

From Open Loop to Closed Loop: A Test-Time Iterative Optimization Framework for Reference-Consistent Image Generation

Baixuan Zhao¹, Xinyu Zhang³, Huayu Zheng¹, Shuaicheng Liu³,
Xionghuo Min¹, Guangtao Zhai¹, and Xiaohong Liu^{1,2*}

¹ Shanghai Jiao Tong University, Shanghai, China

² Shanghai Innovation Institute, Shanghai, China

³ University of Electronic Science and Technology of China, Chengdu, Sichuan, China

Abstract. While controllable image generation has made significant strides by incorporating visual reference conditions, existing methods predominantly operate as open-loop systems. They inject control signals in a strictly feed-forward manner, failing to guarantee strict fidelity to the reference due to the absence of active feedback and error correction mechanisms. To address this fundamental limitation, we propose a novel test-time iterative optimization framework that reformulates reference-consistent generation as a closed-loop dynamic tracking problem. By treating the pre-trained generative model as a control plant, our framework employs a sensor-controller architecture driven by a modified Proportional-Integral-Derivative (PID) algorithm. This mechanism iteratively optimizes the latent control signals at test time based on the sensed discrepancy between the generated output and the reference target. Notably, this approach is entirely training-free, model-agnostic, and integrates seamlessly around existing diffusion pipelines. Extensive evaluations across ID-preserving, pose-controlled, and depth-controlled generation tasks validate the universality of our method. Empirical results demonstrate improvements over computation-matched open-loop baselines, achieving relative performance gains of up to 25.36% for facial similarity, alongside spatial error reductions of up to 27.71% for pose alignment and 28.50% for depth consistency. More broadly, this work offers a new conceptual perspective: it demonstrates that controllable generation can be effectively managed as a dynamic feedback system, bringing the rigorous principles of classical control theory into the optimization of generative models. Code is available at <https://github.com/zzdrill/From-Open-Loop-to-Closed-Loop>.

1 Introduction

Generative AI has reshaped digital content creation [21, 24, 37], with text-to-image diffusion models [4, 26, 29] enabling high-fidelity image synthesis from

* Corresponding author. Email: xiaohongliu@sjtu.edu.cn

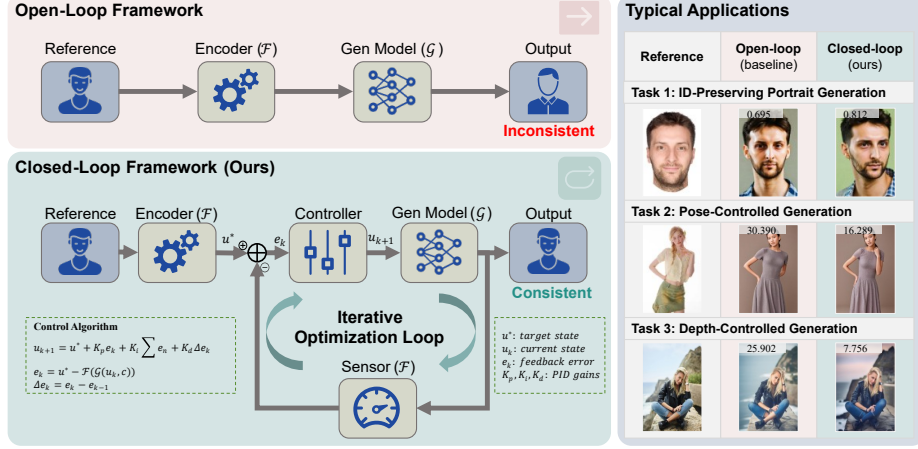


Fig. 1: Overview of the proposed closed-loop optimization framework. We reformulate reference-consistent generation as feedback control over a generative model. The encoder and sensor are the same pre-trained module $\mathcal{F}$, reused in different roles for target extraction and feedback measurement. A modified PID controller iteratively updates the latent control input u_k to reduce the generation error. This closed-loop optimization improves reference consistency over open-loop baselines across ID-, pose-, and depth-controlled generation.

textual descriptions. To reduce the ambiguity of text-only prompts, recent controllable generation methods introduce visual references [47], enabling applications such as ID-preserving generation [20, 32, 41], pose/depth-controlled generation [22, 44, 46], image restoration [13, 33], aesthetic QR codes [10, 34, 35], and virtual try-on [9, 11].

Despite these advances, many real-world applications require high consistency between generated outputs and reference conditions. In ID-preserving portrait generation, even subtle feature drift may change the perceived identity. In pose- or depth-controlled generation, small spatial deviations can lead to visible artifacts such as shifted limbs or distorted geometry. Existing controllable generation methods usually inject reference conditions into the generative model in a feed-forward manner. Once the condition is encoded, the model lacks an explicit mechanism to observe the generated result and correct deviations from the reference. This motivates a control-theoretic question: **Can reference-consistent generation be formulated as a closed-loop dynamic tracking problem?**

Under this formulation, current reference-consistent methods can be viewed as open-loop systems, where reference embeddings or spatial conditions are passed to the generative model without feedback verification. To address this limitation, we propose a training-free test-time optimization framework for reference-consistent image generation, as illustrated in Figure 1. Our framework reuses an off-the-shelf pre-trained module in two roles: as an encoder to extract the target state from the reference, and as a sensor to measure the generated state from the current output. A modified PID controller then updates the latent control input

according to the sensed tracking error, forming an explicit generation-and-check loop. The framework is model-agnostic and can be plugged into existing diffusion pipelines without modifying their weights.

We evaluate the proposed framework on three representative tasks: (1) ID-preserving portrait generation, (2) pose-controlled generation, and (3) depth-controlled generation. The main contributions of this work are summarized as follows:

-Closed-loop formulation. We formulate reference-consistent generation as a closed-loop dynamic tracking problem, providing a control-theoretic perspective for improving alignment between generated outputs and visual references.

-Training-free test-time framework. We introduce a lightweight plug-and-play optimization framework driven by a modified PID controller, which iteratively updates latent control inputs based on sensed tracking errors without additional training or model weight modification.

-Broad applicability and empirical gains. Extensive experiments demonstrate consistent improvements across ID-preserving, pose-controlled, and depth-controlled generation. Under computation-matched comparison with best-of- N open-loop baselines, our framework achieves relative gains of up to 25.36% in ID preservation and spatial error reductions of up to 27.71% and 28.50% for pose and depth control, respectively.

