REFLECTION-GUIDED DYNAMICS RECTIFICATION FOR TEXT-TO-IMAGE GENERATION

Anonymous authors
Paper under double-blind review

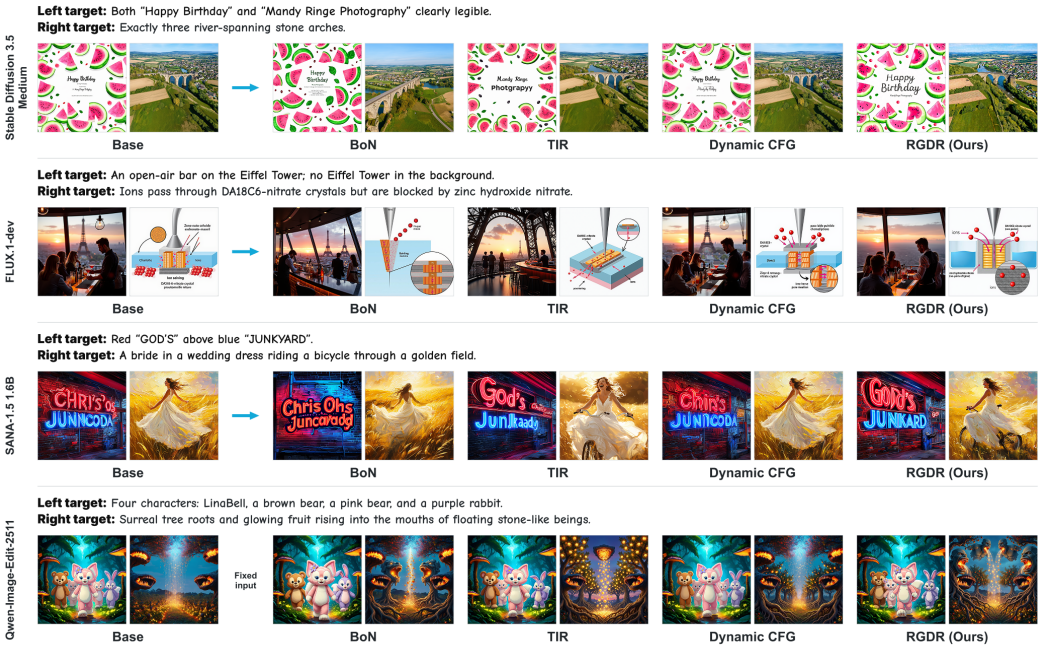


Figure 1: **Reflection-guided Dynamics Rectification (RGDR).** RGDR turns reflections on existing outputs into stepwise corrections along sampling trajectories. Qualitative comparisons with BoN, TIR, and Dynamic CFG across one image-editing model and three text-to-image models illustrate how different methods address residual errors and affect existing content during rectification.

ABSTRACT

Fine-grained rectification of existing generations under a limited additional budget remains a key problem in test-time optimization (TTO) for image generation. As for prior approaches, they could identify residual errors, but could not directly ensure that the model acts on the feedback during reasoning. Such errors stem from multiple sources, such as the initial noise or text prompt, and ultimately induce an execution gap of failing to handle potential deviations along the trajectory generating. However, classic models generally cannot remedy this gap. To narrow it, we propose Reflection-Guided Dynamics Rectification (RGDR), which turns reflections on completed outputs into corrective interventions on the process of sampling dynamics at the trajectory level. Specifically, RGDR could impact every sampling step by adapting to instantaneous intervention. Moreover, RGDR generates flexible candidates while adjudicating between them and the reference. As experiments show, rectification from RGDR starts from the existing base trajectory and reflection, and RGDR is substantially more effective than former trajectory exploration. Under matched budgets for additional image generation, RGDR's gain in the cross-model average final checklist score over the reference is $1.51\times$ that of TIR prompt refinement and $1.81\times$ that of Dynamic CFG.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

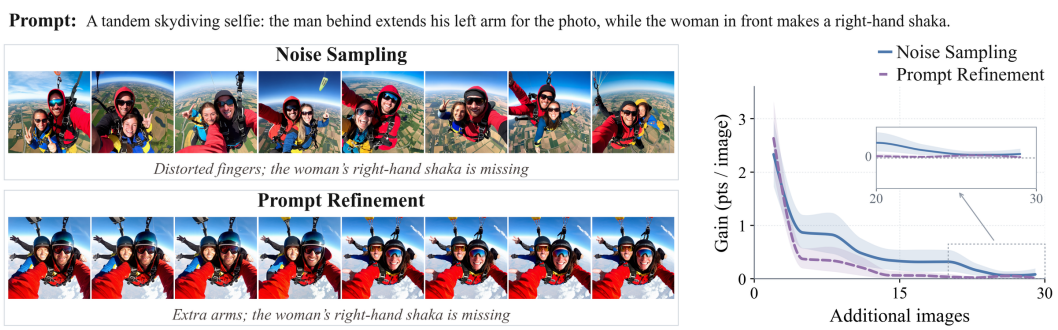


Figure 2: **Limitations of noise sampling and prompt refinement in follow-up rectification.** The left panel shows qualitative results from noise sampling (top) and prompt refinement (bottom). The right panel reports mean score gain per additional image across generation budgets.

1 INTRODUCTION

Research on text-to-image generation is increasingly directed toward achieving superior quality and better alignment with prompts. Specifically, as inference costs accumulate, test-time optimization (TTO) demonstrates its superiority without updating the parameters of a pretrained model. As for prior methods, they have made progress through sampling, search, and feedback-driven refinement, using generative models to produce additional candidates or image-editing models to revise the output (Wan et al., 2026; Jiang et al., 2026). However, a challenge has arisen in optimizing outputs under a limited additional budget, correcting residual errors while preserving prior progress.

We classify the methods commonly into two types. One is the noise level, and the other is the prompt-level scaling. These are two common approaches that improve the quality of image generation by searching for noise initialization and using the text prompt respectively. (Zhuo et al., 2025). Although the previous approaches are still effective, it is no longer possible to continuously achieve higher gains by further adjusting the noise or prompts when the budget is limited. Figure 2 contrasts the behavior of Noise Sampling and Prompt Refinement. The former produces substantial visual variation, whereas the latter repeatedly produces similar outputs, yet key requirements remain unmet in the illustrated examples. Across the evaluated settings, the mean score gain per additional image also generally decreases as the budget grows. Therefore, we believe that the quality of noise initialization and text guidance do indeed affect the correction effect, but the real bottleneck lies in the fact that the way the model transforms the guidance into images through continuous state updates along the generation trajectory is rather fragile. (Ho et al., 2020). If the intended correction does not succeed through updates, residual errors can persist, leaving an execution gap between the intention of the correction and the generative response. This motivates trajectory-level methods, which intervene in the sampling process rather than only searching over its initial conditions.

In terms of prior work, they have explored TTO along generation trajectories, using evaluations or diagnoses of intermediate states to guide subsequent sampling (Ma et al., 2025; Papalampidi et al., 2025; Lv et al., 2025). In parallel, SEGA (Brack et al., 2023) introduces additional semantic guidance, while CFG-Ctrl (Wang et al., 2026a) interprets velocity-field modulation from a control-theoretic perspective. However, they do not fully exploit the information available in follow-up rectification, such as the completed reference revealing residual errors. Therefore, We consider converting the residual-error diagnosis of a completed output into guidance that steers a new trajectory while retaining the reference configuration.

Given the above analysis, proposing a trajectory-level test-time scaling method for rectifying existing generations is crucial. Therefore, we introduce RGDR that generates complete candidates parameterized by η. As the correction field is the difference between velocity predictions conditioned on the desired state (Target) and the current erroneous state (Reference), we establish continuity and bounded deviation of candidates and derive sufficient conditions for corrective progress to outweigh accompanying interference under suitable regularity conditions. Figure 3 illustrates how feedback correction acts throughout generation by comparing trajectories from the same initial noise.

