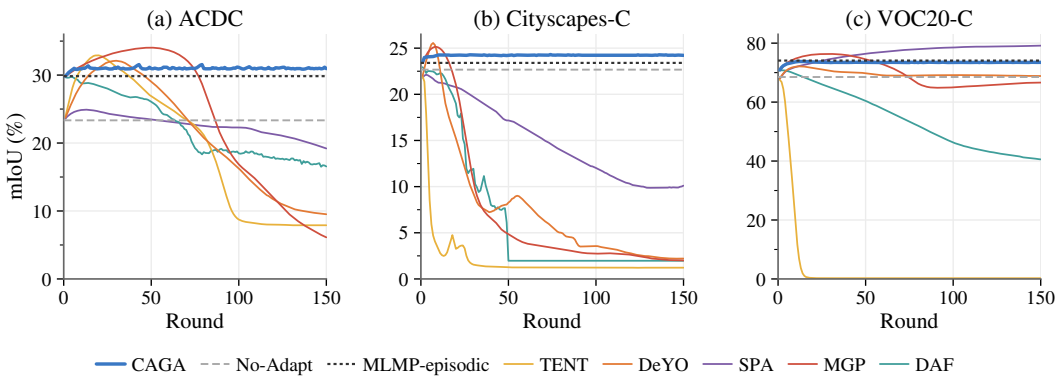


Figure 5: **Long-horizon CTTA with 150 rounds on the ACDC, Cityscapes-C, and VOC20-C benchmarks.** CAGA consistently preserves its adaptation gains throughout long horizon across all three benchmarks. Although individual baselines can outperform CAGA on a particular dataset (e.g., SPA on VOC20-C), their improvements are inconsistent across datasets.

Dietterich, 2019), Pascal-Context59-C (Mottaghi et al., 2014; Hendrycks & Dietterich, 2019), Pascal-Context60-C (Mottaghi et al., 2014; Hendrycks & Dietterich, 2019), COCO-Stuff-C (Caesar et al., 2018; Hendrycks & Dietterich, 2019), and COCO-Object-C (Lin et al., 2014; Hendrycks & Dietterich, 2019) are reported in Appendix A4.

Baselines. We compare CAGA with the non-adapted source model, MLMP episodic adaptation (Noori et al., 2025), and 10 representative TTA/CTTA methods, including TENT (Wang et al., 2021), SAR (Niu et al., 2023), TRIBE (Su et al., 2024), DeYO (Lee et al., 2024), ViDA (Liu et al., 2024), COME (Zhang et al., 2025), PTTA (Ma et al., 2025), SPA (Niu et al., 2025), MGP (Xiong et al., 2026), and the OVSS-specific DAF (Doloriel et al., 2026).

Implementation details. Unless otherwise stated, we use NA-CLIP ViT-L/14 (Hajimiri et al., 2025) as the OVSS backbone, adapt only the affine parameters of the vision encoder's LayerNorm layers, and perform online evaluation with a batch size of 1 without model resets. Following the online evaluation protocol of Tent (Wang et al., 2021), the prediction for each incoming sample is recorded before the corresponding adaptation update. For all baselines, we faithfully reproduce the original methods in the OVSS setting and make necessary adjustments when they improve performance over the default configurations. TENT, DeYO, and SPA are implemented under the same LayerNorm-only Adam framework, with full configurations provided in Appendix A3. CAGA uses a monitoring-window length of $W = 50$ batches (batch size = 1) and a gradient-trend length of $L = 10$ windows, with $\tau = 0.5$, $\gamma = 0.9$, $\gamma_a = 0.95$, and $\rho = 0.01$. These hyperparameters are shared across all datasets without per-dataset tuning. All experiments are conducted on an NVIDIA RTX PRO 6000 Blackwell GPU.

5.2 Main Results

Figure 5 compares long-horizon adaptation trajectories. On ACDC and Cityscapes-C, the proposed CAGA rapidly improves over the non-adapted source model and maintains these gains throughout 150 rounds. In contrast, most adaptation baselines eventually deteriorate after early improvement. Although SPA outperforms CAGA on VOC20-C, it suffers severe degradation on the other datasets, showing inconsistent improvement. CAGA, by contrast, remains consistently effective across benchmarks. More baselines and datasets are reported in Appendix A4.

Table 1 summarizes the performance averaged across all 8 benchmarks. CAGA achieves an average relative gain of 13.60% over the non-adapted model and remains above it on all 8 benchmarks. At the same time, it retains 99.36% of its peak performance on average, with a worst-case retention of 98.10%. Among the evaluated continual methods, CAGA is the only one that maintains a positive round-150 gain on every benchmark, demonstrating consistent effectiveness across datasets.

Table 1: **Performance comparison averaged across 8 benchmarks.** Relative Gain denotes the percentage improvement over No-Adapt at round 150. Mean, Median, Worst, and IQR (interquartile range) are computed across the eight benchmarks. Mean Rel. Gain under Whole Stream averages the relative gain over all 150 rounds. Peak Retention is the ratio between the round-150 and peak mIoU. $>$ NA counts the number of benchmarks on which the round-150 performance is higher than No-Adapt. MLMP-e is episodic and thus has no continual trajectory statistics.

	Relative Gain at Round 150 (%)				Whole Stream	Peak Retention (%)		
Method	Mean ↑	Median ↑	Worst ↑	IQR	Mean Rel. Gain ↑	Mean ↑	Worst ↑	>NA ↑
MLMP-e (NeurIPS'25)	8.26	5.49	1.60	6.53	–	–	–	**8/8**
TENT (ICLR'21)	-94.07	-99.66	-99.96	5.65	-78.42	4.57	0.04	0/8
SAR (ICLR'23)	1.22	0.48	-8.93	5.58	-1.14	94.46	86.85	5/8
TRIBE (AAAI'24)	-4.15	-5.10	-9.36	3.40	-5.16	97.93	89.64	1/8
DeYO (ICLR'24)	-10.79	8.34	-90.25	47.34	-2.74	76.51	8.65	5/8
ViDA (ICLR'24)	-97.61	-99.71	-100.00	0.44	-91.95	2.16	0.00	0/8
COME (ICLR'25)	-87.32	-99.76	-100.00	5.48	-71.71	12.49	0.00	0/8
PTTA (ICML'25)	-47.01	-39.51	-92.20	70.21	-26.83	55.34	8.49	0/8
SPA (ICML'25)	-4.86	4.05	-55.40	11.48	-2.50	90.33	45.71	5/8
MGP (ICML'26)	-30.80	-7.03	-97.12	77.82	-9.45	59.22	2.67	2/8
DAF (arXiv'26)	-40.96	-34.66	-91.31	11.42	-23.30	54.08	8.59	0/8
CAGA (Ours)	**13.60**	**10.50**	**6.40**	9.60	**13.23**	**99.36**	**98.10**	**8/8**

5.3 Ablation Study

Effect of stochastic parameter restoration. Figure 6(a) examines the parameter restoration probability ρ. Disabling stochastic restoration ($\rho = 0$) causes severe long-horizon collapse, whereas all tested nonzero probabilities remain stable and achieve similar round-150 performance (30.95–31.20 mIoU). Thus, restoration is important for long-horizon stability, while CAGA is relatively insensitive to its precise probability within the tested range.

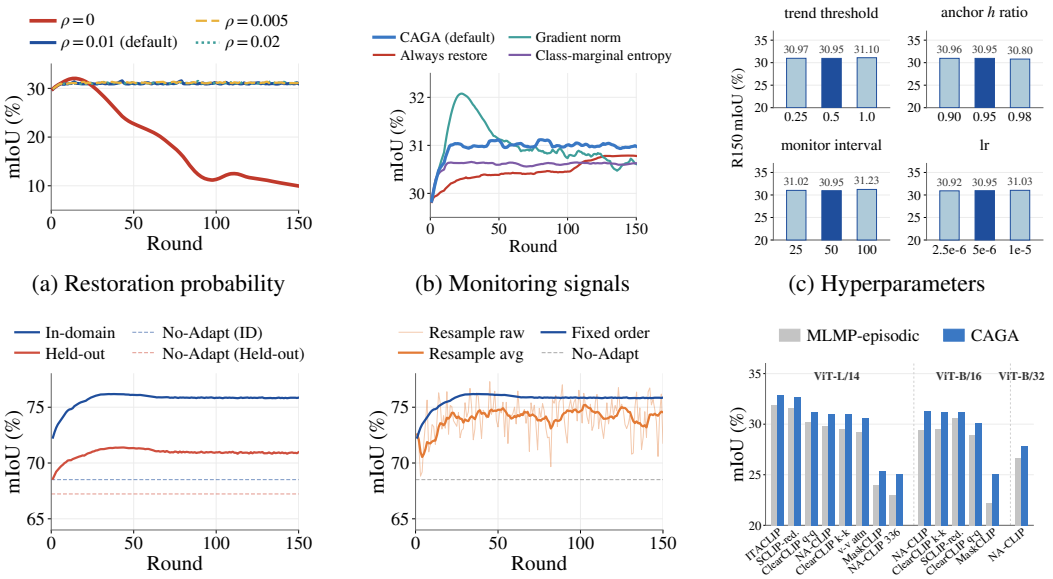
Contribution of the monitoring signals. We next isolate the roles of gradient-norm trend and class-marginal entropy. As shown in Figure 6(b), always restoring suppresses useful adaptation, while monitoring either gradient-norm trend or class-marginal entropy alone yields less favorable long-horizon behavior than their combination. This suggests that the two signals play complementary roles in controlling restoration while preserving long-horizon adaptation gains.