2 Related Works

Controllable Image Generation. Diffusion models have achieved remarkable progress in image synthesis, motivating a wide range of controllable generation methods [4, 17, 18, 29]. Adapter-based approaches, such as ControlNet [44] and IP-Adapter [41], introduce additional conditioning branches for flexible spatial or reference guidance [45]. Recent methods further improve parameter efficiency and multi-modal control [22, 31, 37], while ControlNet++ [19] enhances reference consistency through training-time consistency losses. Despite these advances, most methods remain open-loop at inference time: once the condition is encoded, the model does not actively observe and correct deviations from the reference. Our work addresses this limitation by introducing a closed-loop test-time optimization framework for reference-consistent generation.

Test-Time Optimization and Refinement. Cascaded architectures [23] and post-processing refiners [16, 33] are commonly used to improve resolution and visual fidelity. Prompt optimization methods [38] iteratively refine textual inputs for better semantic coverage. Training-free guidance methods such as UGD [2], FreeDoM [42], and DAS [15, 39] also perform test-time optimization by injecting task-specific guidance or constraint signals into the sampling process. However, these methods are mainly formulated as sampling-time guidance, where optimization is coupled with the denoising trajectory or task-specific energy functions. In contrast, our method targets reference-consistent alignment through an explicit feedback loop. A pre-trained sensor measures the tracking error between

the reference and generated states, and a modified PID controller iteratively updates the latent control input, enabling generation-and-check refinement without additional training or independent post-processing.

Control-Theoretic Interpretation of Diffusion. Another line of work interprets diffusion sampling through the lens of stochastic control [28,30], where the reverse denoising process can be viewed as an implicit regulation mechanism that preserves distributional fidelity. Recent analyses also relate controllable adapters to delayed or low-frequency feedback systems [43]. These studies mainly explain the internal dynamics of diffusion models or existing adapters, but they do not provide an external task-level feedback mechanism for correcting reference-specific semantic drift. Our framework differs by placing a sensor-controller loop outside the generative model, turning reference-consistent generation into an explicit dynamic tracking problem.

3 Preliminaries

3.1 Closed-loop Control System

A closed-loop control system is an automated regulation architecture that continuously adjusts its behavior through active feedback. It typically consists of four essential components: (1) a *target state* that defines the desired condition; (2) a *controller* that calculates actions based on the *tracking error* (the discrepancy between the target and actual state); (3) a *controlled plant* representing the system requiring regulation (e.g., a generative model); and (4) a *sensor* that monitors the actual state of the output to provide continuous feedback signals. Compared to open-loop counterparts, closed-loop architectures demonstrate superior robustness, higher precision, and effective rejection of external disturbances.

3.2 PID Control Algorithm

Generative models often exhibit a black-box nature with extremely high-dimensional complexity, making it intractable to derive explicit transfer functions or state-space equations required by modern optimal control methods like LQR [1] or MPC [6]. In contrast, the Proportional-Integral-Derivative (PID) [5] algorithm is a fundamentally model-free controller that requires no prior knowledge of the plant dynamics, making it an ideal candidate for iterative generative refinement.

PID achieves precise regulation through three synergistic terms: the *proportional* (P) term provides an instantaneous response to the current error, the *integral* (I) term accumulates past errors to eliminate steady-state offsets, and the *derivative* (D) term predicts future trends to suppress overshoot. For discrete-time systems, the standard positional PID control signal is expressed as:

$$u_{k+1} = K_p e_k + K_i T_s \sum_{n=0}^k e_n + K_d \frac{e_k - e_{k-1}}{T_s}, \quad (1)$$

where k denotes the discrete step index; u_{k+1} is the control signal computed for the $(k+1)$ -th step; e_k represents the tracking error between the target state and the system output at step k ; and K_p , K_i , and K_d correspond to the proportional, integral, and derivative gains, respectively. The variable T_s denotes the sampling period.

4 Method

4.1 Framework Overview

Following the principles of control theory, we formulate the reference-consistent generation task as a discrete-time dynamic tracking problem. Our closed-loop framework (illustrated in Figure 1) comprises five key components: *reference*, *encoder*, *controller*, *controlled plant*, and *sensor*.

In this architecture, the reference image I^* defines the target state $u^* = \mathcal{F}(I^*)$. A shared feature extractor $\mathcal{F}$ is employed with dual functionality: as the *encoder*, it maps the reference image to the target state; as the *sensor*, it monitors the generated image I_k at step k to provide feedback on the actual state directly via $\mathcal{F}(I_k)$. The *controller* quantifies the tracking error $e_k = u^* - \mathcal{F}(I_k)$ and computes the updated *control input* u_{k+1} via the control algorithm $\mathcal{C}$. The generative model, integrated with a controllable adapter (denoted as $\mathcal{G}$), acts as the *controlled plant* to synthesize images based on the latent control input u and text prompt c .

As shown in Equation (2), the system operates through successive generation-and-check cycles. In each iteration k , the actual state is observed directly from the generated image as $\mathcal{F}(I_k)$. The controller then updates the control input u_{k+1} to compensate for the sensed discrepancy. Specifically, given the target state $u^* = \mathcal{F}(I^*)$, the iterative optimization loop is formally defined as:

$$\begin{cases} I_k = \mathcal{G}(u_k, c), \\ e_k = u^* - \mathcal{F}(I_k), \\ u_{k+1} = \mathcal{C}(u^*, e_k). \end{cases} \quad (2)$$

The subsequent output $I_{k+1} = \mathcal{G}(u_{k+1}, c)$ thus aligns progressively closer with the reference.

Discussion on Inference Overhead. It is worth noting that each iteration k in Equation (2) executes a complete diffusion denoising process. While this inherently increases the overall inference time compared to open-loop baselines, it trades a manageable latency for a paramount engineering advantage: *zero intrusion*. By treating the base model as a black box, our framework requires no modifications to the original generation pipeline. Implementing this lightweight wrapper merely involves adding a few lines of code to establish the external feedback loop.

