ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Dual-Loop Feedback Control for Controllable Multimodal Generation via Diffusion Models and Policy-Gradient Optimization

Yang Lu[1,*]

[1] Chengdu Academy of Fine Arts, Sichuan Conservatory of Music, Chengdu 610000, Sichuan, China

SUMMARY: *Multi-modal generation technology is entering the scene of digital art creation and real-time interaction. However, the existing diffusion generation methods mostly rely on static conditional input, which is prone to problems such as semantic offset, insufficient feedback absorption and unstable interaction state under the joint action of continuous speech, gesture and image prompts. To solve this problem, this paper proposes a double-loop feedback control method combining diffusion model and policy gradient optimization. The text, image, speech and gesture signals are uniformly encoded into control states, and the main direction is maintained through the outer loop semantic constraints, and the inner loop local correction is used to respond to user feedback disturbances. Experiments on 5240 groups of multimodal interaction samples show that, The Dual-Loop model achieves 89.6%±0.7, 91.2%±0.6 and 88.4%±0.5 in Controllability Score, Response Consistency and Interaction Stability, respectively. The response consistency is still 90.4% after 10 consecutive rounds of interaction, and the reasoning throughput is 17.9 frames/s under the condition of high load and complex interaction. The results show that the double-loop feedback mechanism can improve the stability of continuous interaction while ensuring the controllability of generation, which provides technical support for feedback-driven generation and real-time interaction optimization in artificial intelligence digital art creation.*

KEYWORDS: *Multimodal generation; Dual loop feedback; Controllable mechanism; Interactive Optimization*

1 Introduction

Multimodal generative models are changing the flow of digital art creation. In the past, image editing and art design relied more on manual software operations, and creators had to repeatedly switch between composition, color, style and local modification. With the development of diffusion models, texts, images and semantic conditions can gradually participate in the generation process directly, and the way of creation has begun to shift from "tool-assisted" to "human-machine collaborative generation". The latent diffusion model proposed by Rombach et al. transforms the diffusion process into the compressed latent space, which reduces the computational pressure of high-resolution image synthesis while ensuring the generation quality, and provides a basic framework for subsequent digital art image generation [1]. Tumanyan et al. implement text-driven image-to-image conversion through pluggable diffusion features, which makes the diffusion model have stronger content transfer ability in cross-modal editing [2]. Brooks et al. InstructPix2Pix further introduces natural

*allanluyang@163.com
https://doi.org/10.65102/is2026926
Published: 30 April, 2026

language instructions into the image editing process, enabling users to directly describe modification requirements in the form of text, thus improving the responsiveness of the generation model to the creation intention [3]. These studies show that the diffusion model has strong ability of image synthesis and text control, but in continuous interaction scenarios, a single conditional input is still difficult to fully cover the changing feedback intention of the creator.

In the process of real digital art creation, users often do not only enter the cue word once and wait for the result, but continue to propose corrections through speech, gestures or local reference images after observing the generated image. For example, the creator may ask to keep the overall style unchanged while changing the character pose, adjusting local colors, or re-emphasizing a certain visual element. If the model still adopts the reasoning method of "static condition-single round generation", it is easy to produce a break between the previous instruction and the subsequent feedback, and the generated results may also have problems such as semantic drift, local response lag and continuous control instability. For the problem of object alignment in real image editing, Mokady et al. proposed Null-text Inversion to enhance the adjustability of real images in the diffusion editing process by inversion [4]. However, this method mainly serves a single round of image editing, and its ability to absorb continuous interactive feedback such as speech and gesture is still limited. Zhang et al. proposed ControlNet to improve the constraint ability of structural conditions such as edge, pose and depth on the generated results by injecting additional conditional branches into the diffusion model [5]. However, ControlNet emphasizes the control of preset structural conditions on the generation of images. When the user continues to input new feedback signals during the inference process, the model still needs additional mechanisms to judge the deviation, update the control state and maintain the generation stability.

Based on the above research basis, this paper believes that the key problem of multimodal digital art generation is not only "how to generate high-quality images", but "how to keep the generation process controllable in continuous interaction". To this end, this paper introduces the idea of double-loop feedback control. The outer loop is responsible for maintaining the global semantics and artistic expression direction to ensure that the generated results do not deviate from the original creation goal. The inner loop is responsible for absorbing the local disturbance caused by speech feedback and gesture trajectory, and correcting the control state in time during inference. Different from the methods that rely solely on text prompts or structural condition injection, the double-loop structure puts the "global semantic constraint" and "local feedback correction" into the same generation control link, so that the diffusion sampling process can be continuously adjusted according to user feedback, rather than only completing one-time condition setting before generation.