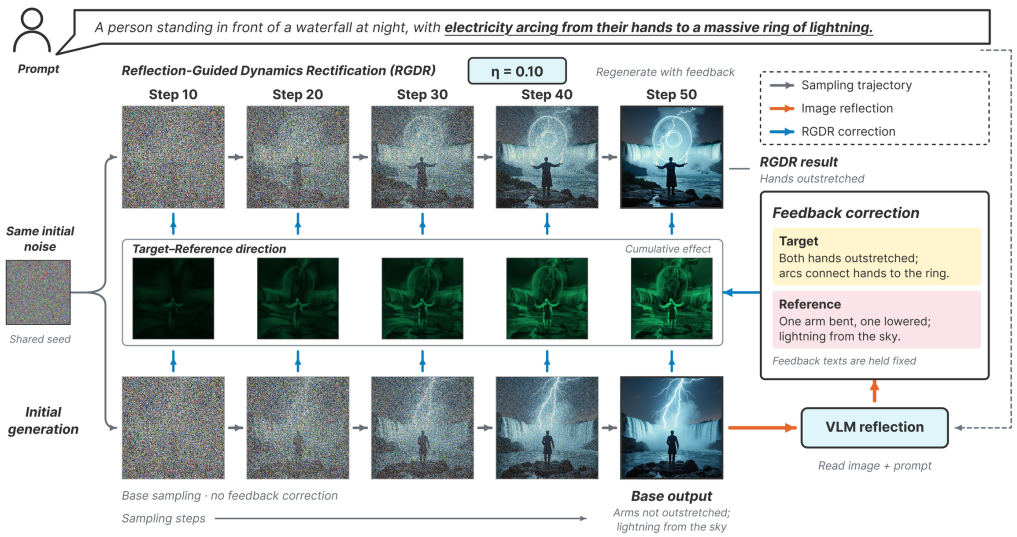


Figure 3: **Stepwise visualization with RGDR.** Reflections on the completed reference provide fixed Target–Reference conditions for constructing the correction field. Heatmaps show the spatial distribution of the correction signal along the generation trajectory.

To show the superiority in reasoning tasks, we evaluate RGDR on tasks drawn from PRISM-Bench (Fang et al., 2025) and SciIR-Bench (Ma et al., 2026) using Stable Diffusion 3.5 Medium, SANA-1.5 1.6B, and FLUX.1-dev. For each task and model, all methods continue from the same reference selected by preceding optimization. Under matched budgets for additional image generation, we compare RGDR with BoN, TIR, and Dynamic CFG, which respectively resample the noise, refine the prompt, and tune the guidance scale. RGDR raises the cross-model average final checklist score by 2.77, 1.61, and 2.09 points over these baselines, respectively. Direction ablations show the advantage of the Target–Reference construction over the tested alternatives, while strength analysis shows that effective correction depends on intervention direction and magnitude, rather than larger perturbations alone.

2 Background and Related Work

2.1 Flow-Based Generation

We formulate trajectory rectification in flow models as an intervention on the sampling velocity field. Flow models describe continuous transformations between noise and data distributions through time-dependent velocity fields. Flow matching (Lipman et al., 2023) and rectified flow (Liu et al., 2023) learn these fields and generate samples using numerical ordinary differential equation (ODE) solvers. Under the time convention from the noise endpoint T to the data endpoint 0, a sampling trajectory conditioned on c satisfies

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = V(t, x_t; c), \qquad x_T \sim \mathcal{N}(0, I_d), \qquad t : T \to 0. \tag{1}$$

Here, $x_t \in \mathbb{R}^d$ is the generative state, $T > 0$ is the sampling horizon, I_d is the d-dimensional identity matrix, and $V(t, x; c) \in \mathbb{R}^d$ is the velocity field incorporating the chosen guidance configuration. The states x_T and x_0 denote the initial noise and generated endpoint, respectively. In latent-space models, a decoder maps x_0 to the final image (Rombach et al., 2022). This framework supports both text-to-image generation and image editing. For example, FLUX.1 Kontext (Black Forest Labs et al., 2025) unifies image generation and conditional editing within a flow-matching model.

2.2 Semantic Guidance and Additive Velocity Fields

Semantic guidance combines or modifies model predictions to change state updates during sampling, thereby steering generation trajectories. Classifier-free guidance (CFG) (Ho & Salimans,

2022) combines conditional and unconditional predictions and scales their difference. SEGA (Brack et al., 2023) introduces independently adjustable concept directions alongside the original prompt guidance, allowing additional semantics to influence sampling updates.

RGDR constructs reflection-guided trajectory corrections using the additive velocity-field formulation

$$V_\alpha(t,x) = V_{\text{base}}(t,x) + \alpha D(t,x), \tag{2}$$

where $V_{\text{base}}(t,x)$ is the base velocity field, $D(t,x) \in \mathbb{R}^d$ is an additional semantic direction field, and α is a scaling coefficient. Both fields are evaluated at the same time and current state.

2.3 Test-Time Optimization for Image Generation

Test-time optimization improves image generation through additional sampling, search, and feedback-driven refinement (Wan et al., 2026). BoN (Ma et al., 2025) selects candidates generated from different noise initializations, while TIR (Khan et al., 2025) refines prompts using the current image and accumulated feedback. ReflectionFlow (Zhuo et al., 2025) and Reflect-DiT (Li et al., 2025) train models to revise outputs using existing images and textual reflections, whereas RAISE (Jiang et al., 2026) coordinates prompt rewriting, noise resampling, and image editing without additional training. These methods use candidate selection or feedback-driven correction to organize subsequent generation and editing around existing outputs.

Trajectory-Level Scaling. Related work uses evaluations and diagnoses of intermediate states to adjust generation trajectories. Search over Paths (Ma et al., 2025) evaluates denoised predictions to select subsequent branches, while Dynamic CFG (Papalampidi et al., 2025) uses latent-space evaluators to select guidance scales at each step. PPAD (Lv et al., 2025) diagnoses intermediate previews and revises subsequent denoising through corrective conditions and rollback. T2S2 (Sakurai et al., 2026) detects prohibited concepts in intermediate clean-image estimates, optimizes a low-rank conditioning residual, and restarts sampling.

3 Method

3.1 Controlled Superposition of the Correction Field

Continuity and bounded deviation of the candidate family. Starting from Eq. (2) in Section 2.2, fix the base field, correction field, and initial noise, and consider

$$\mathrm{d}x_t = [V_{\text{base}}(t,x_t) + \alpha D(t,x_t)]\,\mathrm{d}t, \qquad t : T \to 0. \tag{3}$$

We omit the dependence of the trajectory state on α and denote the endpoint of a complete generation by $\gamma_D(\alpha)$. A strength range $\mathcal{A}$ defines the candidate family $\Gamma_D(\mathcal{A}) = \{\gamma_D(\alpha) : \alpha \in \mathcal{A}\}$, with reference endpoint $x^\star = \gamma_D(0)$.

Theorem 3.1 (Continuity and bounded deviation of the candidate family)**.** *Fix $r > 0$. Suppose that the dynamics admit a unique solution for every $\alpha \in [0,r]$. On a common region containing the relevant trajectories, assume that V_{base} and D are continuous in time and uniformly Lipschitz in the state, and that D is bounded. Then there exists a finite constant K_r, independent of α, α', such that, for all $\alpha, \alpha' \in [0,r]$,*

$$\|\gamma_D(\alpha) - \gamma_D(\alpha')\| \le K_r|\alpha - \alpha'|, \qquad \|\gamma_D(\alpha) - x^\star\| \le K_r\alpha. \tag{4}$$

The two bounds describe continuous dependence of the endpoint on strength and its deviation from the reference, respectively. Candidates converge to the reference endpoint as $\alpha \to 0$. The proof is provided in Appendix B.1.