Hyperparameter sensitivity. We vary the trend threshold τ, anchor entropy ratio γ_a, monitoring interval W, and learning rate one at a time while keeping all other settings fixed. Figure 6(c) shows that all tested configurations remain within 30.80–31.23 mIoU at round 150, with no long-horizon collapse. CAGA is therefore robust to moderate variations of these hyperparameters.

Robustness beyond the fixed test stream. The main experiments repeatedly adapt to a fixed set of images and corruption domains, raising two questions: whether the accumulated adaptation transfers beyond the domains used for updating, and whether the observed long-horizon stability depends on repeatedly revisiting the same images. We examine these on a five-corruption version of VOC20-C.

To evaluate cross-domain transfer, we use {snow, frost, fog, brightness, contrast} as the adaptation corruption set and a disjoint set of {Gaussian noise, shot noise, impulse noise, defocus blur, glass blur} as held-out corruptions. CAGA receives adaptation updates only from the former, while the latter is used exclusively for evaluation and never contributes gradients or parameter updates. As shown in Figure 6(d), performance on the held-out domains remains consistently above the corresponding non-adapted baseline throughout the 150-round stream. This shows that the accumulated adaptation is not restricted to the corruption domains used for updating, but also generalizes to unseen corruptions.

To test whether stability depends on repeatedly observing the same images, we replace the fixed image subset with a newly sampled set of 100 images at every adaptation round while keeping the corruption protocol unchanged. As shown in Figure 6(e), resampling introduces substantially larger round-to-round fluctuations, but performance remains above the non-adapted baseline without



(a) Restoration probability (b) Monitoring signals (c) Hyperparameters

(d) In-domain vs. held-out (e) Image resampling (f) OVSS backbones

Figure 6: **Ablation study of CAGA.** (a) On ACDC, removing stochastic parameter restoration ($\rho = 0$) causes long-horizon collapse, whereas nonzero restoration probabilities remain stable. (b) On ACDC, jointly monitoring the normalized gradient-trend score z_k and class-marginal entropy H_k provides stronger long-horizon behavior than using either signal alone or always restoring. (c) On ACDC, round-150 performance varies little across the tested trend threshold τ, anchor entropy threshold γ_a, monitoring-window length W, and learning rate. (d) On VOC20-C, adaptation on the observed corruption set also improves performance on held-out corruption domains. (e) On VOC20-C, resampling a new image subset at every round increases fluctuations but does not induce progressive long-horizon collapse. (f) On ACDC, CAGA consistently improves round-150 performance over MLMP episodic adaptation across 14 OVSS backbones.

progressive long-horizon collapse. This indicates that CAGA's stability is not solely due to repeatedly adapting to a fixed image subset.

Generalization across OVSS backbones. We evaluate whether CAGA generalizes across different backbones. Figure 6(f) compares CAGA with MLMP episodic adaptation across 14 OVSS backbones built on CLIP (Radford et al., 2021), including NA-CLIP ViT-L/14 at both 224 and 336 input resolution, NA-CLIP ViT-B/16, NA-CLIP ViT-B/32, MaskCLIP ViT-L/14 (Zhou et al., 2022), MaskCLIP ViT-B/16, SCLIP ViT-L/14 (Wang et al., 2024), SCLIP ViT-B/16, ClearCLIP ViT-L/14 (Lan et al., 2024) with both q-q and k-k attention, ClearCLIP ViT-B/16, ITACLIP ViT-L/14 (Aydın et al., 2025), and v-v attention ViT-L/14 (Li et al., 2023). CAGA achieves higher round-150 mIoU across all 14 backbones, indicating that its long-horizon advantage is not specific to a particular OVSS architecture. More robustness and generalization analyses are provided in Appendix A5.

6 Conclusion

We study long-horizon CTTA for OVSS and reveal an adaptation–stability trade-off: methods with strong short-term gains often collapse under prolonged adaptation, while stable methods provide limited improvement. We identify two complementary label-free signals that track collapse progression: gradient-norm trend and class-marginal entropy. Building on these signals, we propose CAGA, a collapse-aware restoration strategy that selectively restores the model toward previously reliable states while preserving useful adaptation. Across 8 benchmarks and 14 OVSS backbones, CAGA consistently maintains adaptation gains over 150 rounds and prevents collapse. These results highlight long-horizon stability as a critical consideration for reliable continual adaptation.

AI USE STATEMENT

In this work, we used generative AI tools to assist with formulating mathematical claims, providing ingredients for and assisting with mathematical proofs, implementing methods, and translating research content. We have not used generative AI tools to develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on methodology or experiments, conduct qualitative or thematic data analysis, or interpret experimental results. Synthetic dataset generation and dataset cleaning or reformatting did not apply to this work.

Additionally, we used generative AI tools to create or modify scientific figures, create or edit software code and other research artifacts, draft parts of the manuscript, support brainstorming and information search, and edit the manuscript for readability. All AI-assisted mathematical arguments and proofs were checked by the authors, and AI-assisted code and figures were manually reviewed and tested for correctness. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, code, figures, and other artifacts produced with the aid of AI.

ETHICS STATEMENT

This work uses existing public datasets and pretrained models and does not involve human subjects or the collection of personal or sensitive data. We do not identify any direct ethical concerns specific to the proposed method. As with other TTA/CTTA methods, deployment in safety-critical settings should be preceded by appropriate validation under the target operating conditions. All datasets and models are used in accordance with their respective licenses and intended research purposes.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the adaptation protocol, optimization settings, monitoring and restoration hyperparameters, baseline configurations, and benchmark construction in the main paper and appendix. Complete long-horizon results and additional implementation details are also included in the appendix. The source code, evaluation scripts, and configuration files required to reproduce the experiments will be released publicly upon acceptance.

REFERENCES

M. Arda Aydın, Efe Mert Çırpar, Elvin Abdinli, Gozde Unal, and Yusuf H. Sahin. ITACLIP: Boosting training-free semantic segmentation with image, text, and architectural enhancements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4113–4123, 2024. doi: 10.1109/CVPR52733.2024.00394.

Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T. Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11580–11590, 2021.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, 2016. doi: 10.1109/CVPR.2016.350.

Chandler Timm C. Doloriel, Yunbei Zhang, Sarthak Kumar Maharana, Muhammad Salman Siddiqui, Tor Kristian Stevik, Fadi Al Machot, Kristian Hovde Liland, and Habib Ullah. Towards continual test-time adaptation of vision-language models in open-vocabulary semantic segmentation, 2026.

Dexin Duan, Rui Xu, Peilin Liu, and Fei Wen. Lifelong test-time adaptation via online learning in tracked low-dimensional subspace. In *Advances in Neural Information Processing Systems*, volume 38, pp. 19024–19053. Curran Associates, Inc., 2025. doi: 10.52202/085713-0640.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2): 303–338, 2010. doi: 10.1007/s11263-009-0275-4.

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *Computer Vision – ECCV 2022*, pp. 540–557. Springer Nature Switzerland, 2022.

Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5061–5071, 2025. doi: 10.1109/WACV61041.2025.00495.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

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

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ClearCLIP: Decomposing CLIP representations for dense vision-language inference. In *Computer Vision – ECCV 2024*, pp. 143–160. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72970-6_9.

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *The Twelfth International Conference on Learning Representations*, 2024.

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022.

Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. CLIP surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

Jiaming Liu, Senqiao Yang, Peidong Jia, Renrui Zhang, Ming Lu, Yandong Guo, Wei Xue, and Shanghang Zhang. ViDA: Homeostatic visual domain adapter for continual test time adaptation. In *The Twelfth International Conference on Learning Representations*, 2024.

Fen Luo, Sen Li, and Zhanqiang Huo. Care: conflict-aware regulation and evolution for continual test-time open-vocabulary semantic segmentation. *Journal of King Saud University Computer and Information Sciences*, 38(7):799, 2026. doi: 10.1007/s44443-026-01252-x.

Jing Ma, Hanlin Li, and Xiang Xiang. PTTA: Purifying malicious samples for test-time model adaptation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 42030–42061. PMLR, 2025.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 891–898, 2014. doi: 10.1109/CVPR.2014.119.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16888–16905. PMLR, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *The Eleventh International Conference on Learning Representations*, 2023.