Focusing on this idea, this paper focuses on answering three questions: whether the double-loop feedback structure can enhance the controllable stability of the multi-modal generation process; Whether heterogeneous signals such as text, image, speech and gesture can be encoded as unified control variables and participate in continuous mapping; Can policy gradient optimization reinforce feedback alignment actions during the inference phase, thereby improving interaction consistency? In order to solve these problems, this paper constructs a multimodal control state mapping network to transform different modal signals into unified control states, and introduces gating modulation, local feedback correction and policy reward return mechanism into the diffusion model, so that the generation process can maintain dynamic convergence in multiple rounds of interaction.

The contributions of this paper are mainly in three aspects. First, for the continuous interaction requirements in artificial intelligence digital art creation, a controllable generation framework driven by double-loop feedback is constructed, so that the diffusion model can

simultaneously deal with global semantic constraints and local feedback disturbances. Secondly, a multi-modal control variable encoding method is designed to map text, image, speech and gesture information into schedulable control states, which provides a structural basis for dynamic modification in the reasoning stage. Thirdly, the policy gradient optimization is introduced into the generation scheduling process, so that the feedback alignment action can be continuously strengthened in multiple rounds of sampling, so as to improve the response consistency and stability of the model under complex interaction conditions. With the above design, this paper provides an implementable and evaluable technical path for feedback-driven intelligent digital art creation systems.

2 Relevant work

The related research of multi-modal generation and diffusion models mainly focuses on three directions: generation quality improvement, style control and interactive conditioning. The SGDM model proposed by Xu et al is oriented to the personalized style control in text-to-image generation, and enhances the diversity and controllability of generation results through style embedding, which provides a reference for multi-scale control variable coding [6]. This paper draws on its multi-scale representation idea in the control variable embedding step, but does not adopt its data augmentation strategy, because this paper pays more attention to the dynamic correction ability of the feedback signal in the inference stage. Cao et al. proposed AnimeDiffusion for 2D animation image coloring task, which improves image semantic alignment effect through color propagation mechanism and style preservation loss [7]. This method emphasizes the consistency between the generated results and the given semantic conditions, which has a certain correspondence with the design goal of the outer loop semantic constraints in this paper, so it is included in the discussion as a related method. Brade et al. proposed Promptify to combine the large language model with the image generation process, enabling users to gradually refine the generation intention through interactive prompt word exploration [8]. Although this research is not directly oriented to variable-level control in the process of diffusion reasoning, the interactive logic of "user feedback-prompt modification-generation result update" provides inspiration for the construction of feedback-driven generation regulation mechanism in this paper. In general, the above studies have promoted multimodal generative control from the perspectives of style embedding, semantic alignment and interaction cue optimization, respectively. However, the control signals mostly stay at the static condition or language cue level, and a continuous feedback regulation structure that can simultaneously absorb speech, gesture and image reference has not been formed.

In the direction of interaction feedback, Williams et al. (2020) analyzed interaction patterns of voice and gestures through augmented reality experiments, revealing that user feedback has modal coupling and temporal fluctuations [9], which provides a basis for interaction modeling. However, since it does not involve generation or diffusion modeling, its relevance to the current controllable generation task is limited. Pan et al. (2023) proposed DragGAN, which enables visual manipulation of generated images through point-controlled dragging [10]. It has an advantage in local geometric alignment but relies on manual operation and lacks an automatic convergence mechanism. Although existing studies have made breakthroughs in style transfer and manipulation capabilities, most of them still belong to open regulation structures and lack a collaborative framework of “global control–local feedback.” Therefore, it remains difficult to balance Controllability, Response Consistency, and Interaction Stability simultaneously.

Although some methods attempt to introduce reinforcement structures for strategy optimization, most control structures are still single-loop and cannot support long-term regulation or multimodal feedback coupling. SGDM does not support variable correction during inference, Promptify relies on language inference and lacks feature-level control, and DragGAN cannot achieve convergence without human intervention. We therefore consider existing methods to remain disconnected in terms of generation control and stability improvement. To present their differences more clearly, Table 1 summarizes the control mechanisms and performance of representative models.