Relative reparameterization of the intervention strength. The theorem characterizes endpoint variation for a fixed direction field, but the relative effect of the raw coefficient α still depends on the norms of the correction and base fields. We therefore reparameterize the additional term relative to the base velocity norm,

$$\begin{aligned} \overline{D}(t,x) &= \frac{\|V_{\text{base}}(t,x)\|}{\|D(t,x)\| + \varepsilon} D(t,x), \qquad \varepsilon > 0, \\ V_{\eta,s}(t,x) &= V_{\text{base}}(t,x) + \eta s(t) \overline{D}(t,x). \end{aligned} \tag{5}$$

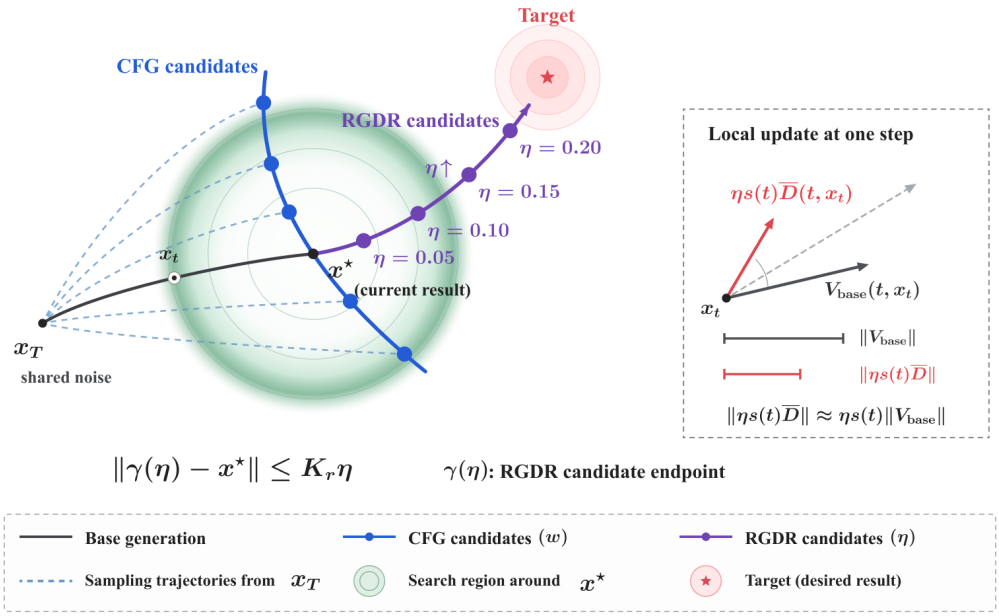


Figure 4: **Continuously controlled candidate space.** The blue and purple curves represent candidate endpoints parameterized by the CFG scale w and RGDR relative strength η.

Here, $\eta \ge 0$ sets the overall relative strength and $s(t) \ge 0$ is a fixed, bounded temporal schedule. By construction,

$$\|V_{\eta,s}(t,x) - V_{\text{base}}(t,x)\| \le \eta s(t)\|V_{\text{base}}(t,x)\|. \tag{6}$$

The correction at each sampling step is thus scaled relative to the current base velocity, with $\eta s(t)$ bounding its relative norm. This gives η a dimensionless interpretation as a relative intervention parameter, rather than merely a multiplier of the raw direction.

Temporal scheduling. The schedule $s(t)$ distributes the correction strength along the generation trajectory. Under the convention $t : T \to 0$, $s(t) = 1$ gives constant intervention, $s(t) = t/T$ places greater strength early in sampling, and $s(t) = 1 - t/T$ places greater strength later. Other nonnegative bounded schedules can also be used. Model-specific schedules and scheduling experiments are provided in Appendices C.2, D.1, and D.2.

With the direction field and schedule fixed, let $\gamma(\eta)$ denote the endpoint of a complete generation under $V_{\eta,s}$. Applying Theorem 3.1 with direction field $s\overline{D}$ and strength η, under the corresponding regularity conditions, shows that candidate endpoints vary continuously with η, have bounded deviations from the reference, and converge to the reference endpoint as $\eta \to 0$. Figure 4 illustrates this candidate space continuously parameterized by relative strength.

3.2 Semantic Construction of the Correction Field

Corrective progress and accompanying interference. Fix the task q, feedback F, and the base field, initial noise, and schedule s from Section 3.1. We write D_F for the direction field D constructed from F, and retain $\gamma(\eta)$ for the endpoint under relative reparameterization. Let $\Phi_q : \mathbb{R}^d \to \mathbb{R}^{m_q}$ be a fixed task representation used for analysis, where m_q is its dimension. For latent-space generation, this mapping includes image decoding. Define

$$z(\eta) = \Phi_q(\gamma(\eta)), \qquad \Delta z(\eta) = z(\eta) - z(0). \tag{7}$$

In this representation space, let the unit vector $e_{\text{fix}} \in \mathbb{R}^{m_q}$ specify the desired correction direction, and let $P_{\text{keep}} \in \mathbb{R}^{m_q \times m_q}$ be the orthogonal projector onto the local subspace sensitive to properties already satisfied by the reference, with $P_{\text{keep}}^\top = P_{\text{keep}}^2 = P_{\text{keep}}$. Let $\beta > 0$ weight the interference

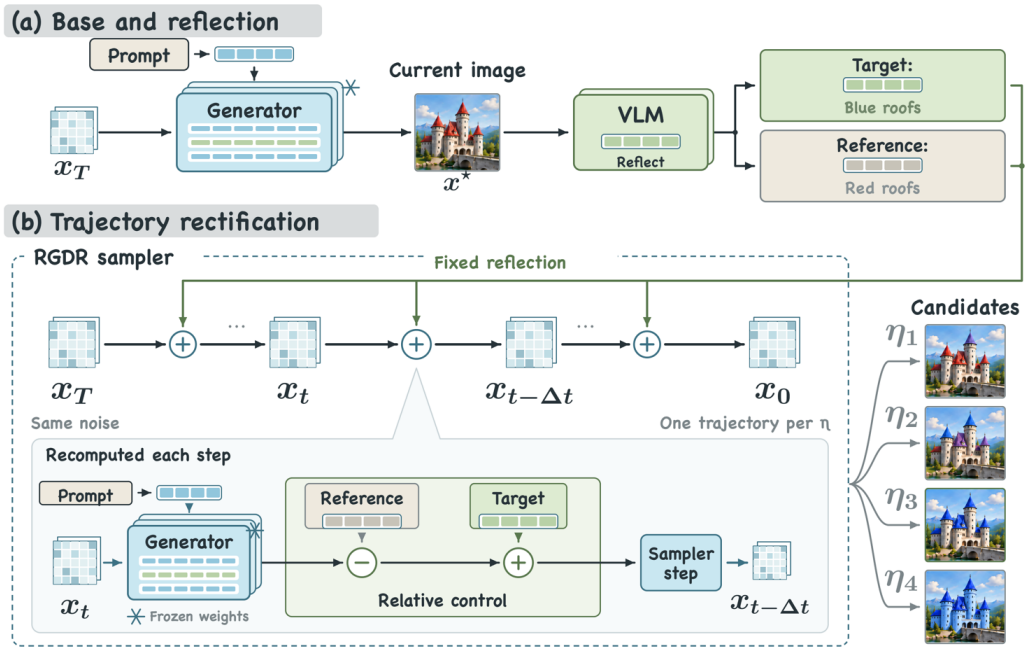


Figure 5: **RGDR pipeline overview.**

term. These analytical objects remain fixed for the current correction and are assumed to capture the progress and interference of interest. Define the net corrective progress

$$\Delta_q(\eta) = \langle e_{\text{fix}}, \Delta z(\eta) \rangle - \beta \| P_{\text{keep}} \Delta z(\eta) \|. \tag{8}$$

The first term measures signed progress along the desired direction; the second measures the magnitude of changes in the subspace associated with already satisfied properties. A positive $\Delta_q(\eta)$ means that corrective progress outweighs weighted accompanying interference. This quantity is used for local task-representation analysis and does not directly imply a higher discrete checklist score; projected changes need not indicate damage to existing properties.