Shuaicheng Niu, Guohao Chen, Peilin Zhao, Tianyi Wang, Pengcheng Wu, and Zhiqi Shen. Self-bootstrapping for versatile test-time adaptation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 46611–46628. PMLR, 2025.

Mehrdad Noori, David Osowiechi, Gustavo Adolfo Vargas Hakim, Ali Bahri, Moslem Yazdanpanah, Sahar Dastani, Farzad Beizaee, Ismail Ben Ayed, and Christian Desrosiers. Test-time adaptation of vision-language models for open-vocabulary semantic segmentation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 74741–74778. Curran Associates, Inc., 2025. doi: 10.52202/085713-2510.

David Osowiechi, Mehrdad Noori, Gustavo Adolfo Vargas Hakim, Moslem Yazdanpanah, Ali Bahri, Milad Cheraghalikhani, Sahar Dastani, Farzad Beizaee, Ismail Ben Ayed, and Christian Desrosiers. WATT: Weight average test time adaptation of CLIP. In *Advances in Neural Information Processing Systems*, volume 37, pp. 48015–48044. Curran Associates, Inc., 2024. doi: 10.52202/079017-1522.

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*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18061–18070, 2022. doi: 10.1109/CVPR52688.2022.01755.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10745–10755, 2021. doi: 10.1109/ICCV48922.2021.01059.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 14274–14289. Curran Associates, Inc., 2022. doi: 10.52202/068431-1038.

Yongyi Su, Xun Xu, and Kui Jia. Towards real-world test-time adaptation: Tri-net self-training with balanced normalization. In *Thirty-Eighth AAAI Conference on Artificial Intelligence*, pp. 15126–15135. AAAI Press, 2024.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *Computer Vision – ECCV 2024*, pp. 315–332. Springer Nature Switzerland, 2024.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7191–7201, 2022. doi: 10.1109/CVPR52688.2022.00706.

Haoyu Xiong, Chengchao Wang, Zhongqiang Wang, He Huang, Qiuxia Yang, Zhengpeng Zhao, and Yuanyuan Pu. Blocking the leakage: Manifold-aware gradient projection for long-horizon test-time adaptation. In *International Conference on Machine Learning*, 2026.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15922–15932, 2023. doi: 10.1109/CVPR52729.2023.01528.

Qingyang Zhang, Yatao Bian, Xinke Kong, Peilin Zhao, and Changqing Zhang. COME: Test-time adaption by conservatively minimizing entropy. In *The Thirteenth International Conference on Learning Representations*, 2025.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In *Computer Vision – ECCV 2022*, pp. 696–712. Springer Nature Switzerland, 2022. doi: 10.1007/978-3-031-19815-1_40.

Appendix

A1 Vanishing Corrective Gradients near Saturated Collapse

We provide the derivation supporting the discussion in Section 3.1. Consider an entropy-minimization objective augmented with marginal-diversity regularization,

$$L_\lambda = \frac{1}{M}\sum_{m=1}^{M} H(\mathbf{p}_m) - \lambda H(\bar{\mathbf{p}}), \qquad \bar{\mathbf{p}} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{p}_m, \tag{10}$$

where $\mathbf{p}_m = \mathrm{softmax}(\mathbf{z}_m)$ and $\lambda < \infty$.

Proposition A.1. Suppose all predictions approach a common one-hot solution, $\mathbf{p}_m \to \mathbf{e}_k$ for every m. Then

$$\nabla_{\mathbf{z}_1,\ldots,\mathbf{z}_M} L_\lambda \to 0. \tag{11}$$

If the logit-to-parameter Jacobian remains bounded, this further implies $\nabla_\theta L_\lambda \to 0$.

Proof. For the entropy of prediction m,

$$H(\mathbf{p}_m) = -\sum_{i=1}^{C} p_{m,i} \log p_{m,i}, \tag{12}$$

the softmax Jacobian gives

$$\frac{\partial H(\mathbf{p}_m)}{\partial z_{m,j}} = -p_{m,j}\left(\log p_{m,j} + H(\mathbf{p}_m)\right). \tag{13}$$

As $\mathbf{p}_m \to \mathbf{e}_k$, we have $H(\mathbf{p}_m) \to 0$. For the dominant class $j = k$, $p_{m,k} \to 1$ and $\log p_{m,k} \to 0$. For any $j \neq k$, $p_{m,j} \to 0$ and $p_{m,j} \log p_{m,j} \to 0$. Therefore,

$$\nabla_{\mathbf{z}_m} H(\mathbf{p}_m) \to 0. \tag{14}$$

We next consider the marginal-entropy term. Since $\bar{\mathbf{p}} = M^{-1}\sum_m \mathbf{p}_m$, its derivative with respect to $z_{m,j}$ is

$$\frac{\partial H(\bar{\mathbf{p}})}{\partial z_{m,j}} = -\frac{1}{M} p_{m,j}\left(\log \bar{p}_j - \sum_{i=1}^{C} p_{m,i} \log \bar{p}_i\right). \tag{15}$$

Under the common one-hot limit, $\mathbf{p}_m \to \mathbf{e}_k$ and $\bar{\mathbf{p}} \to \mathbf{e}_k$. For any $i \neq k$, $\bar{p}_i \geq p_{m,i}/M$, and hence

$$0 \geq p_{m,i} \log \bar{p}_i \geq p_{m,i} \log p_{m,i} - p_{m,i} \log M. \tag{16}$$

Since $p_{m,i} \log p_{m,i} \to 0$ and $p_{m,i} \to 0$, the squeeze theorem gives

$$p_{m,i} \log \bar{p}_i \to 0. \tag{17}$$

For $i = k$, $p_{m,k} \to 1$ and $\bar{p}_k \to 1$, so $p_{m,k} \log \bar{p}_k \to 0$ as well. Therefore,

$$\sum_{i=1}^{C} p_{m,i} \log \bar{p}_i \to 0. \tag{18}$$

For $j = k$, $\log \bar{p}_k \to 0$, so Eq. (15) converges to zero. For $j \neq k$, both $p_{m,j} \log \bar{p}_j$ and $p_{m,j}\sum_i p_{m,i} \log \bar{p}_i$ vanish. Thus,

$$\nabla_{\mathbf{z}_m} H(\bar{\mathbf{p}}) \to 0. \tag{19}$$

Eq. (14) and (19) show that both components of L_λ have vanishing logit gradients for any finite λ, proving Eq. (11).

Finally, by the chain rule,

$$\nabla_\theta L_\lambda = \sum_{m=1}^{M} \left(\frac{\partial \mathbf{z}_m}{\partial \theta}\right)^\top \nabla_{\mathbf{z}_m} L_\lambda. \tag{20}$$

Therefore, if the logit-to-parameter Jacobians remain bounded, $\nabla_\theta L_\lambda \to 0$. □

Algorithm 1 Collapse-Aware Gated Anchoring (CAGA)

Require: Monitoring-window length W, trend length L, thresholds τ, γ, γ_a, restoration probability ρ

Initialize snapshot buffer $\mathcal{B}$ and permanent anchor θ_{A} with the source parameters
restore $\leftarrow$ False
for each incoming batch **do**
  Perform the base adaptation update if accepted
  if restore and the adaptation update is accepted **then**
    Independently restore each parameter element toward θ_{tgt} with probability ρ
  end if
  if a monitoring window of W batches is complete **then**
    Compute window statistics g_k, H_k, and gradient-trend score z_k
    Store the current parameters as a snapshot in $\mathcal{B}$
    anchor_updated $\leftarrow$ False
    if the anchor-update criterion is satisfied **then**
      Update the permanent anchor θ_{A} to the current snapshot
      anchor_updated $\leftarrow$ True
    end if
    if $z_k \leq \tau$ or anchor_updated **then**
      restore $\leftarrow$ False
    else
      restore $\leftarrow$ True
      if $H_k < \gamma H_k^{\max}$ **then**
        $\theta_{\mathrm{tgt}} \leftarrow \theta_{\mathrm{A}}$ ▷ deep restoration
      else
        Select a recent snapshot using the adaptive lag ℓ_k
        $\theta_{\mathrm{tgt}} \leftarrow$ selected snapshot ▷ shallow restoration
      end if
    end if
  end if
end for

Interpretation. Proposition A.1 does not imply that marginal-diversity regularization is ineffective or that single-class collapse is its global optimum. The diversity term changes the objective preference and can delay class concentration. Rather, once predictions become highly saturated, both the entropy and diversity terms provide increasingly weak local corrective gradients. This explains why modifying the adaptation loss alone may be insufficient after collapse has progressed, motivating the parameter-space restoration used by CAGA.

A2 Detailed CAGA Restoration Procedure

This section provides additional details of the collapse-aware restoration mechanism described in Section 4.2. Algorithm 1 summarizes the overall procedure, followed by details of gradient-trend estimation, prediction-state estimation, anchor maintenance, restoration-target selection, and stochastic parameter restoration.

