Adaptive Auxiliary Prompt Blending for Target-Faithful Diffusion Generation

Supplementary Material

7. Theoretical Extension: Log-Concave Setting.

To provide theoretical insight into why our adaptive coefficient outperforms fixed interpolation, we analyze the idealized case where the target distribution satisfies log-concavity and a transport-information inequality [20, 25]. Under these conditions, we formally establish that pointwise adaptive projection yields a provably tighter upper bound on the squared Wasserstein-2 distance compared to any fixed coefficient strategy.

Proposition 1 (Extension to the log-concave case)**.** *Consider two distributions: q_{γ_t}, which uses a fixed coefficient γ_t at timestep t, and q_{proj}, which adaptively projects γ_t^* to minimize local score error. Suppose p_T is k-strongly log-concave and satisfies the transport-information inequality [3, 20].*

$$W_2^2(q, p_T) \leq \frac{1}{k^2} J(q\|p_T), \tag{18}$$

where $k > 0$ is the strong log-concavity constant, and $J(q\|p) := \mathbb{E}_q\left[\|\nabla \log q - \nabla \log p\|_2^2\right]$ denotes the Fisher divergence [20]. Then

$$\begin{aligned} \forall \gamma_t \in \mathbb{R}, \quad W_2^2(q_{\text{proj}}, p_T) &\leq \tfrac{1}{k^2} J(q_{\text{proj}}\|p_T) \\ &\leq \tfrac{1}{k^2} J(q_{\gamma_t}\|p_T). \end{aligned} \tag{19}$$

Hence, in the idealized log-concave case, pointwise projection leads to a smaller Fisher divergence J and correspondingly tighter squared 2-Wasserstein bound compared to any fixed interpolation strategy.

Remark. While this result relies on the assumption of global log-concavity—which natural image distributions typically do not satisfy—it provides a theoretical motivation for minimizing the local score error pointwise.

Proof. The Fisher divergence between a distribution q (at time t) and the target p_T is defined as the expected squared error between their score functions:

$$J(q\|p_T) = \mathbb{E}_{x\sim q}\left[\|\nabla_x \log q(x) - \nabla_x \log p_T(x)\|_2^2\right]. \tag{20}$$

In our framework, we aim to approximate the target score $\nabla \log p_T$ using the blended model $\tilde{s}_\theta$. For clarity, let $s_T(x_t) := \nabla_{x_t} \log p_T(x_t)$ denote the true target score (approximated by the learned rare score in practice).

Consider the comparison at a specific timestep t. The score for our projection model is $s_{\text{proj}}(x_t) = \tilde{s}_\theta(x_t; w, \gamma_t^*(x_t))$. The score for the baseline model with a scalar coefficient is given by $s_{\gamma_t}(x_t) = \tilde{s}_\theta(x_t; w, \gamma_t)$, where $\gamma_t \in \mathbb{R}$ denotes the coefficient applied at timestep t and is fixed across all timesteps.

By definition, our adaptive coefficient $\gamma_t^*(x_t)$ is the solution that minimizes the squared error to the target score $s_T(x_t)$ at each spatial point x_t:

$$\gamma_t^*(x_t) = \arg\min_{\gamma_t \in \mathbb{R}} \|\tilde{s}_\theta(x_t; w, \gamma_t) - s_T(x_t)\|_2^2. \tag{21}$$

Since $\gamma_t^*(x_t)$ is the minimizer over $\mathbb{R}$ for every individual x_t, the error is guaranteed to be lower than or equal to the error yielded by any scalar γ_t:

$$\forall \gamma_t \in \mathbb{R}, \quad \|s_{\text{proj}}(x_t) - s_T(x_t)\|_2^2 \leq \|s_{\gamma_t}(x_t) - s_T(x_t)\|_2^2. \tag{22}$$

Since this pointwise inequality holds for all x_t in the support of q_t, taking the expectation over $x_t \sim q_t$ preserves the inequality:

$$\begin{aligned} &\mathbb{E}_{x_t\sim q_t}\left[\|s_{\text{proj}}(x_t) - s_T(x_t)\|_2^2\right] \\ &\qquad \leq \mathbb{E}_{x_t\sim q_t}\left[\|s_{\gamma_t}(x_t) - s_T(x_t)\|_2^2\right]. \end{aligned} \tag{23}$$

This implies that our method achieves a minimized instantaneous Fisher divergence compared to any scalar choice γ_t:

$$J(q_{\text{proj}}\|p_T) \leq J(q_{\gamma_t}\|p_T). \tag{24}$$

Via the transport-information inequality ($W_2^2 \leq \frac{1}{k^2} J$), this lower Fisher divergence implies a tighter upper bound on the squared Wasserstein-2 distance to the target distribution at each timestep. □

8. Derivation of Closed-Form Adaptive Coefficient

We provide a detailed derivation of the closed-form solution for our adaptive coefficient γ_t^* presented in Eq. (13) of the main paper.

Proposition 2 (Optimal Adaptive Coefficient)**.** *For a given (x_t, t) with $w > 0$, assume that $\|s_\theta(x_t, \tilde{c}_A) - s_\theta(x_t, \tilde{c}_T)\|_2^2 > 0$. Then the optimal adaptive coefficient γ_t^* that minimizes the score-space alignment loss $\mathcal{L}(\gamma_t) = \|\tilde{s}_\theta(x_t; w, \gamma_t) - s_\theta(x_t, \tilde{c}_T)\|_2^2$ has the closed-form solution:*

$$\gamma_t^*(x_t) = \frac{1-w}{w} \cdot \frac{\langle s_\theta(x_t, \tilde{c}_T) - s_\theta(x_t), s_\theta(x_t, \tilde{c}_A) - s_\theta(x_t, \tilde{c}_T)\rangle}{\|s_\theta(x_t, \tilde{c}_A) - s_\theta(x_t, \tilde{c}_T)\|_2^2}. \tag{25}$$

Proof. Recall our blended score function from Eq. (8):

$$\tilde{s}_\theta(x_t;w,\gamma_t) = s_\theta(x_t) + w\left((1-\gamma_t)s_\theta(x_t,\tilde{c}_T) + \gamma_t s_\theta(x_t,\tilde{c}_A) - s_\theta(x_t)\right), \tag{26}$$

where $s_\theta(x_t)$ is the unconditional score, $s_\theta(x_t,\tilde{c}_T)$ is the target-conditioned score, and $s_\theta(x_t,\tilde{c}_A)$ is the auxiliary anchor-conditioned score.

