Measurement Plasticity: Sensor-Level Adaptation for Vision–Language Models

Boyeong Im [1] **Wooseok Lee** [1] **Yoojin Kwon** [1] **Hyung-Sin Kim** [1]

Abstract

We propose Multi-View Physical-prompt (*MVP*) for Test-Time Adaptation (TTA), a forward-only framework that moves TTA from tokens to photons by treating the camera exposure triangle (i.e., ISO, shutter speed, and aperture) as physical prompts. At inference, *MVP* acquires selected multiple physical views using a source-affinity score, evaluates digitally augmented variants of each retained view and filters the lowest-entropy predictions, and aggregates predictions with hard voting. This selection-then-vote design is simple, calibration-friendly, and requires no gradients or model modifications. On ImageNet-ES and ImageNet-ES-Diverse, *MVP* outperforms digital-only TTA on both Auto-Exposure and a combination with conventional sensor control. *MVP* remains effective under reduced parameter candidates that lower capture latency, demonstrating its practicality.

1. Introduction

Foundation models (FMs) are increasingly deployed where inputs differ from the static web corpora on which they are trained, motivating training-free continual adaptation to changing inputs. Most existing approaches perform adaptation inside the model via weights updates, adapters, prompt tuning, or retrieval. (Wang et al.; Niu et al.; Hu et al.; Shu et al., 2022; Karmanov et al., 2024). In sensor-mediated environments, however, there is another natural locus of plasticity: the measurement process. Vision–language models (VLMs) deployed in the physical world do not receive images from the web. Their inputs are produced by sensors, and camera settings such as ISO, shutter speed, and aperture determine what information reaches the encoder. When a scene is under-exposed, saturated, or noisy, downstream adaptation can only operate on an already degraded measurement. This suggests a complementary question for continual adaptation: instead of adapting the model to the input, can we adapt how the input is measured?

[1]Graduate School of Data Science, Seoul National University. Correspondence to: Hyung-Sin Kim <hyungkim@snu.ac.kr>.

Presented at the ICML 2026 Workshop "Continual Adaptation at Scale: Towards Sustainable AI". Copyright 2026 by the author(s).

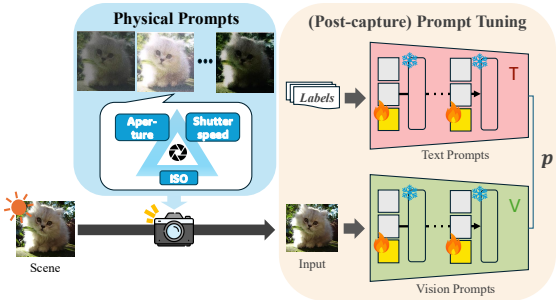


Figure 1. **Conceptual Figure of the Physical Prompts.** In the causal chain of scene → measurement → representation, physical prompts are controlled by sensor parameters (i.e., ISO, shutter speed) to minimize irrecoverable loss during measurement, which cannot be compensated from post-capture images.

We study this question through measurement-level adaptation for VLMs. Our framework, Multi-View Physical-prompt for Test-Time Adaptation (*MVP*), treats the camera exposure triangle as a set of physical prompts (Figure 1). At inference time, *MVP* acquires multiple physical views of a scene under different sensor settings, selects views whose image features are closest to source-domain statistics, filters uncertain augmented views, and aggregates predictions using zero-temperature hard voting. *MVP* is forward-only and requires no gradient updates or model modifications, making it a lightweight alternative to model-side adaptation.

This perspective reframes physical prompting as input-side continual adaptation. Instead of changing the VLM's parameters, *MVP* changes the distribution of measurements presented to the frozen model, shifting inputs closer to the regime where their representations are reliable. In practice, this multi-view framework is uniquely suited for static, precision-critical settings, such as CCTV monitoring, autonomous inspection, and computer-assisted surgery. Experiments on ImageNet-ES (Baek et al., 2024) and ImageNet-ES-Diverse (Baek et al., 2025) show that *MVP* improves robustness over digital-only Test-Time Adaptation (TTA) on Auto-Exposure captures and over pipelines that combine conventional sensor control (Baek et al., 2025) with TTA, while remaining effective under reduced candidate sets that lower capture latency. These results suggest that sustainable adaptation for multi-modal FMs should consider not only where plasticity resides inside the model, but also whether plasticity can occur at the sensor–model interface.

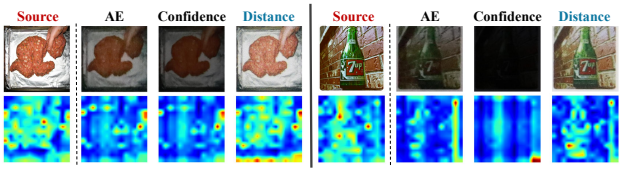


Figure 2. **Visualization of various camera parameters.** From the identical scene, we compare the attention maps of the source (ImageNet) and varying parameter samples (ImageNet-ES-Diverse).

Our contribution is to identify measurement plasticity as a first-class mechanism for continual adaptation in sensor-mediated AI. *MVP* motivates future adaptation methods that operate jointly over models, prompts, memories, and the controllable processes that produce their inputs.

2. Related Work

Post-capture VLM adaptation. VLMs (Radford et al., 2021; Jia et al., 2021) achieve strong zero-shot performance via large-scale vision–language pre-training. To improve downstream transfer, prompt tuning methods learn prompts in the language branch (Zhou et al., 2022b;a) or across both modalities (Khattak et al., 2023), but require annotated downstream data. Test-time prompt tuning removes this requirement by adapting prompts per test sample (Shu et al., 2022), with calibration-aware variants (Yoon et al., 2024; Sharifdeen et al., 2025) and memory-based extensions (Ma et al., 2023; Zhang et al., 2024; Xiao et al., 2025). PromptAlign (Abdul Samadh et al., 2023) aligns test visual-token statistics with source-domain statistics, which motivates our use of source affinity for view selection. Forward-only TTA methods reduce test-time cost by avoiding backpropagation (Zanella & Ben Ayed, 2024; Sui et al., 2025; Farina et al., 2024; Karmanov et al., 2024). However, these methods still operate after capture: they adapt prompts, features, caches, or predictions for a fixed physical measurement.

Sensor-level shift and adaptation. Sensor-level covariate shifts arise from the interaction between lighting and camera parameters (i.e., ISO, shutter speed). ImageNet-ES (Baek et al., 2024) and ImageNet-ES-Diverse (Baek et al., 2025) expose shifts via controlled light–sensor variations for static scenes, showing robustness gaps that digital-only adaptation cannot fully close. Lens (Baek et al., 2025) addresses this issue by changing capture, using a post-hoc, camera-agnostic controller that selects sensor settings per scene based on model confidence. Yet confidence-only single-view selection is prone to overconfident misclassification, leaving open VLM adaptation to sensor-driven input distribution. *MVP* addresses this gap through measurement plasticity: instead of only adapting the model after an image is captured, it adapts the sensor–model interface that determines what visual evidence reaches the VLM.

3. Method

Modern TTA methods for VLMs operate *after capture*, adjusting how the model interprets a fixed image. With Auto-Exposure (AE), however, the captured scene may already have lost information (e.g., saturation or low photon count). Thus, digital TTA faces a hard practical bound: once photons are not measured, downstream adaptation often cannot recover them. This motivates a *sensor-aware* view of TTA that respects the causal chain: scene $\rightarrow$ measurement $\rightarrow$ representation. We therefore treat camera sensor parameters as *physical prompts*: upstream control variables that determine which photons reach the sensor.

3.1. Source-Affinity Selection

Confidence-based sensor control (Baek et al., 2025) can be brittle: a shifted capture may produce high confidence while inducing unreliable VLM features. We instead select measurement settings by source affinity. Following PromptAlign (Abdul Samadh et al., 2023), *MVP* scores each candidate capture by the distance between its visual-token statistics and pre-computed source-domain statistics. Figure 2 visualizes the intuition: sensor parameters change what the model attends to, and source-affinity-selected views yield attention maps closer to the source than AE or confidence-selected captures. Intuitively, source-affinity-selected views let the model "see" the scene through source-like visual evidence. Supporting diagnostics on feature distance and attention similarity are included in Appendix A.

