IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 29 December 2025, accepted 9 January 2026, date of publication 20 January 2026, date of current version 17 February 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3655782

RESEARCH ARTICLE

Adaptive Classifier-Free Guidance for Robust Image-to-Image Translation

BONGGUK SON AND SANGRYUL JEON
School of Computer Science and Engineering, Pusan National University, Busan 46241, Republic of Korea

Corresponding author: Sangryul Jeon (srjeonn@pusan.ac.kr)

This research was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) under the ITRC (Information Technology Research Center) grant (IITP-2026-RS-2023-00259967), the Artificial Intelligence Convergence Innovation Human Resources Development program (IITP-2026-RS-2023-00254177), and the Leading Generative AI Human Resources Development program (IITP-2026-RS-2024-00360227), funded by the Korean government.

ABSTRACT Text-guided image-to-image translation aims to edit a source image according to a textual prompt while preserving its structure. However, existing approaches often rely on a fixed classifier-free guidance scale and a single prompt input, leading to unstable results and a poor balance between semantic fidelity and structural preservation. In this work, we propose a unified framework that improves both the *controllability* and *stability* of text-driven diffusion editing without requiring fine-tuning or paired training data. Our method introduces two key components: (1) an **adaptive guidance scheduler** that dynamically modulates the classifier-free guidance scale over timesteps based on the input image and prompt, and (2) a **prompt ensemble** mechanism that generates and ranks multiple semantically aligned prompt variants to mitigate prompt sensitivity. Together, these components form a plug-and-play framework that significantly improves editing consistency and visual quality. Extensive experiments on NuScenes, AFHQ, and CelebA demonstrate that our method consistently outperforms existing approaches across diverse scenarios.

INDEX TERMS Computer vision (CV), diffusion model, text-guided image editing, classifier-free guidance, image synthesis.

I. INTRODUCTION

The advent of text-to-image models [1], [2], [3] has significantly advanced image generation [4], [5] and editing capabilities. In particular, editing real-world images [7], [8], [9], [10], [11], [12] based on textual descriptions has been a persistent goal. However, existing methods often struggle with this task, showing limitations in balancing two competing objectives: preserving the structure of the source image and fully reflecting the semantics of the target prompt. This trade-off is typically managed by a fixed Classifier-Free Guidance (CFG) scale. However, relying on a static scale prevents the model from adapting to the varying difficulty of different editing tasks. As illustrated in Fig 1, a fixed guidance scale often fails to achieve a proper balance, resulting in either content distortion—such as the disappearance of key objects like vehicles or buildings—or insufficient semantic changes where the target attributes are not fully realized. These failure cases highlight the clear need for an adaptive mechanism that can dynamically tune prompt influence per input, enabling consistent and balanced image editing across diverse scenarios.

Existing research on this editing problem can be broadly categorized into two main approaches. The first is a training-free approach [7], [8], [9], [14], [15], which operates on pre-trained diffusion backbones without requiring fine-tuning. These methods typically function by manipulating internal representations, such as adjusting attention maps or altering latent codes. However, their reliance on a fixed CFG scale prevents them from effectively managing the structure-semantic trade-off, as the optimal guidance strength naturally varies for each input. The second approach involves fine-tuning diffusion models [11], [22], [24] on large-scale paired datasets. While these models can achieve

The associate editor coordinating the review of this manuscript and approving it for publication was Turgay Celik.

strong alignment with instructions similar to their training data, their performance is constrained by the availability of high-quality paired datasets. Furthermore, they typically apply a fixed CFG scale during inference, reintroducing the same fundamental limitations and causing performance degradation on out-of-distribution prompts.

To address these limitations, we propose a unified framework that enhances the controllability and robustness of text-driven image-to-image translation. As illustrated in Fig 2, instead of relying on rigid hyperparameters, our method introduces two key components designed to adapt to the specific needs of each edit. First, we propose an **Adaptive Guidance Scheduler (AdaCFG)** that dynamically modulates the CFG scale across diffusion timesteps. Unlike standard pipelines, our scheduler predicts an input-specific guidance trajectory, injecting strong semantic influence during the early steps to shape global content and gradually reducing it in later steps to preserve fine structural details. Second, to mitigate the inherent sensitivity of diffusion models to prompt phrasing, we introduce a **Prompt Ensemble** strategy. By leveraging a Large Language Model (LLM) [13] to generate semantically consistent prompt variants and aggregating their outputs, we ensure robust editing performance without the need for manual prompt engineering.

Our contributions can be summarized as the development of a robust, unified framework that effectively resolves the trade-off between semantic fidelity and structural preservation. By synergizing input-aware temporal guidance with inference-time prompt augmentation, our method acts as a lightweight, plug-and-play module that can be integrated into existing diffusion frameworks. Extensive experiments demonstrate that our approach consistently outperforms state-of-the-art baselines across diverse benchmarks, including NuScenes [28], AFHQ [29], and CelebA-HQ [30], yielding results that are both visually coherent and semantically faithful.

The remainder of this paper is organized as follows. Section II reviews related work in text-guided image editing and adaptive guidance techniques. Section III details the proposed methodology, including the architecture of the Adaptive Guidance Scheduler and the Prompt Ensemble mechanism. Section IV provides comprehensive experimental results, including implementation details, quantitative and qualitative comparisons with state-of-the-art methods, an ablation study, and a user study. Finally, Section V discusses the limitations of the proposed approach and concludes the paper.

II. RELATED WORK

A. DIFFUSION MODELS

Diffusion models [4], [5], [6] are probabilistic generative models that learn to generate data by progressively reversing a diffusion process. They are founded on two complementary random processes: the forward process and the reverse process. In the forward process, Gaussian noise is progressively added to a clean image x_0 according to a variance schedule β_t. At an arbitrary timestep t, the noisy image x_t can be expressed in a closed form as:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, I)$, $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ denotes the cumulative noise schedule [4]. The reverse process aims to reconstruct the original data by denoising x_t using a learned neural network $\epsilon_\theta(x_t, t)$, which predicts the added noise. Recent advances have demonstrated the remarkable capabilities of these AI models across broad domains [39], [40], [41], [42], establishing them as the foundation for state-of-the-art image synthesis and editing.

B. IMAGE-TO-IMAGE TRANSLATION

Image-to-Image (I2I) translation aims to modify an input image according to a target domain or textual instruction while preserving spatial consistency. Early works utilized GAN-based frameworks [29], [30] for specific domains. Recent research has expanded into more specialized areas, such as face transformation, utilizing heterogeneous prototype learning and multi-domain normalization techniques [45], [46]. With the rise of diffusion models, text-guided editing has become a dominant paradigm. Training-free approaches, such as P2P [8] and PnP [7], operate on pre-trained backbones by manipulating attention maps or latent codes to inject semantic changes. While flexible, these methods often rely on a single prompt and fixed hyperparameters. Fine-tuning-based methods [11], [24] improve alignment by training on paired datasets but are limited by data availability. Crucially, high-quality generated images from these models can serve as effective data augmentation sources for downstream computer vision tasks, such as salient object detection (SOD) [43], [44], highlighting the importance of robust structure preservation in editing capability.

C. CLASSIFIER-FREE GUIDANCE

Classifier-Free Guidance (CFG) [6], [19] is a widely adopted technique for conditional sampling, enabling prompt-controlled generation without an external classifier. It extrapolates the noise prediction by linearly combining conditional and unconditional estimates:

$$\tilde{\epsilon}_\theta(x_t, c) = \epsilon_\theta(x_t, \emptyset) + w \cdot (\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t, \emptyset)) \tag{2}$$

where $w \geq 1$ is the guidance scale. Increasing w strengthens semantic alignment but often degrades image fidelity. While recent studies have explored optimizing CFG [33], [34], [35], [36], [37] for inference efficiency or spatial consistency [38], most editing frameworks still rely on a fixed scale w across all timesteps. This rigid setting limits the model's ability to adapt to the varying semantic complexity and structural constraints of different images. In contrast, our work introduces an adaptive scheduling mechanism specifically optimized for

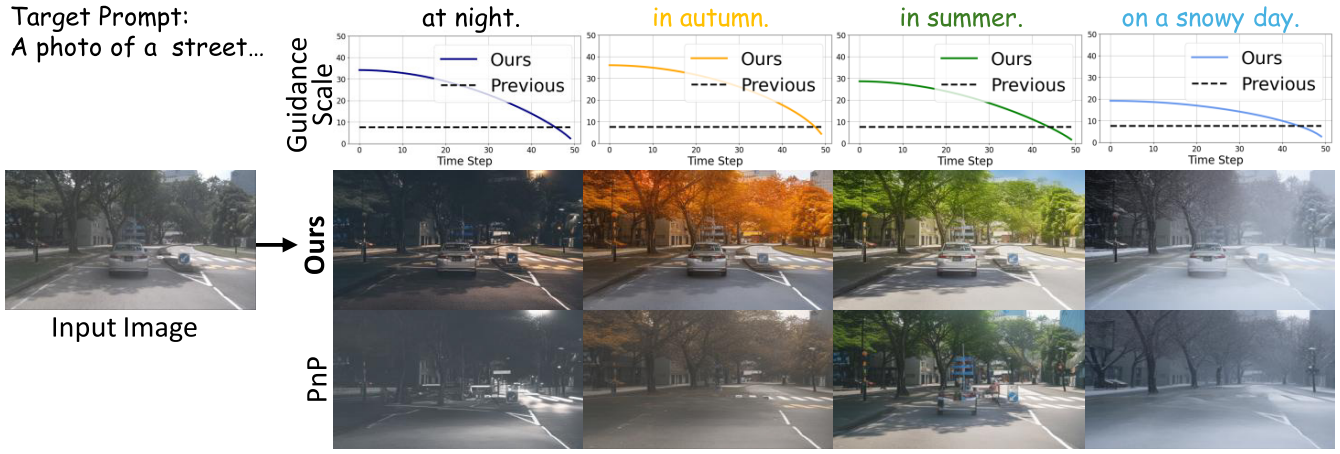

FIGURE 1. Visualization of translated images with our AdaCFG model. Given an input image and target text prompt (left), the baseline PnP method (bottom row) with a fixed guidance scale often results in insufficient semantic transformation or the disappearance of key objects. In contrast, our model dynamically adapts the guidance scale (top), which allows for strong semantic edits while ensuring that important objects are maintained, structures are preserved, and textures are enriched (middle).

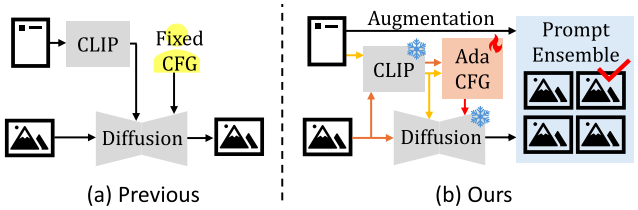

FIGURE 2. Comparison between previous approaches and our proposed framework. (a) Existing methods apply a fixed scale throughout the diffusion process, which often leads to either weak semantic expression or structural degradation.(b) Our method introduces two key innovations. (1) an Adaptive CFG (AdaCFG) module that predicts input-specific and timestep-aware guidance strengths, and (2) a Prompt Ensemble strategy that generates and ranks multiple semantically aligned prompt variants.

the image editing task, balancing the trade-off between semantic reflection and structure preservation dynamically.

III. METHOD

Our objective is to enhance the controllability and stability of text-driven diffusion editing without requiring fine-tuning of the diffusion backbone or large-scale paired training data. Unlike CFG modulation strategies [33], [34], [35], [36], [37] primarily optimized for general image generation, our framework is specifically designed to address the unique challenges of *image editing*, where the model must balance two conflicting goals: preserving the structural integrity of the source image while faithfully reflecting the semantic changes of the target prompt. To achieve this, we propose a unified framework comprising two core mechanisms: an **Adaptive Guidance Scheduler (AdaCFG)** and a **Prompt Ensemble** strategy.