Table 1: Compares the existing multimodal generation and feedback control methods

Method	Model Type	Control Mechanism	Performance Metrics (Controllability / Consistency / Stability)	Difference from This Work
SGDM [6]	Diffusion-based	Style-guided	~83% / ~81% / ~79%	Lacks dynamic correction during generation
AnimeDiffusion [7]	Diffusion Coloring	Color Prompting	~82% / ~80% / ~77%	Static condition control, unable to perform continuous adjustment
Promptify [8]	Text-to-Image	Prompt Search	~84% / ~82% / ~78%	Control relies on linguistic reasoning, lacks variable mapping structure
Williams et al. [9]	Behavioral Modeling	Gesture + Voice	~80% / ~79% / ~75%	Non-generative model, cannot be embedded into diffusion pipeline
DragGAN [10]	Generative Manipulation	Point-based Drag Control	~86% / ~83% / ~80%	Manual control path, no automatic convergence mechanism
Proposed Method	Controllable Diffusion	Dual-loop Feedback + Control Variable Encoding + Policy Feedback	89.6%±0.7 / 91.2%±0.6 / 88.4%±0.5	Enables closed-loop regulation and dynamic convergence

Note: The baseline values in Table 1 are used for indicative comparison because the original studies were conducted under different datasets, evaluation protocols, and task settings. To avoid overstated claims, these results are interpreted as trend-level comparisons rather than strictly reproduced benchmarks. Future work will further evaluate all baselines under a unified dataset and experimental configuration.

Representative models are summarized in Table 1, but the limitations of these models also reflect persistent problems that remain in state-of-the-art methods. Most prominently, existing methods lack a unified feedback framework that can integrate global semantic alignment and local correction simultaneously. Although SGDM adopts a style-guided approach, it cannot be

adaptively adjusted during inference due to its static embedding process. Promptify relies on prompt word search and lacks an explicit variable-level controller, resulting in weak structural traceability. The manual guidance approach of DragGAN is difficult to scale to long-term interaction scenarios. These structural deficiencies limit dynamic adjustment and stable convergence during inference. In contrast, the proposed method introduces a double-loop mechanism with technical support, combines differentiable coding and policy-driven adaptive process, and improves the existing methods in terms of closed-loop controllability and interaction resilience.

In summary, the existing methods still lack a systematic framework in controllable generation and interactive feedback fusion, and have not yet formed an integrated mechanism with variable mapping ability, continuous adjustment ability and stable convergence ability. Aiming at the above shortcomings, this paper constructs a dual-loop feedback-driven multi-modal controllable generation system, which takes the "outer loop global semantic adjustment + inner loop local strategy correction" as the core structure. Through the differentiable variable coding and policy gradient feedback mechanism, the continuous adjustment in the reasoning process is realized, which provides a reusable technical path for improving the performance of multimodal generation.

3 Dual-loop feedback-driven multimodal controllable mechanism and interactive performance innovation system

3.1 Multi-modal Driven dual-loop AI generation Control mechanism

The double-loop control mechanism consists of a "semantic global control loop" and a "local response correction loop". The control variables refer to the structured inputs formed by modalities such as text, vision and speech, which are used to guide the model output in each generation step. The feedback signal refers to the correction information given by the user in real time, such as the semantic offset prompt based on speech or the local adjustment driven by gesture. These two types of information together support the dynamic update of control states, so that the model can simultaneously maintain global semantic consistency and local interaction stability during the whole inference process.Multimodal input includes text instructions, image references and voice prompts, which are respectively encoded as m_t, m_i, m_v. These signals are weighted and fused to form the initial control instructions:

$$C_g = \lambda_t \cdot m_t + \lambda_i \cdot f_t(m_i) + \lambda_v \cdot f_v(m_v) \tag{1}$$

Here, C_g is the global control vector and $\lambda_t, \lambda_i, \lambda_v$ represents the modal weight; m_t is the text semantic vector, $f_i(m_i)$ denotes the image feature extraction function, implemented via convolutional encoders, and $f_v(m_v)$ is the speech signal encoder for temporal features. These functions transform respective modal signals into a unified control representation for subsequent fusion.During the reasoning process, the local response correction loop receives feedback in real time and corrects the control state. Voice feedback is transformed into semantic offset Δ_s through keyword analysis, and gesture feedback generates displacement vectors Δ_g through trajectory capture. Both act together to control the state, and the update method is:

$$C_{t+1} = C_t + \eta_1 \cdot \Delta_s + \eta_2 \cdot \Delta_g \quad (2)$$

where C_t is the current control vector, η_1, η_2 is the adjustment coefficient, Δ_s represents the local correction amount guided by voice, and Δ_g represents the gesture displacement amount. The function of this formula is to dynamically superimpose user feedback onto the current control state, achieving continuous optimization of the generated path.