Theorem 3.2 (Dominance of corrective progress induced by the feedback direction)**.** *Let $r > 0$. Suppose that z is twice continuously differentiable on $[0, r]$, with*

$$\sup_{\eta \in [0,r]} \| z''(\eta) \| \le M < \infty. \tag{9}$$

If the task response induced by D_F satisfies

$$\kappa_q(D_F) := \langle e_{\text{fix}}, z'(0^+) \rangle - \beta \| P_{\text{keep}} z'(0^+) \| > 0, \tag{10}$$

where $z'(0^+)$ is the right derivative at zero strength, then, for every $\eta \in [0, r]$,

$$\Delta_q(\eta) \ge \kappa_q(D_F)\eta - \frac{(1+\beta)M}{2}\eta^2. \tag{11}$$

Consequently, when $M > 0$, we have $\Delta_q(\eta) > 0$ for all strengths satisfying

$$0 < \eta < \min\left\{ r, \frac{2\kappa_q(D_F)}{(1+\beta)M} \right\}. \tag{12}$$

When $M = 0$, the same conclusion holds for all $0 < \eta \le r$.

The result permits accompanying interference, provided that first-order corrective progress at the reference exceeds its weighted contribution. The second-order response bound determines the strength interval over which this advantage holds. The proof and the dynamical expansion of the first-order response are provided in Appendices B.2 and B.3.

Target–Reference semantic construction. To emphasize the requested change and reduce unrelated intervention, we extract from feedback F two conditions describing the desired state, c_{target}, and the current erroneous state, $c_{\text{reference}}$. Using the frozen model's conditional velocity predictions V_θ, we construct

$$D_F(t,x) = V_\theta(t,x;c_{\text{target}}) - V_\theta(t,x;c_{\text{reference}}). \tag{13}$$

Both conditions focus on the attributes, objects, or relations to be corrected and retain the necessary context without restating unrelated requirements that are already satisfied. This construction aims to increase the first-order margin between corrective progress and interference in the first-order direction condition; it does not automatically satisfy that sufficient condition, and its effect is examined through generative responses and experiments.

3.3 RGDR Algorithm

Combining the controlled superposition in Section 3.1 with the semantic construction in Section 3.2, RGDR extracts Target and Reference conditions from reflections on the completed reference output. At each sampling step, it recomputes the correction field, applies relative reparameterization, and adds the correction to the base velocity field. With the generation configuration, feedback, and schedule fixed, candidates at different relative strengths are sampled independently from the same initial noise and evaluated together with the reference for selection. Model parameters remain unchanged throughout. Algorithm 1 gives the complete procedure, and Figure 5 illustrates the pipeline.

Algorithm 1 Reflection-Guided Dynamics Rectification (RGDR)

Input Task q; reference endpoint $x^\star$; fixed initial noise x_T;
base field V_{base}; frozen conditional field V_θ; feedback F; schedule s;
relative strength set $\mathcal{E}$; stabilizer $\varepsilon > 0$;
evaluator $\widehat{S}_{\text{sel}}(q,x)$ scoring the image corresponding to endpoint x

Output Selected endpoint $\widehat{x}$

Extract Target and Reference conditions from F
$D_F(t,x) \leftarrow V_\theta(t,x;c_{\text{target}}) - V_\theta(t,x;c_{\text{reference}})$
Evaluate both terms at the same (t,x)
$\overline{D}_F \leftarrow \|V_{\text{base}}\| D_F / (\|D_F\| + \varepsilon)$
$\mathcal{P} \leftarrow \{(0, x^\star)\}$
for all $\eta \in \mathcal{E}$ **do**
    Restart from the fixed initial noise x_T
    Integrate $V_{\text{base}} + \eta s(t)\overline{D}_F$ over the full interval from T to 0
Recompute the fields at each current state
    $\mathcal{P} \leftarrow \mathcal{P} \cup \{(\eta, x_0)\}$
end for
$(\widehat{\eta}, \widehat{x}) \in \arg\max_{(\eta,x)\in\mathcal{P}} \widehat{S}_{\text{sel}}(q,x)$
return $\widehat{x}$

4 Experiments

4.1 Setup

To evaluate fine-grained rectification of verifiable residual errors, we build an evaluation set from PRISM-Bench and SciIR-Bench, covering compositional requirements and scientific image reasoning (Fang et al., 2025; Ma et al., 2026). References are generated by Stable Diffusion 3.5 Medium, SANA-1.5 1.6B, and FLUX.1-dev, or edited from SD3.5 Medium outputs by Qwen-Image-Edit-2511. All parameters remain frozen. Doubao Seed-2.1 Pro supplies feedback and scores outputs against fixed checklists on a 0–100 scale. We report Final, Gain, Rel. gain, and Improved for overall fulfillment and follow-up improvement, and Repair and Preserve for error correction and retention. Details are provided in Appendices C.1, C.2, and C.5.

4.2 Main Experiments

Comparison of follow-up rectification methods. We compare continued BoN sampling (Ma et al., 2025), TIR prompt refinement (Khan et al., 2025), a guidance-scale baseline adapted from

Table 1: **Comparison of follow-up rectification on PRISM-Bench and SciIR-Bench tasks.** Budget denotes the number of additional complete images per reference requiring correction. Bold indicates the best value in each column for a given model.

Method	Budget (images)	Final ↑ (0–100)	Gain ↑ (points)	Rel. gain ↑ (%)	Improved ↑ (%)	Repair ↑ (%)	Preserve ↑ (%)
Stable Diffusion 3.5 *Medium*							*Base* = 56.90
BoN	10	58.25	+1.35	2.38	12.0	4.11	96.14
TIR	10	58.14	+1.24	2.18	13.5	6.04	**98.39**
Dynamic CFG	10	59.05	+2.15	3.79	23.0	6.99	97.87
RGDR (Ours)	10	**60.64**	**+3.74**	**6.58**	**28.0**	**12.96**	96.44
SANA-1.5 1.6*B*							*Base* = 46.55
BoN	10	48.09	+1.54	3.31	13.0	3.40	**99.08**
TIR	10	49.96	+3.41	7.32	25.0	9.71	96.61
Dynamic CFG	10	48.97	+2.42	5.19	22.5	6.55	99.07
RGDR (Ours)	10	**50.92**	**+4.37**	**9.38**	**31.5**	**13.07**	94.46
FLUX.1-*dev*							*Base* = 70.30
BoN	10	71.57	+1.27	1.80	11.5	7.92	97.65
TIR	10	73.29	+2.99	4.25	15.5	10.23	**97.86**
Dynamic CFG	10	71.93	+1.63	2.32	16.5	8.78	97.65
RGDR (Ours)	10	**74.67**	**+4.36**	**6.21**	**26.5**	**19.15**	96.78
Qwen-Image-Edit-2511							*Base* = 72.77
BoN	5	74.84	+2.07	2.85	24.55	13.99	97.53
TIR	5	76.17	+3.41	4.68	30.91	20.32	97.25
Dynamic CFG	5	75.79	+3.03	4.17	29.09	16.27	**97.95**
RGDR (Ours)	5	**76.99**	**+4.23**	**5.81**	**33.64**	**23.58**	96.48

Dynamic CFG (Papalampidi et al., 2025), and RGDR on the follow-up rectification evaluation sets for three text-to-image models and an image-editing model. The comparison examines whether residual errors in otherwise satisfactory generations or edited outputs can be further corrected. All methods share references and receive the matched additional-image budgets in Table 1. Comparison protocols are detailed in Appendix C.3.