A2.1 Overall procedure

CAGA makes a monitoring decision after every W batches. Once restoration is triggered, the selected restoration target remains active until the next monitoring decision.

A2.2 Gradient-trend estimation

Let g_k denote the mean gradient norm within monitoring window k. Given the latest L window-level gradient norms, CAGA fits a linear trend using the monitoring-window index as the time axis and denotes its slope by s_k.

Because the magnitude of s_k can vary substantially across datasets and adaptation objectives, we normalize it using a robust scale estimate. Let $\mathcal{S}_{<k}$ denote the history of previous slope values and define

$$\hat{\sigma}_{\mathrm{MAD},k} = c_{\mathrm{MAD}}\,\mathrm{MAD}(\mathcal{S}_{<k}), \qquad c_{\mathrm{MAD}} = \frac{1}{\Phi^{-1}(0.75)}, \tag{21}$$

where Φ^{-1} is the inverse CDF of the standard normal distribution. The constant $c_{\mathrm{MAD}} \approx 1.4826$ makes the scaled MAD consistent with the standard deviation under a Gaussian reference.

The normalized trend score is then

$$z_k = \frac{s_k}{\hat{\sigma}_{\mathrm{MAD},k}}. \tag{22}$$

Positive values indicate increasing gradient norm, while larger values correspond to stronger upward trends relative to the model's recent adaptation history. Restoration is considered only when

$$z_k > \tau, \tag{23}$$

where τ is the trend threshold.

A2.3 Prediction-state estimation

Let $\bar{\mathbf{p}}_k$ denote the class marginal accumulated within window k, and define

$$H_k = H(\bar{\mathbf{p}}_k). \tag{24}$$

CAGA maintains the historical maximum marginal entropy

$$H_k^{\max} = \max_{j \le k} H_j. \tag{25}$$

The current model is regarded as entering a concentrated prediction regime when

$$H_k < \gamma H_k^{\max}, \tag{26}$$

where γ is the relative entropy threshold. This criterion depends on relative degradation rather than the absolute entropy scale.

A2.4 Reliable-anchor update

Gradient norm alone cannot identify a reliable state because gradients may also become small after predictions have saturated. CAGA therefore updates its permanent anchor only when both optimization and prediction conditions are satisfied.

Let $g_{\min}$ denote the lowest eligible window-averaged gradient norm observed so far. The current snapshot replaces the permanent anchor when

$$g_k < g_{\min} \qquad \text{and} \qquad H_k \ge \gamma_a H_k^{\max}, \tag{27}$$

where γ_a controls the entropy-based anchor-health criterion.

The first condition favors a locally stable adaptation state, whereas the second prevents saturated low-gradient states from becoming restoration targets. Whenever the anchor is updated, the corresponding window becomes the new reference point for subsequent rollback-depth computation.

A2.5 Shallow and deep restoration

When $z_k > \tau$, CAGA determines the restoration target according to the current prediction state.

If Eq. (26) holds, CAGA performs *deep restoration* toward the permanent anchor θ_{A}. This corresponds to the regime in which class concentration has already become substantial.

Otherwise, CAGA performs *shallow restoration* toward a recent snapshot between the permanent anchor and the current model state. Let $u_k \in [0, 1]$ denote the percentile rank of the current trend score z_k among recent triggered trend scores, and let d_k denote the number of monitoring windows since the permanent anchor was last updated. CAGA first determines the allowable rollback range

$$c_k = \min(|\mathcal{B}| - 1,\ \lceil d_k u_k \rceil), \tag{28}$$

Table 2: **Optimization configurations of CAGA and the evaluated baselines.** Unless noted otherwise, the listed settings are shared across datasets.

Method	Optimizer / LR	Steps	Key settings
CAGA	Adam, 5×10^{-6}	1	$\tau = 0.5$, $\gamma = 0.9$, $\gamma_a = 0.95$, $\rho = 0.01$
TENT	Adam, 1×10^{-5}	1	entropy minimization
SAR	Adam + SAM, 1×10^{-5}	1	$\rho_{\mathrm{SAM}} = 0.05$
SPA	Adam, 1×10^{-5}	1	patch size 14, frequency mask 0.2, noise ratio 0.4
TRIBE	Adam, 1×10^{-5}	1	$\eta = 0.005$, $\gamma = 0$, $h_0 = 0.05$, $\lambda = 0.5$
DeYO	Adam, 1×10^{-5}	1	entropy margin 0.5, $e_0 = 0.4$, PLPD 0.2
ViDA	Adam, 1×10^{-5}	1	EMA 0.99, ViDA EMA 0.8, ranks 4/16
PTTA	Adam, 1×10^{-5}	1	queue 1000, one neighbor, $e_{\mathrm{margin}} = 1.198$
COME	Adam, 1×10^{-5}	1	–
MGP	Adam, 5×10^{-6}	1	buffer 32, max rank 32, residual threshold 0.75
DAF	SGD	–	$\alpha_{\mathrm{cls}} = 1$, $\alpha_{\mathrm{safs}} = 0.5$, $\lambda_{\mathrm{ent}} = 1$, $\lambda_{\mathrm{div}} = 2$
MLMP-episodic	Adam, 1×10^{-3}	1	reset after every sample, 18-layer UAML

where $\mathcal{B}$ denotes the recent-snapshot buffer. The final rollback lag is

$$\ell_k = \max(1, \lceil c_k u_k \rceil). \tag{29}$$

Thus, restoration becomes progressively deeper when either the detected gradient trend is stronger or the model has drifted for longer since the permanent anchor was updated. For deep restoration, $\theta_{\mathrm{tgt}} = \theta_{\mathrm{A}}$; for shallow restoration, θ_{tgt} is the recent snapshot selected by ℓ_k.

A2.6 Stochastic parameter restoration

After selecting the restoration target θ_{tgt}, CAGA follows the stochastic restoration mechanism of CoTTA (Wang et al., 2022). For each adapted parameter element, a binary mask is sampled as

$$m_i \sim \mathrm{Bernoulli}(\rho), \tag{30}$$

and the parameters are updated by

$$\theta \leftarrow (1 - m) \odot \theta + m \odot \theta_{\mathrm{tgt}}, \tag{31}$$

where ρ denotes the restoration probability. Restoration is applied after each accepted adaptation update while the restoration state is active. At the next monitoring decision, the trigger condition and restoration target are reevaluated.

A3 Experimental and Baseline Configurations

Shared evaluation protocol. Unless otherwise stated, all experiments use NA-CLIP ViT-L/14 with the text encoder frozen and only the affine parameters of visual LayerNorm layers adapted. We evaluate online under an evaluate-before-adapt protocol with batch size one, without resetting the model, for 150 adaptation rounds. All reported experiments use seed 0.

For ACDC, we use the full validation split with 406 images per round and traverse the four conditions in the fixed order fog, night, rain, and snow. For Cityscapes-C, VOC20-C, VOC21-C, Pascal-Context59-C, Pascal-Context60-C, COCO-Stuff-C, and COCO-Object-C, we use a fixed subset of 100 validation images and apply the 15 ImageNet-C corruptions at severity level 5, resulting in 1500 test samples per round. Unless otherwise stated, inputs are evaluated using 224×224 sliding windows with stride 112; the initial resize is 1120×560 for ACDC and Cityscapes and 224×224 for all other datasets.

CAGA uses seven text templates and multi-level visual features from 18 layers. The same CAGA configuration is used across all eight benchmarks without per-dataset tuning.

Baseline-specific details. Most baseline runs use the settings above throughout the eight benchmarks. DAF uses a batch size of 1 and learning rate 1.25×10^{-4} on ACDC, Cityscapes-C, and VOC20-C, and a batch size of 8 with learning rate 10^{-3} on the remaining five corruption benchmarks. ViDA uses learning rate 2×10^{-7} on ACDC and 10^{-5} elsewhere. The ACDC SAR run uses an

earlier argument configuration than the remaining datasets. Our MLMP-episodic evaluation uses one optimization step per sample to match the online adaptation budget of the continual methods, rather than the ten-step setting used in its default implementation.