Our objective is to minimize the score-space alignment loss from Eq. (12):

$$\mathcal{L}(\gamma_t) = \|\tilde{s}_\theta(x_t;w,\gamma_t) - s_\theta(x_t,\tilde{c}_T)\|_2^2. \tag{27}$$

First, we simplify the blended score by distributing w:

$$\begin{aligned}\tilde{s}_\theta(x_t;w,\gamma_t) &= s_\theta(x_t) + w(1-\gamma_t)s_\theta(x_t,\tilde{c}_T) \\ &\quad + w\gamma_t s_\theta(x_t,\tilde{c}_A) - w s_\theta(x_t) \\ &= (1-w)s_\theta(x_t) + w(1-\gamma_t)s_\theta(x_t,\tilde{c}_T) + w\gamma_t s_\theta(x_t,\tilde{c}_A).\end{aligned} \tag{28}$$

Substituting into the loss:

$$\begin{aligned}\mathcal{L}(\gamma_t) &= \|(1-w)s_\theta(x_t) + w(1-\gamma_t)s_\theta(x_t,\tilde{c}_T) \\ &\quad + w\gamma_t s_\theta(x_t,\tilde{c}_A) - s_\theta(x_t,\tilde{c}_T)\|_2^2 \\ &= \|(1-w)s_\theta(x_t) - \gamma_t w s_\theta(x_t,\tilde{c}_T) + w\gamma_t s_\theta(x_t,\tilde{c}_A) \\ &\quad + w s_\theta(x_t,\tilde{c}_T) - s_\theta(x_t,\tilde{c}_T)\|_2^2 \\ &= \|(1-w)s_\theta(x_t) + (w-1)s_\theta(x_t,\tilde{c}_T) \\ &\quad + w\gamma_t\left(s_\theta(x_t,\tilde{c}_A) - s_\theta(x_t,\tilde{c}_T)\right)\|_2^2.\end{aligned} \tag{29}$$

For notational convenience, define:

$$\mathbf{r} := (1-w)(s_\theta(x_t) - s_\theta(x_t,\tilde{c}_T)), \tag{30}$$

$$\mathbf{d} := s_\theta(x_t,\tilde{c}_A) - s_\theta(x_t,\tilde{c}_T). \tag{31}$$

Then the loss becomes:

$$\mathcal{L}(\gamma_t) = \|\mathbf{r} + w\gamma_t\mathbf{d}\|_2^2. \tag{32}$$

$$\begin{aligned}\mathcal{L}(\gamma_t) &= \langle \mathbf{r} + w\gamma_t\mathbf{d}, \mathbf{r} + w\gamma_t\mathbf{d}\rangle \\ &= \|\mathbf{r}\|_2^2 + 2w\gamma_t\langle\mathbf{r},\mathbf{d}\rangle + w^2\gamma_t^2\|\mathbf{d}\|_2^2.\end{aligned} \tag{33}$$

This is a quadratic function in γ_t. Taking the derivative with respect to γ_t:

$$\frac{\partial\mathcal{L}}{\partial\gamma_t} = 2w\langle\mathbf{r},\mathbf{d}\rangle + 2w^2\gamma_t\|\mathbf{d}\|_2^2. \tag{34}$$

Setting the derivative to zero:

$$\begin{aligned}2w\langle\mathbf{r},\mathbf{d}\rangle + 2w^2\gamma_t^*\|\mathbf{d}\|_2^2 &= 0 \\ \gamma_t^* &= -\frac{\langle\mathbf{r},\mathbf{d}\rangle}{w\|\mathbf{d}\|_2^2}.\end{aligned} \tag{35}$$

Recall that $\mathbf{r} = (1-w)(s_\theta(x_t) - s_\theta(x_t,\tilde{c}_T))$ and $\mathbf{d} = s_\theta(x_t,\tilde{c}_A) - s_\theta(x_t,\tilde{c}_T)$:

$$\begin{aligned}&\gamma_t^*(x_t) \\ &= \frac{\langle(w-1)(s_\theta(x_t) - s_\theta(x_t,\tilde{c}_T)), s_\theta(x_t,\tilde{c}_A) - s_\theta(x_t,\tilde{c}_T)\rangle}{w\|s_\theta(x_t,\tilde{c}_A) - s_\theta(x_t,\tilde{c}_T)\|_2^2} \\ &= \frac{(w-1)\langle s_\theta(x_t) - s_\theta(x_t,\tilde{c}_T), s_\theta(x_t,\tilde{c}_A) - s_\theta(x_t,\tilde{c}_T)\rangle}{w\|s_\theta(x_t,\tilde{c}_A) - s_\theta(x_t,\tilde{c}_T)\|_2^2} \\ &= \frac{w-1}{w}\cdot\frac{\langle s_\theta(x_t) - s_\theta(x_t,\tilde{c}_T), s_\theta(x_t,\tilde{c}_A) - s_\theta(x_t,\tilde{c}_T)\rangle}{\|s_\theta(x_t,\tilde{c}_A) - s_\theta(x_t,\tilde{c}_T)\|_2^2} \\ &= \frac{1-w}{w}\cdot\frac{\langle s_\theta(x_t,\tilde{c}_T) - s_\theta(x_t), s_\theta(x_t,\tilde{c}_A) - s_\theta(x_t,\tilde{c}_T)\rangle}{\|s_\theta(x_t,\tilde{c}_A) - s_\theta(x_t,\tilde{c}_T)\|_2^2}.\end{aligned} \tag{36}$$

which matches Eq. (13) in the main paper.

To confirm this is a minimum, we check the second derivative:

$$\frac{\partial^2\mathcal{L}}{\partial\gamma_t^2} = 2w^2\|\mathbf{d}\|_2^2 = 2w^2\|s_\theta(x_t,\tilde{c}_A) - s_\theta(x_t,\tilde{c}_T)\|_2^2 > 0, \tag{37}$$

confirming that γ_t^* is indeed a global minimum of the quadratic loss function. □

9. Toy Example.

We empirically validate our method on a toy problem, generating samples from a 2D Gaussian *target* distribution using interpolation with an *auxiliary anchor* distribution. This simplified setup highlights the effectiveness of adaptive blending compared to fixed blending strategies.

The target data distribution $p_{\text{data}}(x \mid \tilde{c}_T)$ is defined as $\mathcal{N}((0,3),1.5I)$. We train a conditional diffusion model on a mixture of two distributions: an *auxiliary anchor* distribution $\mathcal{N}((0,-6),I)$, positioned distant from the target, and a *target-prior* distribution $\mathcal{N}((0,3),1.5I)$ that is identical to the target distribution. This setup represents an idealized scenario in which the target prior provides the same information as the target distribution itself—corresponding to accurate score estimation under sufficient model capacity [31, 33]. The auxiliary anchor samples constitute 80% of the training data, while the target-prior samples account for the remaining 20%. Additionally, following the standard classifier-free guidance (CFG) training scheme [9], we include an unconditional branch for 10% of the training data, where the conditioning input is randomly dropped.