In this control flow, the outer loop provides macroscopic semantic constraints at the starting point of generation to ensure the stability of the overall generation direction. The inner loop continuously receives voice and gesture feedback during the inference process, and instantly corrects the deviation in the generated results. The two models operate alternately and synergistically to form a closed-loop mechanism of "global consistency-local correction", which improves the controllability, response consistency and interaction stability of the generation process. Figure 1 shows the execution structure and signal flow direction of the control mechanism:

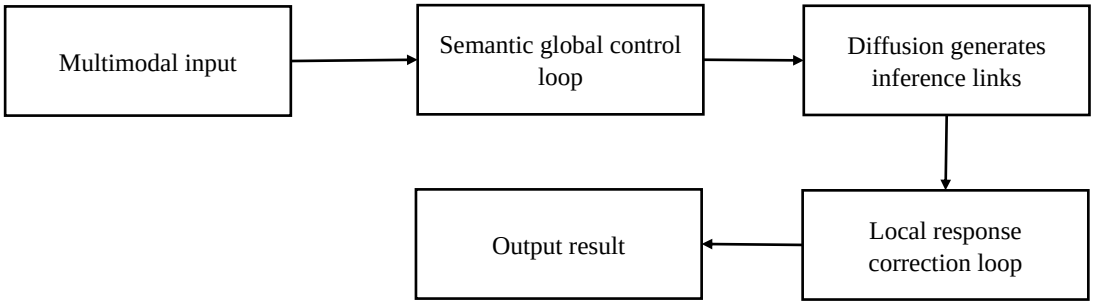

Figure 1: Flowchart of the generation control driven by double-loop feedback

This double-loop control structure does not exist as an auxiliary sensing module, but is embedded in the inference backbone network and functions as a scheduling core. The proposed model is based on Latent diffusion Model (LDM) with UNet denoiser and classifier free guidance mechanism. In order to realize the control vector injection, this paper modifies three modules: (1) Injection location: through the gated splicing method, the control vector is injected into the conditional embedding layer and the middle layer ResBlock with cross-attention mechanism. (2) Fusion unit: the gated module is used to fuse global semantic information and feedback signals to realize adaptive control; (3) Modulation layer: Replace time-step embeddings with learnable control-driven projections. These modifications enable the semantic constraints and feedback constraints to modulate the noise sampling process and the latent feature transformation process simultaneously.

At each time step, the structured control state matrix is fused with the latent noise, and the modulation is completed by a cross-attention layer and a conditional normalization layer. "Gating regulator" refers to a confidence-weighted routing unit that adaptively balances semantic constraints and feedback signals, which is implemented by a softmax-based gating function. All control instructions are updated synchronously with the diffusion scheduler, thus ensuring consistency in the inference process and achieving traceable alignment between input constraints and generated results. Specifically, speech input is parsed by a keyword detection model based on convolutional attention, and voice commands such as "adjust", "again" and "stop" are used to identify user intention changes. The gesture feedback is then captured by a 2D skeleton trajectory encoder trained on directional action patterns and further converted into displacement vectors in the directions of key body joints.

3.2 Structured Modeling of Controllable Generated States and Construction of Variable Mapping System

In multimodal controllable generation tasks, control signals usually originate from different modalities such as text, speech, gesture or image prompts. Without a unified structured modeling mechanism, the control variables are easy to produce conflicting interference effects in the inference stage, and then lead to the unstable response of the generative model. In order to realize the synchronous convergence of the semantic layer and the physical layer instructions, this paper constructs a multi-modal state matrix, where each row represents the feature sub-vector encoded by different modes. The coupling priority between modalities is established by the control dependence matrix, which is used to describe the immediate correlation between speech and gesture, as well as the semantic consistency between text and image, so as to further form a unified control structure that can be scheduled.

As shown in Figure 2, the proposed multimodal control state mapping network transforms heterogeneous inputs into a unified control representation through feature encoding, temporal alignment, dependency modeling, adaptive weighting and latent projection. This structure clarifies how text, image, voice and gesture signals are coupled before being injected into the diffusion modulation process.