A. ADAPTIVE GUIDANCE SCHEDULER

Existing approaches rely on a fixed classifier-free guidance scale, which fails to adapt to the input image and prompt, often leading to prompt-sensitive and inconsistent outputs.

We address this by introducing an *Adaptive Guidance Scheduler (AdaCFG)*, which dynamically modulates the guidance strength at each diffusion timestep based on both the input image I_{src} and the target prompt T_{target}. To effectively balance semantic alignment with structural preservation, we design a monotonically decreasing schedule over diffusion steps. This allows for strong semantic influence in the crucial early stages [31]—when the image is still highly stochastic—while gradually reducing the guidance in later steps to better preserve the structural details of the input.

Specifically, we parameterize the cosine-based scheduler [32] with two adaptive components: the *initial guidance strength* ω_0, and the *decay velocity* β, which controls the steepness of the decay. The resulting adaptive guidance schedule is defined as:

$$\omega(t) = \omega_0 \cdot \left(\cos\left(\frac{\pi}{2} \cdot \frac{t}{T}\right)\right)^{2\beta} \tag{3}$$

where T is the total number of diffusion steps and $t \in [0, T]$ is the current timestep. The parameter ω_0 determines the overall intensity of the edit, while β governs the rate at which guidance strength decays over time.

The necessity of learning both parameters is illustrated in Fig 1: a major transformation such as "night" requires a higher initial strength ω_0 and steeper decay β to inject strong semantics early without corrupting fine details later. In contrast, a subtle edit such as "snow" benefits from a lower initial strength and a more gradual decay. By predicting both ω_0 and β, AdaCFG tailors the entire guidance trajectory to the specific nature of each edit, aligning with the known dynamics of diffusion models, where early steps shape global semantics and later steps refine fine-grained structure.

1) FORMULATION

To implement AdaCFG, we employ a lightweight MLP $\Psi(\cdot)$ that jointly predicts the pair (ω_0, β) from the concatenated

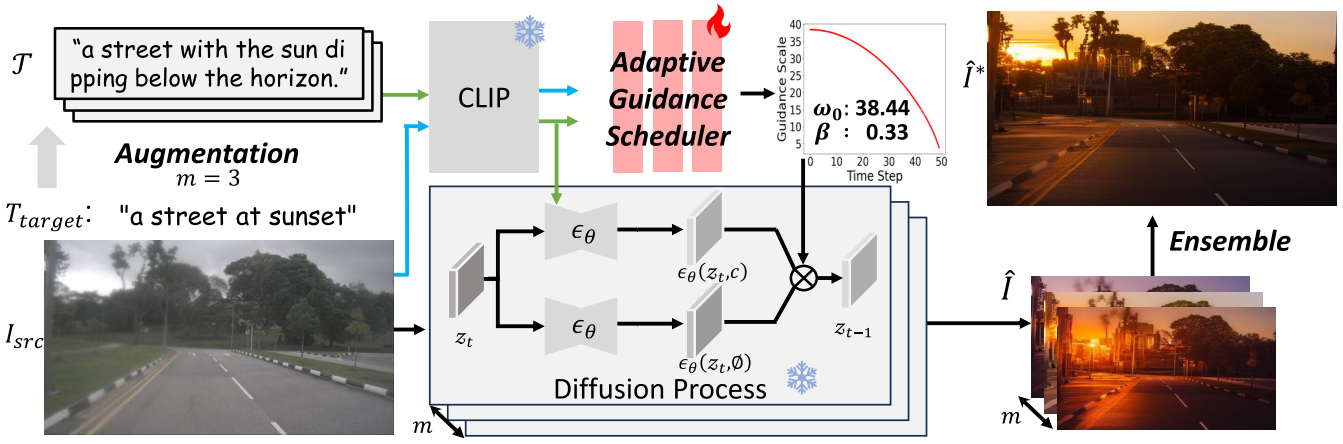

FIGURE 3. Overall pipeline. Given a source image I_{src} and a target prompt T_{target}, a large language model (LLM) is first used to infer a semantic category and generate a diverse prompt set $\mathcal{T} = \{T_1, T_2, \ldots, T_m\}$ within that category. Each prompt $T_i \in \mathcal{T}$ is paired with I_{src} to generate an output image $\hat{I}_i$ using the diffusion model ϵ_θ. During generation, classifier-free guidance (CFG) is applied, where both the maximum guidance scale and its decay over timesteps are adaptively predicted per input. The resulting image set $\hat{I} = \{\hat{I}_1, \ldots, \hat{I}_m\}$ is evaluated using CLIP (for semantic alignment with T_{target}) and DINO (for structural similarity to I_{src}). The image $\hat{I}^* \in \hat{I}$ that best balances both criteria is selected as the final output.

CLIP [26] embeddings of the input image and target prompt:

$$(\omega_0, \beta) = \Psi\left(\text{CLIP}(I_{\text{src}}) \,\|\, \text{CLIP}(T_{\text{target}})\right) \tag{4}$$

Here, $\|$ denotes vector concatenation, and the function Ψ serves as the core component of the adaptive scheduler. By leveraging CLIP's semantic representations, AdaCFG produces edit-specific, timestep-aware guidance curves in a fully differentiable and lightweight manner.

2) TRAINING OBJECTIVE

The adaptive guidance scheduler is trained to predict the pair (ω_0, β) that balances semantic alignment and structural preservation throughout the diffusion process. We define the total loss as a weighted sum of three terms:

- Semantic Alignment Loss $\mathcal{L}_{\text{sem}}$: Encourages the output image $\hat{I}_i$ to align with the intended prompt T_i, measured via CLIP-based cosine similarity:

$$\mathcal{L}_{\text{sem}} = 1 - \cos\left(\text{CLIP}(\hat{I}_i), \text{CLIP}(T_i)\right) \tag{5}$$

- Structural Fidelity Loss $\mathcal{L}_{\text{str}}$: Promotes preservation of spatial structure by minimizing the cosine similarity between DINO [27] last features of the source and generated images:

$$\mathcal{L}_{\text{str}} = 1 - \cos\left(\text{DINO}(\hat{I}_i), \text{DINO}(I_{\text{src}})\right) \tag{6}$$

- Negative Prompt Penalty $\mathcal{L}_{\text{neg}}$: Discourages undesirable visual artifacts by penalizing similarity to negative prompt T_{neg}:

$$\mathcal{L}_{\text{neg}} = \cos\left(\text{CLIP}(\hat{I}_i), \text{CLIP}(T_{\text{neg}})\right) \tag{7}$$

The overall training objective is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{str}}\mathcal{L}_{\text{str}} + \lambda_{\text{neg}}\mathcal{L}_{\text{neg}} \tag{8}$$

where $\lambda_{\text{sem}}, \lambda_{\text{str}}, \lambda_{\text{neg}}$ are weighting coefficients.

B. PROMPT ENSEMBLE WITH AUGMENTATION

While AdaCFG dynamically adjusts guidance over time, it assumes the prompt itself is fixed and reliable. In practice, however, diffusion models are often highly sensitive to the phrasing of the input prompt.

To address the sensitivity to prompt variations, we propose a *prompt ensemble* framework that generates and evaluates multiple semantically consistent prompts derived from the original user instruction. Concretely, using a large language model (LLM) [13], we generate a set of m augmented prompts $\mathcal{T} = \{T_1, T_2, \ldots, T_m\}$, where each T_i is a semantically relevant variant of T_{target}. These prompts are designed to explore the local semantic neighborhood while preserving the core intent.

For each $T_i \in \mathcal{T}$, the image-to-image model produces a corresponding output image $\hat{I}_i = \mathcal{F}(I_{\text{src}}, T_i)$, where $\mathcal{F}$ is the conditional generation function. To select the most appropriate result $\hat{I}^*$, we simply adopt the same scoring criteria used in the adaptive guidance scheduler: semantic alignment with the prompt, structural fidelity to the source image, and a penalty for undesirable attributes. These metrics are computed for each candidate $\hat{I}_i$, and the final output is selected as the one with the highest overall score:

$$\hat{I}_i^* = \arg\max_{\hat{I}_i \in \{\hat{I}_1, \ldots, \hat{I}_m\}} \mathit{Score}(\hat{I}_i). \tag{9}$$

Detailed formulation of the scoring mechanism and the selection process is provided in the Appendix A.

IV. EXPERIMENTS

A. IMPLEMENTATION DETAILS

We train and evaluate our method across three datasets: NuScenes [28], AFHQ [29], and CelebA-HQ [30]. To ensure a rigorous evaluation, we constructed specific subsets for training, validation, and testing from the official splits of

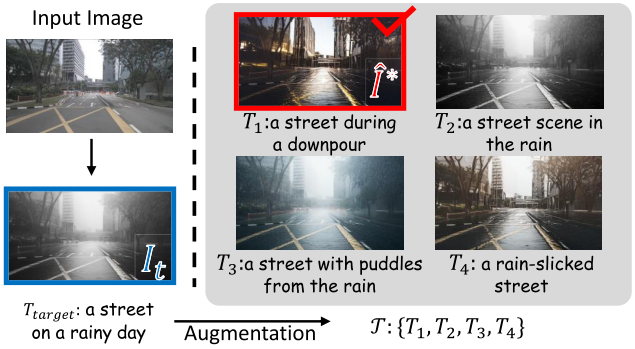

FIGURE 4. Illustration of our prompt ensemble framework. Given a source image and a target prompt, we augment T_{target} to the diverse yet semantically consistent variations $\mathcal{T} = \{T_1, T_2, T_3, T_4\}$. Compared to the baseline result (blue box), the selected output (red box) better captures the desired rainy atmosphere ("a street during a downpour") while preserving key structural details.

each dataset. For NuScenes, we randomly sampled 2,000 front-view street images from the official training split to construct our training set, while utilizing 100 images each from the validation and test splits for evaluation. Similarly, for CelebA-HQ and AFHQ, we sampled 1,000 images from the training set, and 100 images each for validation and testing. All images are resized to 512 × 512 before being processed by the diffusion model. For each dataset, we define a set of representative translation targets appropriate to its domain: for NuScenes, we use 10 weather- and time-based attributes such as "rainy", "night", or "sunset"; for AFHQ, we target 6 stylized domains such as "sculpted face" and "robotic face"; and for CelebA-HQ, we use 6 appearance domain such as "thick beard" or "red hair".

Our AdaCFG model is implemented as stacked linear layers consisting of 6 sequential residual blocks where each block is composed of two MLP layers, and trained with the Adam optimizer setting learning rate to 2.0×10^{-7}. We set the number of augmented prompts to 5 for prompt ensemble ($m = 5$) using GPT-4o [13]. When plugging-in our model to PnP, we adopt Stable Diffusion v2.1-base as a backbone network, and fix the attention injection strength at 0.9 to strongly preserve structural features. For I-P2P, we use their default pre-trained backbones and follow its standard setup where the image CFG scale is set to one-tenth of the text CFG scale. The weighting coefficients in Eq.(7) are set to $\lambda_{sem} = 1.0$, $\lambda_{str} = 0.1$, and $\lambda_{neg} = 0.8$. These values were chosen to balance the different scales of similarity scores derived from DINO (structure) and CLIP (semantic/negative) encoders. All experiments are conducted on a single NVIDIA A6000 GPU. More implementation details and experimental results are included in the supplementary material.

B. BASELINES

We evaluate our method against a diverse set of baselines, categorizing them into four groups: 1) **Latent-based methods** such as SDEdit [9] and CycleDiff [16]; 2) **Attention-based training-free methods** like P2P+NTI [8], [18], P2P-zero [15], and PnP [7]; 3) **Fine-tuning-based models** including I-P2P [11], HQ-Edit [24], and IP-Adapter [17]; and 4) **Adaptive guidance strategies** including S-CFG [38] and CFG++ [37]. Since S-CFG and CFG++ are originally designed for text-to-image generation, we adapted them for the editing task by employing DDIM inversion to retrieve the initial latents before applying their respective guidance sampling.