Fig. 2 shows generated samples under two different γ_t strategies. (a) visualizes the training distributions, with auxiliary anchor samples (orange) clearly separated from the target-prior samples (purple) that overlap with the target region. (b) shows samples generated using fixed γ_t interpolation $p_{\text{lerp}}(x \mid \tilde{c}_A,\tilde{c}_T;\gamma_t = 0.8)$, which linearly combines

	Models	Single Object					Multi Objects			Avg
		Property	Shape	Texture	Action	Complex	Concat	Relation	Complex	
SDXL	SDXL [22]	60.0	56.9	71.3	47.5	58.1	39.4	35.0	47.5	52.0
	LMD [17]	23.8	35.6	27.5	23.8	35.6	33.1	34.4	33.1	30.9
	RPG [36]	33.8	54.4	66.3	31.9	37.5	21.9	15.6	29.4	36.4
	ELLA [11]	31.3	61.6	64.4	43.1	66.3	42.5	50.6	51.9	51.5
	R2F (SDXL) [21]	71.3	77.5	73.8	54.4	70.6	50.6	36.0	52.8	60.9
	Ours (SDXL)	**85.6**	**83.1**	**86.9**	**70.0**	**80.0**	**64.4**	**55.0**	**63.8**	**73.6**
IterComp	IterComp [39]	63.8	66.9	61.3	65.6	61.9	41.3	29.4	53.1	55.4
	R2F (IterComp) [21]	78.1	**77.5**	79.4	66.9	63.9	41.5	36.6	53.4	62.2
	Ours (IterComp)	**90.6**	**77.5**	**88.1**	**86.3**	**84.4**	**75.0**	**66.3**	**65.6**	**79.2**
SD3.0	SD3.0 [6]	49.4	76.3	53.1	71.9	65.0	55.0	51.2	70.0	61.5
	R2F (SD3) [21]	89.4	79.4	81.9	80.0	72.5	70.0	58.8	73.8	75.7
	Ours (SD3)	**96.9**	**89.4**	**87.5**	**85.6**	**80.0**	**82.5**	**65.6**	**85.0**	**84.1**

Table 4. Text-to-image alignment on RareBench with SDXL, SD3.0, and IterComp, comparing our method integrated into each model. Results demonstrate consistent robustness across diverse pre-trained diffusion backbones.

the anchor and target-prior score functions. (c) demonstrates our adaptive interpolation $p(x \mid \tilde{c}_A, \tilde{c}_T; \gamma_t^*)$, where γ_t^* is dynamically optimized at each denoising step. (d) reports 2-Wasserstein distance between the generated distributions and the target $\mathcal{N}((0,3), 1.5I)$ as a function of γ_t. The blue curve shows that fixed interpolation achieves its minimum distance around $\gamma_t \approx 0.8$. In contrast, the adaptive method (red dashed line) consistently achieves a lower distance, confirming the benefit of step-wise optimization.

The key insight from this toy example is that the optimal interpolation parameter γ_t^* varies across the denoising process. Fixed interpolation requires manual tuning and remains suboptimal, while our adaptive approach automatically balances target fidelity and anchor stability at each step, leading to a more accurate approximation of the target distribution.

10. Full Algorithm

Let $s_\theta(x_t, t, \cdot)$ denote a pretrained score estimator (either unconditional or conditional) evaluated at state x_t and diffusion/flow timestep t. We write $s_\theta(x_t, t) \equiv s_\theta(x_t, t, \varnothing)$ for the unconditional score, and $s_\theta(x_t, t, \tilde{c})$ for the score conditioned by prompt $\tilde{c}$. Classifier-Free Guidance (CFG) combines unconditional and conditional scores linearly with a guidance scale w. Following the main text, the blended conditional score is defined as

$$s_\theta(x_t, t, \tilde{c}) \;=\; (1-\gamma_t)\, s_\theta(x_t, t, \tilde{c}_T) \;+\; \gamma_t\, s_\theta(x_t, t, \tilde{c}_A), \tag{38}$$

and the final guided score is given by

$$\tilde{s}_\theta(x_t; t, w, \gamma_t) \;=\; s_\theta(x_t, t) + w\,(s_\theta(x_t, t, \tilde{c}) - s_\theta(x_t, t))\,. \tag{39}$$

Using Tweedie's identity under the Gaussian corruption model, the image-domain denoising loss is equivalent to a score-space ℓ_2 loss, which yields a closed-form per-timestep optimum γ_t (see main text, Eq. (9) – (13)).

Algorithm 1 Adaptive Auxiliary Prompt Blending (AAPB) Generation

Require: Target prompt $\tilde{c}_T$, Anchor prompt $\tilde{c}_A$, pre-trained score model s_θ, guidance scale w, timesteps T
Ensure: Generated image x_0

$x_T \sim \mathcal{N}(0, \mathbf{I})$
for $t = T$ **down to** 1 **do**
  $s_u \leftarrow s_\theta(x_t, t)$ ▷ Unconditional score
  $s_T \leftarrow s_\theta(x_t, t, \tilde{c}_T)$ ▷ Target-conditioned score
  $s_A \leftarrow s_\theta(x_t, t, \tilde{c}_A)$ ▷ Anchor-conditioned score
  $\gamma_t^* \leftarrow \frac{1-w}{w} \cdot \frac{\langle s_T - s_u, s_A - s_T \rangle}{\|s_A - s_T\|_2^2}$ ▷ Eq. (13)
  $\tilde{s}_\theta \leftarrow s_u + w\,((1-\gamma_t^*)s_T + \gamma_t^* s_A - s_u)$
  $x_{t-1} \leftarrow \text{SamplerStep}(x_t, \tilde{s}_\theta, t)$
end for
return x_0

11. Robustness across Various Diffusion Models.

Tab. 4 shows the robustness across three different pre-trained diffusion models on RareBench, including SDXL, IterComp [39], and SD3.0. Our method yields substantial performance improvements compared to R2F across all models, achieving consistent average gains of $+12.7$, $+17.0$, and $+8.4$ on SDXL, IterComp, and SD3.0, respectively. These results demonstrate that our method generalizes effectively across different architectures and training settings, confirming its robustness and broad applicability.

As illustrated in Fig. 7, our AAPB framework consistently preserves the semantic meaning of the input prompt across diverse pre-trained diffusion backbones, including SDXL, IterComp, and SD3.0. While baseline models of-

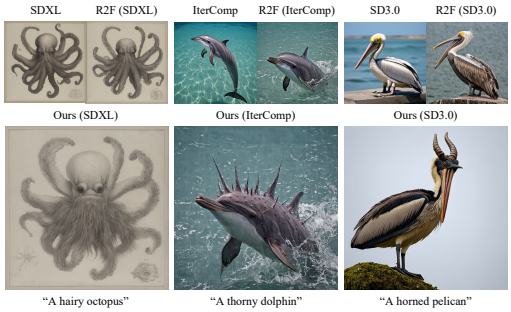

Figure 7. Qualitative comparison on different pre-trained diffusion baselines, SDXL, SD3.0, and IterComp. Comparing our method integrated with these pre-trained models.