Table 1 shows that RGDR achieves the highest Final, Gain, Improved, and Repair across all four models. Averaged equally across models, its Gain exceeds that of BoN and TIR by 2.62 and 1.41 points, respectively. Under matched additional-image budgets, stepwise corrections along generation trajectories therefore realize more residual requirements than further noise or prompt adjustments. Figure 1 shows how targeted corrections can preserve the structure of existing scenes. In the bridal cycling example, RGDR adds the bicycle and riding action while retaining the side-view figure, wedding dress, and golden field, whereas TIR changes the subject's orientation and composition. In the text examples, modifications concentrate on the text region while retaining the surrounding border or sign setting. These comparisons address the difficulty of realizing requested changes while controlling visual variation illustrated in Figure 2. RGDR also achieves 1.81 times the mean Gain of the adapted Dynamic CFG baseline, supporting the error-specific additional direction illustrated in Figure 4. Reflection-guided trajectory correction thus provides greater rectification gains than adjusting native guidance strength alone. The same advantage on previously edited outputs shows that RGDR can serve as an additional rectification mechanism within existing generation and editing workflows.

4.3 Ablation Studies

Correction direction construction. Why use Target–Reference to construct the correction direction D_F in RGDR? Motivated by the analysis of corrective progress and accompanying interference in Theorem 3.2, we compare TR Correct, Operation, TR Wrong, and Gaussian on the rectification subsets of all four models. All configurations share references, relative reparameterization, and the evaluation protocol. The constructions of these four methods are detailed in Appendix C.4.

TR Correct achieves the highest Gain in all three task categories in Table 2, with an overall Gain of 4.34 points and Repair of 15.28%. The chimney example in Figure 6 illustrates the semantic effects of different directions. TR Correct changes three chimneys to two; Gaussian mainly alters minor de-

Table 2: **Correction direction ablation averaged across four models.** Gain is measured in score points and Repair in percent. Bold indicates the best value in each column.

	Category-wise Gain ↑			Overall	
Direction	Human	Open	Scientific & Text	Gain ↑	Repair (%) ↑
TR Correct	**7.56**	**5.71**	**3.18**	**4.34**	**15.28**
Operation	2.09	2.75	2.65	2.60	7.95
TR Wrong	3.87	1.56	0.45	1.15	4.36
Gaussian	0.37	0.43	1.82	1.31	4.16

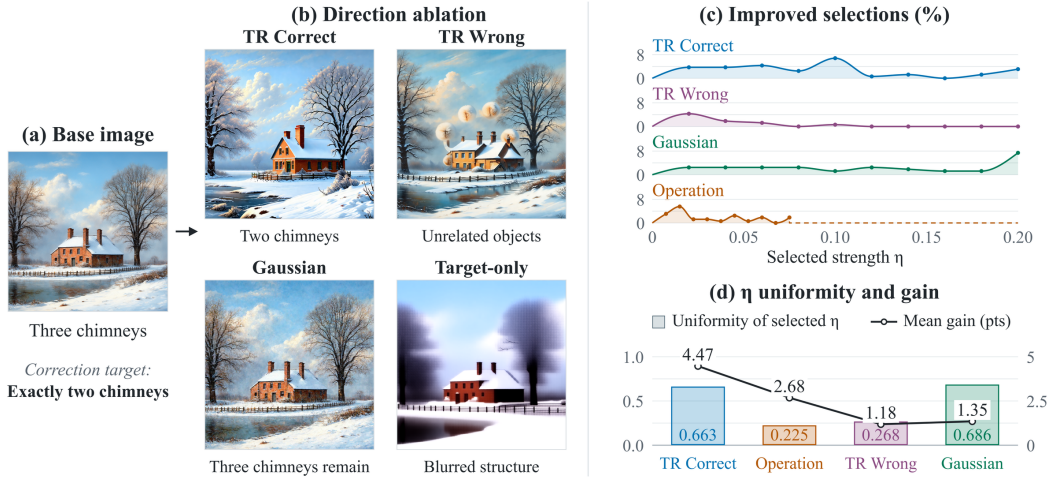


Figure 6: **Qualitative comparison of correction directions and strength analysis.** The left panels show image responses under different directions. The right panels show the selected-strength distributions of improved outputs, together with selected-strength uniformity and mean gain.

tails while leaving the incorrect count unchanged; and TR Wrong introduces unrelated objects. On the separate paired visualization set in Figure 6(c,d), improved selections for TR Correct span multiple strengths from weak interventions up to $\eta = 0.20$, while achieving the highest mean Gain. Gaussian also has broadly distributed selected strengths, but its mean Gain is only about one-third of that of TR Correct. Improved selections for Operation concentrate at small strengths, whereas TR Wrong yields none above $\eta = 0.10$. These results show that the value of a direction lies not in larger perturbations alone, but in producing useful, target-specific corrections at different intervention strengths. TR Correct provides improved candidates across a broad range of strengths while achieving higher rectification gains, empirically supporting the Target–Reference design motivated by Theorem 3.2.

5 Conclusion

In this paper, we introduce Reflection-guided Dynamics Rectification (RGDR) to address the gap of fine-grained refinement of generated images, as a model may not carry out a desired correction along its sampling trajectory, even when feedback correctly identifies the remaining error. RGDR is a training-free approach in trajectory level that turns feedback on a completed output into a corrective intervention during a new sampling process. Our theoretical analysis characterizes the continuity of the resulting candidate family under suitable regularity conditions and provides a sufficient condition under which the corrective effect outweighs collateral changes. Through massive experiments, RGDR achieves higher scores of average final checklist than BoN, TIR, and Dynamic CFG. It also attains the highest repair rate on all four evaluated models. Overall, these findings indicate that reflecting on a completed output and applying that feedback to the sampling dynamics can provide an effective way to rectify existing generations without updating model parameters. We hope our work could be extended to a wider range of models and generation tasks in the future.

AI USE STATEMENT

We used generative AI tools to assist with manuscript polishing, the writing of mathematical proofs, the refinement of existing hypotheses, code development, and figure generation. We did not use these tools to originate the core research ideas or claims. The use of generative Al tools to generate synthetic datasets is not applicable to this work, In particular, AI-assisted proofs and hypothesis refinements were checked for logical consistency, code was reviewed and verified for correctness, and figures were inspected for accuracy and scientific plausibility. We take responsibility for the final content of this work, including all text, claims, code, and figures produced with the assistance of generative AI tools.

REPRODUCIBILITY STATEMENT

To make our results reproducible, we disclose in the appendix all relevant details, including theoretical proofs, data-splitting protocols, normalization procedures applied during training only, checkpoint-selection criteria, inversion and decoding settings, and the random seeds used.

REFERENCES

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *arXiv preprint arXiv:2506.15742*, 2025. URL https://arxiv.org/abs/2506.15742.

Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. SEGA: Instructing text-to-image models using semantic guidance. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1102.

Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. ReNO: Enhancing one-step text-to-image models through reward-based noise optimization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 125487–125519, 2024. doi: 10.52202/079017-3987.