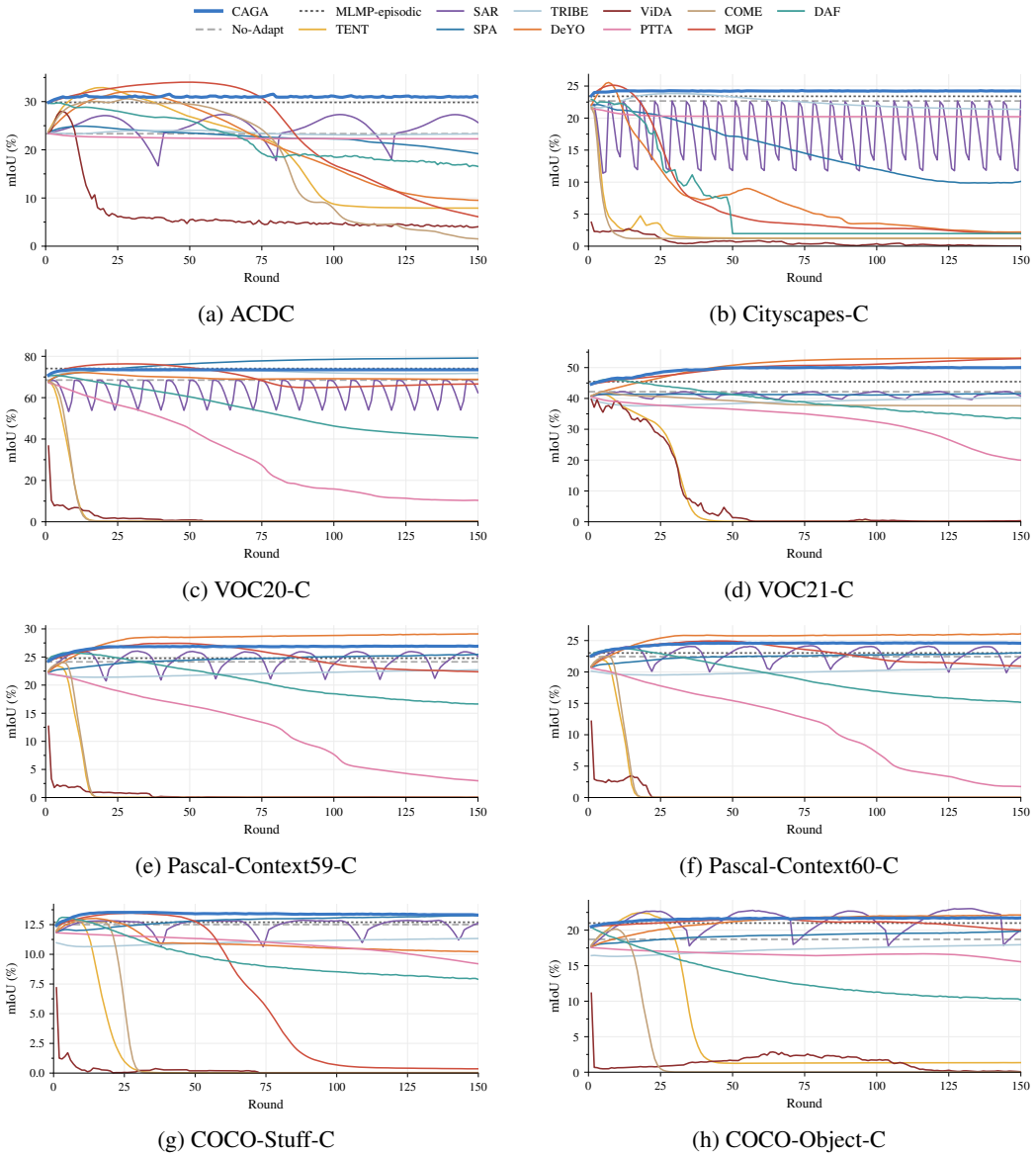
A4 Complete Long-Horizon Results

Figure 7 reports the complete long-horizon trajectories of all evaluated methods on all eight benchmarks. For ACDC, Cityscapes-C, and VOC20-C, these plots extend the representative comparisons in the main paper with the additional baselines. The remaining five benchmarks provide the full results underlying the cross-dataset statistics reported in Table 1.

Across the eight benchmarks, CAGA exhibits consistently stable long-horizon behavior and remains above the corresponding non-adapted model at round 150. Individual baselines can achieve higher performance on particular datasets, but their behavior is substantially less consistent across distribution shifts. In particular, several strongly adapting methods achieve competitive or superior peak performance before deteriorating substantially under prolonged adaptation. These complete trajectories support the cross-dataset adaptation–stability analysis in the main paper.

Qualitative analysis. Figure 8 provides a qualitative example on the ACDC snow condition at round 150. Several methods that deteriorate under prolonged adaptation also exhibit severe prediction concentration: TENT predicts only 3 classes, while ViDA and COME collapse almost entirely to a single dominant class. MGP similarly concentrates its predictions into 5 classes. In contrast, CAGA retains 17 predicted classes with a top-1 class proportion of 30%, remaining close to the source model's 18 classes and 25%. This example qualitatively illustrates the class-concentration behavior observed during long-horizon collapse and the semantic diversity preserved by CAGA.

Per-corruption performance at round 150. To provide a more fine-grained view of long-horizon performance, Tables 3–10 report the performance of all evaluated methods on each individual domain or corruption at round 150. For ACDC, the rows correspond to the four real adverse conditions, while the remaining benchmarks report the 15 ImageNet-C corruptions at severity level 5. These results complement the dataset-level averages by showing how the long-horizon behavior varies across individual distribution shifts.



(a) ACDC (b) Cityscapes-C

(c) VOC20-C (d) VOC21-C

(e) Pascal-Context59-C (f) Pascal-Context60-C

(g) COCO-Stuff-C (h) COCO-Object-C

Figure 7: **Complete long-horizon results on all eight OVSS benchmarks.** We report mIoU over 150 adaptation rounds for all evaluated methods. The first three benchmarks extend the representative main-paper comparisons with additional baselines, while the remaining five provide the complete results used in the cross-dataset evaluation.

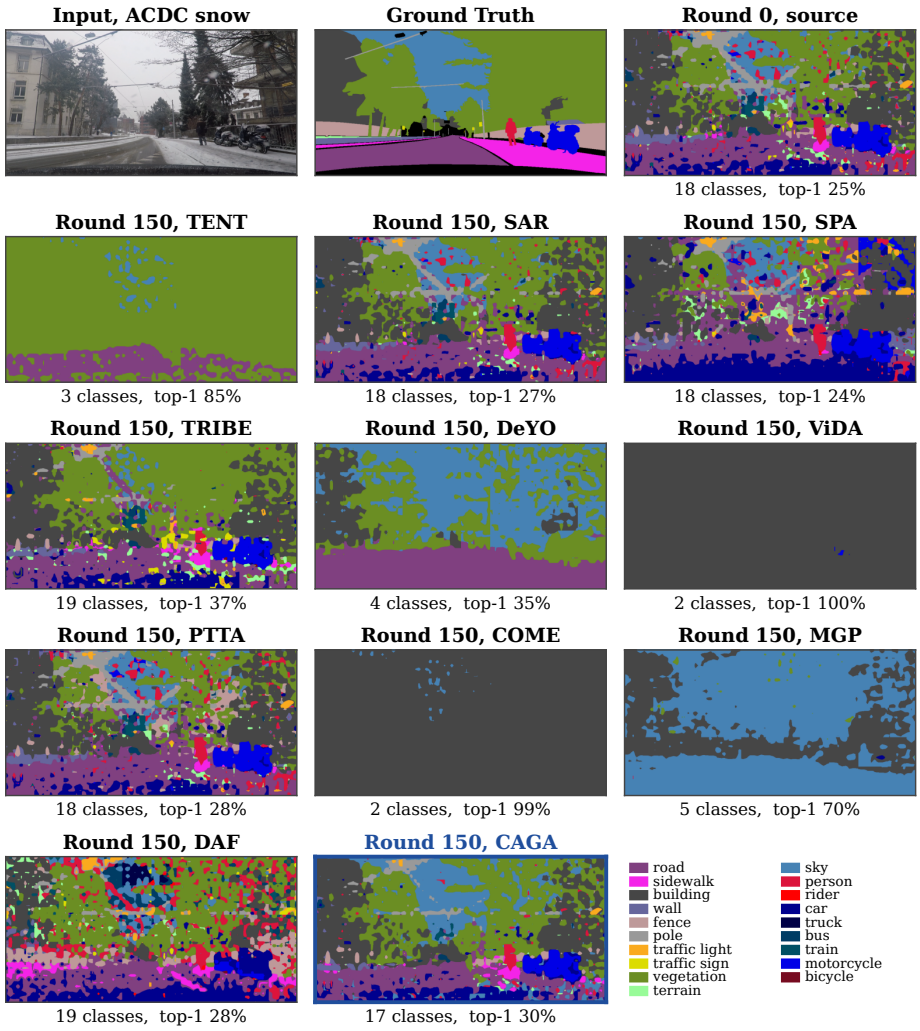


Figure 8: **Qualitative comparison on ACDC snow after 150 adaptation rounds.** We show the input image, ground truth, source prediction, and round-150 predictions of continual adaptation methods. The number of predicted classes and the proportion of pixels assigned to the most frequent class (top-1) are reported below each prediction. Several methods exhibit strong class concentration after prolonged adaptation, whereas CAGA preserves substantially greater semantic diversity.

Table 3: **Per-corruption round-150 mIoU on ACDC. Bold** and Underline denote the best and second-best results in each row, respectively. Green/Red values are above/below the corresponding No-Adapt performance; equal values are shown in black.