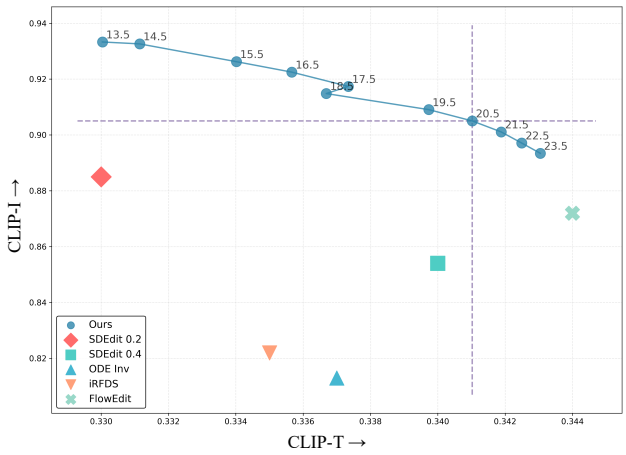

Figure 8. Quantitative comparison on varying classifier-free guidance scale w on the FlowEdit dataset.

ten exhibit incomplete attribute binding or geometric distortion in rare or complex prompts, our method accurately conveys the intended semantics while maintaining visual coherence and structural realism. This demonstrates that adaptive prompt blending effectively generalizes across architectural variations, allowing the model to maintain semantic faithfulness even under different backbone configurations.

12. Varying Classifier-Free Guidance Scale

For rare concept generation, Fig. 15 shows qualitative results for varying classifier-free guidance scales w. When w is too small (e.g., $w = 1.0$ or 2.0), the generated images lack visual fidelity and semantic alignment with the input prompt. Conversely, overly large scales (e.g., $w = 9.0$ or 10.0) lead to noticeable artifacts and structural distortions. We adopt $w = 7.0$ as our default setting, which is also the standard guidance scale used in SD3. Our adaptive coefficient operates on top of this baseline to further refine the balance between fidelity and semantic consistency.

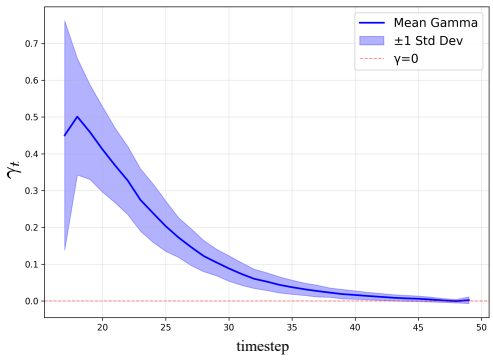

Figure 9. Evolution of the adaptive coefficient γ_t^* across diffusion timesteps in image editing (FlowEdit). The coefficient exhibits clear saturation in later steps as structural guidance from the source image becomes dominant.

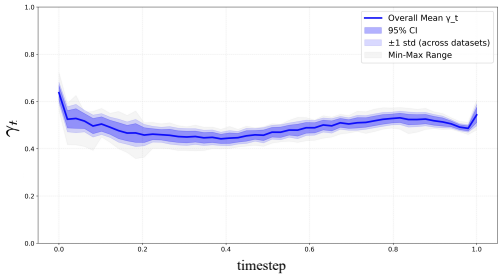

Figure 10. Evolution of the adaptive coefficient γ_t^* across diffusion timesteps in rare concept generation (RareBench). Unlike image editing, γ_t^* maintains stable values throughout the denoising process, reflecting continuous need for anchor stabilization.

For image editing, Fig. 8 presents a quantitative comparison across different classifier-free guidance scales w in terms of text–image alignment (CLIP-T) and image–image fidelity (CLIP-I). When w is near 13.5, the model exhibits strong structural preservation of the original image but weak semantic alignment with the target prompt. Conversely, when w becomes near 23.5 semantic consistency improves marginally, yet structural distortions emerge and the overall gain over SDEdit 0.2 remains limited. We therefore set $w = 20.5$ as a balanced configuration that jointly maximizes both CLIP-T and CLIP-I scores, achieving an optimal trade-off between fidelity and semantic accuracy.

13. Adaptive Coefficient Progress.

Fig. 9 presents the mean adaptive coefficient γ_t^* over all samples in image editing, showing a shift from stronger anchor (source prompt) reliance in early noisy steps to more target-conditioned guidance in later stages. Following the FlowEdit configuration, we report γ_t^* only after step 17, as the early steps do not reflect meaningful guidance behavior. Geometrically, γ_t^* can be viewed as the projection

of the target score residual $(s_\theta(x_t, \tilde{c}_T) - s_\theta(x_t))$ onto the anchor–target direction. As the diffusion progresses, their dot product approaches zero, implying near-orthogonality between target and anchor gradients. This indicates that the target score moves toward a self-consistent manifold region requiring less auxiliary correction. Hence, γ_t^* naturally decays as the model adaptively balances anchor and target influences, producing stable and controllable generations.

In contrast, Fig. 10 shows that γ_t^* in rare concept generation maintains relatively stable values throughout the denoising process (mean ≈ 0.5), without the saturation observed in image editing. This difference stems from the fundamental distinction in task structure: image editing benefits from strong structural guidance provided by the source image, which progressively dominates the generation process and reduces the need for anchor support in later steps. However, rare concept generation operates without such structural constraints—the model must synthesize images purely from textual descriptions. Consequently, the anchor continues to provide essential semantic stabilization throughout the entire denoising trajectory, preventing drift toward high-density regions while maintaining consistent guidance toward the target concept.

14. Additional Analysis for Anchor Sensitivity Analysis with Anchor Quality Metric

Building upon Sec. 4.2, we now analyze what constitutes an effective anchor. Although Eq. (13) guarantees optimal blending for any given anchor, different anchor construction strategies (Tab. 3) exhibit surprisingly competitive performance. This observation suggests that the practical effectiveness of an anchor is not determined solely by its directional alignment with the target–unconditional axis, but also by its overall deviation within the score space.

Substituting γ_t^* into Eq. (12) yields the minimum attainable loss at timestep t:

$$\mathcal{L}_t^*(s_A) = \|\mathbf{r}\|^2 - \frac{\langle \mathbf{r}, \mathbf{d} \rangle^2}{\|\mathbf{d}\|^2}, \tag{40}$$

where $\mathbf{r} = (1-w)(s_u - s_T)$ is the fixed residual and $\mathbf{d} = s_A - s_T$ is the anchor displacement.