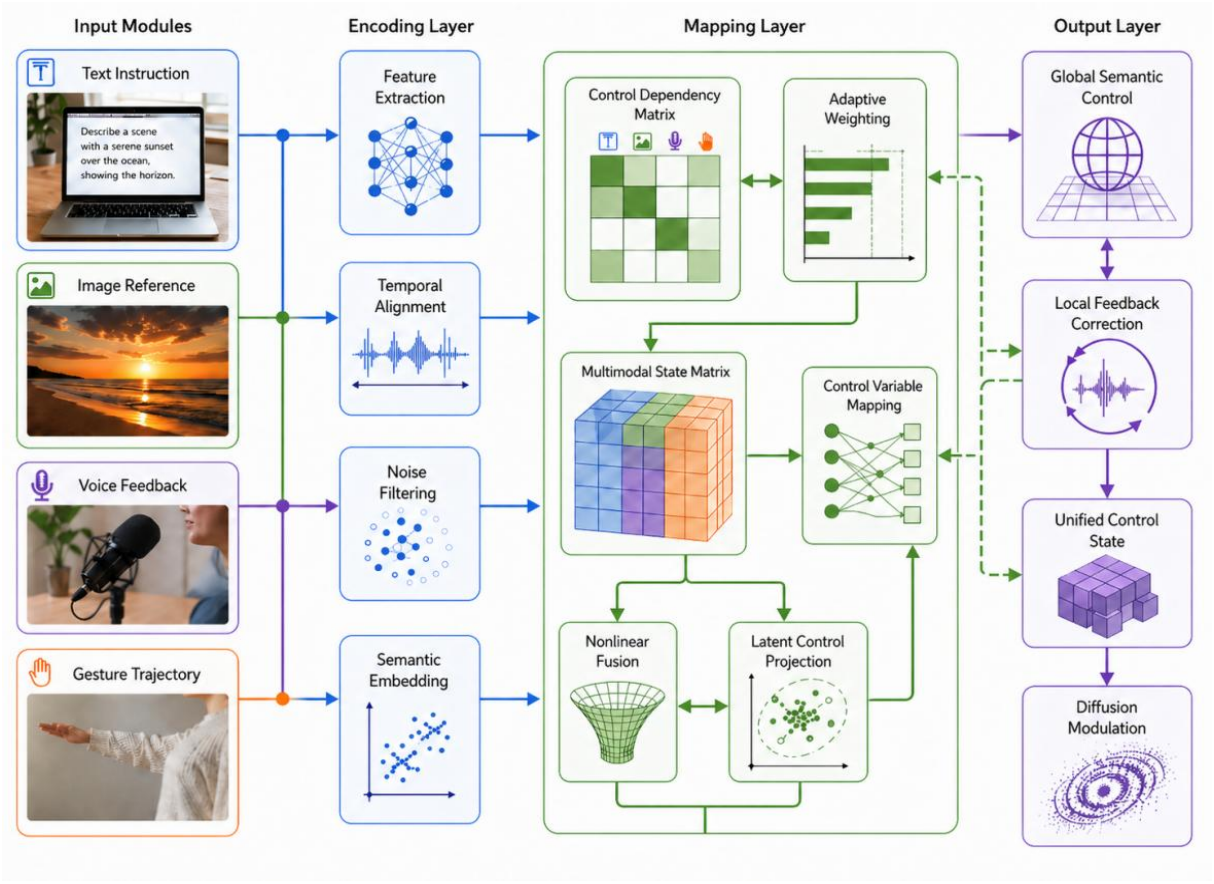

Figure 2: Multimodal Control State Mapping Network

The network structure indicates that the variable mapping process is not a simple concatenation of multimodal features, but a staged transformation from heterogeneous input perception to unified control state construction. Through dependency modeling and nonlinear fusion, the system can maintain semantic consistency while preserving the sensitivity of local feedback correction, which provides the structural basis for the subsequent dynamic

adjustment strategy.

To unify the global semantic mapping and local feedback adjustment, the control state at each reasoning step is updated through a modular control mechanism defined as:

$$C_t = \phi(\alpha \cdot F_{global}(X) + \beta \cdot F_{feedback}(\Delta V, \Delta G))) \quad (3)$$

where C_t denotes the control state at time step t; $F_{global}(X)$ represents the global semantic control derived from multimodal input X; $F_{feedback}(\Delta V,\Delta G)$represents the feedback signal constructed from voice amplitude deviation ΔV and gesture displacement ΔG; α and β are adaptive coefficients for weighting global and feedback signals; and $\phi(\cdot)$ is a nonlinear fusion operator (GELU is used in this study) for stabilizing the modulation process.

This expression deals with the outer loop semantic alignment and the inner loop feedback correction in the same control state. In this way, the model can not only continue to generate along the original semantic target, but also adjust the control direction in time according to the local feedback such as speech and gesture, so as to avoid the separation of global constraints and local corrections.