4.2 Controller Formulation

Classical dynamic systems are characterized by their output states being dependent on both instantaneous inputs and cumulative historical states. Conventional PID algorithms are specifically designed for such dynamics, regulating the system by calculating control signals solely based on the tracking error. In contrast, image generation models behave as static (memoryless) systems: the output depends exclusively on the current control input u_k and prompt c , with no inherent mechanism to retain internal states across inference iterations.

Consequently, a traditional PID controller—where the control signal diminishes as the error approaches zero—is fundamentally unsuitable for generative tasks, as the model requires a persistent baseline signal to maintain the presence of the target subject. To address this mismatch, we propose a *Modified PID Algorithm* that introduces the target state as a persistent bias term at each iteration. Given that the condition input space and the sensory feedback space are aligned in our tasks, we directly utilize the target state $u^* = \mathcal{F}(I^*)$ as the persistent base control input.

By reformulating the control task as a residual adjustment relative to u^* , the system is naturally initialized by setting $u_1 = u^*$. This initial pass generates the base image I_1 , yielding the first tracking error e_1 . With the sampling period normalized to unit steps, the modified control law for all $k \geq 1$ is formulated as:

$$\begin{cases} e_k = u^* - \mathcal{F}(\mathcal{G}(u_k, c)), \\ u_{k+1} = u^* + K_p e_k + K_i \sum_{n=1}^k e_n + K_d (e_k - e_{k-1}), \end{cases} \quad (3)$$

where $e_0 \equiv 0$ is defined for the initial derivative calculation. In this formulation, the PID terms act as a residual compensator, refining the base signal u^* through historical feedback to keep u_{k+1} grounded in the target condition while driving the output toward asymptotic convergence.

4.3 Task Formulations

Our closed-loop framework serves as a universal wrapper that can be instantiated across various controllable generation tasks. The versatility of the framework relies on defining appropriate state spaces and sensors for specific conditions.

ID-preserving portrait generation. In this task, the control objective is to maintain high fidelity to the subject’s identity. We formulate the state space in the latent facial embedding domain, $\mathbb{R}^d$. The sensor $\mathcal{F}_{id}$ is defined as a facial recognition encoder. Given the target reference embedding $u_{id}^* = \mathcal{F}_{id}(I^*)$, the tracking error at step k is quantified as the element-wise difference: $e_k = u_{id}^* - \mathcal{F}_{id}(I_k) \in \mathbb{R}^d$. This error vector drives the controller to iteratively optimize the identity-conditioned base signal along each feature dimension.

Pose-controlled generation. To achieve precise structural alignment, the state space is defined in the 2D spatial coordinate domain, $\mathbb{R}^{H \times W \times 2}$. The sensor $\mathcal{F}_{pose}$

functions as a pose estimator, extracting structural keypoint representations. Given the target reference pose map $u_{pose}^* = \mathcal{F}_{pose}(I^*)$, the tracking error is formulated as the spatial deviation: $e_k = u_{pose}^* - \mathcal{F}_{pose}(I_k) \in \mathbb{R}^{H \times W \times 2}$. The resulting control signals are subsequently added as spatial residuals to update the explicit pose guidance map in each iteration.

Depth-controlled generation. For depth-constrained tasks, the state space operates in the dense pixel-wise geometric domain, $\mathbb{R}^{H \times W}$. The sensor $\mathcal{F}_{depth}$ acts as a monocular depth estimator. Given the target reference depth map $u_{depth}^* = \mathcal{F}_{depth}(I^*)$, the tracking error evaluates the pixel-level deviations against the estimated depth of the current output: $e_k = u_{depth}^* - \mathcal{F}_{depth}(I_k) \in \mathbb{R}^{H \times W}$. The controller iteratively compensates for these geometric deviations, ensuring the output structure progressively converges to the target spatial configuration.

5 Experiments

5.1 Experimental Setup

Implementation Details. All experiments are conducted on an NVIDIA GeForce RTX 4090 GPU. As shown in Table 1, PID coefficients, resolutions, and iteration numbers are task-specific: ID-preserving uses a fixed 1024^2 resolution for 20 iterations, while pose/depth control uses adaptive aspect-ratio scaling with the shorter edge set to 1024 pixels for 15 iterations.

Table 1: Parameter configurations.

Task	K_p	K_i	K_d	Res.	Iter.
ID	0.30	0.05	0.01	1024^2	20
Pose	0.05	0.02	0.01	Adaptive	15
Depth	0.01	0.005	0.005	Adaptive	15

Datasets. For ID-preserving generation, we evaluate on: (1) *Web100*, a multi-reference set containing 130 non-celebrity identities with 2–5 photos each; (2) *CelebA300*, comprising 300 random identities from CelebA-HQ [14]. For pose and depth control, we utilize a filtered subset of 477 single-person photos from a public repository¹.

Metrics. We evaluate our framework across three primary dimensions. To avoid sensor-evaluator coupling, we use different models for feedback and evaluation: InsightFace² ResNet100@Glint360K / ResNet50@WebFace600K for ID, OpenPose / DWPose for pose [7, 40], and MiDaS / ZoeDepth for depth [3, 27], respectively. Specifically, the first model in each pair is used as the feedback sensor, while the second one is used only for evaluation.

Consistency. We measure the alignment between the generated output and the target reference. For ID-preserving tasks, we compute Facial Similarity via the cosine similarity between identity embeddings:

$$S_{face} = \cos(\mathbf{v}^*, \mathbf{v}^{gen}), \quad (4)$$

¹ https://huggingface.co/datasets/raulc0399/open_pose_controlnet, accessed on March 5, 2026.

² <https://github.com/deepinsight/insightface>, accessed on March 5, 2026.

where $\mathbf{v}$ denotes the identity embedding vector.

For the pose-controlled generation task, the consistency metric is evaluated using the Mean Per Joint Position Error (MPJPE). We formulate this as the average L_2 distance between corresponding structural keypoints:

$$E_{pose} = \frac{1}{M} \sum_{m=1}^M \|\mathbf{p}_m^* - \mathbf{p}_m^{gen}\|_2, \quad (5)$$

where M represents the total number of matched keypoints, and $\mathbf{p}_m^*, \mathbf{p}_m^{gen} \in \mathbb{R}^2$ denote the 2D spatial coordinates of the m -th joint in the target and generated pose maps, respectively.