C. QUALITATIVE EVALUATION

As shown in Fig 5, for "daytime sky" and "snow scene", the baselines often distort key objects like buildings and vehicles, whereas our method maintains the original structure while accurately reflecting the target attributes. This is also evident in Fig 6, where AdaCFG achieves more powerful style transfers for prompts like "sculpted", "robotic" and "cat". Similarly, in Fig 6, for attribute edits like "thick beard", "white hair", and "red hair", our method preserves the subject's identity and structure much more effectively.

D. USER STUDY

We conducted a user study on Amazon Mechanical Turk (AMT) using a two-alternative forced choice (2AFC) protocol on the NuScenes dataset. We collected 1,854 responses from 291 valid evaluators who were asked to choose the image with better overall quality (considering both semantics and structure) between a baseline's output and the same baseline enhanced with our AdaCFG. The results in Fig 7 show a decisive preference for our method. AdaCFG was chosen over PnP in 70.6% of comparisons and over I-P2P in 79.2% of cases. Chi-square tests confirm that these preferences are statistically significant ($p < 0.001$), demonstrating that our approach generates results that are not only quantitatively superior but also significantly more appealing to human evaluators. Even when baseline performance is strong, user preferences may be split; however, AdaCFG clearly improves results in more challenging cases, highlighting its robustness. Further details on the experimental protocol and statistical analysis are provided in Appendix B.

E. QUANTITATIVE EVALUATION

To quantitatively evaluate editing performance, we employ three complementary metrics that capture different aspects of edit quality. Following prior works, we assess semantic alignment using the CLIP-based cosine similarity [26] between the generated image and the target prompt, and structural preservation using the cosine similarity between DINO [27] features of the input and edited images. In addition, we adopt the recently proposed *Alignment Score* from HQ-Edit [24], which leverages GPT-4o [13] to holistically evaluate whether the edited image accurately reflects the prompt while maintaining the core content of the source. Unlike CLIP or DINO scores, which individually capture only one side of the trade-off, this LLM-based score offers a

FIGURE 5. Qualitative comparisons on the NuScenes dataset. Plugging-in our AdaCFG to I-P2P and PnP yields results that better reflect the target prompt while preserving the original structure.

TABLE 1. Quantitative comparisons on Nuscenes, AFHQ and CelebA-HQ datasets. We report CLIP [26] and DINO [27] scores to separately assess semantic alignment and structural preservation. The metric marked with † provides a unified evaluation of overall edit quality using GPT-4o [13], based on the source image, target prompt, and edited output. This follows the evaluation protocol introduced in HQ-Edit [24], with detailed prompt formats provided in the supplementary material.

	Nuscenes [28]			AFHQ [29]			CelebA-HQ [30]		
	CLIP ↑	DINO ↑	Align.† ↑	CLIP ↑	DINO ↑	Align.† ↑	CLIP ↑	DINO ↑	Align.† ↑
S-CFG [38]	18.4	33.1	14.4	20.7	51.7	32.4	19.2	61.8	19.4
CFG++ [37]	15.2	18.6	19.8	21.2	18.6	50.8	19.0	39.1	35.7
SDEdit [9]	21.6	46.3	58.5	20.7	71.5	63.7	18.7	78.0	70.6
CycleDiff [16]	22.4	55.5	69.3	25.9	44.8	85.7	23.9	59.0	80.2
HQ-Edit [24]	21.9	33.9	57.8	22.6	57.8	65.8	20.2	60.1	67.8
P2P+NTI [8], [18]	23.2	22.0	40.4	25.5	19.9	75.5	24.5	32.0	53.4
IP-Adapter [17]	20.4	31.9	44.5	22.9	34.5	61.3	21.2	38.7	72.3
P2P-zero [15]	20.3	57.1	54.2	19.0	76.4	43.6	18.5	72.1	51.5
I-P2P [11]	19.9	79.3	76.1	24.3	65.8	72.2	21.5	88.8	76.7
+AdaCFG	22.1	78.2	**86.2**	26.5	45.8	**83.6**	23.4	88.7	**84.7**
gain	+2.2	-1.1	**+10.1**	+2.2	-20.0	**+11.4**	+1.9	-0.1	**+8.0**
PnP [7]	25.0	51.0	85.5	24.5	63.8	79.3	23.8	71.9	85.8
+AdaCFG	24.8	59.5	**92.4**	26.4	53.2	**90.7**	24.8	74.2	**87.9**
gain	-0.2	+8.5	**+6.9**	+1.9	-10.6	**+11.4**	+1.0	+2.3	**+2.1**

unified and more interpretable metric for assessing the overall balance between structure and semantics. Further details on the measurement protocol for the Alignment Score, including prompt format and evaluation setup, are provided in the Appendix A.

As shown in Table 1 and Fig. 8, our method consistently improves editing quality across all datasets and settings. While Table 1 provides detailed metrics, Fig. 8 is included to specifically visualize the critical trade-off relationship between semantic alignment (CLIP ↑) and structure preservation (DINO ↑). This figure clearly illustrates that our AdaCFG-augmented variants (represented by red stars) achieve a superior balance, positioning them in the top-right region relative to most other baselines. This positioning

FIGURE 6. Qualitative comparisons on the AFHQ and CelebA-HQ datasets. Applying our AdaCFG to PnP enables more faithful semantic edits while better preserving the subject's original identity.

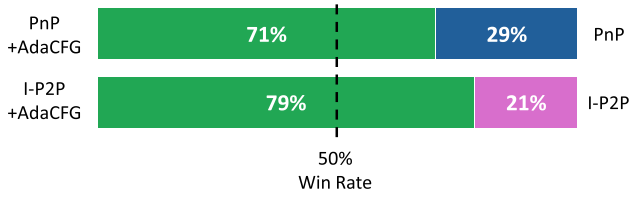

FIGURE 7. User Study. Participants were shown image pairs and asked to choose the more faithful and visually coherent edit, given the same source image and target prompt.

indicates that our method attains high performance in both semantic fidelity and structural consistency simultaneously, effectively resolving the trade-off that limits existing approaches.

This balanced improvement is numerically evident in Table 1. Notably, recent adaptive guidance methods such as S-CFG [38] and CFG++ [37], while innovative for text-to-image generation, show suboptimal performance in an editing context. As shown in Table 1, these methods yield much lower Alignment scores (e.g., 14.4 for S-CFG on NuScenes), as general-purpose guidance often fails to address the specific structural and semantic demands of image editing. For the

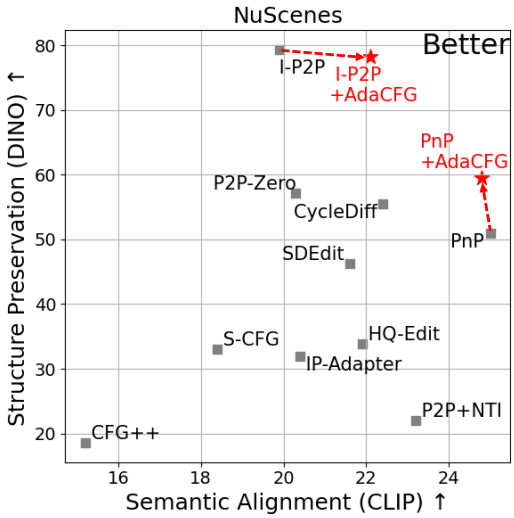

FIGURE 8. Trade-off analysis on balancing semantic alignment and structure preservation. Our AdaCFG-augmented variants (red stars) consistently occupy the top-right region, indicating superior balance between semantic fidelity and structural preservation.

I-P2P baseline, integrating AdaCFG yields substantial gains, especially in the Alignment Score: +10.1 on NuScenes, +11.4 on AFHQ, and +8.0 on CelebA-HQ. Notably, CLIP similarity also improves +2.2, while DINO similarity remains stable, reflecting a controlled trade-off that favors semantic fidelity without catastrophic structural loss. In the PnP baseline, AdaCFG enhances DINO scores substantially (up to +8.5 on NuScenes), while also boosting the overall Alignment Score by +6.9 to +11.4. Interestingly, this improvement comes with only marginal or no change in CLIP score, suggesting that our method strengthens structural consistency without sacrificing semantic quality.

We attribute the lowered DINO scores to the inherent trade-off between structure preservation and necessary semantic transformation. A clear example is observed in the AFHQ dataset for prompts like "a sculpted dog face" or "a robotic dog face" (Fig 6). These edits demand substantial structural modifications, reflected in a drop in DINO score (−10.6). However, this structural sacrifice leads to a significant improvement in overall editing quality, as evidenced by a large increase in Alignment Score (+11.4). This underscores that DINO alone can be a misleading metric—meaningful semantic edits naturally entail structural deviation, but the holistic quality of the output is ultimately better.

F. ABLATION STUDY

1) ABLATION OF EACH COMPONENT

To assess the contribution of each component in our framework, we conduct ablations on the adaptive guidance scheduler and prompt ensemble. We analyze the effect of learning each scheduler parameter—initial strength ω_0 and decay rate β—by selectively disabling them. We also vary

TABLE 2. Ablation study. We assess the contribution of each component in our framework — the initial guidance strength w_0, decay velocity β, and the number of augmented prompts m for ensemble.

init str. w_0	*velocity* β	*aug.* m	CLIP ↑	DINO ↑	Align. ↑
✗	✗	5	23.6	60.5	88.3
✓	✗		23.6	**66.3**	90.1
✓	✓	1	24.3	52.6	85.7
✓	✓	3	24.6	58.0	90.0
✓	✓	5	**24.8**	59.5	**92.4**
✓	✓	7	**24.8**	60.7	91.4

the number of augmented prompts m to examine the trade-off between robustness to prompt phrasing and inference cost.

As shown in Table 2, adjusting the initial guidance strength ω_0 improves overall editing quality: enabling ω_0 raises the Alignment score from 88.3 to 90.1. Without decay velocity β, the model applies a fixed and aggressive decay, preserving structure well (highest DINO score of 66.3), but underperforms in semantic fidelity. When both ω_0 and β are predicted, the model achieves better balance, leading to improvements in CLIP (23.6→24.8) and a notable gain in Alignment (90.1→92.4), confirming the importance of dynamic, input-aware scheduling. We also ablate the number of augmented prompts m. Our model without prompt ensemble framework ($m = 1$) yields a lower Alignment score (85.7). The best result is achieved at $m = 5$, and increasing to $m = 7$ offers marginal benefit.

2) COMPUTATIONAL EFFICIENCY

Our method employs a Prompt Ensemble mechanism that generates $m = 5$ semantically augmented prompts. It is important to note that the prompt generation by the LLM is a one-off offline process performed prior to the inference stage; thus, it incurs negligible marginal cost and carbon footprint during the actual editing phase. While the ensemble setup increases the number of forward passes per edit compared to single-prompt baselines, the actual runtime increase is sublinear.

Crucially, the execution times reported in Table 3 account for the entire end-to-end pipeline, including the candidate filtering process (i.e., CLIP/DINO scoring and final selection). Even with this selection overhead, the sequential implementation of our method incurs lower execution time than the expected 5× overhead due to shared computational resources. Furthermore, since each augmented prompt operates independently, our framework naturally supports parallel execution. When run across 5 GPUs, each prompt-conditioned diffusion run is dispatched to a separate GPU, leading to a near-single-prompt latency (∼1.1×) while substantially improving editing robustness and output quality. This makes our method particularly suitable for deployment in cloud-based high-quality editing services where multiple GPUs are available, prioritizing achieving the highest possible visual fidelity and structural consistency over real-time interactive feedback.