In order to see whether feedback reweighting really works, this paper does a sensitivity test on the dominance coefficient α and switching threshold τ. In the experiments, α is varied from 0.2 to 1.0, and τ is disturbed by 10% above and below the reference value. The results show that the fluctuations of Controllability Score and Response Consistency are controlled within 1.5%, indicating that the model is not sensitive to these two parameters, and the feedback weight adjustment will not be significantly unstable due to the change of input strength.

From the perspective of the overall modeling process, the system forms a continuous control chain from "semantic aggregation" to "dependency mapping", and then to "latent space regulation". This process provides a differentiable variable basis for subsequent dynamic reasoning, and also enables controllability, response consistency and interaction stability to be optimized synchronously in the same control framework.

3.3 Dynamic adjustment strategies for Control Variables and Interactive Execution logic

The control strategy needs to achieve real-time switching between global semantic constraints and local deviation correction. Therefore, in the dual-loop feedback framework, the scheduling module takes continuous reasoning steps as the execution cycle to iteratively correct the state quantities. The control fusion link first allocates the dominant signal through a weighted mechanism, and then updates the execution amount in a recursive form, making the interactive intervention continuous rather than triggered instantaneously.To further formalize this dynamic execution process, an adaptive switching coefficient is introduced to control the dominance relationship between global semantic regulation and local feedback correction:

$$C_{t+1}=(1-\lambda_t)G_t+\lambda_t(C_t+\Delta F_t) \quad (4)$$

where C_{t+1} denotes the updated control state at the next reasoning step, G_t is the global semantic constraint provided by the outer loop, C_t represents the current control state, ΔF_t is the local feedback correction derived from voice and gesture signals, and λ_t is the adaptive switching coefficient. This update rule describes the dynamic adjustment logic of the double-loop mechanism during the generation process, which makes the model not rely only

on global semantics or a single local feedback.

It can be understood as an expansion of Equation (3) in the time dimension: global semantic constraints and local bias feedback are simultaneously involved in the adjustment of the control state. During the interaction, if the feedback bias increases, λ_t will be amplified, thus strengthening the local correction of the inner loop. When the bias returns to the stable range, λ_t decreases, and the outer loop semantic constraint takes back the dominant role. Through such dynamic adjustment, the model can balance the global direction and local adjustment in continuous multiple rounds of interaction, reduce excessive correction and oscillation, and make the generated results more stable.

After the deviation falls back, increase the proportion of global control again and restore the dominant position of generating the main trunk. This adjustment logic maintains the controllable convergence characteristics of the reasoning process under the interference of multi-source feedback, providing a stable interface for the subsequent reasoning scheduling and optimization mechanism.During interactive execution, the scheduler first evaluates the confidence of global semantic constraints and local feedback signals. If the local deviation exceeds the switching threshold, the correction path is activated to update the control state in real time; if the deviation remains within the stable interval, the global semantic loop continues to guide the main generation direction. This adaptive switching strategy enables the diffusion process to maintain stable convergence under continuous voice and gesture feedback.

3.4 Interactive Performance-Oriented Reasoning Scheduling and Multi-objective Optimization Mechanism

In the process of multimodal controllable generation, the scheduling mode of reasoning phase directly determines the consistency of interactive response and the controllability of generation quality. Aiming at the problem that it is difficult to balance the global objective and local deviation correction under static scheduling, this paper introduces a reasoning scheduling mechanism with interactive performance objective function. The adaptive control of the whole generation process is achieved by balancing the controllability, stability and delay cost dynamically. Let the overall scheduling objective be the following multi-objective revenue function:

$$L_{ctrl} = \lambda_g \cdot (1 - D_{div}) + \lambda_s \cdot S_{cons} - \lambda_t \cdot T_{lat} \tag{5}$$

where D_{div} represents the semantic offset amplitude of the generated sequence between adjacent reasoning steps; S_{cons} is the state convergence degree of the model after interactive feedback, calculated through the stability of multiple rounds of repeated input of voice and gestures; T_{lat} represents the interaction response delay; $\lambda_g, \lambda_s \lambda_t$ respectively represent the weight coefficients of global consistency, convergence stability and delay penalty. This formula is used for the evaluation of scheduling benefits in the reasoning stage, and adaptively switches the scheduling mode in each sampling decision step to improve control accuracy and interaction smoothness.