For the depth-controlled generation task, consistency is assessed via the Mean Absolute Error (MAE) of corresponding pixels in the depth maps:

$$E_{depth} = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W |d_{i,j}^* - d_{i,j}^{gen}|, \quad (6)$$

where H and W represent the height and width of the image, while d denotes the estimated depth value at a specific pixel coordinate.

Diversity. To quantify facial variation and address the generic “pose-copying” issue in ID tasks, we propose Structure Diversity (D_{sd}). This metric measures the structural deviation of the generated image from the reference set. Specifically, we first apply coordinate normalization $\mathcal{N}(\cdot)$ to eliminate scale and translation differences. We then calculate the mean Euclidean distance over all K facial landmarks, averaged across the N reference images:

$$D_{sd} = \frac{1}{N} \sum_{n=1}^N \left(\frac{1}{K} \sum_{j=1}^K \|\mathcal{N}(\mathbf{p}_j^{gen}) - \mathcal{N}(\mathbf{p}_{j,n}^{ref})\|_2 \right). \quad (7)$$

General Quality. Semantic similarity is evaluated via CLIP-I [25] and DINO [8], while perceptual generation quality is assessed via Q-Align [36].

5.2 Qualitative Evaluation

ID-Preserving Generation. To evaluate the effectiveness of our approach, we integrate the proposed closed-loop optimization with four state-of-the-art ID-preserving generation baselines: PuLID [12], PhotoMaker v2 (PMv2) [20], InstantID [32], and IP-Adapter-FaceID-Portrait (IPA)³. As illustrated in Fig. 2, the visual comparisons indicate that the closed-loop integration (+CL.) consistently improves identity fidelity across diverse base architectures. Supported by the facial similarity scores, our method facilitates a closer alignment with the reference subjects. Furthermore, as highlighted by the local zoom-in patches, the optimization process recovers structural proportions and better preserves fine-grained facial details (e.g., eye characteristics and subtle skin textures) that are

³ <https://huggingface.co/h94/IP-Adapter-FaceID>, accessed on March 5, 2026.

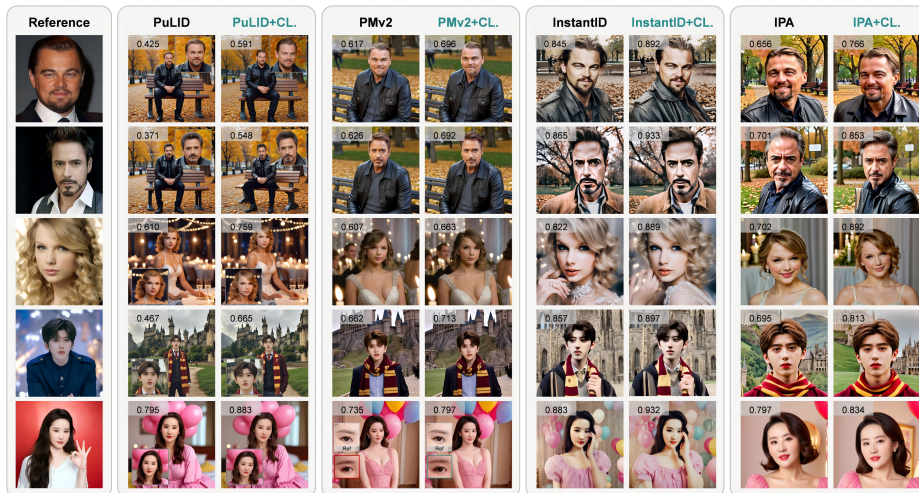


Fig. 2: Qualitative comparisons of ID-preserving portrait generation. We integrate our closed-loop optimization (+CL.) into four representative baselines. The results demonstrate noticeable improvements in both identity fidelity and fine-grained detail preservation compared to the standard open-loop generations.

occasionally smoothed out by standard open-loop methods. Overall, by establishing an effective closed-loop system, our approach significantly enhances identity consistency without compromising the underlying photorealistic generation quality, proving its plug-and-play adaptability across fundamentally different adapter designs.

We further evaluate the robustness of our framework through a *recurrent generation experiment*, where the output of the current iteration serves as the reference image for the subsequent generation. While standard feed-forward methods typically suffer from rapid identity degradation due to error accumulation across iterations, our closed-loop approach effectively mitigates this issue. As shown in Fig. 3, the feedback mechanism maintains high facial similarity across multiple rounds, demonstrating enhanced stability against identity drift.

Spatial Control Tasks. Beyond identity preservation, we evaluate the applicability of our closed-loop system on spatial control tasks, specifically pose- and depth-guided generation. As illustrated in Figures 4 and 5, standard open-loop baselines can exhibit structural artifacts under strict geometric constraints, such as misplaced limb joints or blurred depth boundaries. By integrating the iterative feedback mechanism, the system monitors and reduces these spatial deviations. Furthermore, these visual improvements are quantitatively supported by the reductions in spatial error metrics (i.e., MPJPE for pose and MAE for depth) reported below each image pair, indicating the effectiveness of the proposed spatial tracking approach.

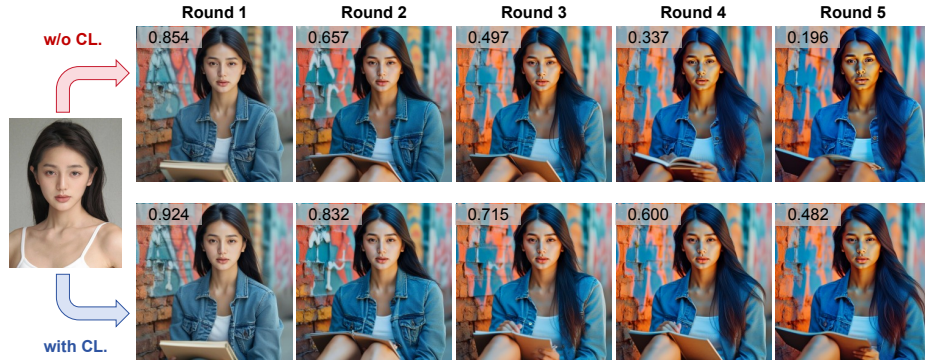


Fig. 3: Recurrent generation experiment. The output of the previous round serves as the reference for the next. Our closed-loop optimization effectively mitigates identity degradation across iterations. Facial similarity ($\uparrow$) are provided below each image.