Rongyao Fang, Aldrich Yu, Chengqi Duan, Linjiang Huang, Shuai Bai, Yuxuan Cai, Kun Wang, Si Liu, Xihui Liu, and Hongsheng Li. FLUX-Reason-6M & PRISM-Bench: A million-scale text-to-image reasoning dataset and comprehensive benchmark. *arXiv preprint arXiv:2509.09680*, 2025. doi: 10.48550/arXiv.2509.09680. URL https://arxiv.org/abs/2509.09680.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. doi: 10.48550/arXiv.2207.12598. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020. doi: 10.48550/arXiv.2006.11239. URL https://arxiv.org/abs/2006.11239.

Liyao Jiang, Ruichen Chen, Chao Gao, and Di Niu. RAISE: Requirement-adaptive evolutionary refinement for training-free text-to-image alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2026.

Mohammed Abdul Hafeez Khan, Yash Jain, Siddhartha Bhattacharyya, and Vibhav Vineet. Test-time prompt refinement for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 6565–6575, October 2025.

Subin Kim, Sangwoo Mo, Mamshad Nayeem Rizve, Yiran Xu, Difan Liu, Jinwoo Shin, and Tobias Hinz. Rethinking prompt design for inference-time scaling in text-to-visual generation. *arXiv preprint arXiv:2512.03534*, 2025. doi: 10.48550/arXiv.2512.03534. URL https://arxiv.org/abs/2512.03534.

Anthony W. Knapp. *Basic Real Analysis*. Published by the author, East Setauket, NY, digital second edition, 2016. URL https://www.math.stonybrook.edu/~aknapp/download/b2-realanal-inside.pdf.

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Arsh Koneru, Yusuke Kato, Kazuki Kozuka, and Aditya Grover. Reflect-DiT: Inference-time scaling for text-to-image diffusion transformers via in-context reflection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15657–15668, October 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Zheqi Lv, Junhao Chen, Qi Tian, Keting Yin, Shengyu Zhang, and Fei Wu. Multimodal LLM-guided semantic correction in text-to-image diffusion. *arXiv preprint arXiv:2505.20053*, 2025.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Scaling inference time compute for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2523–2534, June 2025.

Zhiyuan Ma, Zhengfeng Shi, Yuning An, Peize Li, Jiabao Wei, Ruijie Li, Junhao Xiao, Jianjun Li, and Bowen Zhou. SciIR: A large-scale training dataset and benchmark for scientific image reasoning generation. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2606.30124.

Jiadong Pan, Zhiyuan Ma, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Self-Reflective Reinforcement Learning for Diffusion-based Image Reasoning Generation. *arXiv preprint arXiv:2505.22407*, 2025. URL https://arxiv.org/abs/2505.22407.

Pinelopi Papalampidi, Olivia Wiles, Ira Ktena, Aleksandar Shtedritski, Emanuele Bugliarello, Ivana Kajic, Isabela Albuquerque, and Aida Nematzadeh. Dynamic classifier-free diffusion guidance via online feedback. *arXiv preprint arXiv:2509.16131*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Jinya Sakurai, Shueicheng Yan, and Xun Xu. Test-Time Scaling for Safe Text-Guided Image Generation via Intermediate Clean Estimates. *arXiv preprint arXiv:2608.03284*, 2026. URL https://arxiv.org/abs/2608.03284.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55810–55827, 2025.

Gerald Teschl. *Ordinary Differential Equations and Dynamical Systems*, volume 140 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012. ISBN 9780821883280.

Cong Wan, Ying He, Zhongzhan Huang, and Hefeng Wu. Test-time scaling in multimodal foundation models: A comprehensive survey of generation and reasoning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 7751–7767, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.383. URL https://aclanthology.org/2026.findings-acl.383/.

Hanyang Wang, Yiyang Liu, Jiawei Chi, Fangfu Liu, Ran Xue, and Yueqi Duan. CFG-Ctrl: Control-based classifier-free diffusion guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11437–11447, June 2026a.

Tianyu Wang, Zhiyuan Ma, Qian Wang, Xinyi Zhang, Xinwei Long, and Bowen Zhou. RL-RIG: A Generative Spatial Reasoner via Intrinsic Reflection. *arXiv preprint arXiv:2602.19974*, 2026b. URL https://arxiv.org/abs/2602.19974.

Zhengyuan Yang, Jianfeng Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. Idea2Img: Iterative Self-Refinement with GPT-4V(ision) for Automatic Image Design and Generation. arXiv preprint arXiv:2310.08541, 2023. URL https://arxiv.org/abs/2310.08541. First released in 2023.

Xiangcheng Zhang, Haowei Lin, Haotian Ye, James Zou, Jianzhu Ma, Yitao Liang, and Yilun Du. Inference-time scaling of diffusion models through classical search, 2025. URL https://arxiv.org/abs/2505.23614.

Le Zhuo, Liangbing Zhao, Sayak Paul, Yue Liao, Renrui Zhang, Yi Xin, Peng Gao, Mohamed Elhoseiny, and Hongsheng Li. From reflection to perfection: Scaling inference-time optimization for text-to-image diffusion models via reflection tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15329–15339, October 2025.

A Extended Related Work

Noise optimization and multi-round reflection. Beyond the noise sampling and prompt refinement methods discussed in the main text, ReNO (Eyring et al., 2024) optimizes initial noise using reward gradients. Idea2Img (Yang et al., 2023) iteratively updates prompts based on diagnoses of generated images, while PRIS (Kim et al., 2025) aggregates recurring errors across candidates to guide prompt redesign. Other work enhances multi-round reflection through training. SRRL (Pan et al., 2025) and RL-RIG (Wang et al., 2026b) use reinforcement learning to improve multi-round image generation and reflection-based editing, respectively.

Particle and tree search along generation trajectories. Feynman–Kac Steering (Singhal et al., 2025) explores generation trajectories through particle resampling, whereas Classical Search (Zhang et al., 2025) combines local and tree search to expand sampling paths. These methods complement the trajectory selection, guidance-scale adjustment, and semantic diagnostic interventions discussed in the main text.

B Mathematical Proofs

We retain the definitions and assumptions of the main text. Vector norms are Euclidean, and matrix norms are the corresponding induced operator norms.

B.1 Proof of Theorem 3.1

Let L_V, L_D be the uniform state-Lipschitz constants of the base and direction fields, respectively, and choose a finite $C_D \geq 0$ such that $\|D(t,x)\| \leq C_D$. For $\alpha, \alpha' \in [0, r]$, let $x_t^{\alpha}, x_t^{\alpha'}$ be the solution trajectories starting from the same initial noise x_T, and write $e(t) = \|x_t^{\alpha} - x_t^{\alpha'}\|$. Their integral equations and the Lipschitz assumptions give

$$e(t) \leq \int_t^T \left[(L_V + rL_D)e(u) + C_D|\alpha - \alpha'|\right] \mathrm{d}u. \tag{B.1}$$

Set $L_r = L_V + rL_D$. Applying Grönwall's inequality (Teschl, 2012, Lemma 2.7) after the change of time variable $\tau = T - t$ yields

$$e(t) \leq C_D|\alpha - \alpha'| \int_t^T e^{L_r(u-t)} \mathrm{d}u. \tag{B.2}$$

Taking $t = 0$ and defining $K_r = C_D \int_0^T e^{L_r u} \mathrm{d}u < \infty$, we obtain

$$\|\gamma_D(\alpha) - \gamma_D(\alpha')\| \leq K_r|\alpha - \alpha'|. \tag{B.3}$$

Setting $\alpha' = 0$ and using $x^\star = \gamma_D(0)$ then gives $\|\gamma_D(\alpha) - x^\star\| \leq K_r\alpha$. This completes the proof. The integral expression for K_r also covers $L_r = 0$ and $C_D = 0$.