3.2. Multi-view Physical-prompt for TTA

Selecting physical views in feature-space. Figure 3 shows the overall pipeline of *MVP*. Instead of AE capture in the conventional work, *MVP* obtains the most source-aligned views from multiple M captures of a single scene under different camera settings. Each physical capture $v_i \in \mathcal{V} = \{v_1, v_2, ..., v_M\}$ is expanded into a set of N digital augmentations $\{v_{i,1}, ..., v_{i,N}\}$, including the original. For each capture v_i, we derive N' confident augmentations by selecting the top α fraction of N total augmentations based on confidence. Then, for each layer l, we extract the feature-level mean and variance of image token embedding $\mu_{i,l} = \frac{1}{N'}\sum_{n=1}^{N'} \mu_{i,l,n}$; $\sigma^2_{i,l} = \frac{1}{N'}\sum_{n=1}^{N'} \sigma^2_{i,l,n}$ from the frozen visual encoder using N' views. These statistics are then compared against the pre-computed source-domain statistics $(\mu_{s,l}, \sigma^2_{s,l})$. Since CLIP's pretraining data is non-public, we use ImageNet as a proxy source dataset and additionally evaluate LAION-based statistics (Appendix E). To measure the similarity of physical views with source image in feature-space, a source-affinity score is defined as:

$$S_i = -\frac{1}{L}\sum_{l=1}^{L}\left(\|\mu_{i,l} - \mu_{s,l}\|_2^2 + \|\sigma^2_{i,l} - \sigma^2_{s,l}\|_2^2\right), \quad (1)$$

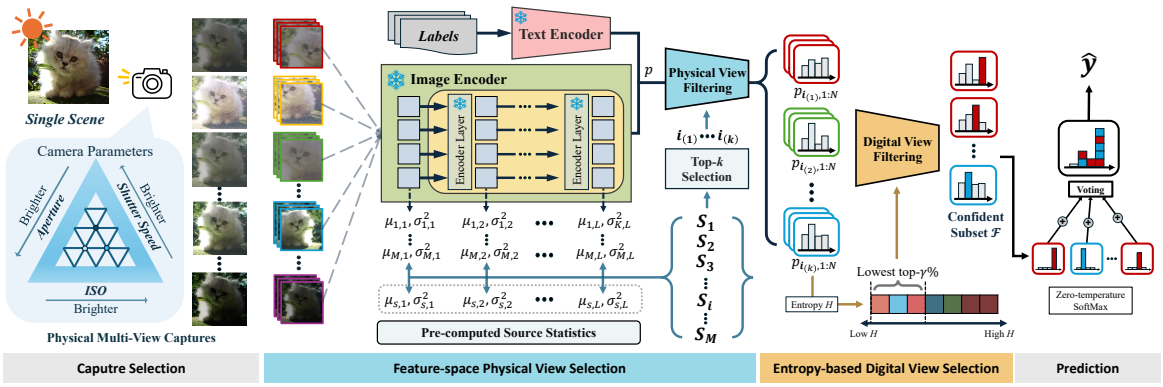


Figure 3. **Overall Framework.** Given a single scene, multiple physical captures are obtained by varying camera parameters along the exposure triangle—ISO, shutter speed, and aperture—forming a set of physical multi-views, as controllable physical prompts.

where L denotes the number of layers of the visual encoder. We select the top-k settings maximizing S_i, corresponding to most source-aligned physical views. This replaces prompt optimization with gradient-free *physical-prompt selection*, preserving gray-box compatibility and low computations.

Entropy-based filtering for augmented views. Even among the selected top-k physical parameters, some augmented samples may produce uncertain or noisy predictions due to local variations in illumination or visual context. To mitigate the negative impact of unreliable samples, we measure how uncertain the model is to each augmented view using entropy $H_{i,n} = -\sum_c p_{i,n}(c) \log p_{i,n}(c)$, where $p_{i,n}(c)$ denotes the predicted probability of class c for the n-th digital augmentation of the i-th physical capture. Across all $k \times N$ augmented views obtained from the selected k physical captures, we retain only the bottom $\gamma\%$ in entropy as the most confident subset $\mathcal{F}$. The predictions of these entropy-filtered views are then aggregated using a hard voting scheme across all selected views:

$$\hat{y} = \arg\max_{y \in \mathcal{C}} \sum_{(i,n) \in \mathcal{F}} \mathbf{1}\left[\arg\max_{c \in \mathcal{C}} p_{i,n}(c) = y\right], \quad (2)$$

where $\mathcal{C}$ is the set of class labels. This aggregation mitigates the overconfidence issues commonly observed in averaging softmax probabilities across multi-views, while preserving the robustness from physical multi-view observations.

Physical multiviews as augmentations. Rather than collapsing sensor control to a single capture, *MVP* votes over the top-k source-affinity-selected physical views. Because ISO, shutter speed, and aperture change the measurement itself, these captures provide a physical augmentation axis that post-capture crops, flips, or photometric perturbations (Buslaev et al., 2020) cannot fully emulate. This introduces useful degrees of freedom for observing the same scene through meaningfully different measurements and yields more robust hard-voting predictions than single-setting inference. LDA diagnostics supporting this distinction are provided in Appendix B.

4. Experiments

4.1. Experimental Setup

For each lighting condition, every sample is photographed for 5 times using the AE setting, forming the *auto-exposure set*, and captured with 27 controlled combinations of sensor parameters—three levels each of ISO, aperture, and shutter speed—forming the *parameter-controlled* set.

Datasets and baselines. We evaluate our method on sensor-level shift benchmarks, ImageNet-ES (Baek et al., 2024) and ImageNet-ES-Diverse (Baek et al., 2025), both constructed from Tiny-ImageNet (Le et al., 2015). We compare *MVP* to representative approaches across three categories; prompt learning methods, prompt-based TTA methods, and training-free TTA methods. We also include Lens (Baek et al., 2025) as the sensor-control baseline, which selects the most confident physical capture per scene. Further details are provided in the Appendix C and D.

Evaluation protocols. For each scene, we evaluate all baselines under three sensor-control settings: (1) Auto-Exposure (AE): Without any sensor control, the results are obtained by averaging all five AE captures. (2) AE with digital photometric augmentation: For the same AE captures, additional photometric augmentations (*i.e.,* hue, saturation, brightness) (Buslaev et al., 2020) are incorporated into the augmentation process within each TTA pipeline to simulate diverse exposure variations. (3) Lens-based selection: following the procedure in (Baek et al., 2025), we first identify the most confident physical view among 27 candidates using CLIP-based MaPLe inference, then apply each baseline method to the selected view.

4.2. Results

Main evaluation. We investigate the effectiveness of *MVP* against post-capture TTA baselines. Table 1 summarizes the results across three sensor-control protocols. Under AE, the zero-shot CLIP performance degrades significantly across lighting conditions, with existing TTAs

Table 1. **Performance comparison of vision-language adaptation methods.** All experiments are conducted using the ViT-B/16 backbone. For *with Lens*, all tuning-based methods are applied after sensor control via Lens (Baek et al., 2025). Results are reported by average across light environments and detailed per-light environment results are provided in the Appendix.

	Method	Tiny-ImageNet	ImageNet-ES			ImageNet-ES-Diverse		
			AE	AE + Photo Aug.	with Lens	AE	AE + Photo Aug.	with Lens
Pretrained	CLIP	87.6	48.98	-	82.45	37.65	-	61.13
Prompt Learning	CoOp	89.2	51.82	-	84.90	40.47	-	65.15
	MaPLe	89.0	52.57	-	84.40	39.47	-	64.68
Prompt-based TTA	TPT	89.5	55.66	49.66	84.45	41.20	40.82	64.90
	PromptAlign	90.0	55.45	50.27	84.75	41.51	41.04	64.90
	C-TPT	90.5	53.55	45.92	84.55	39.88	39.09	64.72
	O-TPT	89.3	52.70	44.56	84.50	39.14	38.47	64.63
Training free TTA	MTA	90.6	56.56	54.64	84.70	41.70	41.33	65.08
	TDA	90.0	58.17	57.29	84.80	40.78	40.83	64.85
	ZERO	90.6	57.05	56.77	84.80	39.91	39.80	64.67
***MVP* (ours)**		-	**87.85**			**67.28**		

offering only marginal gains. This results indicate that post-capture adaptation alone cannot recover information loss at sensor-level. Integrating digital photometric augmentations, which simulate various lightings and exposures, also proved ineffective at improving performance, confirming that digital perturbations cannot fully reproduce the physical effects of sensor changes. In contrast, ***MVP* outperforms all AE-based TTAs**, achieving significant gains of at least **29.68 pp** and **25.58 pp** on ImageNet-ES and ImageNet-ES-Diverse, respectively. This highlights the critical need for TTA strategies that explicitly incorporate sensor-level diversity at capture time. Our method also yields a robust performance gain of up to **3.4 pp** over pipelines combining Lens with TTA methods, validating the efficacy of multiple physical parameters. These findings emphasize that *MVP* mitigates the sensitivity to suboptimal sensor-setting and enhances robustness against sensor-level shifts.

Evaluation with limited parameter settings. While the number of combinations of possible camera parameter grows exponentially, capturing all camera configurations is infeasible given the high cost of acquisition. To reduce the redundancy of physical capturing, we further utilize Candidate Selection Algorithms (CSA) (Baek et al., 2025) as a pre-selection strategy that reduces the candidate sensor parameter space to a smaller set of M discrete grids. As shown in Table 2, *MVP* consistently outperforms baselines across different CSA. This demonstrates robustness and scalability under reduced-capture regimes, thereby supporting its practical use in real-world deployments where capture cost is non-negligible. Details on the CSA and full results of different M are in Appendix L and Tables 11 to 13.