5.3 Quantitative Evaluation

Optimization Dynamics. Figure 6 illustrates the convergence behavior of the proposed closed-loop system. Across all three tasks (identity, pose, and depth), the metrics exhibit a consistent two-phase trajectory: rapid improvement during early iterations (typically 1–5) to rectify discrepancies, followed by a stable fine-tuning phase. The system generally reaches a steady state within 20 iterations for identity-preserving tasks and 15 for spatial control tasks. These trends indicate that the iterative feedback effectively guides generation toward the target constraints without noticeable divergence.

ID-Preserving Generation. Table 2 reports the quantitative results for ID-preserving generation. Each method group compares the original open-loop baseline, a stronger best-of-20 open-loop variant, and our closed-loop optimization. Compared with the best-of-20 baselines, our method consistently improves facial similarity across all models and datasets. Specifically, PuLID obtains the largest gains, with relative improvements of 25.36% on CelebA300 and 12.98% on Web100. PMv2 also benefits substantially, improving by 14.70% and 14.38% on the two datasets, respectively. For stronger identity-preserving baselines, our framework still brings clear gains, improving InstantID by 7.47% / 5.52% and IPA by 7.62% / 6.87% on CelebA300 / Web100. Beyond identity fidelity, our method maintains competitive CLIP-I, DINO, Q-Align, and Structure Diversity scores compared with both the original and best-of-20 open-loop baselines. These results indicate that closed-loop feedback effectively enhances identity consistency without causing a notable degradation in semantic alignment, perceptual quality, or structural diversity.

Spatial Control Tasks. For spatial control tasks, closed-loop optimization consistently improves geometric precision over strong open-loop baselines. As shown in Table 3, compared with the best-of-15 variant, applying our method to ControlNet reduces pose error (MPJPE) by 27.71% and depth error (MAE)

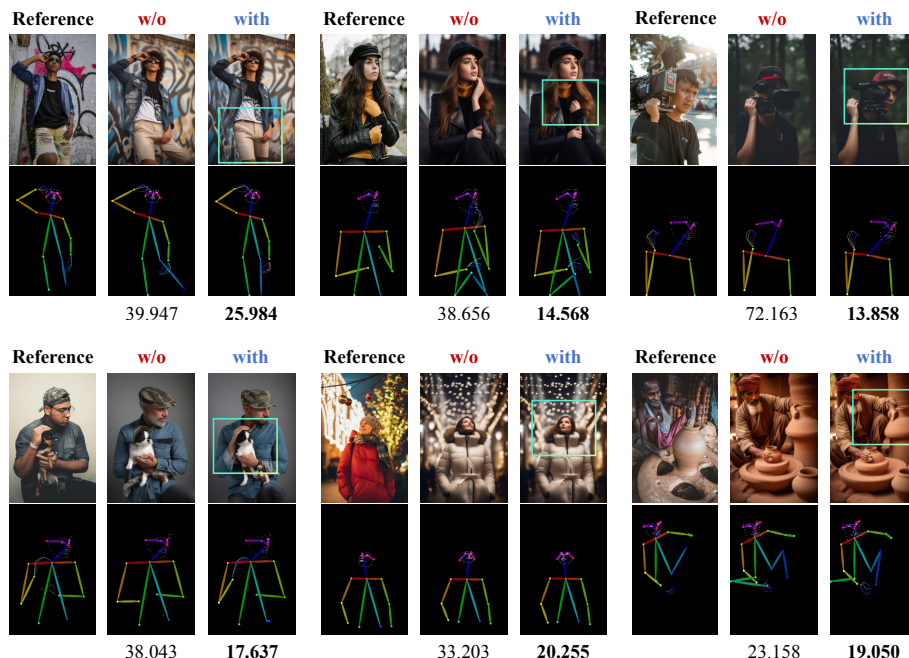


Fig. 4: Qualitative comparisons for pose-controlled generation. Compared to the standard open-loop baseline (**w/o**), our closed-loop optimization (**with**) drastically rectifies structural misalignments and complex articulations (highlighted in boxes). Extracted skeletal maps and the corresponding Mean Per Joint Position Error (MPJPE, $\downarrow$) are provided below each result to quantify the geometric improvements.

by 28.50%. The framework also generalizes to ControlNext, achieving error reductions of 13.46% for pose control and 14.42% for depth control. These results demonstrate that iterative feedback correction improves alignment with geometric control signals beyond repeated open-loop sampling. Meanwhile, DINO, CLIP-I, and Q-Align scores remain comparable in most cases, indicating that the enhanced spatial consistency does not come at the cost of semantic alignment or perceptual quality.

5.4 Ablation Study

Impact of PID Components. Table 4 ablates the P, I, and D terms using 50 random samples per task. The P term alone brings clear improvements over the open-loop baseline on the main alignment metrics. Adding the I term further improves final alignment, yielding competitive or second-best results across tasks. The D term mainly accelerates convergence, as the PD variant requires the fewest iterations to reach 95% of the final gain. Overall, the full PID controller achieves the best final performance, including the highest identity similarity, the lowest pose MPJPE, the lowest depth MAE, and the best CLIP-I scores for



Fig. 5: Qualitative comparisons for depth-controlled generation. The open-loop baseline (**w/o**) frequently struggles with ambiguous geometric boundaries. In contrast, our closed-loop system (**with**) ensures precise depth alignment and sharp foreground-background separation (highlighted by arrows/boxes). The Mean Absolute Error (MAE, $\downarrow$) evaluated on the extracted depth maps is reported below.

identity and depth. These results show that P, I, and D play complementary roles in error correction, steady-state refinement, and convergence acceleration.

Robustness to Prompt Complexity. Table 5 evaluates generation under four prompt levels: *no prompt*, *simple*, *moderate*, and *complex*. Our closed-loop framework (**+CL.**) consistently outperforms the open-loop baseline across all conditions. Notably, even without any prompt, our method achieves the highest identity similarity and significantly reduces spatial errors. This confirms that our iterative feedback effectively overrides textual ambiguities, strictly anchoring the generation to the visual reference.