TABLE 3. Inference time comparison. The table compares the execution time of baseline models (I-P2P, PnP) with and without AdaCFG. We show both sequential and parallel implementations of our prompt ensemble approach. The parallel version distributes each of the $m = 5$ prompt-conditioned diffusion runs across separate GPUs, significantly reducing total latency.

	aug. m	Execution Time (s)	Time Difference (s)	Time Ratio
I-P2P	1	7.665	-	-
+ AdaCFG (seq.)	5	35.707	28.042	× 4.658
+ AdaCFG (par.)	**5**	**8.119**	**0.454**	× **1.059**
PnP	1	10.936	-	-
+ AdaCFG (seq.)	5	38.677	27.741	× 3.537
+ AdaCFG (par.)	**5**	**11.398**	**0.462**	× **1.042**

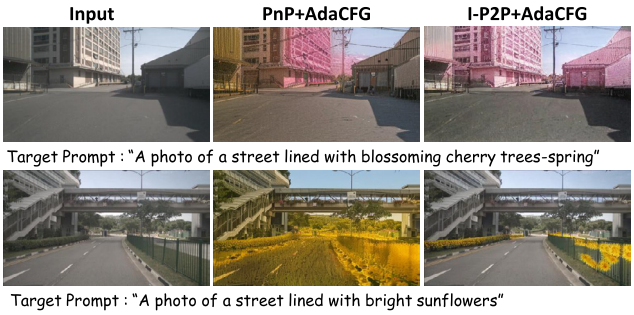

FIGURE 9. Failure cases of our method with different backbones. When the target object is absent (top: trees) or requires significant geometric generation (bottom: sunflowers), the model tends to apply textures to existing surfaces (buildings, roads) rather than inserting new 3D objects. Note that while I-P2P allows slightly more flexibility than PnP, both struggle with object hallucination.

V. CONCLUSION AND LIMITATIONS

In this work, we have presented a robust and unified framework for text-guided image editing that effectively resolves the trade-off between structural preservation and semantic alignment. By introducing the **Adaptive Guidance Scheduler (AdaCFG)** and a robust **Prompt Ensemble** strategy, our method dynamically tailors the diffusion process to the specific needs of each input, overcoming the limitations of fixed guidance scales used in prior arts. Extensive experiments on NuScenes, AFHQ, and CelebA-HQ demonstrate that our approach consistently outperforms existing state-of-the-art methods in both quantitative metrics and human preference.

A. CONTRIBUTIONS AND IMPACT

A key advantage of our framework is its efficiency; it achieves highly realistic editing capabilities by training only a lightweight adapter, eliminating the need for computationally expensive fine-tuning of the entire diffusion backbone. This efficiency, combined with our model's robustness to diverse prompts, opens up significant possibilities for practical applications. In particular, our results on the NuScenes dataset suggest that our method can serve as a powerful tool for *semantic data augmentation* in autonomous driving—generating diverse weather and lighting conditions to train

Algorithm 1 PnP Instruction

Given categories and prompts as shown below, create a list of five alternative sentences for each category that preserve the meaning of the given prompt and category.
{ 'clear day': 'a photo of a street on a clear day',
'foggy day': 'a photo of a street on a foggy day',
'rainy day': 'a photo of a street on a rainy day',
'snowy day': 'a photo of a street on a snowy day',
'night': 'a photo of a street at night',
'sunset': 'a photo of a street at sunset',
'spring': 'a photo of a street in spring',
'summer': 'a photo of a street in summer',
'autumn': 'a photo of a street in autumn',
'winter': 'a photo of a street in winter' }
The output should be in JSON format, with each category containing a list of prompts.

Algorithm 2 I-P2P Instruction

Given categories and prompts as shown below, create a list of five alternative sentences for each category that preserve the meaning of the given prompt and category. Then, convert each alternative sentence into a simple imperative expression that instructs to change the image accordingly.
{ 'clear day': 'a photo of a street on a clear day',
'foggy day': 'a photo of a street on a foggy day',
'rainy day': 'a photo of a street on a rainy day',
'snowy day': 'a photo of a street on a snowy day',
'night': 'a photo of a street at night',
'sunset': 'a photo of a street at sunset',
'spring': 'a photo of a street in spring',
'summer': 'a photo of a street in summer',
'autumn': 'a photo of a street in autumn',
'winter': 'a photo of a street in winter' }
The output should be in JSON format, with each category containing a list of prompts.

Algorithm 3 Score

Your task is to provide a single score from 0 to 100 that evaluates the second image. Your score should be based on a holistic assessment of the following three criteria:
1. Faithfulness to the Edit Text This is the most important criterion. How well does the second image implement the changes described in the EDIT TEXT? If the image does not sufficiently reflect the text, rate very low (<50). Completeness: All changes mentioned in the text must be present in the second image. Accuracy: No changes should be made that were not described in the text. A perfect score requires the edit to be exactly as instructed.
2. Preservation of Core Content Does the second image preserve the important elements and identity of the first image that were not supposed to be changed by the EDIT TEXT? For stylistic edits, the original content and details must be preserved. Rate low (<80) if the content is unnecessarily changed. The core identity of the image must be maintained. Rate very low (<50) if the second image becomes unrelated to the first without reason.
3. Quality of the Second Image Is the second image clear, realistic, and free of technical flaws? The image quality should be high. Rate low (<70) for noticeable issues like digital artifacts, distortions, blurriness. Provide a few lines for explanation and give the final response in a json format as such:
{"Explanation": "", "Score": ""}

more robust perception systems. Furthermore, in the creative design industry, our prompt ensemble mechanism offers users a controllable way to explore diverse visual concepts without manual prompt engineering.

B. LIMITATIONS AND FAILURE CASES

Despite these advancements, our method is not without limitations. A primary challenge arises from the inherent trade-off between structural preservation and semantic manipulation. Since the guidance scale is applied globally to the latent features, our method prioritizes maintaining the spatial layout of the source image. Consequently, tasks requiring significant geometric changes—such as inserting new objects at specific locations or hallucinating missing semantic categories—remain difficult.

Figure 9 illustrates these failure modes. In the top row, the prompt requests "blossoming cherry trees" in a scene where no trees exist. Due to the absence of a target object to map the transformation onto, the model incorrectly applies the pink floral texture to the building façade. Similarly, the bottom row shows the result for "street lined with bright sunflowers." Instead of erecting 3D sunflower geometries, the model paints the road surface with a sunflower texture. This occurs because the strong structural constraint prevents the generation of vertical objects that would occlude or break the flat geometry of the road. Future work will focus on addressing these issues by exploring local guidance mechanisms or object-aware attention control to enable precise object insertion without compromising global structural integrity.

C. ETHICAL CONSIDERATIONS

While our method enhances the quality and accessibility of image editing, we acknowledge the potential risks associated with generative AI, such as the creation of deepfakes or the manipulation of visual information for malicious purposes. The ability to alter identity attributes or environmental contexts requires responsible deployment. We emphasize the importance of developing accompanying detection technologies and watermarking standards to distinguish synthesized content from reality, ensuring that these advancements contribute positively to the digital ecosystem.

TABLE 4. Quantitative comparison using objective metrics (LPIPS, SSIM) and masked evaluations via SAM [47]. FG: Foreground, BG: Background. ↑ indicates higher is better, ↓ indicates lower is better.

	Nuscenes [28]		AFHQ [29]			CelebA-HQ [30]		
	LPIPS ↓	SSIM ↑	CLIP(FG) ↑	LPIPS(BG) ↓	SSIM(BG) ↑	CLIP(FG) ↑	LPIPS(BG) ↓	SSIM(BG) ↑
S-CFG [38]	81.4	30.8	21.0	21.1	80.5	18.9	24.7	75.3
CFG++ [37]	73.7	40.9	21.0	19.6	84.3	18.8	21.9	83.4
SDEdit [9]	35.2	64.0	21.1	7.2	91.1	18.7	7.1	93.0
CycleDiff [16]	28.6	71.1	25.3	9.9	90.8	22.5	10.8	91.5
HQ-Edit [24]	66.0	39.6	22.0	18.5	81.6	20.1	24.6	78.6
P2P+NTI [8], [18]	74.0	46.0	25.0	27.8	76.2	22.6	32.0	72.0
IP-Adapter [17]	64.4	40.4	23.3	22.3	78.7	19.6	31.7	73.7
P2P-zero [15]	30.6	72.6	20.1	4.7	94.4	17.7	7.2	93.8
I-P2P [11]	28.1	66.8	23.9	7.3	92.3	22.0	7.6	92.2
+AdaCFG	37.3	62.0	25.6	16.1	86.3	22.3	5.2	94.2
gain	+9.2	-4.8	+1.7	+8.8	-6.0	+0.3	-2.4	+2.0
PnP [7]	43.8	61.6	24.7	8.0	92.1	22.7	9.9	91.2
+AdaCFG	49.4	51.7	25.8	13.3	90.6	23.0	12.6	91.6
gain	+5.6	-9.9	+1.1	+5.3	-1.5	+0.3	+2.7	+0.4

TABLE 5. Ablation study on varying the scheduling function $\omega(t)$. The table shows the results of different guidance scheduling strategies. Our monotonically decreasing schedule $\omega(t)$ clearly outperforms the alternative functions $\omega'(t)$ and $\omega''(t)$ across all metrics, validating our setting.

Scheduling function	CLIP ↑	DINO ↑	Align. ↑
$\omega(t)$ - ***mono. decreasing***	**24.8**	**60.7**	**92.4**
$\omega'(t)$ - *mono. increasing*	24.1	59.5	88.45
$\omega''(t)$ - *sine wave*	24.5	56.2	88.07

APPENDIX A
MORE IMPLEMENTATION DETAILS

A. ARCHITECTURE DETAILS

Our AdaCFG model is designed as a lightweight Multilayer Perceptron (MLP) composed of 6 sequential residual blocks. Each block contains two linear layers with a hidden dimension of $d = 768$ and utilizes ReLU activation.

1) OUTPUT FORMULATION

The final output of the stack is projected to the guidance scale through a linear layer W. To ensure stable training and positive guidance values, we apply a bounded activation function followed by scaling. Specifically, given the network output h, the final adaptive guidance g is computed as:

$$g = \text{ReLU}\left(2 \cdot (\sigma(h \cdot \lambda) - 0.5)\right) \cdot \gamma_{\text{init}} + 1 \tag{10}$$

where σ denotes the sigmoid function, $\lambda = 0.1$ is a scaling factor for the raw output, and $\gamma_{\text{init}} = 100.0$ represents the initial guidance scale. We chose this value to ensure a sufficiently wide dynamic range without numerical instability, as optimal scales can sometimes exceed 50.

2) TRAINING

The model is trained using the Adam optimizer with a learning rate of 2.0×10^{-7}. Since our module maps high-dimensional features (1,536 dimensions) to a low-dimensional output (2 dimensions), we employed this conservative learning rate to prevent instability caused by potentially large gradients.

B. DETAILED SCORING MECHANISM FOR PROMPT ENSEMBLE

To automatically select the optimal result from the m candidate images generated by the prompt ensemble, we employ a unified scoring function $S(\cdot)$. This function is designed to align strictly with our training objectives, ensuring that the selected image maximizes semantic fidelity and structural preservation while minimizing artifacts.