Runtime latency. Figure 4 illustrates that *MVP* delivers a more effective trade-off between performance and runtime latency (ms)—including capture time, adaptation, and inference— per sample compared to baselines. Specifically, AE attains the lowest latency but the weakest accuracy, consistent with its optimization for human perception rather than model compatibility. Without CSA ($M = 27$ physical views), both *MVP* and Lens exhibit the expected performance–latency trade-off relative to AE, delivering $\geq$23.3 pp higher top-1 accuracy at roughly 4$\times$ latency, reflecting the cost of capturing 27 physical views per scene. Notably, when CSA is applied ($M = 6$), *MVP* (red triangle) maintains high performance even at low latency which is similar to AE settings. Compared to Lens-applied TTA settings (blue markers), *MVP* achieves the highest accuracy at lower latency than competing baselines, demonstrating its strong efficiency–adaptability balance. We also include qualitative comparisons that visualize the final selected views of MVP and ZERO with Lens, and ablations on MVP's design choices, in Appendix H and I.

Table 2. **Performance Across Different Capture Scenarios (CSA).** Without CSA, the capture latency is the full ImageNet-ES setting – 2.41 seconds per scene. For CSA1 / CSA2 / CSA3, the number of captures, M, are 12 / 6 / 21, respectively.

Method	CSA 1 (1.06 sec)		CSA 2 (0.37 sec)		CSA 3 (0.91 sec)	
	IN-ES	IN-ES -Diverse	IN-ES	IN-ES -Diverse	IN-ES	IN-ES -Diverse
Lens	84.75	61.79	84.38	61.76	85.00	54.38
+ TPT	84.88	62.22	84.55	62.49	85.05	62.29
+ PromptAlign	84.95	62.03	84.32	62.56	85.25	62.37
+ C-TPT	84.85	61.90	84.68	61.83	85.10	61.91
+ O-TPT	84.77	61.66	84.42	61.27	85.03	61.77
+ MTA	85.37	62.44	85.10	62.74	85.33	62.62
+ TDA	85.30	61.71	84.70	62.08	85.28	62.33
+ ZERO	85.43	61.49	84.60	61.96	85.62	61.69
***MVP* (ours)**	**87.27**	**63.79**	**86.65**	**63.79**	**87.87**	**64.41**

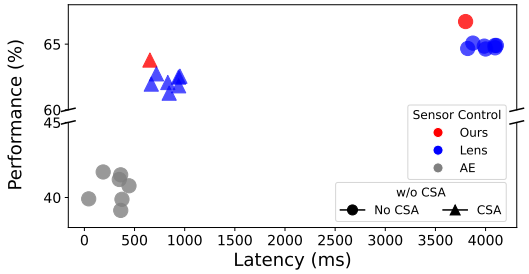


Figure 4. **Latency (ms) of adaptation and inference.** Experiments were conducted on NVIDIA RTX 4090 GPU and Intel(R) Xeon(R) Gold 6326 CPU. Color denotes the sensor control strategy, and markers distinguish TTA variant under the same strategy.

5. Conclusion

In this paper, we introduce *MVP*, a novel framework that integrates sensor control into the TTA paradigm by interpreting the camera setting as a *physical prompt* for VLMs. By treating the sensor as a controllable prompt, our approach enables the model to decide *what to capture*, extending conventional post-capture TTA into the sensor domain. Experiments show that our approach consistently outperforms baselines with high efficiency, indicating that robust real-world VLM generalization requires adapting not only how to interpret visual inputs, but also what to measure.

References

Abdul Samadh, J., Gani, M. H., Hussein, N., Khattak, M. U., Naseer, M. M., Shahbaz Khan, F., and Khan, S. H. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. *Advances in Neural Information Processing Systems*, 36: 80396–80413, 2023.

Baek, E., Park, K., Kim, J., and Kim, H.-S. Unexplored faces of robustness and out-of-distribution: Covariate shifts in environment and sensor domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22294–22303, 2024.

Baek, E., Han, S., Gong, T., and Kim, H.-S. Adaptive camera sensor for vision models. *arXiv preprint arXiv:2503.02170*, 2025.

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., and Kalinin, A. A. Albumentations: fast and flexible image augmentations. *Information*, 11(2): 125, 2020.

Farina, M., Franchi, G., Iacca, G., Mancini, M., and Ricci, E. Frustratingly easy test-time adaptation of vision-language models. *Advances in Neural Information Processing Systems*, 37:129062–129093, 2024.

Hu, E. J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Karmanov, A., Guan, D., Lu, S., El Saddik, A., and Xing, E. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14162–14171, 2024.

Khattak, M. U., Rasheed, H., Maaz, M., Khan, S., and Khan, F. S. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19113–19122, 2023.

Le, Y., Yang, X., et al. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Ma, X., Zhang, J., Guo, S., and Xu, W. Swapprompt: Test-time prompt adaptation for vision-language models. *Advances in Neural Information Processing Systems*, 36: 65252–65264, 2023.

Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., and Tan, M. Towards stable test-time adaptation in dynamic wild world. In *The Eleventh International Conference on Learning Representations*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., and Dosovitskiy, A. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Sharifdeen, A., Munir, M. A., Baliah, S., Khan, S., and Khan, M. H. O-tpt: Orthogonality constraints for calibrating test-time prompt tuning in vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19942–19951, 2025.

Shu, M., Nie, W., Huang, D.-A., Yu, Z., Goldstein, T., Anandkumar, A., and Xiao, C. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35: 14274–14289, 2022.

Sui, E., Wang, X., and Yeung-Levy, S. Just shift it: Test-time prototype shifting for zero-shot generalization with vision-language models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 825–835. IEEE, 2025.

Wang, D., Shelhamer, E., Liu, S., Olshausen, B., and Darrell, T. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*.

Xiao, Z., Yan, S., Hong, J., Cai, J., Jiang, X., Hu, Y., Shen, J., Wang, Q., and Snoek, C. G. Dynaprompt: Dynamic test-time prompt tuning. *arXiv preprint arXiv:2501.16404*, 2025.

Yoon, H. S., Yoon, E., Tee, J. T. J., Hasegawa-Johnson, M., Li, Y., and Yoo, C. D. C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. *arXiv preprint arXiv:2403.14119*, 2024.

Zanella, M. and Ben Ayed, I. On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning? In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pp. 23783–23793, 2024.

Zhang, J., Huang, J., Zhang, X., Shao, L., and Lu, S. Historical test-time prompt tuning for vision foundation models. *Advances in Neural Information Processing Systems*, 37: 12872–12896, 2024.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16816–16825, 2022a.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.

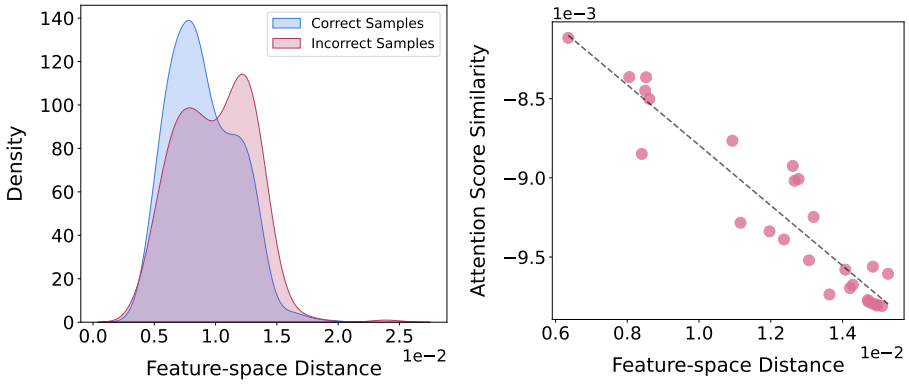
Appendix

A. Additional Motivation for Source-Affinity Selection

We provide additional diagnostic evidence supporting the use of source-affinity selection in MVP. These analyses are used only for motivation and do not require labels or attention maps at inference time.

Feature distance reveals failures of confidence-based selection. Confidence-based sensor control selects the physical capture with the highest model confidence (Baek et al., 2025). However, high confidence does not necessarily imply that the selected capture is represented in a source-like feature space. To examine this failure mode, we take the top-1 physical view selected by Lens on ImageNet-ES-Diverse (Baek et al., 2025), compute its feature-space distance to the pre-computed source-domain statistics, and split samples by whether the resulting prediction is correct. As shown in Figure 5a, incorrect predictions exhibit a bimodal distribution, with a larger mode shifted toward higher feature-space distances, whereas correct predictions are more concentrated at smaller distances. This suggests that confidence-only selection can choose views that are confidently predicted but representation-shifted.

Feature distance tracks source-like attention. We further examine whether feature-space distance is related to the model's visual evidence. For candidate physical views of the same scene, we compare their attention-score similarity to the corresponding source image. Figure 5b shows that views with smaller feature-space distance tend to have higher attention-score similarity. Thus, source-affinity is not only a global feature-statistic criterion: it also correlates with whether the frozen VLM attends to visual evidence in a source-like manner. This supports our design choice of extending the distribution-alignment principle of PromptAlign (Abdul Samadh et al., 2023) upstream, using source affinity to select physical measurements before post-capture aggregation.