To understand how anchor choice affects this loss, we decompose the displacement into components parallel and orthogonal to $\mathbf{r}$:

$$\mathbf{d} = \mathbf{d}^{\parallel} + \mathbf{d}^{\perp}, \qquad \mathbf{d}^{\parallel} = \frac{\langle s_A - s_T,\, s_u - s_T \rangle}{\|s_u - s_T\|^2}(s_u - s_T). \tag{41}$$

Since $\langle \mathbf{r}, \mathbf{d}^{\perp} \rangle = 0$, the numerator of Eq. (40) depends only on $\mathbf{d}^{\parallel}$, while the denominator grows with both components:

$$\mathcal{L}_t^*(s_A) = \|\mathbf{r}\|^2 - \frac{\langle \mathbf{r}, \mathbf{d}^{\parallel} \rangle^2}{\|\mathbf{d}^{\parallel}\|^2 + \|\mathbf{d}^{\perp}\|^2}. \tag{42}$$

Strategy	$\|\mathbf{d}^{\parallel}\|$	$\|\mathbf{d}^{\perp}\|$	**Total** $\downarrow$	**T2I** $\uparrow$
Arbitrary	42.2	47.9	90.1	71.0
Objects	32.5	26.6	59.1	83.1
Human Generated	30.2	28.4	58.6	82.6
GPT-4o	27.3	28.0	55.3	**87.9**
LLaMA3	26.8	27.0	**53.8**	81.0

Table 5. Anchor quality comparison across different anchor generation strategies. Lower displacement indicates anchors closer to the target with fewer artifacts. T2I scores are evaluated using GPT-4o. All correlations are negative.

This reveals a key trade-off: while $\mathbf{d}^{\parallel}$ provides the guidance signal, excessive $\|\mathbf{d}^{\perp}\|$ dilutes its effectiveness by inflating the denominator.

Motivated by this insight, we propose a simple anchor quality metric:

$$\text{Disp}(s_A) = \|\mathbf{d}^{\parallel}\| + \|\mathbf{d}^{\perp}\|. \tag{43}$$

As shown in Tab. 5, this displacement metric exhibits a consistent trend with T2I alignment performance. Strategies with moderate displacement (e.g., GPT-4o: 55.3, Human Generated: 58.6, Objects: 59.1) achieve strong T2I scores (82.6–87.9), while anchors with excessive displacement (e.g., Arbitrary: 90.1) show significantly degraded performance (71.0), likely due to orthogonal drift that introduces spurious features from distant regions of the score space.

Crucially, however, geometry is not the sole determinant. While LLaMA3 achieves the lowest displacement (53.8), it trails GPT-4o (55.3) in T2I accuracy (81.0 vs. 87.9). This suggests that the alignment between anchor and target semantics—beyond geometric displacement alone—also contributes to performance. GPT-4o's superior language understanding may produce anchors that better preserve task-relevant categorical structure (e.g., substituting "hairy frog" with "hairy animal" rather than generic "object"), enabling more effective guidance even at comparable displacement levels. However, this does not require the anchor itself to be semantically detailed or rare; rather, the anchor should maintain appropriate semantic alignment with the target's core attributes.

In summary, effective anchors should satisfy two complementary criteria: (1) maintain a well-calibrated geometric displacement from the target score—close enough to avoid drift into spurious modes, yet sufficiently separated to provide stabilization in low-density regions; and (2) preserve basic semantic alignment with task-relevant attributes, even when substituting rare concepts with frequent ones. By ensuring anchors meet these criteria, AAPB's adaptive coefficient γ_t^* can reliably modulate their influence across timesteps, enabling even simple, frequent anchors to perform competitively without requiring rich semantic detail.

Models	RareBench Multi		
	Concat	Relation	Complex
SD3.0	55.0	51.2	70.0
R2F+	74.4	63.7	64.8
Ours (R2F+)	83.8	83.6	80.3

Table 6. Quantitative comparison on RareBench-Multi between R2F+ and our method built upon the R2F+ baseline.

Models	Rare Concept			Image Editing	
	SD3	R2F	Ours	FlowEdit	Ours
Peak Memory (GB)	31.52	31.76	32.22	19.97	19.97
GPU Time (sec)	27.14	26.12	37.94	7.41	7.54

Table 7. GPU time and memory required to generate an image.

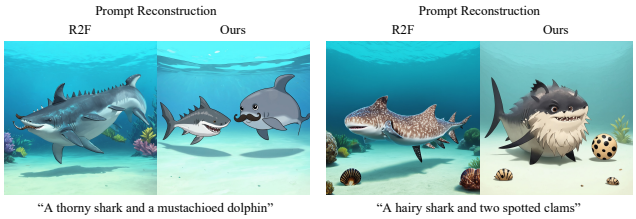

Figure 11. Comparison of prompt reconstruction between R2F and our method. R2F often collapses multiple rare concepts into a single entity, leading to entangled generations. In contrast, our method explicitly preserves each rare concept by directly pairing it with its frequent counterpart, resulting in disentangled and faithful generations.

15. Results on R2F+ Baseline

The R2F+ [21] framework extends R2F to region-controlled generation, where LLMs extract per-object rare–frequent concept pairs with bounding boxes and visual detail levels. Our Adaptive Auxiliary Prompt Blending (AAPB) integrates seamlessly into R2F+'s diffusion pipeline without retraining, operating directly on per-timestep denoising trajectories. During alternating region guidance, the adaptive coefficient γ_t dynamically modulates frequent-anchor influence to preserve both semantic fidelity and local structure across object-wise generations.

Tab. 6 presents a quantitative comparison between R2F+ and our method built upon the same R2F+ baseline. Across all RareBench-Multi categories, our framework consistently outperforms the original R2F+ in text–image alignment accuracy, achieving notable improvements of $+9.4$ on the *Concat*, $+19.9$ on the *Relation*, and $+16.5$ on the *Complex*. These results demonstrate that our adaptive auxiliary prompt blending mechanism effectively complements the alternating rare–frequent concept scheduling in R2F+.

16. GPU Time and Memory Analysis.

Tab. 7 reports computational costs measured on an NVIDIA RTX 6000 Blackwell GPU. For rare concept generation, the peak memory usage remains comparable (33.22GB vs. 31.52GB), indicating that AAPB introduces minimal memory overhead. The longer computation time of AAPB is due to the evaluation of three score functions per timestep (unconditional, target, and anchor) instead of two (unconditional, conditional). For image editing, the overhead is minimal (7.54s vs. 7.41s) with identical peak memory, as FlowEdit already performs three-branch scores (unconditional, source, and target). In this case, AAPB simply reuses the source branch as an adaptive anchor, preserving efficiency while enhancing edit faithfulness. In return, AAPB delivers a clear performance advantage, significantly surpassing state-of-the-art baselines in both semantic alignment and structural preservation.

17. User Study

We conducted a user study with eleven participants on the RareBench dataset. For each prompt, images generated by Ours, R2F, and SD3 were shown simultaneously to enable direct side-by-side comparison across methods [32], facilitating a more accurate assessment of semantic alignment and visual consistency. To prevent bias, model names were anonymized and the display order was randomized for every prompt. Participants scored each result using the criteria in Tab. 8, with the $\{1, 2, 3, 4, 5\}$ scale linearly mapped to $\{0, 25, 50, 75, 100\}$ for analysis (see Fig. 14).

As summarized in Tab. 9, participants frequently noted that our method delivered clearer attribute grounding and reduced compositional inconsistencies, particularly in multi-object scenes. The most pronounced qualitative gains appeared in the Concat and Relation categories, where users consistently favored our results. We attribute this to more explicit entity separation and improved preservation of individual attributes, which enhances visual clarity in complex settings. This finding is consistent with the disentangled prompt reconstruction results in Fig. 11, demonstrating that our binary alignment strategy better supports multi-entity generation.