Robustness to Random Seeds. To assess stability against the inherent stochasticity of diffusion models, we compute performance statistics across multiple random seeds (Table 6). Our closed-loop method consistently yields superior mean performance (μ) while drastically reducing standard deviation (σ). For example, the depth error variance is more than halved. This confirms that dynamic optimization effectively mitigates stochastic fluctuations, ensuring a highly stable generation process.

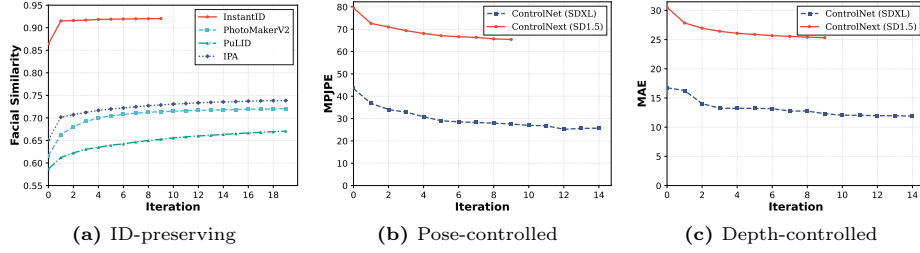


Fig. 6: Optimization dynamics across three generation tasks: (a) ID-preserving, (b) pose-controlled, and (c) depth-controlled generation. The plots illustrate the progression of Facial Similarity ($\uparrow$), MPJPE ($\downarrow$), and MAE ($\downarrow$) over iterations, demonstrating stable and monotonic convergence across diverse baselines.

Table 2: Quantitative comparison on ID-preserving generation. Each method group includes the original open-loop baseline, its best-of-20 variant, and our closed-loop optimization (+CL.). **Bold** indicates the best result within each method. The superscript gain reports the relative improvement of +CL. over the best-of-20 baseline in Facial Similarity. [Sim.: Facial Similarity, Q-Al.: Q-Align, Div.: Structure Diversity.]

Method	CelebA300					Web100				
	Sim. $\uparrow$	CLIP-I $\uparrow$	DINO $\uparrow$	Q-Al. $\uparrow$	Div. $\uparrow$	Sim. $\uparrow$	CLIP-I $\uparrow$	DINO $\uparrow$	Q-Al. $\uparrow$	Div. $\uparrow$
PuLID [12]	0.550	0.733	0.282	4.980	0.059	0.584	0.779	0.341	4.971	0.107
PuLID-best20 [12]	0.560	0.737	0.285	4.972	0.054	0.593	0.785	0.342	4.964	0.101
PuLID +CL.	0.702 ^{+25.36%}	0.743	0.287	4.975	0.047	0.670 ^{+12.98%}	0.794	0.344	4.966	0.105
PhotoMaker [20]	0.548	0.768	0.329	4.974	0.0738	0.589	0.823	0.412	4.928	0.131
PMv2-best20 [20]	0.551	0.770	0.332	4.969	0.070	0.598	0.824	0.410	4.933	0.128
PMv2 +CL.	0.632 ^{+14.70%}	0.771	0.330	4.963	0.0741	0.684 ^{+14.38%}	0.827	0.407	4.965	0.130
InstantID [32]	0.812	0.774	0.604	4.314	0.023	0.718	0.814	0.493	4.035	0.101
InstantID-best20 [32]	0.817	0.780	0.605	4.305	0.017	0.724	0.812	0.489	4.037	0.094
InstantID +CL.	0.878 ^{+7.47%}	0.786	0.601	4.313	0.022	0.764 ^{+5.52%}	0.818	0.491	4.433	0.100
IPA	0.708	0.756	0.529	4.946	0.114	0.672	0.798	0.492	4.973	0.149
IPA-best20	0.722	0.757	0.530	4.941	0.106	0.684	0.803	0.494	4.963	0.142
IPA +CL.	0.777 ^{+7.62%}	0.774	0.532	4.950	0.108	0.731 ^{+6.87%}	0.808	0.486	4.977	0.151

5.5 Coefficient Sensitivity Analysis

We further analyze the sensitivity of the PID coefficients by fixing the derivative gain K_d and performing a two-dimensional grid search over K_p and K_i . As shown in Fig. 7, the closed-loop results remain consistently better than the open-loop baseline across a broad coefficient range. For ID-preserving generation, most coefficient settings lead to higher facial similarity, suggesting that the feedback loop can robustly correct identity drift. For pose- and depth-controlled generation, the controller achieves lower MPJPE and MAE under most settings, indicating improved geometric alignment with the reference conditions. Although different tasks favor slightly different coefficient regions, the performance surfaces show stable effective areas rather than sharp isolated optima. This indicates that the proposed framework does not rely on a narrow hyperparameter choice, but instead exhibits stable effectiveness within a practical range of controller gains.

Table 3: Quantitative comparison on pose- and depth-controlled generation. Each method group includes the original open-loop baseline, its best-of-15 variant, and our closed-loop optimization (+CL.). **Blue bold** indicates the best result within each method group. The superscript reduction reports the relative error decrease of +CL. over the best-of-15 baseline for MPJPE and MAE. [Q-Al.: Q-Align.]

Method	Pose-Controlled Generation				Depth-Controlled Generation			
	MPJPE↓	DINO↑	CLIP-I↑	Q-Al.↑	MAE↓	DINO↑	CLIP-I↑	Q-Al.↑
ControlNet [44]	39.193	0.6297	0.7891	4.936	16.983	0.781	0.840	4.956
ControlNet-best15	36.521	0.626	0.788	4.928	15.797	0.775	0.838	4.954
ControlNet +CL.	26.401 ^{-27.71%}	0.6300	0.7887	4.940	11.295 ^{-28.50%}	0.785	0.843	4.963
ControlNext [22]	79.542	0.572	0.742	4.951	30.544	0.549	0.733	4.930
ControlNext-best15	75.827	0.569	0.738	4.945	29.598	0.543	0.733	4.928
ControlNext +CL.	65.623 ^{-13.46%}	0.575	0.739	4.952	25.331 ^{-14.42%}	0.548	0.734	4.924

Table 4: Ablation study on different PID component combinations. **Blue bold** and **bold** indicate the best and second-best results, respectively. 95%Gain denotes the average number of iterations required to reach 95% of the final performance gain.