Given a set of generated candidates $\mathcal{I} = \{\hat{I}_1, \hat{I}_2, \ldots, \hat{I}_m\}$ corresponding to the augmented prompts $\mathcal{T} = \{T_1, T_2, \ldots, T_m\}$, we calculate three component scores for each candidate $\hat{I}_i$:

1) **Semantic Score ($\mathcal{S}_{sem}$):** Measures how well the image reflects the specific prompt variant. We use the cosine similarity between the CLIP image embedding and the text embedding of the target prompt T_{target}:

$$\mathcal{S}_{sem}(\hat{I}_i) = \cos(\text{CLIP}(\hat{I}_i), \text{CLIP}(T_{target})) \tag{11}$$

2) **Structural Score ($\mathcal{S}_{str}$):** Measures the preservation of the source image's spatial structure. We compute the cosine similarity between the DINO feature embeddings of the generated image and the source image I_{src}:

$$\mathcal{S}_{str}(\hat{I}_i) = \cos(\text{DINO}(\hat{I}_i), \text{DINO}(I_{src})) \tag{12}$$

3) **Penalty Score ($\mathcal{S}_{neg}$):** Measures the presence of undesirable artifacts. We calculate the similarity to a predefined negative prompt T_{neg} (e.g., "blurry, distorted, low quality"):

$$\mathcal{S}_{neg}(\hat{I}_i) = 1 - \cos(\text{CLIP}(\hat{I}_i), \text{CLIP}(T_{neg})) \tag{13}$$

The final aggregate score $S(\hat{I}_i)$ is a weighted combination of these components, formulated to reward positive attributes

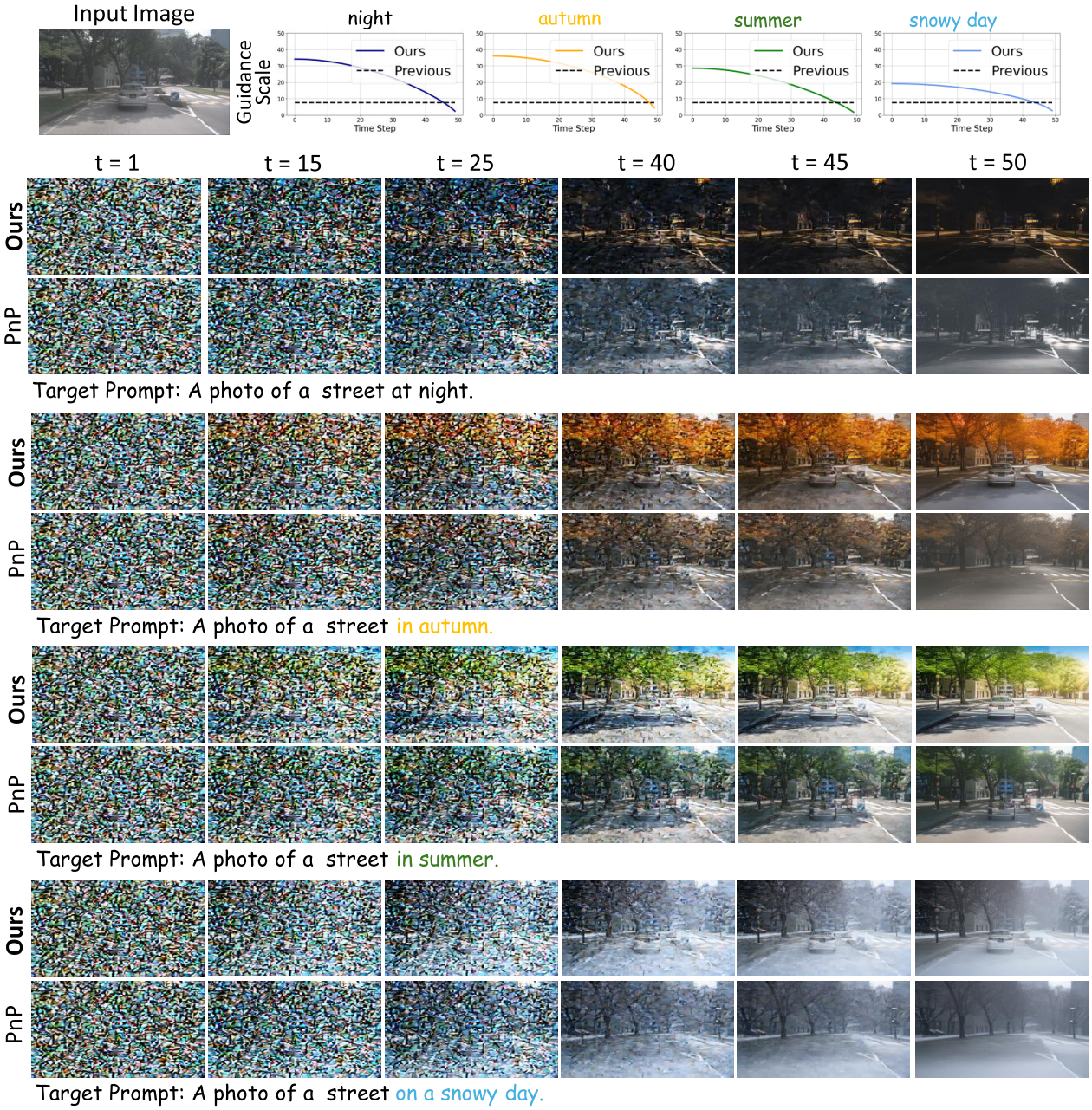

FIGURE 10. Visualization of the diffusion generation process for "A photo of a street in autumn." The top panel displays the input image and a plot comparing the CFG scale of our AdaCFG against a fixed-scale previous method. The main panels show the image generation sequence over time. In our result, the strong initial guidance scale leads to an early change in the color of the noise, infusing the target "autumn" semantics from the beginning. As the guidance scale decreases in later timesteps, our method successfully preserves the structure of the vehicles. In contrast, the previous method with a fixed, weaker guidance fails to achieve a sufficient semantic shift and also distorts the vehicle structures.

and penalize negative ones:

$$S(\hat{I}_i) = \lambda_{sem} \cdot \mathcal{S}_{sem}(\hat{I}_i) + \lambda_{str} \cdot \mathcal{S}_{str}(\hat{I}_i) + \lambda_{neg} \cdot \mathcal{S}_{neg}(\hat{I}_i) \tag{14}$$

where λ_{sem}, λ_{str}, and λ_{neg} correspond to the weighting coefficients used in the training objective (Eq. 7). Finally, the optimal output $\hat{I}^*$ is selected by maximizing this score:

$$\hat{I}^* = \underset{\hat{I}_i \in \mathcal{I}}{\operatorname{argmax}}\, S(\hat{I}_i) \tag{15}$$

This selection process ensures that the final output is not only semantically consistent with the user's intent but also structurally robust and free from visual degradation.

C. EXPERIMENTAL SETTINGS

We define specific transformation domains for each dataset to evaluate the generality of our approach. For *NuScenes*, we perform translations across 10 environmental and seasonal categories: *Rain*, *Snow*, *Clear*, *Fog*, *Night*, *Sunset*,

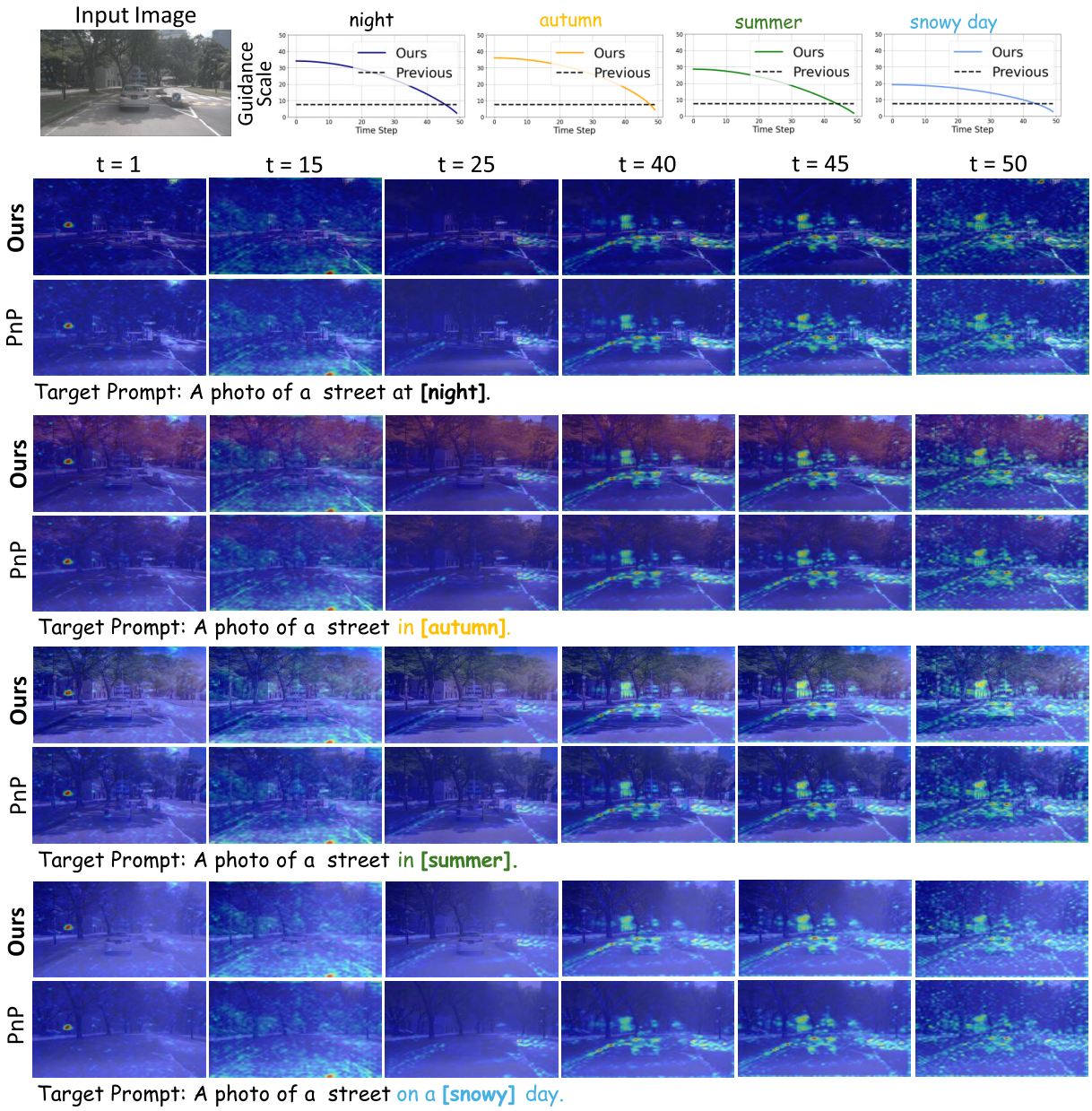

FIGURE 11. **Visualization of cross-attention maps across diffusion timesteps. We visualize the average cross-attention maps aggregated from all U-Net Up-blocks, focusing on the core editing token (marked in [brackets]). The top row displays the guidance scale schedule for each prompt. Comparing the baseline (PnP) and our method (Ours), the spatial attention distribution remains noticeably consistent across timesteps. This behavior is inherent to the PnP framework, which utilizes *attention injection* to preserve the spatial layout of the source image. This suggests that our method's superior editing capability does not arise from altering *where* the model attends (spatial attention), but rather from dynamically modulating *how strongly* the semantic guidance is applied to those attended regions via the adaptive schedule.**

Spring, *Summer*, *Autumn*, and *Winter*. For *AFHQ*, we use 6 style and species domains: *Cartoon*, *Cat*, *Sculpture*, *Robot*, *Crochet*, and *Art*. For *CelebA-HQ*, the edits focus on 6 attributes: *Beard*, *Blonde*, *Red Hair*, *White Hair*, *Tanned*, and *Sad expression*.

For each category, we evaluate the performance using 100 source images, resulting in a total of 1,000, 600, and 600 evaluation pairs for NuScenes, AFHQ, and CelebA-HQ, respectively. Prompt ensembles for all categories are generated using GPT-4o, ensuring semantically rich and diverse formulations.