18. LLM Instruction for Rare Concept Generation

Tab. 11 and Tab. 12 detail the full LLM prompt and the in-context examples for AAPB, respectively. Unlike R2F, which decomposes each rare concept into a *step-wise sequence* with detailed annotations (including an explicit *Visual Detail Level* for heuristic scheduling), our instruction employs a direct binary pairing between rare and frequent concepts. R2F generates an expanded, hierarchical reasoning trace before concatenating them into the final sequence.

<Task>
Evaluate how well each image matches the given text prompt. Focus on whether the objects in the image and their attributes (e.g., color, shape, texture), spatial arrangement, and actions align with the text.

<Rating Scale>
5: The image perfectly matches the text prompt with no noticeable mistakes.
4: The image matches most of the prompt, with only minor inconsistencies.
3: The image reflects the prompt partially, but some important details are missing or incorrect.
2: The image shows only a few elements from the prompt, and many key parts are missing or wrong.
1: The image does not match the prompt at all or fails to convey the main idea.

Table 8. User study task and rating criteria for evaluating text–image alignment.

Models	Single Object					Multi Objects			Avg
	Property	Shape	Texture	Action	Complex	Concat	Relation	Complex	
SD3 [6]	46.3	58.3	46.0	43.5	53.1	40.8	46.7	60.5	49.4
R2F (SD3) [21]	71.9	66.2	71.4	63.7	69.3	42.6	44.6	67.3	62.1
Ours (SD3)	**76.7**	**68.5**	**71.9**	**65.0**	**70.1**	**59.8**	**49.7**	**68.7**	**66.3**

Table 9. Qualitative comparison based on our user study. For each prompt, participants were shown multiple images generated by different models in random order and without model names (Fig. 14). Participants selected the preferred result according to the evaluation criteria define in Tab. 8.

Figure 12. Qualitative results for face-accessory image editing. We extract the source prompt using GPT and augment the target prompt by adding "glasses" as an accessory. Our method preserves facial details more faithfully than FlowEdit.

In contrast, our approach streamlines this process by focusing solely on semantic substitution: it replaces the rare noun phrase with its frequent counterpart in a single pass and directly constructs the rare-frequent pair.

Furthermore, R2F decomposes a prompt into multiple sub-prompts (e.g., "Two spotted sea creatures" and "A hairy aquatic creature"), which are sequentially substituted to form the final sentence. In contrast, our formulation enforces a direct rare–frequent pairing (e.g., "A hairy animal and two spotted objects"), enabling the entire prompt to be reconstructed in a single pass. In Fig. 11, we compare the reconstruction strategy of R2F with our binary alignment approach. Our binary alignment approach resulting in improved disentanglement and fidelity in multi-entity generations.

19. Image Editing for Face Accessories

Fig. 12 presents image editing results for adding facial accessories on the SFHQ dataset [1]. We extract the source prompt using GPT and augment the target prompt by appending "glasses" as an accessory. Compared to FlowEdit, our method preserves fine facial details—particularly in the hair and ear regions—while generating the glasses faithfully.

20. Quantitative Image Quality Analysis

For quantitative image quality evaluation, we employ three widely used metrics: LAION-Aesthetic [29], PickScore [14], and ImageReward [35]. Tab. 10 presents a comparison among AAPB, R2F, and SD3.0. Our method achieves the highest scores on ImageReward and PickScore, indicating superior human preference alignment and semantic consistency, while SD3.0 attains a slightly higher aes-

Models	LAION-aesthetic	ImageReward	PickScore
SD3.0	**6.317±0.190**	0.915±0.151	21.771±0.462
R2F	6.215±0.220	1.048±0.266	22.155±0.239
Ours	6.191±0.166	**1.123±0.328**	**22.177±0.222**

Table 10. Aesthetic and preference evaluation on *RareBench*. LAION-aesthetic measures visual appeal, ImageReward models human preference, and PickScore evaluates text–image alignment. Our method achieves the best preference and alignment scores.

Figure 13. Failure results on RareBench where both ours and R2F exhibit attribute-object mismatches. This behaviors aligns with previously reported limitations in CLIP's ability to bind compositional concpets faithfully [16].

thetic score. This minor trade-off in visual appeal is acceptable, as AAPB prioritizes rare concept fidelity and prompt-faithful generation over conventional aesthetics—aligning well with the objective of our task.

21. Limitations

Although our approach substantially improves rare-concept alignment, it still inherits the inherent limitations of CLIP-based text encoders, which struggle to preserve compositional bindings when multiple attributes and objects coexist in a prompt (Fig. 13). As also reported by Lewis et al. [16], CLIP often fails to maintain correct attribute–object associations in multi-component scenarios, leading to occasional entanglement or attribute leakage in our generations. Overcoming this limitation would likely require representation-level advances (e.g., more compositionally robust vision–language encoders) or additional architectural modules explicitly designed for attribute binding, which we leave for future exploration.

22. More Visualization Results

Fig. 16 presents uncurated generations of our model on RareBench across multiple random seeds. Most results exhibit strong alignment with the input prompts while preserving naturalness and visual quality. Furthermore, Fig. 17 provides additional qualitative results on the FlowEdit benchmark, where leveraging the blended score yields higher anchor-image fidelity compared to the baseline FlowEdit method.

<**System Prompt**>
You are a helper language model for a text-to-image generation program that aims to create images based on input text. The program often struggles to accurately generate images when the input text contains rare concepts that are not commonly found in reality. To address this, when a rare concept is identified in the input text, you should replace it with relevant yet more frequent concepts.

<**User Prompt**>
Extract rare concepts from the input text and replace them with relevant yet more frequent ones. Perform the following process step by step:

a. Identify and extract any rare concepts from the provided input text. If the text contains one or more rare concepts, extract them all. If there are no rare concepts present, do not extract any concepts. The extracted rare concepts should not overlap.

b. Given the rare concepts extracted at a., replace each extracted rare concept with a more frequent concept. Specifically, replace the main noun subject with a more frequent noun subject that is likely to appear in a similar context. Ensure that the replaced frequent noun subject retains the properties of the original main noun subject as much as possible.

c. Generate a text sequence that starts from the text with replaced frequent concepts and ends with the text with the original rare concepts.

The output should follow the format of the examples below:

<**In-context Examples**>
Input: {USER PROMPT}
Output:

Table 11. Full LLM instruction for AAPB to generate target-to-anchor concept mappings for rare concept generation.