Corruption	No-Adapt	MLMP-episodic	TENT	SAR	SPA	TRIBE	DeYO	ViDA	PTTA	COME	MGP	DAF	**CAGA**
Fog	23.89	33.20	9.65	29.84	21.15	24.20	12.10	5.48	22.50	1.67	7.03	18.92	**34.28**
Night	22.09	24.37	5.59	20.63	16.02	21.54	3.23	2.54	21.15	1.36	4.85	13.49	**25.18**
Rain	23.85	31.67	8.53	26.37	20.57	23.81	11.50	4.49	23.13	1.56	6.74	15.52	**32.91**
Snow	23.53	30.12	7.84	25.85	19.08	23.59	11.24	3.62	22.30	1.44	5.88	18.28	**31.44**
Average	23.34	29.84	7.90	25.67	19.20	23.29	9.52	4.03	22.27	1.51	6.12	16.55	**30.95**

Table 4: **Per-corruption round-150 mIoU on Cityscapes-C. Bold** and Underline denote the best and second-best results in each row, respectively. Green/Red values are above/below the corresponding No-Adapt performance; equal values are shown in black.

Corruption	No-Adapt	MLMP-episodic	TENT	SAR	SPA	TRIBE	DeYO	ViDA	PTTA	COME	MGP	DAF	**CAGA**
Gaussian noise	14.48	14.05	1.21	**15.52**	5.17	14.09	2.39	0.01	13.89	1.18	2.03	1.97	14.73
Shot noise	19.17	19.23	1.21	**19.96**	6.36	18.37	2.47	0.01	18.20	1.18	2.02	1.97	19.85
Impulse noise	14.75	14.43	1.21	**15.79**	5.84	14.74	2.40	0.01	14.26	1.18	2.00	1.97	15.24
Defocus blur	25.21	26.61	1.22	24.77	10.83	23.98	2.14	0.01	21.89	1.18	1.73	1.97	**27.48**
Glass blur	26.50	27.79	1.20	26.33	11.15	24.83	2.11	0.01	23.17	1.18	1.70	1.97	**29.06**
Motion blur	25.94	27.37	1.21	25.88	12.82	25.21	2.10	0.01	22.22	1.18	1.52	1.97	**28.62**
Zoom blur	14.22	13.71	1.21	**15.00**	6.04	13.42	2.25	0.01	13.55	1.18	1.31	1.97	14.12
Snow	20.82	22.28	1.22	20.71	8.16	20.11	2.34	0.01	19.07	1.18	1.98	1.97	**22.56**
Frost	**16.91**	16.20	1.21	16.62	4.44	13.94	2.23	0.01	14.40	1.18	1.68	1.97	16.67
Fog	24.30	24.95	1.24	23.30	13.02	21.59	2.02	0.01	22.02	1.18	2.45	1.97	**26.45**
Brightness	31.01	33.76	1.24	31.20	16.28	29.96	2.19	0.01	27.75	1.18	3.04	1.97	**35.02**
Contrast	**17.78**	16.38	1.20	16.83	8.75	14.49	2.06	0.01	14.63	1.18	1.50	1.97	16.63
Elastic transform	31.54	34.02	1.23	31.76	17.23	30.50	2.07	0.01	27.94	1.18	2.75	1.97	**35.13**
Pixelate	31.60	34.11	1.24	31.64	15.14	30.54	2.05	0.01	27.93	1.18	2.22	1.97	**35.19**
JPEG compression	25.81	26.14	1.24	25.02	10.37	24.81	2.35	0.01	22.42	1.18	2.07	1.97	**26.48**
Average	22.67	23.40	1.22	22.69	10.11	21.37	2.21	0.01	20.22	1.18	2.00	1.97	**24.22**

Table 5: **Per-corruption round-150 mIoU on VOC20-C. Bold** and Underline denote the best and second-best results in each row, respectively. Green/Red values are above/below the corresponding No-Adapt performance; equal values are shown in black.

Corruption	No-Adapt	MLMP-episodic	TENT	SAR	SPA	TRIBE	DeYO	ViDA	PTTA	COME	MGP	DAF	**CAGA**
Gaussian noise	63.80	68.32	0.25	58.38	**72.12**	69.38	66.98	0.25	8.94	0.25	60.70	34.89	67.25
Shot noise	67.64	72.58	0.25	60.79	**78.27**	72.92	65.09	0.25	11.13	0.25	63.51	37.64	72.03
Impulse noise	63.81	69.62	0.25	60.19	**76.58**	71.22	65.77	0.25	11.39	0.25	62.03	35.36	69.71
Defocus blur	71.59	77.43	0.25	64.69	**77.47**	72.82	68.60	0.25	9.75	0.25	70.51	41.71	74.88
Glass blur	69.30	74.91	0.25	63.66	**79.34**	72.06	68.34	0.25	11.79	0.26	68.76	42.78	72.21
Motion blur	74.37	**80.09**	0.25	66.59	79.64	73.48	72.19	0.25	10.15	0.26	69.88	42.58	77.46
Zoom blur	55.70	61.79	0.25	54.77	**67.57**	62.91	55.66	0.25	5.39	0.28	48.59	30.53	58.89
Snow	70.27	75.30	0.25	63.27	**81.28**	73.56	70.48	0.25	10.31	0.25	67.99	43.56	74.54
Frost	63.43	71.20	0.25	58.21	**76.80**	68.37	62.83	0.25	9.68	0.25	60.74	37.29	71.68
Fog	70.29	74.60	0.25	65.35	**84.90**	72.25	71.47	0.25	10.46	0.27	69.59	43.66	76.03
Brightness	72.42	78.33	0.25	69.22	**85.00**	75.53	75.27	0.26	11.58	0.25	74.10	45.84	79.28
Contrast	71.60	76.44	0.25	62.56	**83.29**	75.25	72.06	0.25	10.97	0.26	70.21	45.30	76.58
Elastic transform	62.02	69.85	0.25	57.47	**77.93**	63.84	69.30	0.25	10.14	0.25	68.05	36.63	70.34
Pixelate	77.34	82.43	0.25	66.71	**84.52**	75.80	76.22	0.25	11.83	0.25	73.89	46.73	81.88
JPEG compression	74.11	78.42	0.25	64.04	**82.49**	76.48	72.55	0.25	11.71	0.25	70.80	43.88	79.50
Average	68.51	74.09	0.25	62.39	**79.15**	71.72	68.85	0.25	10.35	0.26	66.62	40.56	73.48

Table 6: **Per-corruption round-150 mIoU on VOC21-C. Bold** and Underline denote the best and second-best results in each row, respectively. Green/Red values are above/below the corresponding No-Adapt performance; equal values are shown in black.

Corruption	No-Adapt	MLMP-episodic	TENT	SAR	SPA	TRIBE	DeYO	ViDA	PTTA	COME	MGP	DAF	**CAGA**
Gaussian noise	39.01	42.55	0.09	38.00	38.23	35.90	**53.56**	0.32	18.06	35.91	51.32	31.53	47.13
Shot noise	39.51	43.05	0.07	38.04	41.68	35.93	**54.08**	0.41	18.79	34.65	53.45	31.86	49.10
Impulse noise	40.19	44.64	0.09	39.43	40.40	38.41	**54.32**	0.20	15.86	35.08	53.30	31.76	48.82
Defocus blur	43.40	45.96	0.07	41.79	38.32	40.62	**52.99**	0.44	21.52	37.59	50.75	35.21	50.96
Glass blur	42.83	45.44	0.07	42.50	39.68	40.09	54.68	0.30	21.89	39.09	**54.75**	34.55	50.67
Motion blur	43.74	46.20	0.11	41.70	41.05	40.76	52.78	0.28	24.88	39.26	**53.21**	34.57	50.74
Zoom blur	34.23	37.86	0.07	32.04	31.25	33.76	38.74	0.33	15.10	27.82	**39.35**	25.39	38.87
Snow	41.65	44.83	0.07	39.32	40.98	39.12	53.10	0.39	23.23	36.84	**53.57**	34.61	49.68
Frost	41.81	46.48	0.07	40.45	44.54	41.70	**52.90**	0.32	17.30	35.02	52.11	32.15	50.23
Fog	44.29	47.71	0.12	42.61	43.03	44.33	**55.29**	0.37	20.67	39.34	55.07	36.23	52.18
Brightness	44.87	48.75	0.23	45.11	44.60	44.43	55.42	0.54	21.44	41.82	**56.44**	36.66	53.77
Contrast	42.78	46.01	0.07	41.69	44.47	38.73	**54.78**	0.39	20.89	40.22	54.49	36.22	52.14
Elastic transform	41.28	45.12	0.09	41.35	45.31	41.56	**54.98**	0.22	17.60	39.63	54.90	30.70	49.75
Pixelate	47.01	49.38	0.20	43.83	44.03	44.47	54.53	0.33	22.29	41.94	**55.58**	37.09	52.93
JPEG compression	45.89	47.57	0.15	43.97	46.03	44.49	54.10	0.24	20.11	39.86	**55.33**	34.52	53.17
Average	42.17	45.44	0.10	40.79	41.57	40.29	**53.08**	0.34	19.98	37.60	52.91	33.54	50.01

Table 7: **Per-corruption round-150 mIoU on Pascal-Context59-C. Bold** and Underline denote the best and second-best results in each row, respectively. Green/Red values are above/below the corresponding No-Adapt performance; equal values are shown in black.