Method	P	I	D	Identity			Pose			Depth		
				Sim.↑	CLIP-I↑	95%Gain↓	MPJPE↓	CLIP-I↑	95%Gain↓	MAE↓	CLIP-I↑	95%Gain↓
Open loop	×	×	×	0.629	0.739	-	108.867	0.762	-	30.612	0.816	-
Closed loop (ours)	✓	×	×	0.694	0.763	5.58	78.903	0.763	3.66	24.222	0.818	3.98
	✓	✓	×	0.723	0.7736	6.80	67.400	0.771	4.70	23.345	0.8189	4.44
	✓	×	✓	0.695	0.765	5.52	76.514	0.766	3.42	23.645	0.8194	3.88
	✓	✓	✓	0.724	0.7738	6.58	66.231	0.769	4.62	23.213	0.821	4.20

5.6 Analysis of Computational Efficiency

We analyze the computational overhead for ID-preserving and pose-controlled generation, with depth control showing a similar cost profile to pose control. For ID-preserving generation, each iteration takes 4.2 seconds on average with a peak GPU memory usage of 15,286 MB. For pose-controlled generation, each iteration takes 11.6 seconds with a peak memory usage of 14,148 MB. Since the per-iteration cost is comparable to standard single-round inference, the total latency scales approximately linearly with the maximum number of iterations.

Table 5: Ablation on text prompts. **Blue bold** indicates closed-loop improvements. **Table 6:** Stability across seeds. [μ : Mean, σ : Std.]

Prompt Level	Identity		Pose		Depth	
	Base	+CL.	Base	+CL.	Base	+CL.
No prompt	0.651	0.707	29.952	22.066	22.598	14.553
Simple	0.592	0.667	26.240	21.810	21.638	12.610
Moderate	0.596	0.659	27.064	22.491	22.970	13.269
Complex	0.593	0.666	26.677	21.324	21.902	12.576

Method	Identity		Pose		Depth	
	μ	σ	μ	σ	μ	σ
Base	0.658	0.021	23.618	9.050	20.808	5.348
+CL. (ours)	0.706	0.009	18.220	5.629	10.942	1.975

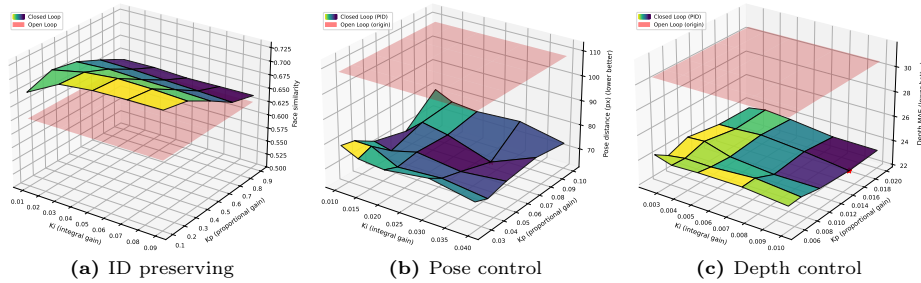


Fig. 7: Coefficient sensitivity analysis. With K_d fixed, we perform a 2D grid search over K_p and K_i . The closed-loop controller consistently improves ID similarity and reduces pose/depth errors over a broad coefficient range.

6 Conclusion

In this work, we propose a test-time iterative optimization framework that brings automatic control principles into controllable image generation. By introducing a closed-loop feedback mechanism during inference, our method uses a modified PID controller to iteratively update latent control signals and reduce reference deviations. This simple yet effective feedback design improves reference consistency in both identity fidelity and spatial structure, while requiring no model retraining. Extensive experiments across diverse tasks and backbone models demonstrate the effectiveness and plug-and-play versatility of the proposed framework. Beyond a specific optimization method, this work offers a new perspective: controllable generation can be viewed as a dynamic control problem, where classical control theory provides interpretable tools for improving generative models.

Limitations and Future Work:

-Manual Parameter Tuning: The current PID controller uses manually selected hyperparameters. Future work will explore adaptive control strategies to automate coefficient tuning.

-Inference Latency: The iterative nature of closed-loop optimization increases inference latency and computational cost. Future research will focus on accelerating convergence and reducing the required number of iterations.

-Dependence on Sensor Accuracy: The feedback loop relies on external sensors, such as identity encoders and pose/depth estimators. Future work will investigate robust control strategies to mitigate sensor noise and improve generation quality under imperfect feedback.

Acknowledgment

The work was supported in part by the National Natural Science Foundation of China under Grant 62301310, 62572317 and 62225112.

References

1. Åström, K.J., Murray, R.M.: Feedback systems: An introduction for scientists and engineers. Princeton university press (2010)
2. Bansal, A., Chu, H.M., Schwarzschild, A., Sengupta, R., Goldblum, M., Geiping, J., Goldstein, T.: Universal guidance for diffusion models. In: International Conference on Learning Representations. vol. 2024, pp. 51304–51323 (2024)
3. Bhat, S.F., Birkel, R., Wofk, D., Wonka, P., Müller, M.: ZoeDepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288 (2023)
4. Black Forest Labs: FLUX. <https://github.com/black-forest-labs/flux> (2024), accessed on March 5, 2026
5. Borase, R.P., Maghade, D., Sondkar, S., Pawar, S.: A review of PID control, tuning methods and applications. International Journal of Dynamics and Control **9**(2), 818–827 (2021)
6. Camacho, E.F., Alba, C.B.: Model predictive control. Springer science & business media (2013)
7. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(1), 172–186 (2019)
8. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660 (2021)
9. Choi, Y., Kwak, S., Lee, K., Choi, H., Shin, J.: Improving diffusion models for virtual try-on. arXiv e-prints pp. arXiv–2403 (2024)
10. Cui, X., Wu, G., Gan, Z., Zhai, G., Liu, X.: Face2QR: A unified framework for aesthetic, face-preserving, and scannable QR code generation. arXiv preprint arXiv:2411.19246 (2024)
11. Dam, P., Jeong, J., Tran, A., Kim, D.: Time-efficient and identity-consistent virtual try-on using a variant of altered diffusion models. In: European Conference on Computer Vision. pp. 35–51. Springer (2024)
12. Guo, Z., Wu, Y., Chen, Z., Chen, L., Zhang, P., He, Q.: PuLID: Pure and lightning ID customization via contrastive alignment. In: Advances in Neural Information Processing Systems (2024)
13. Jiang, H., Luo, A., Fan, H., Han, S., Liu, S.: Low-light image enhancement with wavelet-based diffusion models. ACM Transactions on Graphics (TOG) **42**(6), 1–14 (2023)
14. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
15. Kim, S., Kim, M., Park, D.: Test-time alignment of diffusion models without reward over-optimization. arXiv preprint arXiv:2501.05803 (2025)
16. Li, C., Wu, H., Hao, H., Zhang, Z., Kou, T., Chen, C., Bai, L., Liu, X., Lin, W., Zhai, G.: G-Refine: A general quality refiner for text-to-image generation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 7375–7384 (2024)
17. Li, G., Cen, K., Zhao, B., Xin, Y., Luo, S., Zhai, G., Zhang, L., Liu, X.: LayerT2V: A unified multi-layer video generation framework. arXiv preprint arXiv:2508.04228 (2026)
18. Li, G., Yang, Y., Song, C., Liu, X., Zhang, C.: FlowDirector: Training-free flow steering for precise text-to-video editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7805–7815 (2026)