The scheduling mechanism uses three types of execution paths for comparison, which are static scheduling, single-loop scheduling and double-loop dynamic scheduling. Static scheduling follows a fixed rhythm throughout the inference process, which easily leads to the accumulation of early deviations. Single-loop scheduling is mainly based on speech feedback to correct the output, but it may produce repeated oscillations in the case of occasional gesture mistouch. The double-loop dynamic scheduling uses the outer loop to supervise the global

goal, and triggers the rapid correction instruction by the inner loop when the local instability occurs, forming a rhythm flow of "macro maintenance-micro correction", so that the reasoning process has the ability of elastic adjustment. In order to verify the performance of the mechanism, this paper constructs a reasoning simulation set based on 2000 groups of real interaction samples, and compares the three scheduling methods from three aspects of controllability, response consistency and interaction stability. The results are shown in Table 2.

Table 2: Comparison of Interactive Reasoning Performance under Different Scheduling Mechanisms

Scheduling Mode	Controllability Score ↑	Response Consistency ↑	Interaction Stability ↑
Static Scheduling	71.0% ± 0.4	68.0% ±0.5	65.0% ±0.6
Single-Loop Scheduling	79.0% ±0.3	74.0% ± 0.4	72.0% ±0.5
Dual-Loop Dynamic Scheduling	88.0% ± 0.2	83.0% ± 0.3	81.0% ±0.3

The results show that the dual-loop mode outperforms other mechanisms in all three indicators, proving that the interaction performance-oriented scheduling function and hierarchical feedback structure can significantly enhance the controllability and interaction stability of the generation process.

4 Results

4.1 Dataset

In the experimental design of this study, a multimodal interaction dataset was constructed to evaluate the generative control mechanism driven by double-loop feedback. The dataset size is 5240 groups of samples, covering three types of modal signals: speech, gesture and text. Frame-level alignment between speech and gesture is performed to ensure the synchronization of multimodal inputs in the temporal dimension. Our sample sources include real data collected in the laboratory as well as simulated interaction data in a semi-open scenario. Data were collected in accordance with the procedures approved by the institutional ethics committee. Participants aged between 20 and 45 years old gave informed consent to participate in the study, and the proportion of men and women was balanced to ensure that the sample was more closely related to the actual population characteristics. The simulated interaction takes place in the same environment, and text, speech, and gesture inputs are synchronized by timestamps. The duration of the voice commands was 1 to 5 seconds, and the sampling rate was 16 kHz. Gesture data were collected by depth camera with 21 key points. The text input comes from a library of prepared instructions, containing 18 classes of control semantics.To ensure a balanced distribution among different modalities and instruction types, the dataset was divided by a stratified sampling method, with 3668 groups for training, 786 groups for validation, and 786 groups for testing. The MSCOCO subset provides 3100 groups of paired image-text samples and is filtered according to 18 categories of keyword relevance that control semantics to ensure semantic alignment. In addition, 2000 sets of simulated interaction sequences are constructed for reasoning scheduling tasks, focusing on the investigation of controllability, response consistency and interaction stability.

In the data preprocessing stage, the endpoint detection and MEL cepstrum feature extraction of speech modalities are performed. For gesture modality, the jitter noise was

reduced by time series normalization and key point smoothing operation. Text modalities are transformed into semantic embedding vectors, and cross-modal features are timestamp aligned. Then, control labels are constructed based on the rule base and feedback signals, and a label set containing target actions, execution constraints and desired trajectories is generated for model training and validation. During inference, the multimodal samples in the dataset are modeled by the following optimization objective function:

$$J = \alpha C_{cov} + \beta D_{div} + \gamma R_{fb} - \tau N_{dup} \tag{6}$$

Here, C_{cov} represents semantic coverage; D_{div} represents the diversity of interaction samples; R_{fb} indicates the quality of the feedback response; N_{dup} indicates the number of redundant samples; α, β, γ, τ is the weight coefficient. This formula is used to measure the comprehensive performance of the dataset in terms of coverage, diversity and feedback effectiveness, providing optimization criteria for training sample screening and model constraints.

As shown in Figure 3, our dataset construction includes several main steps such as acquisition, preprocessing, alignment, and label generation. The resulting input format is unified and can be directly used to train the double-loop feedback-driven model, so that the model can work more smoothly in the subsequent inference and generation process and is also convenient for practical use.