(a) *(b)*

Figure 5. **Motivating findings.** (a) Incorrect samples among Lens top-1 predictions form a bi-modal distribution. Larger mode is skewed toward the higher feature-space distances compared to the correct samples. (b) A strong correlation between attention score similarity and feature-space distance shows that samples closer in feature space exhibit more similar attention patterns.

B. Physical Multiviews as an Augmentation Axis

We further examine whether physical multiviews provide feature-space variability that is distinct from standard digital augmentations. For each scene, we take the top-5 physical views selected by source affinity and project their digitally augmented visual embeddings using LDA. As shown in Figure 6, embeddings from different sensor settings occupy distinct regions, whereas geometric augmentations remain largely within each sensor-specific region. Adding photometric perturbations (Buslaev et al., 2020) changes appearance but still does not fully reproduce the feature-space variability induced by changing camera parameters. Figure 9 shows additional LDA projection visualizations for each lighting condition in ImageNet-ES and ImageNet-ES-Diverse, along with the corresponding top-5 camera parameters ranked by affinity score. This supports our use of multiple physical captures as a complementary augmentation axis for hard-vote aggregation.

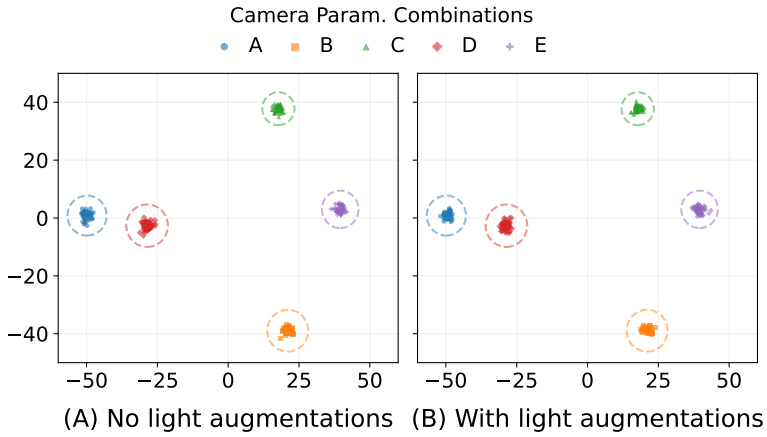


Figure 6. **LDA projection of visual embeddings from five physical views.** (Left) With digital transformation such as Random Crop or Horizontal Flip. (Right) With additional photometric augmentations such as perturbing hue, saturation, and brightness. Visualizations for more samples can be found in the Appendix.

C. ImageNet-ES/ImageNet-ES-Diverse details

ImageNet-ES (Baek et al., 2024) and ImageNet-ES-Diverse (Baek et al., 2025) are benchmarks designed to evaluate robustness under sensor-level covariate shifts induced by variations in lighting conditions and camera parameters. Their test sets consist of 64k and 192k samples, respectively.

The two datasets differ primarily in object presentation: ImageNet-ES captures a monitor displaying the objects, treating them as light-emitting sources, while ImageNet-ES-Diverse captures printed images of the objects, enabling more realistic evaluation under ambient lighting. Furthermore, ImageNet-ES includes only two lighting conditions (e.g. light-on, light-off), whereas ImageNet-ES-Diverse provides six diverse illumination settings.

Each dataset contains (i) an *Auto-Exposure (AE)* subset, where scenes are captured several times using the camera's default exposure setting, and (ii) a *Parameter-Controlled* subset, where the same scenes are recorded under systematically varied ISO, shutter speed, and aperture combinations across multiple lighting environments. The AE set contains five captures per scene using the camera's automatic exposure mode, while the parameter-controlled set contains 27 captures obtained under controlled combinations of ISO, shutter speed, and aperture—three levels each (ISO: {250, 2000, 16000}, Shutter: {1/4″, 1/60″, 1/1000″}, Aperture: f/5.0, f/9.0, f/16). Example samples from both datasets are shown in Fig. 11.

D. Implementation details and full results about main evaluation

Details of *MVP* We implement *MVP* upon MaPLe (Khattak et al., 2023), the multi-modal prompting model with the CLIP ViT-B/16 backbone. Source-affinity score is computed for a depth of three layers of the visual encoder ($L = 3$), while the source dataset statistics are pre-computed on ImageNet. For each physical view, we generate 63 digitally augmented views using random resized cropping and horizontal flipping, forming a batch of $N = 64$ digitally augmented images including the original as in prior TTA methods. We use the fraction $\alpha = 0.3$ to include only confident augmentation samples in source-affinity score aggregation. We select the top-5 parameter settings (i.e., k=5) with the largest source-affinity score, and within their augmented samples, retain the bottom γ=3% in entropy as confident predictions for voting.

Details of baselines We compare *MVP* to representative approaches across three categories. (1) Prompt learning methods, including CoOp (Zhou et al., 2022b) and MaPLe (Khattak et al., 2023), which optimize learnable context tokens for downstream adaptation. We use the publicly released pretrained weights of CoOp and MaPLe without any additional training. (2) Prompt-based TTA methods, including TPT (Shu et al., 2022), C-TPT (Sharifdeen et al., 2025), O-TPT (Yoon et al., 2024), and PromptAlign (Abdul Samadh et al., 2023), which adapt prompt representations at test time. (3) Training-free, forward-only TTA methods, including MTA (Zanella & Ben Ayed, 2024), TDA (Karmanov et al., 2024), and ZERO (Farina et al., 2024). In addition, we include Lens (Baek et al., 2025) as a sensor-control baseline, which selects the single most confident physical observation for each scene based on model confidence. A more detailed discussion of Lens and its

Table 3. **Evaluation with Alternative Proxy Datasets.** Source-affinity scores are computed using either ImageNet or LAION as the proxy dataset. To ensure a fair comparison, all experiments use the same CLIP ViT-B/16 backbone without MaPLe prompts.

Proxy Dataset	ImageNet-ES			ImageNet-ES-Diverse						
	L1	L5	Average	L1	L2	L3	L4	L6	L7	Average
ImageNet	85.3	84.3	84.80	65.3	64.8	65.5	64.2	63.4	62.7	64.32
LAION	82.4	82.5	82.45	63.2	65.3	64.2	63.4	64.0	63.0	63.85

comparison with *MVP* is provided in Section F.

To ensure a fair comparison, all TTA baselines are initialized with the same pretrained MaPLe (Khattak et al., 2023) weights, which place learnable prompt vectors in both the text and visual encoders up to the third layer. Following each method's original design, we restrict gradient updates to the textual prompt vectors (`prompt_learner.ctx`) for all prompt-based TTA baselines except PromptAlign. PromptAlign is the only method allowed to update visual prompt vectors, as its objective explicitly relies on adapting visual features for cross-modal alignment. Consequently, for all other prompt-based TTA methods, no gradients flow into the visual encoder and only the text-encoder input prompts are optimized during test-time adaptation.

Full results We also report the full accuracy results across individual lighting environments for both ImageNet-ES (Baek et al., 2024) and ImageNet-ES-Diverse (Baek et al., 2025). Table 10 summarizes detailed accuracies under the three evaluation settings used in the main paper—AE, AE + photometric augmentations, and Lens-based selection—across all lighting conditions.

E. Effect of Proxy Dataset Choice

Since both ImageNet-ES and ImageNet-ES-Diverse are derived from ImageNet, we additionally use LAION (Schuhmann et al., 2021), a large-scale and diverse web-scale dataset, as the proxy dataset for computing source-affinity scores. To isolate the effect of the proxy dataset, all experiments are conducted using the same CLIP ViT-B/16 backbone without MaPLe initialization. As shown in Table 3, replacing ImageNet with LAION results in a slight performance decrease. Nevertheless, the performance drop only remains within approximately 1.5%p on ImageNet-ES and 0.5%p on ImageNet-ES-Diverse, indicating that the proposed source-affinity mechanism still remains effective under an alternative proxy distribution.

F. Conceptual Difference from Lens

Lens and MVP share a common hardware-driven philosophy: rather than modifying the model itself through additional training or adaptation, both approaches aim to improve performance by providing the model with more informative physical observations.

However, the two approaches operate at different levels. Lens performs confidence-based selection of a single physical observation and directly uses the selected image for prediction. In contrast, MVP employs source-affinity-based selection as the first stage of a broader multi-view adaptation framework. Rather than making a prediction from a single selected observation, MVP first identifies multiple source-aligned physical observations and subsequently aggregates them through entropy-based filtering and voting.

Consequently, Lens can be viewed as a view-selection strategy that is compatible with existing test-time adaptation methods, similar to the setting explored in our experiments, as mentioned in D. In MVP, by contrast, view selection constitutes only one component within the overall adaptation pipeline, which further incorporates adaptation and multi-view aggregation to produce the final prediction.

G. Evaluation on different architectures

To verify that the effectiveness of *MVP* does not depend on a specific prompting architecture, we further evaluate our framework on two additional backbones: (1) CLIP ViT-B/16 *without* MaPLe initialization (*i.e.,* no visual prompt tokens), and (2) CLIP ResNet50. All components of *MVP* —including feature-space distance computation, physical-view selection,

and entropy-filtered aggregation—are kept identical to the main setup; only the underlying visual encoder is replaced. For fair comparison, the text prompt is initialized using the standard CLIP template, "a photo of a class", without any learned prompt vectors.