To ensure a fair and consistent evaluation, the number of diffusion timesteps is fixed at 50 for both inversion and inference across all baseline methods. Additionally, except for I-P2P, we consistently apply the same negative prompt: "`ugly, blurry, low resolution, unrealistic, distortion`". These terms were selected to mitigate general quality degradation and

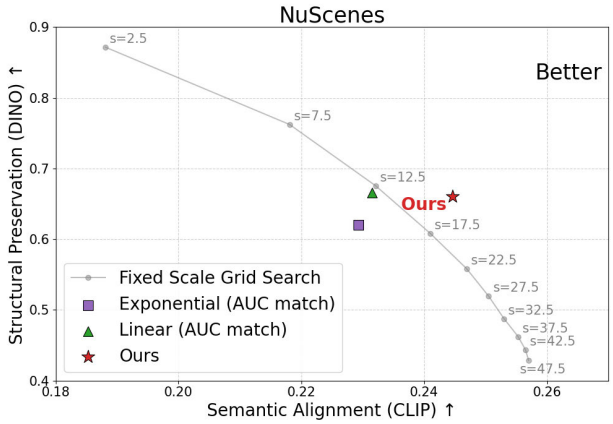

FIGURE 12. Trade-off analysis between Semantic Alignment (CLIP) and Structural Preservation (DINO) on NuScenes. The gray line represents the frontier obtained by grid-searching fixed guidance scales ($s = 2.5 \sim 47.5$). The Linear and Exponential baselines are matched to the same total guidance energy (AUC) as our method. Our method (Star) achieves the best trade-off, verifying its efficiency in allocating guidance budget.

ensure high-fidelity generation regardless of the domain.

To integrate our method with each baseline:

- **PnP:** Prompts are generated following the format shown in Algorithm 1. We set both the *attention injection timestep* and *feature injection timestep* to 0.9 to ensure strong structural preservation even under varying guidance scales from AdaCFG. We apply the aforementioned standardized negative prompt to reduce artifacts. Both the number of inference timesteps and inversion timesteps are set to 50.
- **I-P2P:** Since I-P2P takes instructional prompts, we generate them using the format illustrated in Algorithm 2. The *Image CFG scale* is set to one-tenth of the *Text CFG scale* predicted by AdaCFG, clipped to a minimum of 1.0. Unlike other baselines, we utilize a `null` text as the negative prompt for I-P2P. This is because we empirically observed that applying descriptive negative prompts to the I-P2P baseline leads to significant performance degradation. To ensure a fair comparison and maintain the baseline's optimal performance, we follow this convention for both the original I-P2P and our AdaCFG-integrated version. The number of inference timesteps is also set to 50.

1) LLM-BASED ALIGNMENT SCORE

The Alignment metric is adapted from the protocol proposed in HQ-Edit [24] and utilizes GPT-4o for a comprehensive evaluation. The original purpose in HQ-Edit was to evaluate edits based on both a source prompt and a target prompt, which did not align with our task's objective. Therefore, we designed a new instruction protocol, as shown in Algorithm 3.

For our evaluation, the inputs to GPT-4o are provided in the following order: the source image, the target (edited) image, and the target prompt. GPT-4o is instructed to assess the output based on three distinct criteria. Specifically, the model evaluates **faithfulness to the edit text** by checking how well the edited image reflects the target prompt's semantics, and **preservation of core content** by measuring the stability of the source image's key structure and identity. Finally, it assesses the **quality of the second image** in terms of overall visual realism and the absence of technical artifacts, independent of the editing accuracy. Based on this holistic assessment, GPT-4o provides a detailed explanation of its reasoning followed by a final score ranging from 0 to 100.

APPENDIX B
USER STUDY

As mentioned in the main paper, we conducted a large-scale user study on Amazon Mechanical Turk (AMT) using a two-alternative forced choice (2AFC) protocol.

A. EXPERIMENTAL PROTOCOL

As shown in Fig 13, evaluators were presented with a source image, a target prompt, and a pair of edited images—one from a baseline model and one from the same baseline enhanced with our AdaCFG—and were asked to choose the one with better overall quality. To eliminate position bias, the display order (left vs. right) of the images was fully randomized for each trial. Each participant was assigned a batch of 8 pairwise comparisons. To ensure data quality, we embedded 1 ''gold standard'' question (a pair with an obvious better choice) within each batch as an attention check. Only participants who correctly answered this validity check were included in the final analysis. Furthermore, we enforced a time limit for the selection process, ensuring that the decision time for each pair did not exceed 30 seconds to capture intuitive perceptual preferences.

B. STATISTICAL ANALYSIS

We collected a total of 1,854 valid responses. To evaluate the statistical significance of the results, we conducted a Chi-square goodness-of-fit test against a random guess baseline (50/50) and calculated 95% Confidence Intervals (CI).

- **vs. I-P2P:** Our method achieved a win rate of 79.2% (733/926). The preference was statistically significant ($\chi^2 = 314.90, p < 0.001$), with a 95% CI of [76.5%, 81.8%].
- **vs. PnP:** Our method achieved a win rate of 70.6% (655/928). The preference was statistically significant ($\chi^2 = 157.25, p < 0.001$), with a 95% CI of [67.7%, 73.5%].

These results confirm that the preference for our method is statistically robust and reflects a significant improvement in perceptual quality compared to the baselines.

APPENDIX C
ADDITIONAL QUANTITATIVE RESULTS

To demonstrate the robustness of our method using metrics independent of our optimization objectives, we report LPIPS and SSIM scores in Table 4. For object-centric

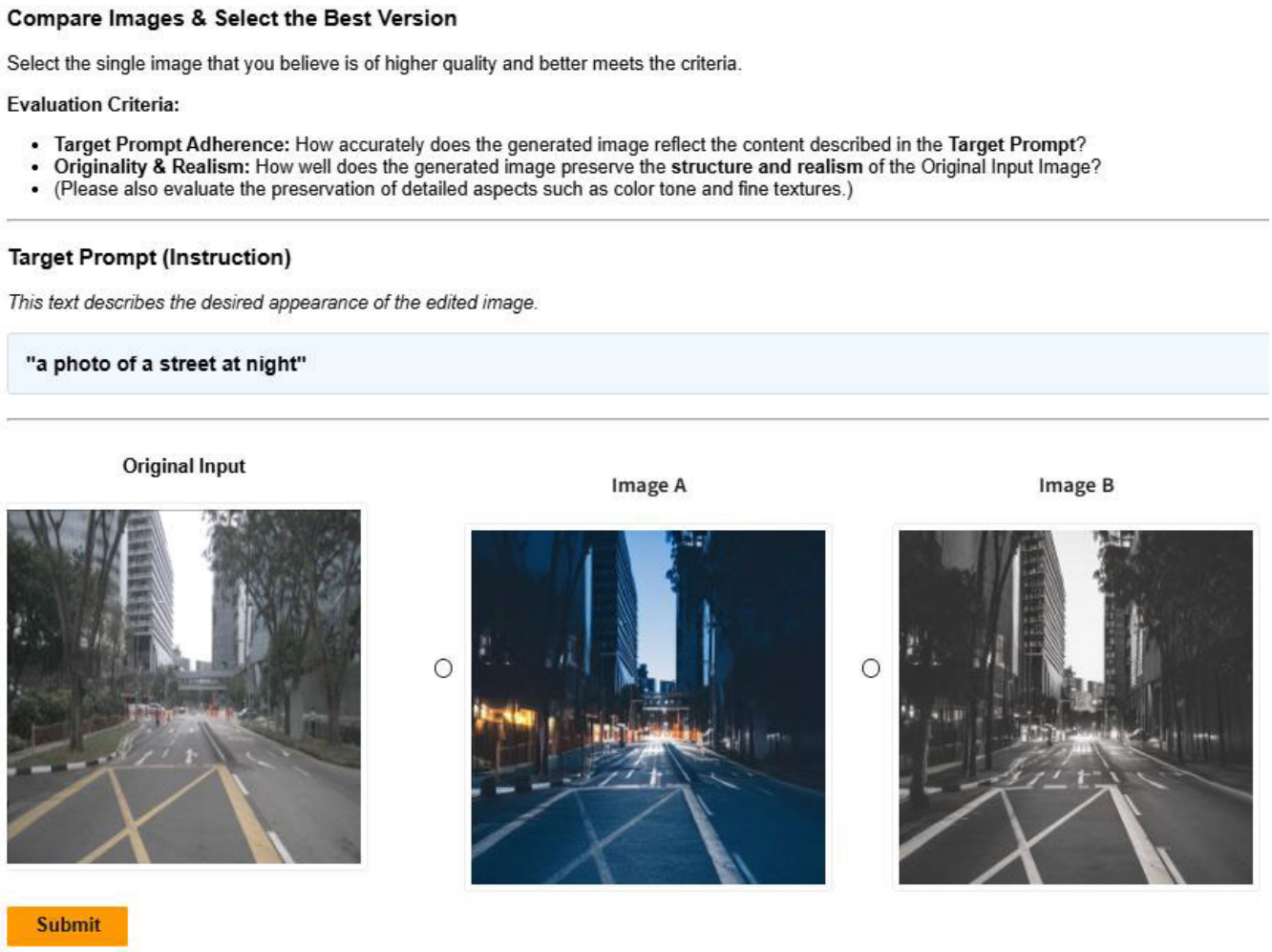

FIGURE 13. Example screenshot from the user study.

datasets (AFHQ, CelebA-HQ), we utilized the Segment Anything Model (SAM) [47] to generate masks, allowing us to evaluate the foreground (FG) and background (BG) separately. CLIP(FG) measures semantic editing success, while LPIPS(BG) and SSIM(BG) measure background preservation.

A. ANALYSIS OF GLOBAL METRICS (NuSCENES)

In the NuScenes dataset, we evaluated LPIPS and SSIM on the entire image due to the complexity of separating foreground and background in driving scenes. As observed in Table 4, methods that strictly prioritize structural invariance (e.g., P2P-zero) achieve high SSIM scores but fail to attain high CLIP scores (Table 1), indicating limited editing capability. In contrast, our method (+AdaCFG) demonstrates a balanced trade-off. While the global LPIPS and SSIM scores show slight deviations compared to the baseline (e.g., I-P2P), this is an expected outcome of effective semantic manipulation. Transformations such as changing weather or time of day inherently alter global lighting and pixel values, which naturally increases LPIPS and decreases SSIM. The significant boost in qualitative editing performance (as seen in visual results) justifies this marginal trade-off.

B. ANALYSIS OF MASKED METRICS (AFHQ & CELEBA-HQ)

For object-centric datasets, the masked evaluation provides a clearer insight into local editing precision.

- **Foreground Editing (CLIP-FG):** Our method consistently achieves higher CLIP(FG) scores across both datasets compared to baselines (I-P2P, PnP). This confirms that AdaCFG effectively injects the target semantics into the desired object.
- **Background Preservation (LPIPS-BG / SSIM-BG):** In the CelebA-HQ dataset, our method applied to I-P2P not only improves semantic alignment but also *improves* background preservation (LPIPS decreased by 2.4, SSIM increased by 2.0). This suggests that AdaCFG helps the model focus the diffusion process on the target region, reducing unintended artifacts in the background. In AFHQ, while there is a slight increase in

FIGURE 14. Qualitative results on the CelebA-HQ dataset. We compare our method (AdaCFG) applied to I-P2P and PnP against the original baselines and other methods across 6 different categories.

LPIPS(BG), it remains competitive, and the substantial gain in CLIP(FG) (+1.7 over I-P2P) demonstrates a superior Pareto efficiency between editing strength and structural fidelity.

Overall, the quantitative results confirm that AdaCFG enhances the semantic alignment of the edited regions without causing catastrophic degradation to the background, outperforming baselines that either fail to edit or destroy the image structure.

APPENDIX D
MORE ABLATION STUDIES
A. VISUALIZATION OF DIFFUSION DYNAMICS

As illustrated in Fig 10, our adaptive scheduler applies a high guidance scale during early timesteps (e.g., $t = 15$, 25), effectively injecting the target semantics—as evidenced by distinct color and noise pattern changes. In contrast, the baseline with fixed CFG fails to reflect the prompt, showing minimal variation at the same stages. At later steps ($t = 45$, 50), where our scheduler reduces the guidance, our model preserves the structure of vehicles and other critical scene elements. Meanwhile, the baseline gradually distorts and eventually erases them, leading to missing objects in the final result. This staged behavior demonstrates how our monotonically decreasing schedule first infuses strong semantic information and then prioritizes structural fidelity.