Corruption	No-Adapt	MLMP-episodic	TENT	SAR	SPA	TRIBE	DeYO	ViDA	PTTA	COME	MGP	DAF	**CAGA**
Gaussian noise	20.83	23.01	0.01	21.92	22.21	21.17	**25.44**	0.01	1.82	0.01	20.91	14.42	24.26
Shot noise	21.99	22.77	0.01	22.09	21.85	20.71	**26.50**	0.01	2.66	0.01	20.36	14.86	24.65
Impulse noise	22.08	23.01	0.01	21.66	22.73	20.86	**26.42**	0.01	2.20	0.02	20.24	14.75	25.15
Defocus blur	25.33	26.01	0.01	27.24	24.88	23.90	**31.03**	0.11	4.14	0.01	23.93	17.69	29.07
Glass blur	25.74	26.77	0.01	27.71	26.27	23.54	**31.55**	0.08	3.51	0.01	20.59	17.81	28.10
Motion blur	26.72	26.58	0.02	28.05	26.29	25.22	**30.65**	0.12	3.36	0.02	23.92	17.93	28.77
Zoom blur	20.80	21.71	0.01	21.30	22.02	19.19	**23.90**	0.01	1.62	0.01	17.34	13.43	23.42
Snow	22.35	22.36	0.01	23.78	24.57	20.81	**25.76**	0.01	2.73	0.01	21.73	15.75	25.61
Frost	22.01	22.27	0.01	22.29	22.55	20.92	**25.94**	0.01	2.76	0.01	20.09	15.31	24.23
Fog	25.05	25.53	0.01	26.59	27.53	23.05	**29.67**	0.09	3.16	0.01	24.61	17.53	27.34
Brightness	28.07	28.44	0.01	29.15	30.77	26.65	**33.29**	0.06	3.61	0.01	27.31	19.45	30.61
Contrast	23.31	24.34	0.01	24.78	26.97	20.39	**30.98**	0.07	3.30	0.01	23.05	16.92	27.10
Elastic transform	22.60	22.80	0.01	23.62	27.97	23.16	**29.79**	0.01	2.47	0.01	18.81	15.87	24.78
Pixelate	28.72	29.44	0.01	29.63	27.92	26.55	**32.86**	0.09	4.38	0.01	27.74	19.91	30.80
JPEG compression	26.49	26.70	0.01	26.88	27.26	24.33	**32.87**	0.08	2.98	0.01	24.97	18.00	29.98
Average	24.14	24.78	0.01	25.11	25.45	22.70	**29.11**	0.05	2.98	0.01	22.37	16.64	26.92

Table 8: **Per-corruption round-150 mIoU on Pascal-Context60-C. Bold** and Underline denote the best and second-best results in each row, respectively. Green/Red values are above/below the corresponding No-Adapt performance; equal values are shown in black.

Corruption	No-Adapt	MLMP-episodic	TENT	SAR	SPA	TRIBE	DeYO	ViDA	PTTA	COME	MGP	DAF	**CAGA**
Gaussian noise	19.46	21.57	0.01	20.16	19.84	19.25	**23.11**	0.01	0.85	0.01	19.67	13.36	22.37
Shot noise	20.62	21.32	0.01	20.71	20.01	18.97	**23.92**	0.01	1.43	0.01	19.14	13.47	23.07
Impulse noise	20.73	21.77	0.01	21.02	20.87	19.03	**23.82**	0.01	1.37	0.01	18.97	13.44	23.37
Defocus blur	23.54	23.94	0.01	23.02	22.54	21.98	**27.64**	0.01	2.32	0.01	21.80	16.05	26.72
Glass blur	24.08	24.97	0.01	23.26	23.54	21.38	**27.89**	0.01	2.08	0.01	20.20	16.02	25.75
Motion blur	24.98	24.78	0.01	24.14	24.02	22.85	**27.37**	0.01	2.11	0.01	22.21	16.57	26.31
Zoom blur	19.58	20.17	0.01	19.61	19.73	17.39	**20.97**	0.01	0.91	0.01	16.80	12.25	20.85
Snow	20.53	20.56	0.01	20.72	21.81	18.91	**23.97**	0.01	1.56	0.01	20.64	14.39	23.39
Frost	20.44	20.42	0.01	20.29	20.65	19.03	**23.84**	0.01	1.60	0.01	18.78	13.96	21.86
Fog	23.07	23.53	0.01	23.05	24.87	20.81	**26.55**	0.01	1.85	0.01	22.76	15.82	24.79
Brightness	25.71	26.05	0.01	25.97	27.80	23.97	**30.16**	0.01	2.12	0.01	25.23	17.60	27.68
Contrast	22.00	23.11	0.01	21.72	24.31	18.55	**27.59**	0.01	2.02	0.01	20.90	15.38	24.74
Elastic transform	21.01	21.17	0.01	22.37	25.65	21.02	**26.33**	0.01	1.62	0.01	17.39	14.55	22.80
Pixelate	26.24	27.07	0.01	25.89	24.92	23.99	**29.32**	0.04	2.61	0.01	25.39	18.10	27.85
JPEG compression	24.68	24.84	0.01	24.72	25.06	22.16	**28.55**	0.01	1.85	0.01	24.05	16.46	27.06
Average	22.44	23.02	0.01	22.44	23.04	20.62	**26.07**	0.01	1.75	0.01	20.93	15.16	24.57

Table 9: **Per-corruption round-150 mIoU on COCO-Stuff-C. Bold** and Underline denote the best and second-best results in each row, respectively. Green/Red values are above/below the corresponding No-Adapt performance; equal values are shown in black.

Corruption	No-Adapt	MLMP-episodic	TENT	SAR	SPA	TRIBE	DeYO	ViDA	PTTA	COME	MGP	DAF	**CAGA**
Gaussian noise	10.81	11.21	0.00	10.85	**12.02**	9.69	9.52	0.00	7.73	0.00	0.14	6.45	11.18
Shot noise	11.34	11.32	0.03	11.58	**12.54**	10.59	9.83	0.00	7.65	0.00	0.25	7.29	12.24
Impulse noise	11.27	11.87	0.01	11.64	**12.27**	10.27	9.13	0.00	7.75	0.00	0.25	6.84	11.47
Defocus blur	12.45	12.29	0.02	12.54	12.54	11.10	10.55	0.00	9.02	0.00	0.32	7.83	**13.99**
Glass blur	12.29	11.98	0.06	12.16	**13.72**	11.45	10.02	0.00	9.14	0.00	0.26	8.08	12.90
Motion blur	13.85	13.59	0.04	13.81	13.96	11.89	11.42	0.00	10.20	0.00	0.33	8.54	**14.41**
Zoom blur	8.83	8.95	0.00	9.07	9.39	8.90	5.84	0.00	6.07	0.00	0.20	5.30	**9.69**
Snow	12.28	12.65	0.04	12.06	12.08	10.97	9.58	0.00	8.74	0.00	0.25	7.47	**12.74**
Frost	11.26	11.64	0.04	11.94	11.92	10.57	8.58	0.00	8.44	0.00	0.29	6.89	**12.10**
Fog	13.90	14.18	0.07	13.36	14.35	12.22	11.22	0.00	10.66	0.00	0.63	9.11	**14.57**
Brightness	14.69	14.91	0.06	14.67	15.01	13.04	12.33	0.00	11.50	0.00	0.57	9.54	**15.51**
Contrast	13.00	13.28	0.08	12.92	**14.58**	11.66	10.66	0.00	9.50	0.00	0.30	8.26	13.88
Elastic transform	12.34	12.71	0.08	12.96	**14.39**	11.53	10.10	0.00	9.56	0.00	0.43	8.21	13.19
Pixelate	15.32	15.32	0.10	15.10	15.25	13.28	12.56	0.00	11.29	0.00	0.68	9.85	**16.28**
JPEG compression	13.93	14.60	0.03	14.56	14.52	12.76	11.99	0.00	10.74	0.00	0.47	8.75	**15.42**
Average	12.50	12.70	0.04	12.61	13.24	11.33	10.22	0.00	9.20	0.00	0.36	7.89	**13.30**

Table 10: **Per-corruption round-150 mIoU on COCO-Object-C. Bold** and Underline denote the best and second-best results in each row, respectively. Green/Red values are above/below the corresponding No-Adapt performance; equal values are shown in black.