For ViT-B/16 *without* MaPLe, we retain the identical procedure to the main experiments for computing source statistics and feature-space distances.

For ResNet50, we replace the layer-wise token statistics used for ViT with block-wise statistics by treating each `Bottleneck block` as an individual layer. Specifically, each block's feature map $[B, C, H, W]$ is reshaped by flattening the spatial dimensions $(H \times W)$ so that each spatial location is regarded as a token. We then compute mean and variance over the resulting (token, channel) representations and measure the feature-space affinity between these block-level statistics and the pre-computed source statistics. The affinity score is computed over the first 7 out of 16 bottleneck blocks, following the observation in (Raghu et al., 2021) that ResNet requires a larger portion of its early blocks to produce representations comparable to those obtained from a much smaller number of early ViT layers.

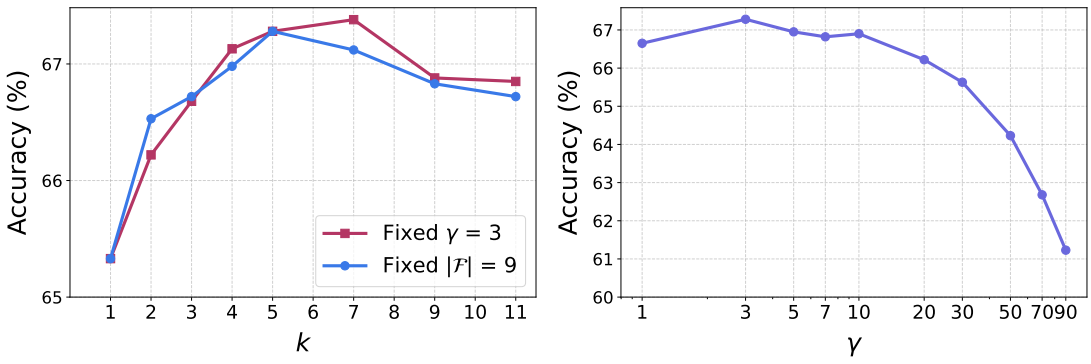
Results are presented in Table 8 and 9. Across both backbones, the average performance of *MVP* consistently outperforms all forward-only and prompt-based TTA baselines, demonstrating that its effectiveness does not depend on the choice of visual backbone or on the presence of tuned prompt vectors.

H. Qualitative analyses.

Figure 10 presents confident, digitally-augmented views selected for the final prediction by ZERO with Lens and *MVP*, respectively. *MVP* avoids being trapped by mistakenly selected overconfident physical view in the capture time by considering various physical settings in terms of the model's perception and delivers optically improved images. It is noteworthy that the lack of the diversity introduced by a limited physical sensor space in a conventional sensor-control approach cannot be mitigated solely by digitally augmented views.

I. Ablation Study

Effect of varying k. We examine how the number of physical views selected by the source-affinity score affects accuracy. Figure 7a shows an inverted U-shaped relationship between k and performance. With a fixed confidence percentile $\gamma = 3\%$ (red), increasing k raises the total number of augmented views that enter the vote; when k is too small, the ensemble has too few voters, leading to unstable predictions. To decouple the effect of k from the number of voters, we also hold the total count of selected augmented views constant at $|\mathcal{F}| = 9$ (blue). In this setting, smaller k allocates more augmentations per physical view, allowing lower-confidence augmentations to participate in the vote. Conversely, once k exceeds the plateau, accuracy declines slightly because additional, lower–source-affinity physical views are admitted and their associated (often overconfident) augmentations introduce noise that dilutes the vote.



(a) Effect of various k *(b)* Effect of various γ

Figure 7. **Ablation on parameter-view selection in *MVP*.** (a) shows the selection of top-k physical parameter settings based on the source-affinity score, while (b) illustrates the confidence percentile thresholding used to filter augmented views for final prediction. We fix $k = 5$ and vary γ from 1% to 90%.

Effect of varying γ. Figure 7b reports the effect of varying the entropy filtering threshold γ, which determines the proportion of digitally augmented views selected for the final aggregation. As shown, the performance peaks at γ = 3% and remains relatively stable up to γ = 10%, after which it sharply degrades, indicating that including too many high-entropy (less confident) views introduces noise into the aggregation process, resulting in unstable predictions.

Table 4. **Ablation Study on Physical Augmentation Design Choices.** Comparison across digital (geometric/photometric) augmentations and physical augmentations using multiple camera parameter views. k physical views are selected from all 27 camera parameter settings in each case. A total of 320 augmented views are produced per scene for k=1 settings, matching the number of those at k=5 setting; $64 \times 5 = 320$ views. Therefore, all comparisons are conducted using the same pool of pre-captured physical observations, with differences arising only from the view selection and aggregation strategy.

Number of top physical views (k)	Digital geo. aug.	Digital photo. aug.	Physical aug.	IN-ES	IN-ES -Diverse
$k = 1$	✓			86.90	66.72
$k = 1$	✓	✓		86.85	66.62
Ours ($k = 5$)	✓		✓	**87.85**	**67.28**

Number of physical views. Table 4 highlights the importance of leveraging multiple physical views rather than relying on a single capture or digital-level augmentations. Aggregating predictions from multiple physical captures consistently outperforms applying digital geometric or additional photometric augmentations to a single physical capture. This demonstrates that using multiple physical views leads to more robust and stable predictions than depending on a single parameter setting. Moreover, even when photometric augmentations are added to the single physical view, the performance still falls short of our approach, underscoring the inherent limitation of digital-level augmentations in simulating sensor-level diversity.

Component-wise Ablation Study. Table 5 presents an ablation study on the key components of our framework. We first evaluate different physical view selection strategies. Compared with random selection, source-affinity-based selection consistently achieves higher performance on both benchmarks. This result highlights the importance of sensor control in physical multi-view adaptation. Rather than relying on arbitrary camera parameter configurations, selecting physical views that are better aligned with the source distribution provides observations that are more suitable for the model, leading to more reliable adaptation.

We further compare source-affinity-based selection against using all available physical views without sensor control. Although leveraging all views already provides strong performance, source-affinity-based selection consistently achieves better results. This indicates that not all camera parameter configurations contribute equally to adaptation and that filtering physical views according to their source affinity produces a more reliable candidate set for downstream prediction.

Finally, we compare two aggregation strategies using the selected and entropy-filtered views. While marginalized aggregation averages prediction probabilities across views, hard voting determines the final prediction based on label-level consensus. As shown in Table 5, hard voting consistently outperforms marginalized aggregation, suggesting that consensus-based aggregation is more robust to noisy or miscalibrated predictions from individual views.

Combining source-affinity-based physical view selection, entropy-based digital view filtering, and hard voting achieves the best overall performance on both benchmarks, validating the effectiveness of each component in the proposed framework.

Source-affinity score measurement layers. We evaluate the overall performance by varying the choice of layers for measuring feature-space source affinity score, as summarized in Table 6. It shows that using early layers, particularly the first to third layers, achieves the best accuracy on both datasets. In contrast, computing the score over all layers or only the later layers leads to a significant performance degradation. By measuring the affinity in early-layer feature space, our method can more reliably identify physical views that the model perceives as closer to its pre-training domain—effectively providing inputs that are "familiar" to the model. In contrast, later layers encode highly semantic representations, making them less transferable across variations in illumination or sensor parameters. As a result, source-affinity scores from these layers provide less reliable cues about their affinity to the source distribution.

Table 5. **Ablation Study on Key Components.** Performance impact of physical view selection and prediction aggregation strategies. "No Sensor Control" applies digital view filtering and hard voting over all physical views. "Random Selection" randomly selects k physical views before digital filtering and hard voting, and the reported results are averaged over three random seeds. "Marginalization" replaces hard voting with probability averaging over source-affinity-selected and entropy-filtered views. The proposed framework combines source-affinity-based physical view selection, digital view filtering, and hard voting, achieving the best performance on both benchmarks.

Method	Physical View Selection	Digital View Filtering	Hard Voting	IN-ES	IN-ES -Diverse
No Sensor Control	All views	✓	✓	87.25	66.37
Random Selection	Random k views	✓	✓	86.52	61.90
Marginalization	Source Affinity	✓		87.80	66.72
Ours	Source Affinity	✓	✓	**87.85**	**67.28**

Table 6. **Ablation on layer depth for source-affinity score.**

Layer	IN-ES	IN-ES-Diverse	Latency (s)
1	87.65	67.25	1.22
1-3 (Ours)	**87.85**	**67.28**	**1.32**
1-12	82.65	18.22	1.40
10-12	82.55	18.85	1.35
12	82.75	20.02	1.40

J. Controlled Comparison of Physical View Selection under Fixed k

Table 7 presents a controlled comparison in which the number of selected physical views (k) and the aggregation strategy over filtered digital view predictions are fixed, while only the physical view selection criterion is varied. This setting allows us to isolate the effects of the two key components of *MVP*: (i) using multiple physical views as a new augmentation axis, and (ii) the proposed source affinity criterion.