Fig 11 shows that cross-attention maps remain consistent between the baseline and our method, as the PnP framework enforces structural preservation via feature injection. This

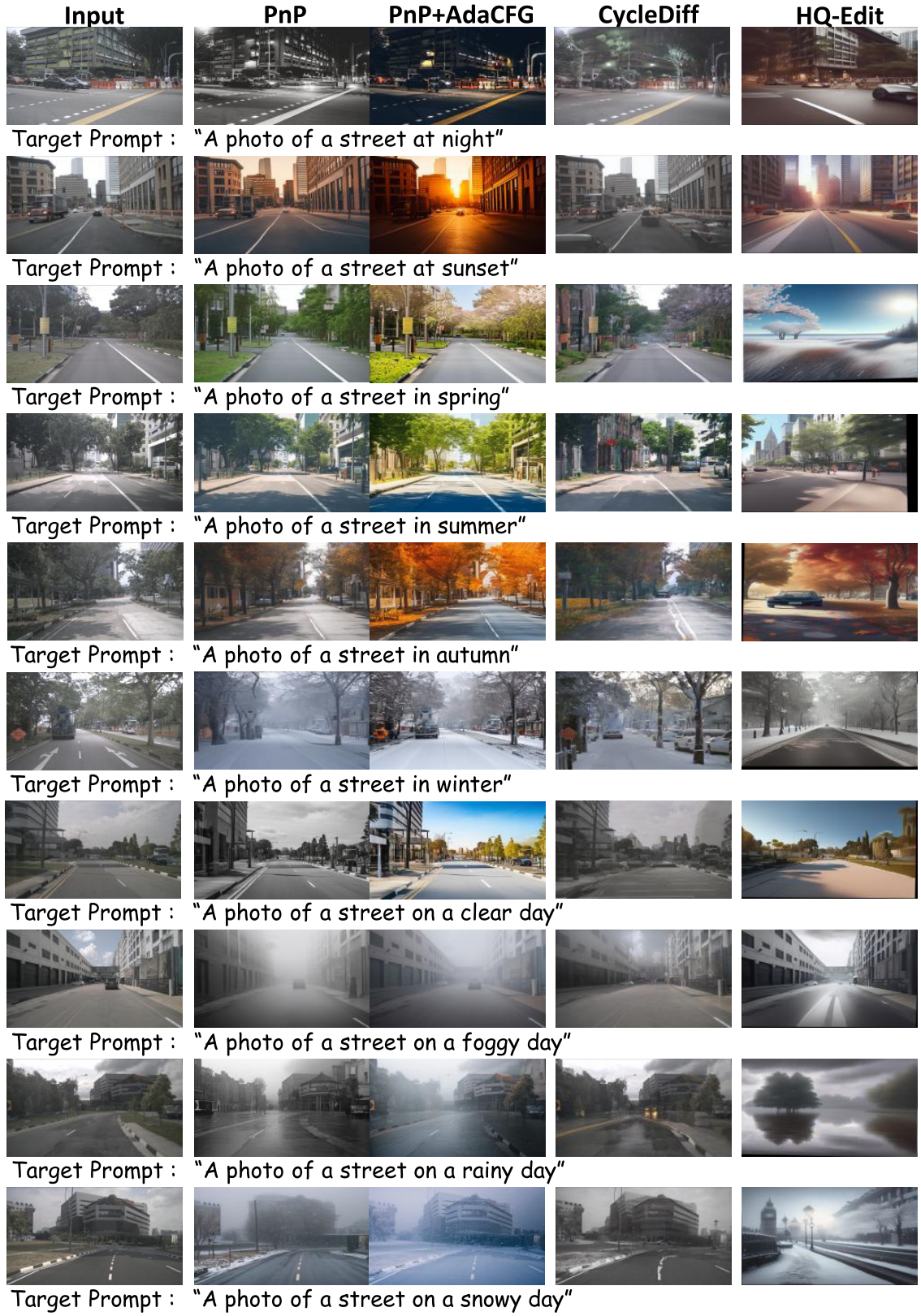

FIGURE 15. Qualitative results on the NuScenes dataset. We compare our method (AdaCFG) applied to PnP against the original baselines and other methods across 10 different categories.

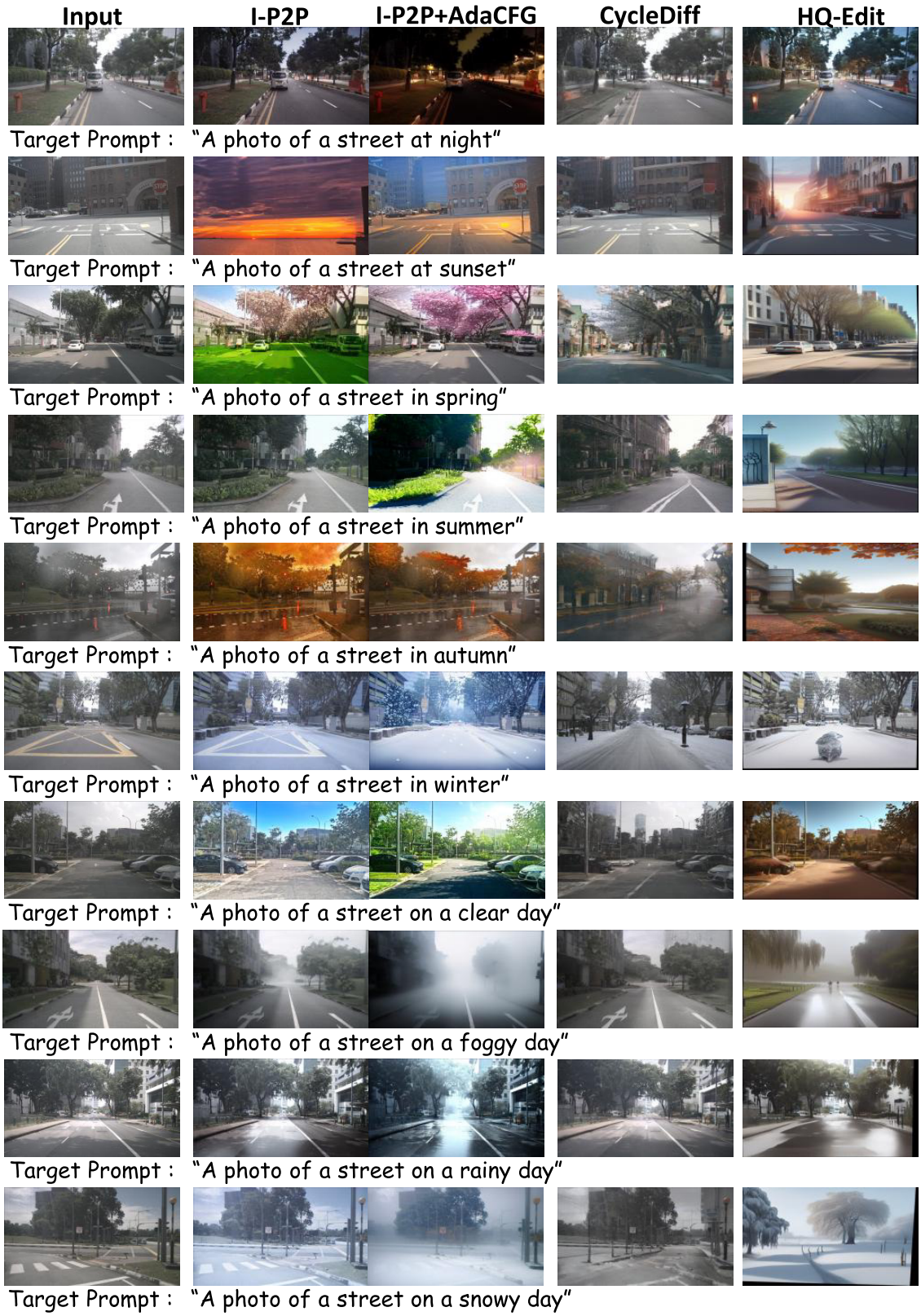

FIGURE 16. Qualitative results on the NuScenes dataset. We compare our method (AdaCFG) applied to I-P2P against the original baselines and other methods across 10 different categories.

FIGURE 17. Qualitative results on the AFHQ dataset. We compare our method (AdaCFG) applied to PnP against the original baselines and other methods across 6 different categories.

confirms that AdaCFG improves editing by modulating guidance intensity during early semantic-planning stages rather than altering the spatial layout.

To validate the effectiveness of this design, we compare our schedule with two alternative strategies on the NuScenes dataset using the PnP baseline:

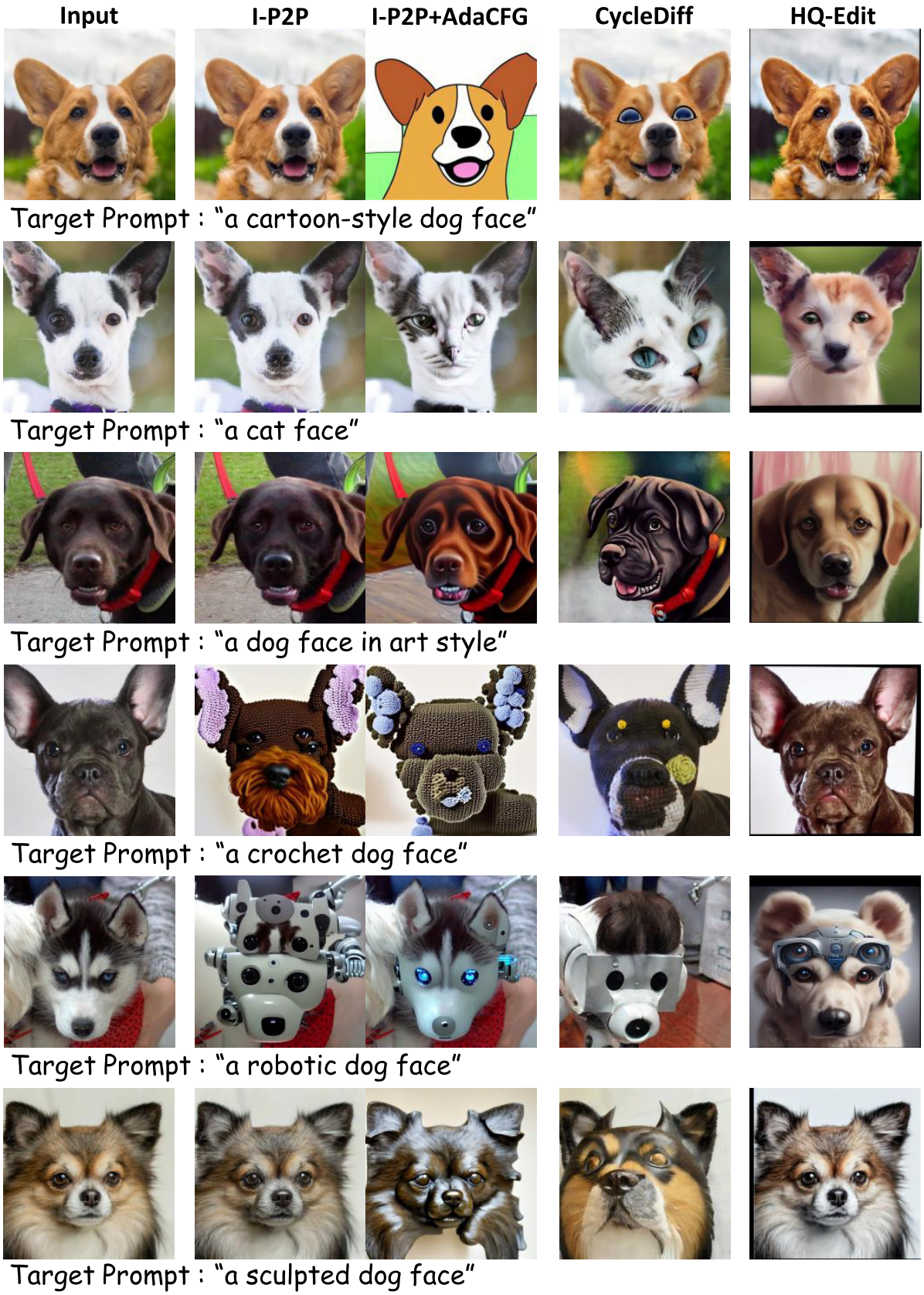

FIGURE 18. Qualitative results on the AFHQ dataset. We compare our method (AdaCFG) applied to I-P2P against the original baselines and other methods across 6 different categories.