Application after relative reparameterization. Suppose additionally that $\|V_{\text{base}}(t,x)\| \leq C_V < \infty$ on the common trajectory region. The reverse triangle inequality and the $2/\varepsilon$-Lipschitz bound for $u \mapsto u/(\|u\| + \varepsilon)$ imply that the reparameterized direction in the main text satisfies

$$\begin{aligned} \|\overline{D}(t,x)\| &\leq C_V, \\ \|\overline{D}(t,x) - \overline{D}(t,y)\| &\leq \left(L_V + \frac{2C_V L_D}{\varepsilon}\right) \|x - y\|. \end{aligned} \tag{B.4}$$

Thus, for a fixed continuous bounded schedule s, the field $s\overline{D}$ satisfies the same types of boundedness and state-Lipschitz conditions, and the preceding comparison applies to $\gamma(\eta)$. The same integral estimate also holds for bounded piecewise-continuous schedules, provided the corresponding solutions exist.

B.2 Proof of Theorem 3.2

Let $w = z'(0^+)$. Applying Taylor's formula with an integral remainder componentwise to z (Knapp, 2016, Theorem 1.36) gives

$$\Delta z(\eta) = \eta w + R(\eta), \qquad R(\eta) = \int_0^\eta (\eta - u) z''(u)\,\mathrm{d}u, \qquad \|R(\eta)\| \le \frac{M}{2}\eta^2. \tag{B.5}$$

Using $\|e_{\mathrm{fix}}\| = 1$, $\|P_{\mathrm{keep}}\| \le 1$, and the triangle inequality,

$$\begin{aligned} \Delta_q(\eta) &= \langle e_{\mathrm{fix}}, \eta w + R(\eta)\rangle - \beta\|P_{\mathrm{keep}}(\eta w + R(\eta))\| \\ &\ge \eta\big(\langle e_{\mathrm{fix}}, w\rangle - \beta\|P_{\mathrm{keep}} w\|\big) - (1+\beta)\|R(\eta)\| \\ &\ge \kappa_q(D_F)\eta - \frac{(1+\beta)M}{2}\eta^2. \end{aligned} \tag{B.6}$$

When $\kappa_q(D_F) > 0$ and $M > 0$, the right-hand side is strictly positive for

$$0 < \eta < \min\left\{r, \frac{2\kappa_q(D_F)}{(1+\beta)M}\right\} \tag{B.7}$$

When $M = 0$, we have $R(\eta) = 0$, so $\Delta_q(\eta) = \kappa_q(D_F)\eta > 0$ for every $0 < \eta \le r$. This completes the proof.

B.3 Dynamical Expression of the First-Order Response

To expand $z'(0^+)$ in the main text, additionally assume that V_{base} and $s\overline{D}_F$ are continuously differentiable in the state, with state derivatives bounded in a neighborhood of the relevant trajectories, and that Φ_q is differentiable at $x^\star$. The fields and their state derivatives are continuous in time. Under these conditions, differentiable dependence on parameters (Teschl, 2012, Theorem 9.16) justifies differentiating the integral equation with respect to η. Let x_t^0 denote the reference trajectory and define

$$p_t = \partial_\eta x_t^\eta|_{\eta=0^+}, \qquad A(t) = \mathrm{D}_x V_{\mathrm{base}}(t, x_t^0). \tag{B.8}$$

Differentiating the sampling dynamics with respect to η and evaluating at $\eta = 0$ yields

$$\frac{\mathrm{d}p_t}{\mathrm{d}t} = A(t)p_t + s(t)\overline{D}_F(t, x_t^0), \qquad p_T = 0. \tag{B.9}$$

The condition $p_T = 0$ follows from using the same initial noise at every strength. Let the state-transition matrix $\Psi(t,u)$ satisfy $\partial_t\Psi(t,u) = A(t)\Psi(t,u)$ and $\Psi(u,u) = I_d$. The variation-of-constants formula (Teschl, 2012, Theorem 3.12), followed by the chain rule, yields

$$\begin{aligned} p_0 &= \int_T^0 \Psi(0,t)s(t)\overline{D}_F(t, x_t^0)\,\mathrm{d}t, \\ z'(0^+) &= \mathrm{D}\Phi_q(x^\star)\int_T^0 \Psi(0,t)s(t)\overline{D}_F(t, x_t^0)\,\mathrm{d}t. \end{aligned} \tag{B.10}$$

This expression propagates each step's correction through the subsequent dynamics to the generated endpoint and then maps it into the task representation. The integration direction agrees with the $T \to 0$ convention in the main text. The additional smoothness assumptions in this subsection are used for this dynamical expansion, not for the preceding proof stated directly in terms of z.

C Experimental Details

C.1 Evaluation Set and Reference Construction

We construct the follow-up rectification evaluation set from PRISM-Bench and SciIR-Bench. To describe its content coverage, we organize tasks into three groups: Human, covering human-centered images and requirements concerning pose, hand gestures, and interactions; Open, covering general scenes and requirements concerning objects, attributes, spatial relations, and environmental conditions; and Scientific & Text, combining scientific diagrams with text-containing images, including

scientific structures and relations, diagram labels, and specified rendered text. The combined Scientific & Text group follows the category-wise reporting in Table 2. PRISM checklists are constructed from the original prompts, whereas SciIR items are organized from the official entries. Tasks and checklists are fixed before method comparison. For each text-to-image model, up to 20 candidates are generated from different noise initializations under the original prompt; the highest-scoring image is the Best@20 reference. Qwen-Image-Edit-2511 uses an SD3.5 Medium output as the fixed source image and selects a Best@5 reference from five preceding editing candidates. Subsequent methods share the corresponding reference, and the original source image remains fixed throughout editing.

C.2 Model Configuration and Correction-Field Implementation

Table S1 lists the sampling configurations. Inference uses BF16 precision and a resolution of 1024×1024. The three text-to-image models use their native `guidance_scale`, whereas Qwen-Image-Edit-2511 uses `true_cfg_scale`. RGDR uses $s(t_j) = j/(N-1)$ for SD3.5 Medium and $s(t_j) = 1$ for the other three models.

Table S1: Sampling and guidance configurations. $j = 0, \ldots, N-1$ indexes the executed sampling steps.

Model	Steps N	Guidance scale	RGDR schedule $s(t_j)$
SD3.5 Medium	40	4.5	$j/39$
SANA-1.5 1.6B	20	4.5	1
FLUX.1-dev	50	3.5	1
Qwen-Image-Edit-2511	40	4.0	1

Each rectification attempt fixes the feedback, initial noise, and base generation configuration. Conditional prediction differences are recomputed at the current state and time of each candidate. Retaining the base field means retaining its prediction rule, rather than replaying cached velocities from the original trajectory.

C.3 Main-Experiment Comparison Protocol

All methods share the reference selected during preceding optimization, and the final output is selected from that reference and the additional candidates. The additional budget per method is ten complete images for each text-to-image model and five for the editing model, as in Table 1 of the main text. Editing retains the original source image. These image budgets are distinct from the number of sampling steps per image.

BoN. Following BoN candidate generation and selection(Ma et al., 2025), we retain the original prompt or editing instruction, generate candidates from different initial noise, and select by checklist score.

TIR. The prompt or editing instruction is revised iteratively using the current image and its evaluation feedback(Khan et al., 2025). The reference seed remains fixed. Each round produces a complete candidate under the revised condition, which informs the next reflection.

Dynamic CFG. We use a constant-schedule adaptation within the guidance-schedule space of Dynamic CFG(Papalampidi et al., 2025), imposing the same guidance scale at every step of a trajectory, $w_j \equiv w$. With the prompt and reference noise fixed, complete candidates are generated for different values of w and then evaluated for selection. SD3.5 Medium and SANA use ten uniformly spaced guidance values in $[2.5, 6.5]$; FLUX uses ten uniformly spaced native guidance-parameter values in $[2.1, 4.9]$.