Regardless of whether $k = 1$ or $k = 5$, the source affinity score of *MVP* consistently outperforms the confidence score used in Lens (Baek et al., 2025), as it does not rely on physical views that the model may judge overconfidently. Instead, it selects physical views that are more aligned with the distribution learned by the model and therefore better understood by it, leading to higher performance.

In addition, Lens (Baek et al., 2025) also benefits from multi-view aggregation, confirming the augmentation effect of using multiple physical views and further supporting the necessity of multi-physical view modeling, as emphasized in our approach.

K. Runtime latency

Fig 8 presents the runtime latency breakdown for baseline TTA methods and *MVP* on ImageNet-ES-Diverse. All methods adopt CSA2 (M=6) to balance between capture cost and overall latency. The *Parameter Selection* stage corresponds to selecting the most confident physical view in Lens and selecting the top-5 physical views with the highest source-affinity scores in *MVP*. Prompt-based TTA methods incur additional forward and backward passes to update prompt vectors during inference, while training-free TTAs and *MVP* do not require such gradient-based operations. *Inference* stage denotes the end-to-end forward pass that produces the final prediction for a single test sample. Latency for parameter selection is averaged across all 27 parameter configurations, while latency for the remaining procedures is computed by measuring per-sample runtime over 1,000 test samples and reporting the average. The total latency is illustrated in Fig 4.

L. CSA details

CSA (Candidate Selection Algorithm) is a pre-selection strategy designed to reduce redundant capture time before performing sensor control (Baek et al., 2025). It provides three approaches for selecting M candidate parameter configurations from the full set of available sensor-control options.

CSA1 performs uniform random sampling across all parameter configurations.

Table 7. CLIP-ViT-B-16 initialized with MaPLe (Khattak et al., 2023) weights

# of physical views	Method	IN-ES	IN-ES -Diverse
Single-view ($k = 1$)	Lens + hard voting	84.30	65.05
	***MVP*(ours)**	**85.45**	**65.33**
Multi-view ($k = 5$)	Lens + hard voting	86.95	66.78
	***MVP*(ours)**	**87.85**	**67.28**

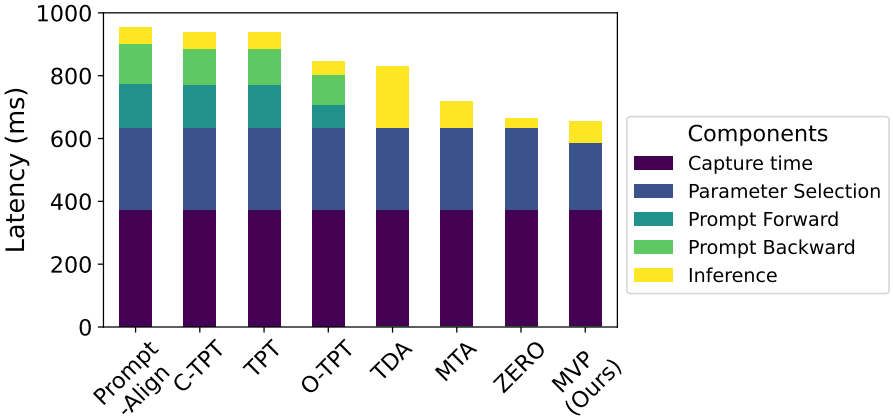


Figure 8. Total latency (ms) of adaptation and inference

CSA2 adopts a grid-based random selection strategy to ensure even coverage of the parameter space. The full 3D (ISO, shutter speed, aperture) parameter space is partitioned into several grids, and one candidate is sampled from each grid until M selections are made. Since parameters within the same grid are close in value, the resulting captures tend to exhibit similar image qualities. The number of grids is pre-defined as 1^3 for $M = 6$ in (Baek et al., 2025), but for fairer exposure variability, we adjust this to 2^3 grids in our evaluation.

CSA3 selects M parameter configurations with the lowest capture cost, determined primarily by shutter speed. When multiple configurations have identical costs, the algorithm breaks ties randomly.

Table 11, 12, and 13 present extended CSA results, including those from Table 2, evaluated across various values of M. For all experiments, we report the average performance over three runs with different random seeds. CSA3 places the strongest emphasis on efficiency among the three strategies, which leads to severe performance degradation especially under low-illumination conditions. Consequently, its accuracy drops sharply as M decreases, and we therefore evaluate CSA3 only for $M = 21$ and $M = 18$.

MVP consistently achieves the highest accuracy across all evaluated values of M.

M. Limitations and future work

While *MVP* operates in a fully forward-only manner and demonstrates strong robustness, it still requires more than k physical captures per scene as well as pre-computed source-domain statistics. Future work includes extending the framework beyond static scenes and the restricted parameter space considered in this work—toward dynamic environments and broader regions of the exposure parameter manifold. Moreover, applying the proposed physical-prompting paradigm to tasks beyond classification (e.g., detection, retrieval, or semantic segmentation) represents a promising direction.

Table 8. Detailed results on ViT-B/16 backbone *without* MaPLe

Auto Exposure										
Method	ImageNet-ES			ImageNet-ES-Diverse						
	L1	L5	Average	L1	L2	L3	L4	L6	L7	Average
CLIP	50.92	47.04	48.98	38.50	37.88	39.38	39.02	36.20	34.92	37.65
TPT	54.65	50.44	52.56	40.80	40.72	42.04	40.74	38.32	36.96	39.93
C-TPT	51.72	47.88	49.80	38.86	38.56	40.26	39.12	37.02	34.76	38.10
O-TPT	51.06	46.84	48.95	38.30	37.40	39.02	38.08	35.82	33.92	37.09
MTA	54.42	50.60	52.51	40.66	39.52	40.82	40.38	38.22	36.78	39.40
TDA	54.44	50.08	52.26	38.74	38.40	39.88	39.42	36.24	35.46	38.02
ZERO	56.38	51.38	53.88	39.80	39.60	40.10	39.48	37.60	35.44	38.67
with Lens										
Method	ImageNet-ES			ImageNet-ES-Diverse						
	L1	L5	Average	L1	L2	L3	L4	L6	L7	Average
CLIP+Lens	82.70	82.20	82.45	63.30	61.40	61.70	60.60	61.00	58.80	61.13
TPT	82.80	82.00	82.40	63.70	62.40	62.20	61.30	61.20	59.40	61.70
C-TPT	82.80	82.20	82.50	63.60	61.90	62.50	60.60	61.40	59.20	61.53
O-TPT	82.70	81.80	82.25	63.10	61.30	62.10	60.60	60.80	59.30	61.20
MTA	82.60	82.10	82.35	63.70	61.90	62.40	60.80	60.70	59.30	61.47
TDA	82.70	82.40	82.55	62.20	61.10	61.20	60.70	60.20	58.70	60.58
ZERO	82.90	82.40	82.65	63.60	61.20	62.10	61.70	59.90	58.80	61.22
***MVP* (ours)**	**85.30**	**84.30**	**84.80**	**65.30**	**64.80**	**65.50**	**64.20**	**63.40**	**62.70**	**64.32**

Table 9. Detailed results on ResNet50 backbone

Auto Exposure										
Method	ImageNet-ES			ImageNet-ES-Diverse						
	L1	L5	Average	L1	L2	L3	L4	L6	L7	Average
CLIP	35.88	31.96	33.92	17.44	17.70	18.32	17.66	16.26	15.52	17.15
TPT	39.50	36.18	37.84	21.08	21.60	21.82	21.36	20.00	18.70	20.76
C-TPT	37.62	33.12	35.37	17.74	18.06	18.74	18.06	16.56	15.74	17.48
O-TPT	36.08	31.68	33.97	18.34	18.24	19.02	18.52	17.26	16.46	17.97
MTA	38.52	34.90	36.71	16.76	16.52	17.36	16.34	15.20	14.22	16.15
TDA	38.62	35.16	36.89	16.16	16.26	16.86	16.32	14.24	13.60	15.57
ZERO	39.34	35.82	37.58	15.82	15.42	15.84	15.96	14.22	13.10	15.06
with Lens										
Method	ImageNet-ES			ImageNet-ES-Diverse						
	L1	L5	Average	L1	L2	L3	L4	L6	L7	Average
CLIP+Lens	73.60	73.30	73.45	40.90	41.50	39.90	38.70	38.60	37.20	39.47
TPT	73.90	73.70	73.80	41.40	41.50	40.40	39.60	39.30	37.40	39.93
C-TPT	73.70	73.40	73.55	41.30	41.20	40.20	39.60	39.30	37.60	39.87
O-TPT	73.70	73.40	73.55	41.40	41.60	40.30	39.40	39.00	37.10	39.80
MTA	74.00	73.70	73.85	41.20	41.00	39.90	39.00	38.40	37.40	39.48
TDA	72.50	73.40	72.95	40.70	39.60	39.80	38.60	37.80	36.00	38.75
ZERO	73.50	73.90	73.70	40.20	39.80	40.00	37.60	38.20	35.20	38.50
***MVP* (ours)**	**76.00**	**74.80**	**75.40**	**43.20**	**41.70**	**41.50**	**38.80**	**39.10**	**36.50**	**40.13**