Input: A banana shaped apple
Output:
Num Rare Concepts: 1
a. Rare concept: A banana shaped apple
b. A banana shaped apple does not exist in reality, while the possibility of a banana shaped red object existing is much higher. Main noun object: apple, Replaced frequent object: red object
c. A banana shaped red object BREAK A banana shaped apple
Final Prompt Sequence: A banana shaped red object BREAK A banana shaped apple

Input: A running dog
Output:
Num Rare Concepts: 0
No rare concept was found in the given input text.
Final Prompt Sequence: A running dog

Input: A horned lion and a hairy frog
Output:
Num Rare Concepts: 2
a. Rare concept: A horned lion
b. A horned lion does not exist in reality, while a horned animal does. Main noun subject: lion, Replaced frequent subject: animal
c. A horned animal BREAK A horned lion
AND
a. Rare concept: A hairy frog
b. A hairy frog does not exist in reality, while a hairy animal does. Main noun subject: frog, Replaced frequent subject: animal
c. A hairy animal BREAK A hairy frog
Final Prompt Sequence: A horned animal AND a horned animal BREAK A horned lion AND a hairy frog

Input: A horned lion is sitting on a tiger striped rock
Output:
Num Rare Concepts: 2
a. Rare concept: A horned lion
b. A horned lion does not exist in reality, while a horned animal does. Main noun subject: lion, Replaced frequent subject: animal
c. A horned animal BREAK A horned lion
AND
a. Rare concept: a tiger striped rock
b. A tiger striped rock does not exist in reality, while a tiger striped object does. Main noun subject: rock, Replaced frequent subject: object
c. a tiger striped object BREAK a tiger striped rock
Final Prompt Sequence: A horned animal AND a tiger striped object BREAK A horned lion AND a tiger striped rock

Table 12. In-context examples of the LLM prompt for AAPB.

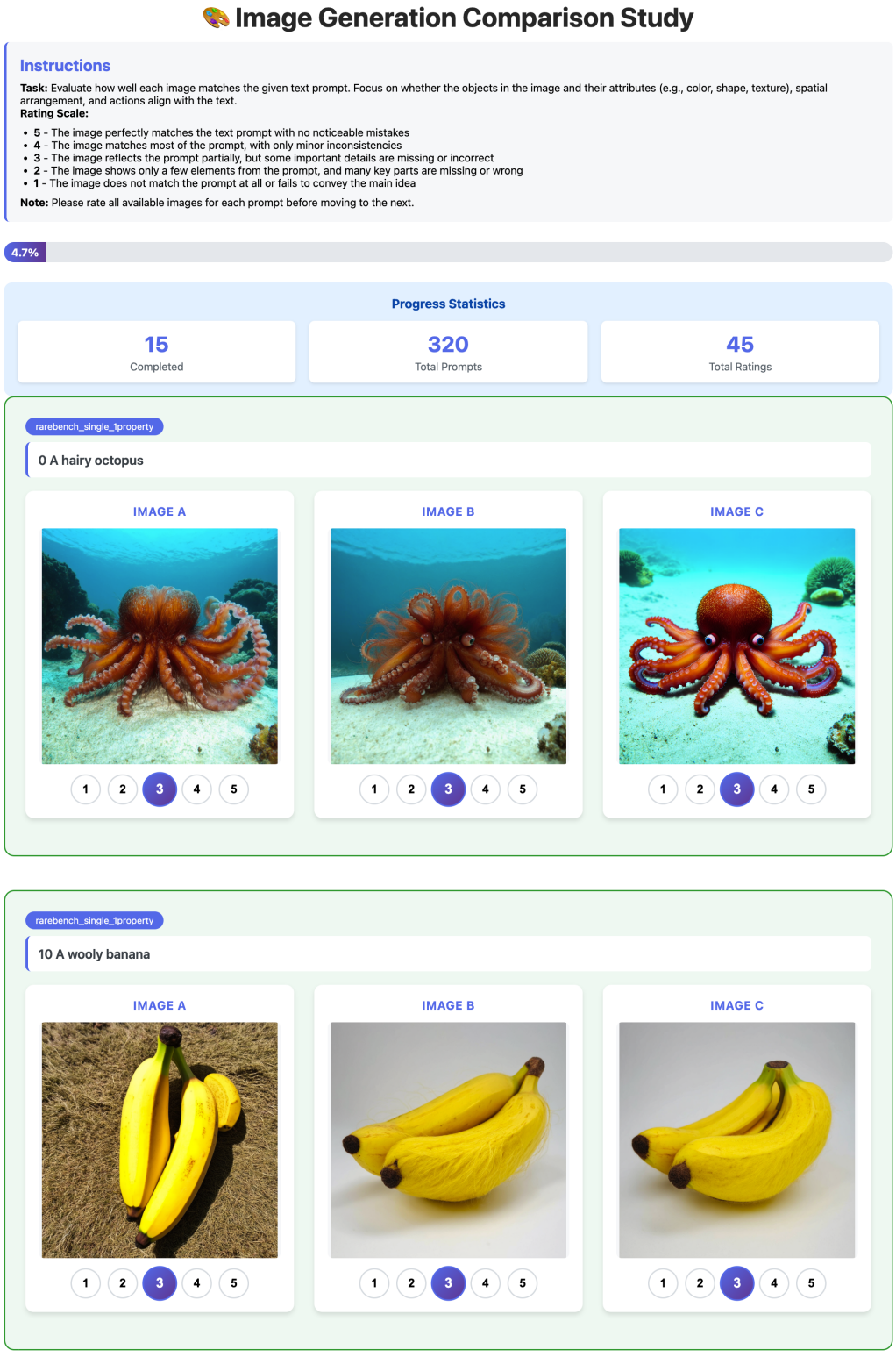

Figure 14. User study interface for evaluating text-to-image alignment. Participants were asked to rate how well each generated image matched the given text prompt based on semantic accuracy and visual consistency. The image order was randomized, and model identities were hidden to avoid bias.