Corruption	No-Adapt	MLMP-episodic	TENT	SAR	SPA	TRIBE	DeYO	ViDA	PTTA	COME	MGP	DAF	**CAGA**
Gaussian noise	16.21	**19.06**	1.17	16.77	17.16	15.40	17.45	0.02	13.00	0.02	14.91	8.83	18.48
Shot noise	16.89	18.31	1.12	17.80	20.10	15.68	**21.16**	0.01	13.38	0.02	18.24	9.55	19.92
Impulse noise	15.92	**19.14**	1.16	17.09	19.09	15.56	19.11	0.02	12.70	0.02	16.25	8.78	18.92
Defocus blur	18.63	19.99	1.45	19.71	19.31	17.29	**22.08**	0.06	14.88	0.02	20.66	10.44	21.74
Glass blur	17.93	19.49	1.40	20.02	20.21	18.42	**22.06**	0.03	16.08	0.02	21.27	10.22	21.77
Motion blur	20.42	22.01	1.32	22.09	20.66	18.61	22.62	0.04	16.88	0.04	21.70	10.87	**22.88**
Zoom blur	13.82	14.99	0.97	14.27	14.87	14.33	**16.19**	0.01	9.60	0.02	11.04	6.90	15.83
Snow	19.06	21.70	1.15	20.19	18.80	18.27	**22.45**	0.02	15.71	0.03	20.99	10.14	21.71
Frost	18.09	20.52	1.04	19.56	18.29	17.86	**22.06**	0.02	14.54	0.02	18.95	9.40	20.71
Fog	21.45	23.86	1.69	22.32	22.23	20.06	24.05	0.03	18.68	0.02	22.26	11.89	**24.29**
Brightness	21.23	23.83	1.96	22.73	22.12	20.32	**25.18**	0.04	18.79	0.02	24.03	11.62	24.36
Contrast	19.71	22.11	1.18	20.61	21.59	18.90	23.23	0.45	16.43	0.03	21.89	10.49	**23.86**
Elastic transform	18.46	21.05	1.19	20.78	20.72	18.41	**23.82**	0.15	15.66	0.02	20.37	10.15	21.75
Pixelate	22.35	24.86	1.67	23.12	21.95	19.80	**25.74**	0.14	19.73	0.02	24.97	12.14	25.05
JPEG compression	20.43	24.07	1.73	23.24	20.73	20.30	**24.77**	0.07	17.23	0.03	22.85	11.33	24.63
Average	18.71	21.00	1.35	20.02	19.86	17.95	**22.13**	0.07	15.55	0.02	20.03	10.18	21.73

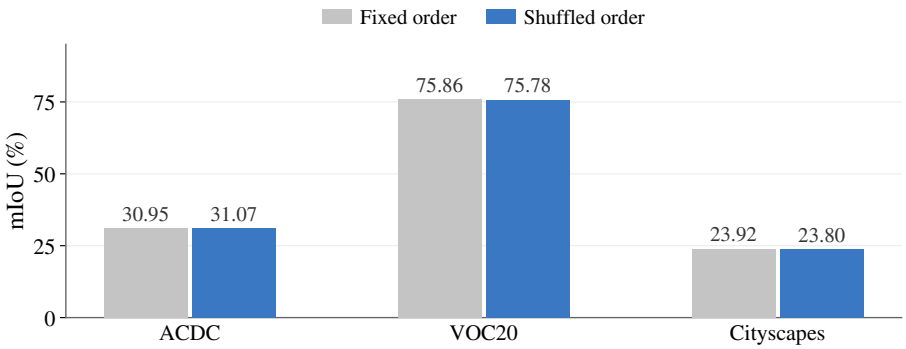


Figure 9: **Robustness to randomized domain ordering.** Changes in round-150 performance when the domain order is randomized. CAGA shows only minor sensitivity to the ordering of the test stream across VOC20-C, Cityscapes-C, and ACDC.

A5 Additional Robustness and Generalization Analyses

We provide additional analyses of CAGA under changes to the test-stream construction and beyond the open-vocabulary setting. Specifically, we examine sensitivity to domain ordering, robustness to an abrupt change in the corruption set, and extension to conventional closed-set semantic segmentation.

Randomized domain ordering. We first examine whether CAGA's long-horizon behavior depends on the fixed domain ordering used in the main experiments. We randomly shuffle the domain order while keeping the underlying samples and adaptation protocol unchanged. As shown in Figure 9, at round 150, performance changes only marginally from 75.86 to 75.78 mIoU on VOC20-C, from 23.92 to 23.80 on Cityscapes-C, and from 30.95 to 31.07 on ACDC. These results indicate that CAGA's long-horizon stability is largely insensitive to the particular ordering of domains in the test stream.

Abrupt domain-set transition. We next consider a more challenging non-stationary stream in which the corruption set changes abruptly during adaptation. For this analysis, we use {snow, frost, fog, brightness, contrast} during the first 75 rounds and switch to the disjoint set {gaussian noise, shot noise, impulse noise, defocus blur, glass blur} from round 76 onward. As shown in Figure 10, performance exhibits an immediate drop after the transition, largely reflecting the different difficulty of the new corruption set rather than adaptation collapse: CAGA remains above the corresponding No-Adapt baseline both before and after the transition and does not exhibit progressive degradation over the remaining rounds. It reaches 68.44 mIoU at round 150, compared with a peak of 76.18 before the transition. This experiment shows that the stability mechanism remains effective even when the test distribution changes abruptly after substantial adaptation has already accumulated.

Extension to closed-set semantic segmentation. Finally, we evaluate whether the same restoration mechanism can be applied to a conventional closed-set segmentation model. We conduct a 150-round continual adaptation experiment on SegFormer under the same long-horizon evaluation setting. As shown in Figure 11, unlike the open-vocabulary models studied in the main paper, the closed-set model exhibits little degradation over the stream: the non-adapted model achieves 50.62 mIoU, and the evaluated adaptation variants remain within a narrow range around this baseline. CAGA likewise remains stable throughout the 150 rounds, but does not provide a clear improvement over the non-adapted model. This result suggests that CAGA is compatible with closed-set segmentation, while the severe adaptation–stability trade-off studied in this work is much more pronounced in the evaluated OVSS setting.

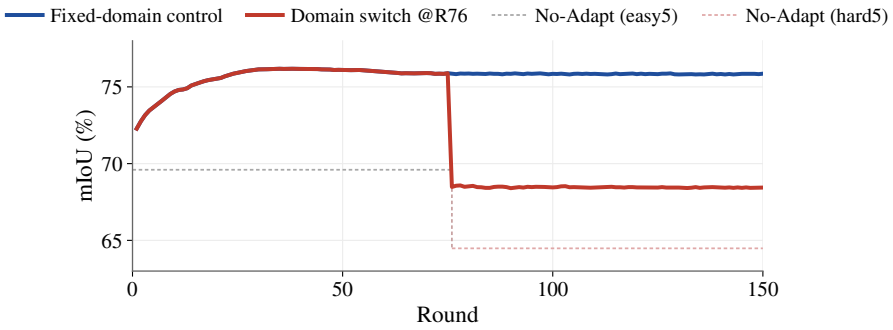


Figure 10: **Robustness to an abrupt domain-set transition.** CAGA is adapted on one five-corruption VOC20-C stream for the first 75 rounds and then exposed to a disjoint corruption set. Performance decreases after the transition to the new distribution but does not exhibit progressive long-horizon collapse.

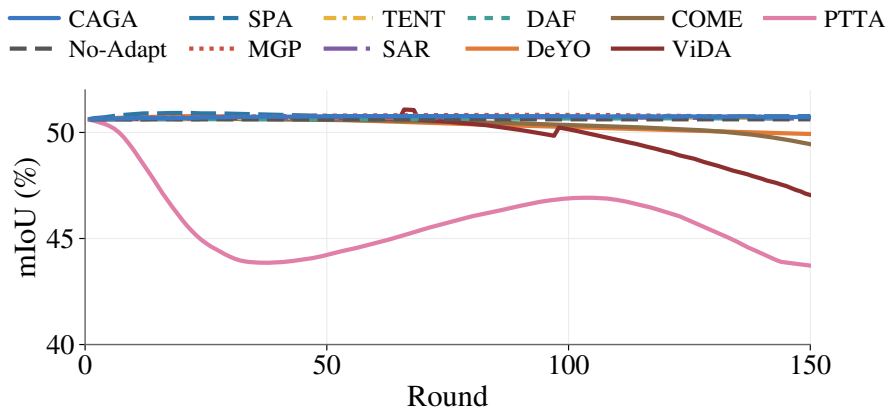


Figure 11: **Extension to closed-set semantic segmentation.** Long-horizon continual adaptation results on SegFormer over 150 rounds. CAGA remains stable throughout the stream, while providing no clear improvement over the non-adapted model.