Table 10. Detailed results of main evaluation

Auto Exposure										
Method	ImageNet-ES			ImageNet-ES-Diverse						
	L1	L5	Average	L1	L2	L3	L4	L6	L7	Average
CLIP	50.92	47.04	48.98	38.50	37.88	39.38	39.02	36.20	34.92	37.65
CoOp	54.48	49.16	51.82	41.36	40.78	42.20	41.72	39.36	37.38	40.47
MaPle	54.98	50.16	52.57	40.56	39.76	41.24	40.24	38.48	36.52	39.47
TPT	58.64	52.68	55.66	42.86	41.34	43.00	41.76	39.62	38.60	41.20
PromptAlign	58.20	52.70	55.45	42.66	41.40	43.70	42.32	39.90	39.06	41.51
C-TPT	56.24	50.86	53.55	41.24	40.16	41.46	40.22	39.02	37.18	39.88
O-TPT	55.12	50.28	52.70	40.38	39.24	40.56	39.98	38.26	36.42	39.14
MTA	59.30	53.82	56.56	43.14	42.04	43.60	42.10	40.20	39.10	41.70
TDA	61.16	55.18	58.17	41.96	41.06	42.64	41.42	39.30	38.30	40.78
ZERO	59.92	54.18	57.05	40.90	40.36	41.82	40.88	38.34	37.16	39.91
Auto Exposure + Photo. Aug.										
Method	ImageNet-ES			ImageNet-ES-Diverse						
	L1	L5	Average	L1	L2	L3	L4	L6	L7	Average
TPT	52.12	47.20	49.66	42.80	40.88	42.96	41.78	38.92	37.60	40.82
PromptAlign	53.18	47.36	50.27	42.98	41.06	42.94	41.60	39.82	37.86	41.04
C-TPT	48.52	43.32	45.92	40.94	39.62	40.64	40.16	37.54	35.64	39.09
O-TPT	47.00	42.12	44.56	40.36	38.90	39.94	39.30	36.88	35.42	38.47
MTA	57.22	52.06	54.64	42.84	41.66	43.56	42.14	39.52	38.26	41.33
TDA	60.44	54.14	57.29	42.12	41.06	42.58	41.52	39.50	38.20	40.83
ZERO	59.56	53.98	56.77	41.20	39.62	41.74	40.26	38.52	37.46	39.80
with Lens										
Method	ImageNet-ES			ImageNet-ES-Diverse						
	L1	L5	Average	L1	L2	L3	L4	L6	L7	Average
CLIP+Lens	82.70	82.20	82.45	63.30	61.40	61.70	60.60	61.00	58.80	61.13
CoOp	84.80	85.00	84.90	68.20	66.10	65.70	64.20	64.40	62.30	65.15
MaPle	85.10	83.70	84.40	65.90	64.80	66.20	64.90	63.60	62.70	64.68
TPT	85.10	83.80	84.45	66.50	65.10	66.30	65.00	64.10	62.40	64.90
PromptAlign	85.70	83.80	84.75	66.90	65.30	65.70	65.50	64.00	62.00	64.90
C-TPT	85.30	83.80	84.55	66.00	64.90	66.10	64.70	63.90	62.70	64.72
O-TPT	85.30	83.70	84.50	66.00	64.80	66.10	64.50	63.80	62.60	64.63
MTA	85.40	84.00	84.70	66.20	65.70	66.50	65.20	64.20	62.70	65.08
TDA	85.80	83.80	84.80	65.90	64.40	66.30	66.00	63.60	62.90	64.85
ZERO	85.20	84.40	84.80	66.00	65.20	66.00	65.10	63.30	62.40	64.67
Ours										
***MVP* (ours)**	87.50	**88.20**	**87.85**	**69.30**	**67.50**	**69.20**	**66.70**	**66.00**	**65.00**	**67.28**

Table 11. **Results with CSA1 across different M.**

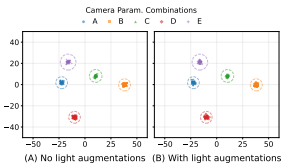
Method	$M = 18$		$M = 15$		$M = 12$		$M = 9$		$M = 6$	
	IN-ES	**IN-ES -Diverse**	**IN-ES**	**IN-ES -Diverse**	**IN-ES**	**IN-ES -Diverse**	**IN-ES**	**IN-ES -Diverse**	**IN-ES**	**IN-ES -Diverse**
Lens	84.73	63.53	84.82	63.52	84.75	61.79	84.48	60.48	83.33	55.40
+ TPT	84.82	63.94	84.98	63.86	84.88	62.22	84.90	61.03	83.77	56.33
+ PromptAlign	84.98	63.74	85.25	63.82	84.95	62.03	84.82	61.01	84.03	56.24
+ C-TPT	84.88	63.56	84.83	63.56	84.85	61.90	84.83	60.66	83.65	53.52
+ O-TPT	84.75	63.39	84.75	63.41	84.77	61.66	84.57	60.34	83.43	51.06
+ MTA	85.18	64.09	85.23	64.09	85.37	62.44	85.12	61.17	84.02	56.51
+ TDA	85.22	63.81	85.37	63.53	85.30	61.71	85.12	60.64	84.23	56.02
+ ZERO	85.43	63.50	85.75	63.38	85.43	61.49	85.23	60.03	84.60	55.59
***MVP* (ours)**	**87.82**	**66.02**	**87.57**	**65.97**	**87.27**	**63.79**	**86.40**	**62.27**	**85.40**	**57.49**

Table 12. **Results with CSA2 across different M.**

Method	$M = 18$		$M = 15$		$M = 12$		$M = 9$		$M = 6$	
	IN-ES	**IN-ES -Diverse**	**IN-ES**	**IN-ES -Diverse**	**IN-ES**	**IN-ES -Diverse**	**IN-ES**	**IN-ES -Diverse**	**IN-ES**	**IN-ES -Diverse**
Lens	84.62	63.31	84.52	63.21	84.28	63.29	84.43	61.10	84.38	61.76
+ TPT	84.52	63.78	84.47	63.66	84.28	63.78	84.48	61.54	84.55	62.49
+ PromptAlign	84.65	63.48	84.52	63.73	84.25	63.67	84.60	61.53	84.32	62.56
+ C-TPT	84.72	63.33	84.65	63.31	84.47	63.41	84.60	60.97	84.68	61.83
+ O-TPT	84.62	63.29	84.50	63.22	84.37	63.27	84.53	60.48	84.42	61.27
+ MTA	84.98	63.89	84.87	63.89	84.88	64.04	84.95	61.80	85.10	62.74
+ TDA	84.83	63.52	85.00	63.23	84.82	63.36	84.38	61.26	84.70	62.08
+ ZERO	85.28	63.19	85.10	63.06	84.70	63.16	84.75	60.86	84.60	61.96
***MVP* (ours)**	**87.68**	**65.68**	**87.48**	**65.44**	**86.90**	**65.47**	**86.92**	**63.09**	**86.65**	**63.79**

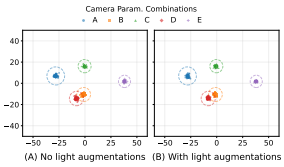
Table 13. **Results with CSA3 across different M.**

Method	$M = 21$		$M = 18$	
	IN-ES	**IN-ES -Diverse**	**IN-ES**	**IN-ES -Diverse**
Lens	84.98	61.67	85.00	54.38
+ TPT	85.05	62.29	85.15	55.32
+ PromptAlign	85.25	62.37	85.40	55.40
+ C-TPT	85.10	61.91	85.10	54.58
+ O-TPT	85.03	61.77	85.05	54.40
+ MTA	85.33	62.62	85.30	55.30
+ TDA	85.28	62.33	85.05	54.93
+ ZERO	85.62	61.69	85.90	54.28
***MVP* (ours)**	**87.87**	**64.41**	**87.70**	**55.92**