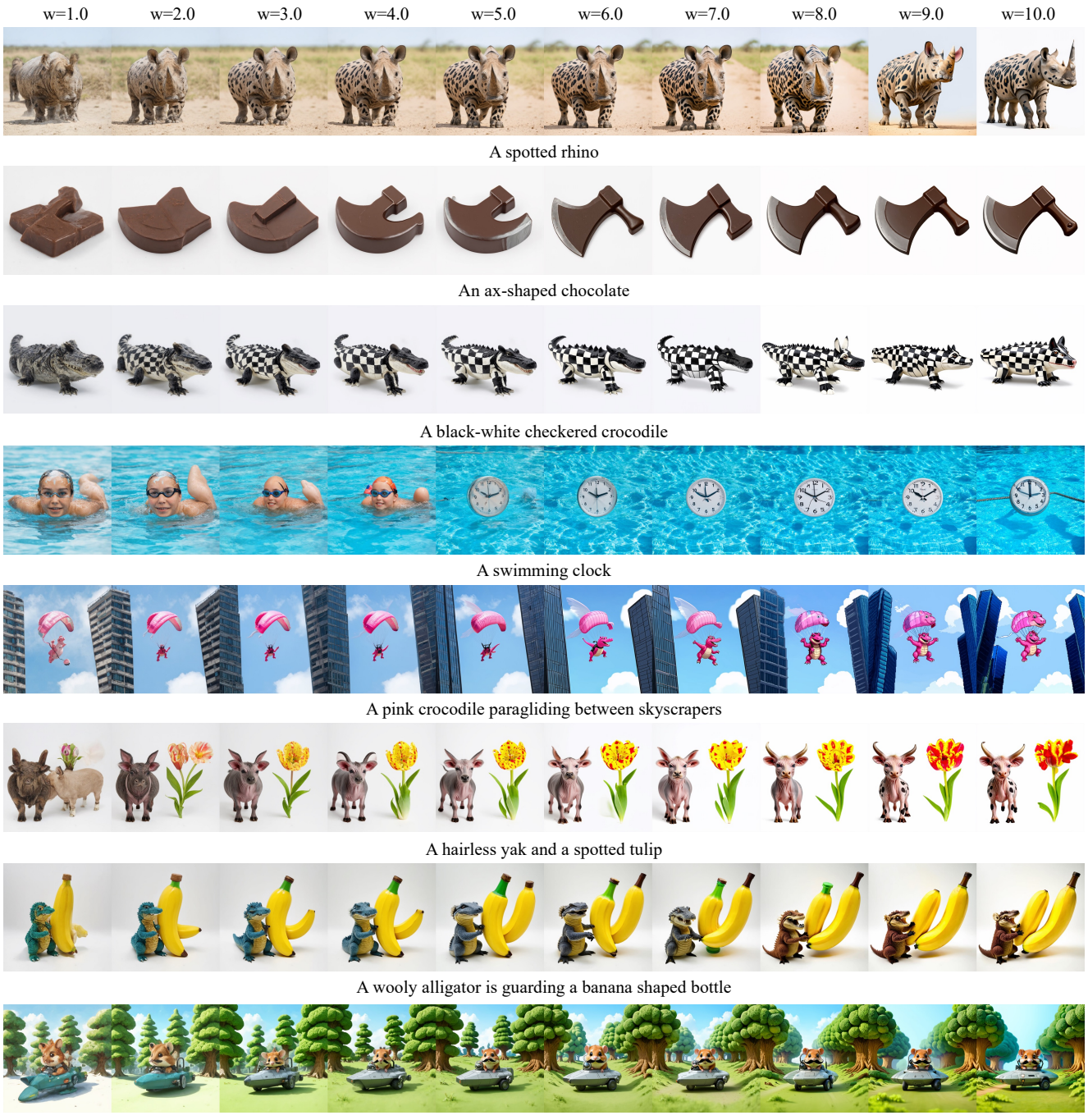

Figure 15. Ablation study with varying classifier-free guidance scale w.

“A mustachioed monkey”

“A donut shaped train”

“A flamingo made of glass”

“A dancing koala”

“A blue elephant spitting fire while floating in the air”

“A stork made of diamonds and a tulip made of diamonds”

“A wrinkled globe is resting beside an ax shaped chocolate”

“A blue ant hides in the forest in military uniform, aiming a gun at a red enemy ant on horseback”

Figure 16. Uncurated qualitative visualizations of our method on RareBench, generated from eight randomly selected prompts across all categories and eight random seeds.

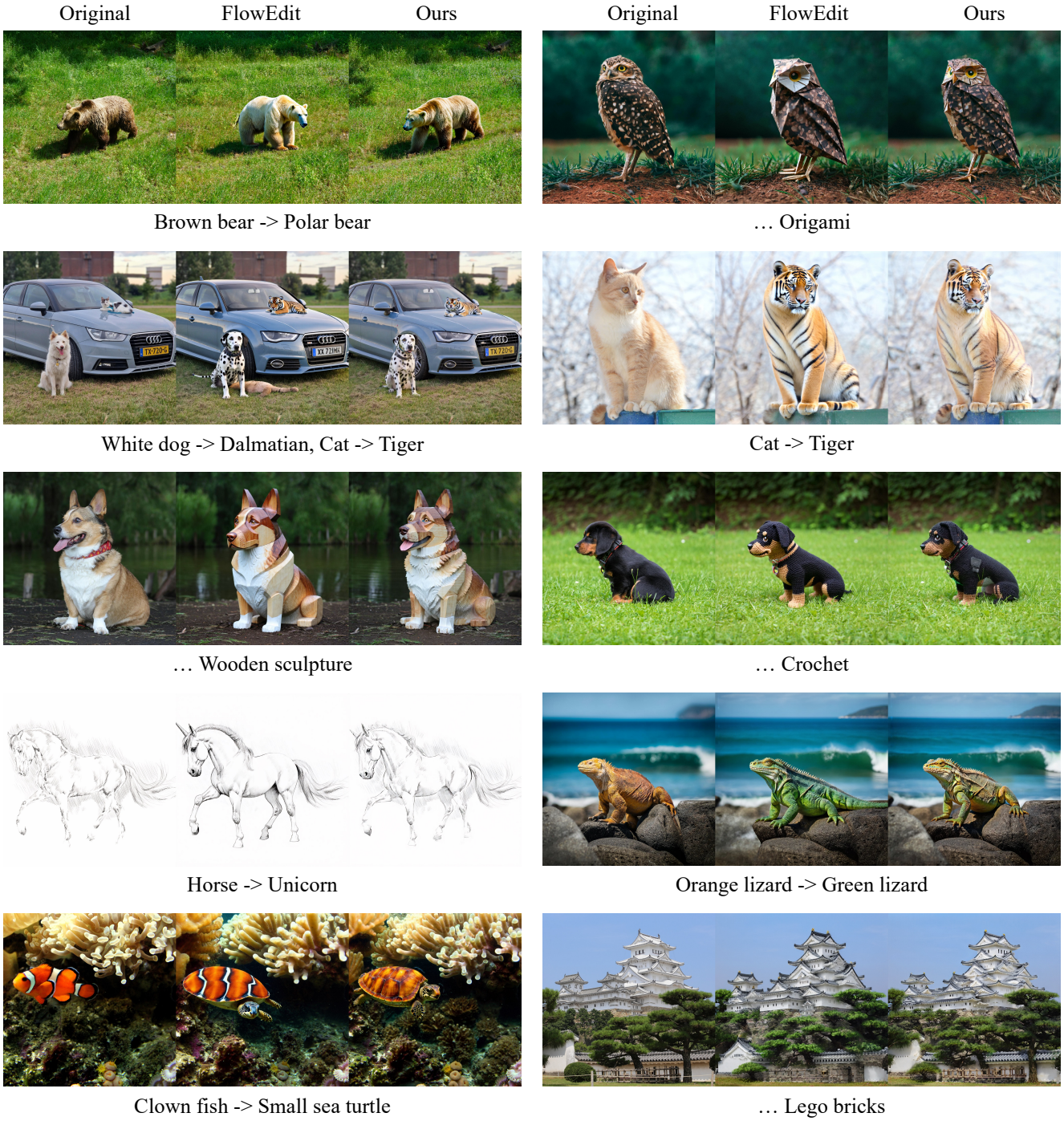

Figure 17. Qualitative results of the image editing.