- A **monotonically increasing schedule** $\omega'(t)$, implemented by reversing the timestep input of our default schedule:

$$\omega'(t) = \omega_0 \cdot \left(\cos\left(\frac{\pi}{2} \cdot \left(1 - \frac{t}{T}\right)\right)\right)^{2\beta} \tag{16}$$

- A **sine-shaped schedule** $\omega''(t)$ that increases guidance early, peaks mid-way, then decreases:

$$\omega''(t) = \omega_0 \cdot \left(\sin\left(\pi \cdot \frac{t}{T}\right)\right)^{2\beta} \tag{17}$$

As reported in Table 5, our default *monotonically decreasing* schedule $\omega(t)$ significantly outperforms the alternatives across all metrics—CLIP alignment, DINO similarity, and LLM-based Alignment score. These results support our hypothesis that early semantic guidance combined with late-stage structural preservation yields the best trade-off. The superiority of our scheduler confirms its role in producing high-quality, balanced edits.

B. PARETO OPTIMALITY AND ISO-ENERGY COMPARISON

To further validate the efficiency of our adaptive scheduler, we compared it against a comprehensive set of baselines, including fixed guidance scales and heuristic decay functions.

First, we conducted a grid search using **fixed guidance scales** ranging from $s = 2.5$ to 47.5. The gray line in Fig. 12 illustrates the inherent trade-off frontier: increasing the scale improves semantic alignment (CLIP) but degrades structural preservation (DINO).

Second, to verify whether our method's performance stems simply from total guidance energy, we compared it against **Linear** and **Exponential** decay schedules. Crucially, these baselines were calibrated to match the *mean Area Under the Curve (AUC)* of our adaptive scheduler (S_{total}), ensuring an "iso-energy" comparison.

As shown in Fig. 12, heuristic schedules like **Linear** (green triangle) and **Exponential** (purple square) fall below or near the fixed-scale frontier. In contrast, our method (**Ours**, red star) resides in the top-right region, surpassing the Pareto frontier formed by fixed scales. Specifically, compared to the AUC-matched Linear schedule, our method achieves significantly higher CLIP scores while maintaining comparable structural integrity. This demonstrates that our adaptive approach allocates the guidance budget more efficiently across timesteps than static or heuristic strategies.

It should be noted that the numerical values in Fig. 12 differ from those reported in Table 2 (specifically for the $m = 1$ setting). This discrepancy arises because the ablation study in Table 2 employs *random* prompts to evaluate robustness, whereas the Pareto analysis in Fig. 12 utilizes *fixed* prompts to ensure a controlled comparison. Consequently, the results under the fixed prompt setting align with the superior performance observed in the plot, unlike the randomized trials in the table.

APPENDIX E
MORE QUALITATIVE RESULTS

The qualitative results for each category, compared against the baselines, can be found in Fig 14, Fig 15, Fig 16, Fig 17, and Fig 18.

REFERENCES

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 10674–10685.

[2] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," 2021, *arXiv:2112.10741*.

[3] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, Z. Wang, J. Kwok, P. Luo, H. Lu, and Z. Li, "PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis," 2023, *arXiv:2310.00426*.

[4] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.

[5] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," 2020, *arXiv:2010.02502*.

[6] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 8780–8794.

[7] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel, "Plug-and-play diffusion features for text-driven image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1921–1930.

[8] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," 2022, *arXiv:2208.01626*.

[9] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "SDEdit: Guided image synthesis and editing with stochastic differential equations," 2021, *arXiv:2108.01073*.

[10] B. Forest Labs et al., "FLUX.1 kontext: Flow matching for in-context image generation and editing in latent space," 2025, *arXiv:2506.15742*.

[11] T. Brooks, A. Holynski, and A. A. Efros, "InstructPix2Pix: Learning to follow image editing instructions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 18392–18402.

[12] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, "Imagic: Text-based real image editing with diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 6007–6017.

[13] A. Hurst et al., "GPT-4o system card," 2024, *arXiv:2410.21276*.

[14] M. Cao, X. Wang, Z. Qi, Y. Shan, X. Qie, and Y. Zheng, "MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, Oct. 2023, pp. 22503–22513.

[15] G. Parmar, K. Kumar Singh, R. Zhang, Y. Li, J. Lu, and J.-Y. Zhu, "Zero-shot image-to-image translation," in *Special Interest Group Comput. Graph. Interact. Techn. Conf. Conf. Proc.*, Los Angeles, CA, USA, Jul. 2023, pp. 1–11.

[16] C. H. Wu and F. De La Torre, "A latent space of stochastic diffusion models for zero-shot image editing and guidance," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, Oct. 2023, pp. 7378–7387.

[17] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "IP-adapter: Text compatible image prompt adapter for text-to-image diffusion models," 2023, *arXiv:2308.06721*.

[18] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, "Null-text inversion for editing real images using guided diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 6038–6047.

[19] J. Ho and T. Salimans, "Classifier-free diffusion guidance," 2022, *arXiv:2207.12598*.

[20] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 36479–36494.

[21] G. Couairon, J. Verbeek, H. Schwenk, and M. Cord, "DiffEdit: Diffusion-based semantic image editing with mask guidance," 2022, *arXiv:2210.11427*.

[22] K. Zhang, L. Mo, W. Chen, H. Sun, and Y. Su, "MagicBrush: A manually annotated dataset for instruction-guided image editing," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 31428–31449.
[23] S. Sheynin, A. Polyak, U. Singer, Y. Kirstain, A. Zohar, O. Ashual, D. Parikh, and Y. Taigman, "Emu edit: Precise image editing via recognition and generation tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2024, pp. 8871–8879.
[24] M. Hui, S. Yang, B. Zhao, Y. Shi, H. Wang, P. Wang, Y. Zhou, and C. Xie, "HQ-edit: A high-quality dataset for instruction-based image editing," 2024, *arXiv:2404.09990*.
[25] S. Zhang, X. Yang, Y. Feng, C. Qin, C.-C. Chen, N. Yu, Z. Chen, H. Wang, S. Savarese, S. Ermon, C. Xiong, and R. Xu, "HIVE: Harnessing human feedback for instructional visual editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2024, pp. 9026–9036.
[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.
[27] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," 2023, *arXiv:2304.07193*.
[28] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. Erin Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuScenes: A multimodal dataset for autonomous driving," 2019, *arXiv:1903.11027*.
[29] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8185–8194.
[30] W. Xia, Y. Yang, J.-H. Xue, and B. Wu, "TediGAN: Text-guided diverse face image generation and manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2256–2265.
[31] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, Q. Zhang, K. Kreis, M. Aittala, T. Aila, S. Laine, B. Catanzaro, T. Karras, and M.-Y. Liu, "EDiff-I: Text-to-image diffusion models with an ensemble of expert denoisers," 2022, *arXiv:2211.01324*.
[32] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8162–8171.
[33] A. Castillo, J. Kohler, J. C. Pérez, J. P. Pérez, A. Pumarola, B. Ghanem, P. Arbeláez, and A. Thabet, "Adaptive guidance: Training-free acceleration of conditional diffusion models," in *Proc. AAAI Conf. Artif. Intell.*, 2025, pp. 1962–1970.
[34] M. Fu, G.-H. Wang, X. Chen, Q.-G. Chen, Z. Xu, W. Luo, and K. Zhang, "TeEFusion: Blending text embeddings to distill classifier-free guidance," 2025, *arXiv:2507.18192*.
[35] M. Kwon, S. S. Kim, J. Jeong, Y. T. Hsiao, and Y. Uh, "TCFG: Tangential damping classifier-free guidance," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jun. 2025, pp. 2620–2629.
[36] A. Jain, Y. Kobayashi, T. Shibuya, Y. Takida, N. Memon, J. Togelius, and Y. Mitsufuji, "Classifier-free guidance inside the attraction basin may cause memorization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jun. 2025, pp. 12871–12879.
[37] H. Chung, J. Kim, G. Yeong Park, H. Nam, and J. Chul Ye, "CFG++: Manifold-constrained classifier free guidance for diffusion models," 2024, *arXiv:2406.08070*.
[38] D. Shen, G. Song, Z. Xue, F.-Y. Wang, and Y. Liu, "Rethinking the spatial inconsistency in classifier-free diffusion guidance," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2024, pp. 9370–9379.
[39] M. Abdelsattar, A. AbdelMoety, and A. Emad-Eldeen, "ResNet-based image processing approach for precise detection of cracks in photovoltaic panels," *Sci. Rep.*, vol. 15, no. 1, Jul. 2025, Art. no. 24356, doi: 10.1038/s41598-025-09101-z.
[40] M. Abdelsattar, A. AbdelMoety, and A. Emad-Eldeen, "Comparative analysis of machine learning techniques for temperature and humidity prediction in photovoltaic environments," *Sci. Rep.*, vol. 15, no. 1, May 2025, Art. no. 15650, doi: 10.1038/s41598-025-98607-7.
[41] A. Rabee, Z. Anwar, A. AbdelMoety, A. Abdelsallam, and M. Ali, "Comparative analysis of automated foul detection in football using deep learning architectures," *Sci. Rep.*, vol. 15, no. 1, Apr. 2025, Art. no. 14236, doi: 10.1038/s41598-025-96945-0.
[42] M. Abdelsattar, A. AbdelMoety, and A. Emad-Eldeen, "Advanced machine learning techniques for predicting power generation and fault detection in solar photovoltaic systems," *Neural Comput. Appl.*, vol. 37, no. 15, pp. 8825–8844, May 2025, doi: 10.1007/s00521-025-11035-6.
[43] J. Chen, H. Zhang, M. Gong, and Z. Gao, "Collaborative compensative transformer network for salient object detection," *Pattern Recognit.*, vol. 154, Oct. 2024, Art. no. 110600.
[44] S. Lee, S. Cho, C. Park, S. Park, J. Kim, and S. Lee, "LSHNet: Leveraging structure-prior with hierarchical features updates for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5642516, doi: 10.1109/TGRS.2024.3469992.
[45] M. Pang, B. Wang, M. Ye, Y.-M. Cheung, Y. Zhou, W. Huang, and B. Wen, "Heterogeneous prototype learning from contaminated faces across domains via disentangling latent factors," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 4, pp. 7169–7183, Apr. 2025.
[46] M. Pang, W. Zhang, Y. Lu, Y.-M. Cheung, and N. Zhou, "A unified multi-domain face normalization framework for cross-domain prototype learning and heterogeneous face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 20, pp. 5282–5295, 2025.
[47] A. M. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Lo, P. Dollár, and R. Girshick, "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 3992–4003.

BONGGUK SON received the bachelor's degree, with a double major in biological science and computer science from Pusan National University, Busan, South Korea, in 2024, where he is currently pursuing the master's degree in artificial intelligence. His research interests include the field of computer vision, with a particular focus on diffusion models for image generation and editing.

SANGRYUL JEON received the B.S. and Ph.D. degrees in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2016 and 2022, respectively. From 2022 to 2023, he was a Postdoctoral Researcher with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley. Since 2023, he has been an Assistant Professor with the Department of Computer Science and Engineering, Pusan National University, Seoul, South Korea. His research interests include 2D/3D computer vision and machine learning, with a particular focus on image–text alignment and representation learning.