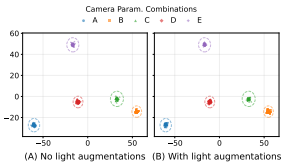
(a) ImageNet-ES L1

(b) ImageNet-ES L1; American bullfrog



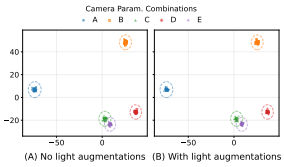
(c) ImageNet-ES L5

(d) ImageNet-ES L5; American alligator



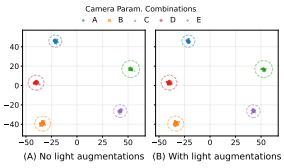
(e) ImageNet-ES-Diverse L1

(f) ImageNet-ES-Diverse L1; goldfish



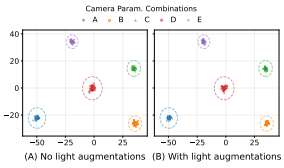
(g) ImageNet-ES-Diverse L2

(h) ImageNet-ES-Diverse L2; southern black widow



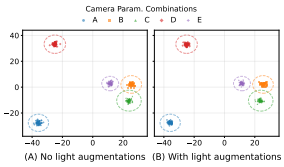
(i) ImageNet-ES-Diverse L3

(j) ImageNet-ES-Diverse L3; goose



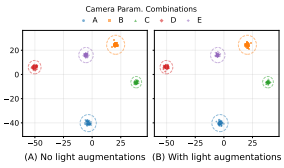
(k) ImageNet-ES-Diverse L4

(l) ImageNet-ES-Diverse L4; koala



(m) ImageNet-ES-Diverse L6

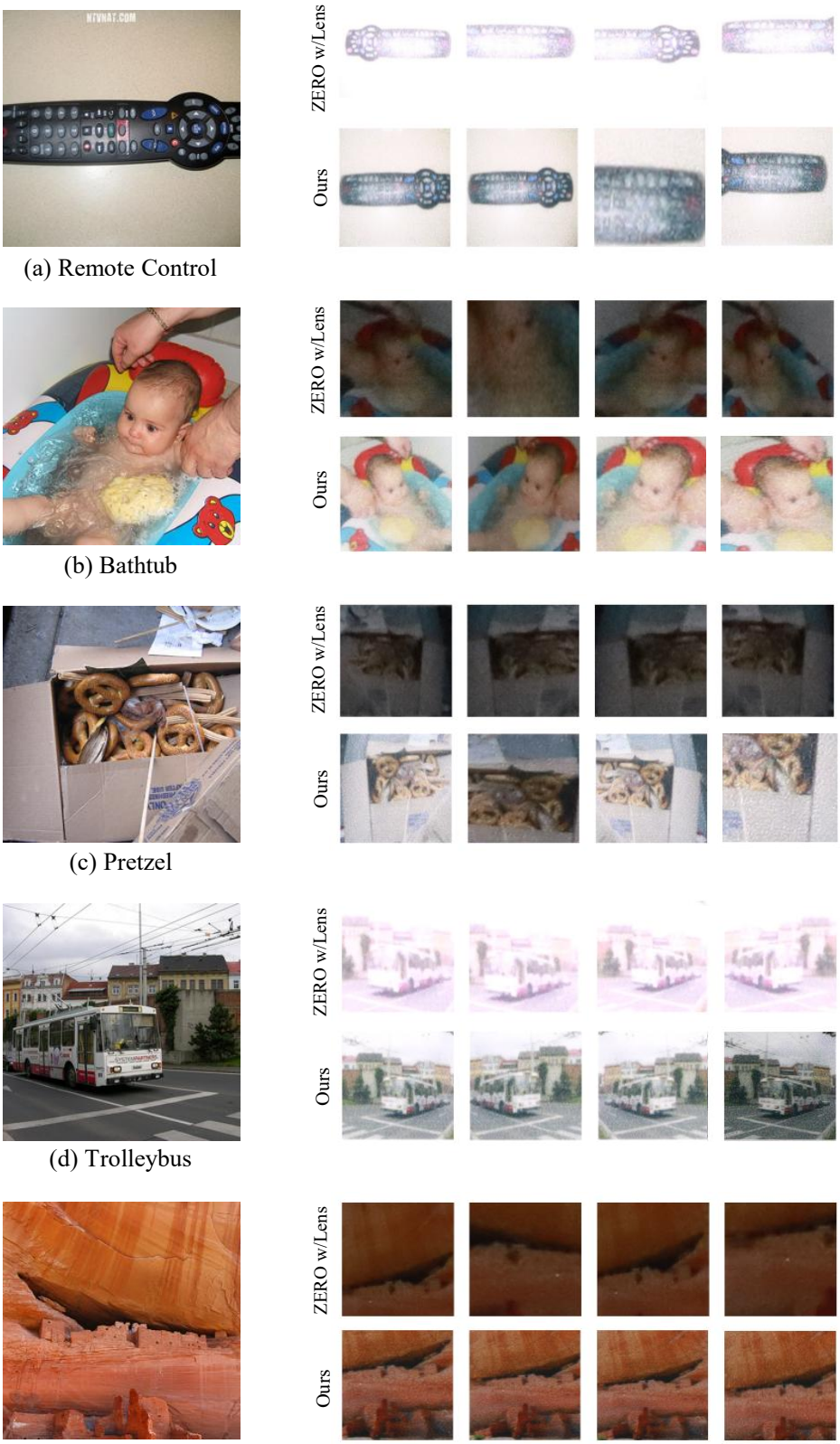
(n) ImageNet-ES-Diverse L6; jellyfish



(o) ImageNet-ES-Diverse L7

(p) ImageNet-ES-Diverse L7; American lobster

Figure 9. Additional LDA Projection Visualizations by Light Condition



(a) Remote Control

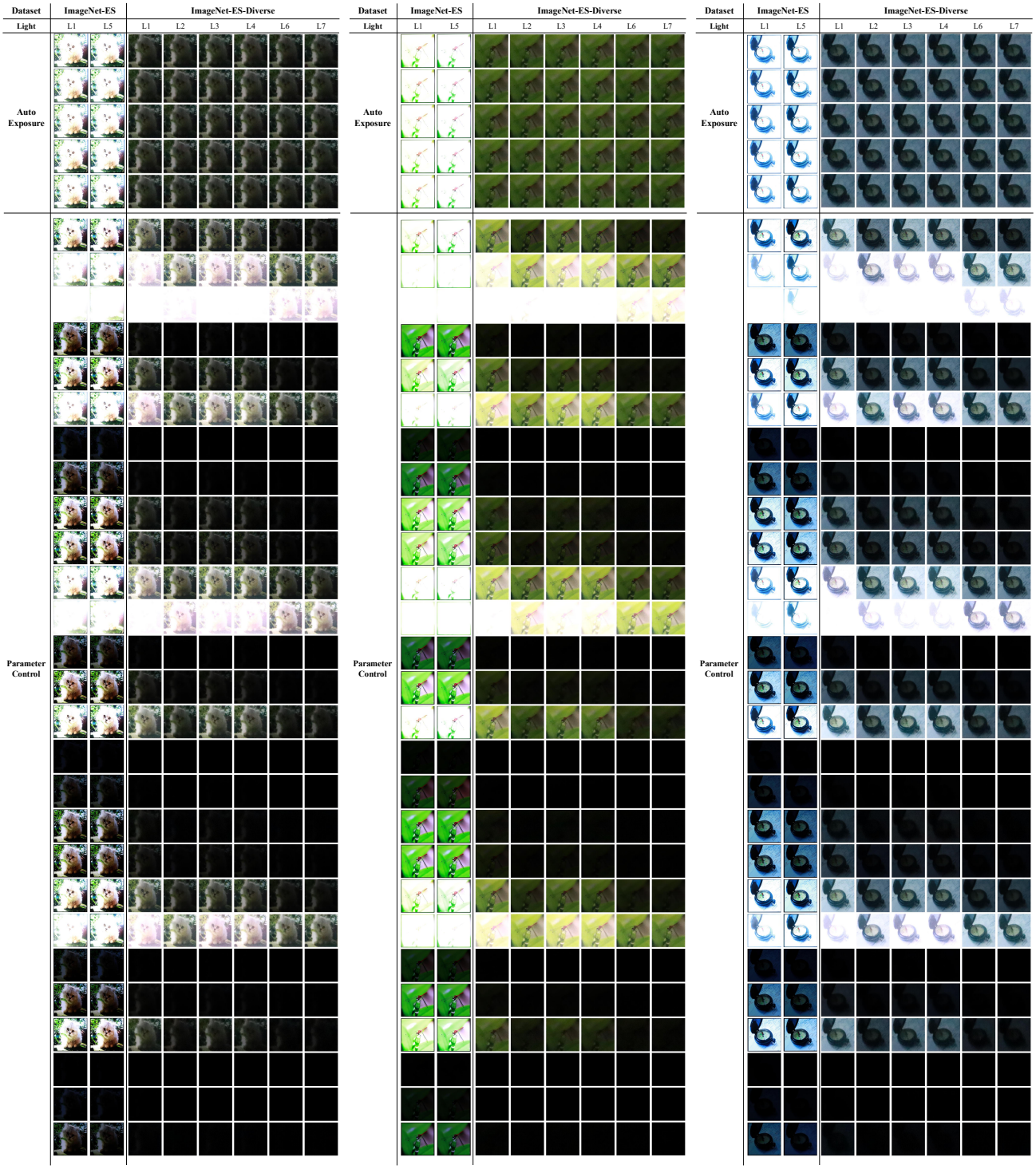
(b) Bathtub

(c) Pretzel

(d) Trolleybus

(e) Cliff Dwelling

Figure 10. Additional qualitative results



(a) Class name: Persian cat

(b) Class name: Dragonfly

(c) Class name: Magnetic compass

Figure 11. Additional qualitative results