RGDR. Target–Reference conditions and reference noise remain fixed, and complete candidates are generated independently at different relative strengths. The text-to-image models use $\mathcal{E} = \{0.02, 0.04, \ldots, 0.20\}$ and the schedules in Table S1. The additional correction field is recomputed at each sampling step.

For the editing branch, Dynamic CFG uses $w \in \{1.5, 2.5, 3.5, 5.0, 6.0\}$, and RGDR uses $\eta \in \{0.01, 0.025, 0.05, 0.075, 0.10\}$ with constant scheduling. Both sets contain five candidates, and each candidate is generated from the same immutable original input image.

C.4 Ablation Directions and VLM Prompts

Doubao Seed-2.1 Pro identifies residual errors from the task requirements and reference image, and expresses the diagnosis as Target, Reference, or Operation conditions. Within an RGDR rectification attempt, Target and Reference texts remain fixed while the conditional velocity predictions are recomputed at the current state.

Target. The Target describes the desired state after the specified error is corrected. It identifies the required attributes, objects, or relations, retaining the context needed to locate the change without restating unrelated requirements already satisfied.

Reference. The Reference describes the current erroneous state associated with the same correction target. It contrasts with the Target around the same objects, attributes, or relations. The two state descriptions separately condition the frozen model; the difference between their conditional velocity predictions defines the TR Correct direction.

Operation. The same correction request is expressed as a direct operation-style instruction, including the content to preserve. This ablation directly uses the model prediction conditioned on that instruction as the additional direction, without a Target–Reference prediction difference.

TR Wrong retains the Target–Reference difference formulation but uses mismatched feedback; Gaussian uses a random direction without task semantics. All directions use relative reparameterization. The archived SD3.5 direction experiment uses $\{0.02, 0.04, \ldots, 0.20\}$ for TR Correct, TR Wrong, and Gaussian, and $\{0.0075, 0.015, \ldots, 0.075\}$ for Operation.

C.5 Evaluation, Selection, and Aggregation

Doubao Seed-2.1 Pro provides feedback and evaluates candidates against fixed checklists; final outputs are selected under the same evaluation protocol. Let task i have K_i checklist items, and let $a_{ik}(I) \in [0, 1]$ be the numerical score of image I on item k. The task score is

$$S_i(I) = \frac{100}{K_i} \sum_{k=1}^{K_i} a_{ik}(I). \tag{C.1}$$

Let b_i, f_i denote the scores of the shared reference and selected output, respectively, and let n be the number of tasks in the aggregate. We report

$$\begin{aligned} \text{Final} &= \frac{1}{n} \sum_i f_i, & \text{Gain} &= \frac{1}{n} \sum_i (f_i - b_i), \\ \text{Rel. gain} &= 100 \frac{\text{Gain}}{\frac{1}{n} \sum_i b_i}, & \text{Improved} &= \frac{100}{n} \sum_i \mathbf{1}\{f_i > b_i + \delta\}. \end{aligned} \tag{C.2}$$

The recorded protocols for the three generative models use an improvement tolerance of 10^{-10} on the $[0, 1]$ scale, equivalent to $\delta = 10^{-8}$ on the 0–100 scale. Rel. gain is the ratio of mean gain to mean reference score, not the mean of per-task relative percentages.

Repair and Preserve use paired pass/fail judgments. For task i, let $N_{ab}^{(i)}$ count items with reference judgment a and output judgment b, where 0 denotes fail and 1 denotes pass. Items marked uncertain at either endpoint are excluded. Define $\mathcal{I}_R = \{i : N_{00}^{(i)} + N_{01}^{(i)} > 0\}$ and $\mathcal{I}_P = \{i : N_{10}^{(i)} + N_{11}^{(i)} > 0\}$. Then

$$\begin{aligned} \text{Repair} &= \frac{100}{|\mathcal{I}_R|} \sum_{i \in \mathcal{I}_R} \frac{N_{01}^{(i)}}{N_{00}^{(i)} + N_{01}^{(i)}}, \\ \text{Preserve} &= \frac{100}{|\mathcal{I}_P|} \sum_{i \in \mathcal{I}_P} \frac{N_{11}^{(i)}}{N_{10}^{(i)} + N_{11}^{(i)}}. \end{aligned} \tag{C.3}$$

Each metric is macro-averaged over tasks with a nonzero denominator. Numerical scores and binary judgments are recorded separately. An uncertain item can contribute its numerical score to S_i, but not to Repair or Preserve. Results are pooled across sources using the relevant valid-task sets; cross-model comparisons assign equal weight to each model.

D Temporal Scheduling Experiments

D.1 Schedule Functions and Comparison Setting

We compare $s(t_j) = 1$ and $s(t_j) = j/(N-1)$ on the three generative models. Both configurations share the Best@20 reference, initial noise, and Target–Reference conditions. Complete candidates are generated at $\mathcal{E} = \{0.02, 0.04, \ldots, 0.20\}$ and evaluated together with the reference for selection. Full-score references are retained directly with zero additional gain.

Let N be the number of sampling steps, $j = 0, \ldots, N-1$ the execution-step index, and t_j the corresponding sampling time. For a relative strength η, the applied coefficient is

$$\eta_j = \eta s(t_j), \qquad s^{(1)}(t_j) = 1, \qquad s^{(2)}(t_j) = \frac{j}{N-1}. \tag{D.1}$$

The second configuration therefore uses $j/39$, $j/19$, and $j/49$ for SD3.5 Medium, SANA, and FLUX, respectively. Schedule values are computed from the execution-step index without assuming uniformly spaced native scheduler timestamps. Both configurations use the same grid of peak strengths; their step-averaged schedule values are 1 and $1/2$, respectively.

D.2 Results and Analysis

Table S2 reports results pooled across the two task sources. SD3.5 Medium achieves a Gain of 3.74 points with $s(t_j) = j/(N-1)$, exceeding $s(t_j) = 1$ by 0.78 points. SANA and FLUX instead obtain higher Gain with $s(t_j) = 1$, by 1.36 and 1.73 points, respectively. The main experiments therefore use $s(t_j) = j/(N-1)$ for SD3.5 Medium and $s(t_j) = 1$ for SANA and FLUX.

Table S2: Pooled temporal-scheduling results. Column headings specify $s(t_j)$; bold indicates the better value for each model.

		Final		Gain	
Model	Reference	1	$j/(N-1)$	1	$j/(N-1)$
SD3.5 Medium	56.90	59.86	**60.64**	2.96	**3.74**
SANA-1.5 1.6B	46.55	**50.92**	49.56	**4.37**	3.01
FLUX.1-dev	70.30	**74.67**	72.94	**4.36**	2.64

Table S3 shows that the same schedule ranking holds separately on PRISM and SciIR. All three models obtain positive follow-up gains under both configurations, but the better choice of $s(t)$ is model-dependent. These results support configuring the temporal distribution of correction for each model.

Table S3: Temporal-scheduling results by task source. Column headings specify $s(t_j)$; bold indicates the better value for each model and source.

		Final		Gain	
Model	Source	1	$j/(N-1)$	1	$j/(N-1)$
SD3.5 Medium	PRISM	81.39	**82.51**	2.90	**4.01**
	SciIR	38.33	**38.78**	3.03	**3.48**
SANA-1.5 1.6B	PRISM	**75.25**	73.79	**4.41**	2.94
	SciIR	**26.59**	25.34	**4.33**	3.08
FLUX.1-dev	PRISM	**88.54**	87.56	**3.50**	2.53
	SciIR	**60.79**	58.31	**5.23**	2.74