19. Li, M., Yang, T., Kuang, H., Wu, J., Wang, Z., Xiao, X., Chen, C.: ControlNet++: Improving conditional controls with efficient consistency feedback. In: European Conference on Computer Vision. pp. 129–147. Springer (2024)
20. Li, Z., Cao, M., Wang, X., Qi, Z., Cheng, M.M., Shan, Y.: PhotoMaker: Customizing realistic human photos via stacked ID embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8640–8650 (2024)
21. Liu, D., Zhao, S., Zhuo, L., Lin, W., Xin, Y., Li, X., Qin, Q., Qiao, Y., Li, H., Gao, P.: Lumina-mGPT: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. arXiv preprint arXiv:2408.02657 (2024)
22. Peng, B., Wang, J., Zhang, Y., Li, W., Yang, M.C., Jia, J.: ControlNext: Powerful and efficient control for image and video generation. arXiv preprint arXiv:2408.06070 (2024)
23. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
24. Qin, Q., Zhuo, L., Xin, Y., Du, R., Li, Z., Fu, B., Lu, Y., Li, X., Liu, D., Zhu, X., et al.: Lumina-Image 2.0: A unified and efficient image generative framework. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20031–20042 (2025)
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
26. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: DALL-E 2: Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125 1(2), 3 (2022)
27. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(3), 1623–1637 (2020)
28. Richter, L., Berner, J., Bötte, L.: An optimal control perspective on diffusion-based generative modeling. arXiv preprint arXiv:2211.01364 (2023)
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
30. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (ICLR) (2021)
31. Tan, Z., Liu, S., Yang, X., Xue, Q., Wang, X.: OMNI-control: Minimal and universal control for diffusion transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14940–14950 (2025)
32. Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A., Li, H., Tang, X., Hu, Y.: InstantID: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519 (2024)
33. Wang, R., Zheng, Y., Zhang, Z., Li, C., Liu, S., Zhai, G., Liu, X.: Learning hazing to dehazing: Towards realistic haze generation for real-world image dehazing. arXiv preprint arXiv:2503.19262 (2025)
34. Wu, G., Liu, X., Jia, J., Cui, X., Zhai, G.: Text2QR: Harmonizing aesthetic customization and scanning robustness for text-guided QR code generation. In: Pro-

- ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8456–8465 (2024)
35. Wu, G., Zheng, H., Luo, S., Zhai, G., Liu, X.: AnimateQR: Bridging aesthetics and functionality in dynamic QR code generation. *Advances in Neural Information Processing Systems* **38**, 3472–3490 (2026)
 36. Wu, H., Zhang, Z., Zhang, W., Chen, C., Li, C., Liao, L., Wang, A., Zhang, E., Sun, W., Yan, Q., Min, X., Zhai, G., Lin, W.: Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090* (2023), equal Contribution by Wu, Haoning and Zhang, Zicheng. Project Lead by Wu, Haoning. Corresponding Authors: Zhai, Guangtai and Lin, Weisi.
 37. Xin, Y., Qin, Q., Luo, S., Zhu, K., Yan, J., Tai, Y., Lei, J., Cao, Y., Wang, K., Wang, Y., et al.: Lumina-DiMOO: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308* (2025)
 38. Yang, X., Wang, Z., Hu, A.: Batch-instructed gradient for prompt evolution: Systematic prompt optimization for enhanced text-to-image synthesis. *arXiv preprint arXiv:2406.08713* (2024)
 39. Yang, Y., Liu, Z., Jia, J., Gao, Z., Li, Y., Sun, W., Liu, X., Zhai, G.: DiffStega: Towards universal training-free coverless image steganography with diffusion models. *arXiv preprint arXiv:2407.10459* (2024)
 40. Yang, Z., Zeng, A., Yuan, C., Li, Y.: DWPose: Effective whole-body pose estimation with two-stage distillation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 4210–4220 (2023)
 41. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: IP-adapter: Text-compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721* (2023)
 42. Yu, J., Wang, Y., Zhao, C., Ghanem, B., Zhang, J.: FreeDoM: Training-free energy-guided conditional diffusion model. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 23174–23184 (2023)
 43. Zavadski, D., Feiden, J.F., Rother, C.: ControlNet-XS: Rethinking the control of text-to-image diffusion models as feedback-control systems. In: *European Conference on Computer Vision*. pp. 343–362. Springer (2024)
 44. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 3836–3847 (2023)
 45. Zhang, X., Wei, X., Hu, W., Wu, J., Wu, J., Zhang, W., Zhang, Z., Lei, Z., Li, Q.: A survey on personalized content synthesis with diffusion models. *Machine Intelligence Research* **22**(5), 817–848 (2025)
 46. Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., Wong, K.Y.K.: Uni-ControlNet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems* **36**, 11127–11150 (2023)
 47. Zheng, H., Li, G., Zhao, B., Luo, S., Jiang, H., Zhai, G., Liu, X.: A^2 -Edit: Precise reference-guided image editing of arbitrary objects and ambiguous masks. *arXiv preprint arXiv:2603.10685* (2026)