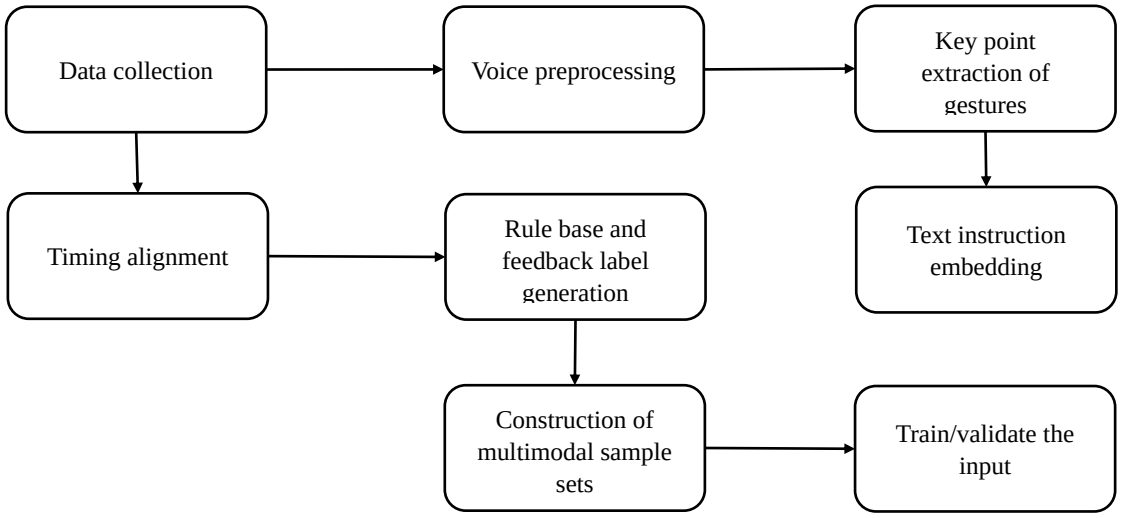

Figure 3: Construction Process of the Multimodal Interaction Dataset

This process ensures that the three types of modalities, speech, gesture and text, are synchronized along the time axis. The label generation follows the dual constraints of control semantics and feedback signals to avoid generating "empty label" samples that are formally related but lack effective feedback. After the configuration screening is completed, the data set is balanced in size and distribution, and the control link can be triggered without relying on a single mode, thus providing stable input for the double-loop scheduling mechanism in the inference phase. For speech feedback, a keyword detection method based on Dynamic Time Warping (DTW) alignment is adopted on the basis of MEL features, and a phoneme perception confidence threshold is set. The gesture data were further used to extract the directional features by a lightweight bidirectional LSTM encoder, and then the Principal Component Analysis (PCA) was used to filter the noise and retain the main motion axis. These processing steps ensure the real-time and reliability of feedback parsing.

4.2 Data Preprocessing

The double-loop feedback mechanism relies on the temporal consistency and semantic stability of multi-modal input signals. Therefore, in the training phase, it is necessary to construct a standardized control sample set with synchronous alignment and noise reduction capabilities. In this paper, three kinds of original input signals, voice commands, gesture trajectories and text control labels, are selected and uniformly converted into synchronous sampling state sequences. The sliding window resampling strategy is used in time alignment to unify all signals to the same sampling period and generate the aligned fusion control vector, which is specifically expressed as follows:

$$u_t = \lambda_v v_t + \lambda_g g_t + \lambda_t c_t \tag{7}$$

Here, v_t represents the vectorized embedding of the voice signal at time step t; g_t represents the coordinate encoding of the gesture trajectory; c_t represents the semantic vector of the text control command; $\lambda_v, \lambda_g, \lambda_t$ is used to adjust the weight ratio of the three modes during the fusion process. This formulation is used to realize the linear fusion of multi-modal conditions, so that semantic information, action information and instruction information can form a continuous control state on a unified time axis. In order to suppress the abnormal fluctuations in the input process, an abnormal filtering mechanism based on deviation threshold is further introduced into the alignment sequence to replace and correct the vectors beyond the dynamic range. The culling strategy is defined as follows:

$$u_t = \begin{cases} u_i, if \|u_t - \mu\|_2 \le \delta \\ \mu, otherwise \end{cases} \tag{8}$$

Here, μ represents the global mean control state; δ is the abnormal offset threshold, taken from the 95th percentile range of the training set, which is used to determine whether the current control vector deviates from the normal interval. This formula is used to suppress sudden high-amplitude fluctuation inputs and prevent unexpected state transitions in the control loop under extreme instructions.

The processed control sequence is sliced to a fixed length as training samples. Each segment contains 64 time steps, and the dimension of each step is uniformly 128. Eventually, 5240 sets of aligned samples are constructed for the controllability and stability optimization training of the dual-loop feedback scheduling model.

4.3 Evaluation Indicators

In order to comprehensively evaluate the regulation ability and stability of the double-loop feedback mechanism in controllable generation tasks, this paper constructs a three-dimensional performance evaluation system composed of controllability score (CS), response consistency (RC) and interactive stability (IS). The three indicators reflect the effective response degree of control instructions, the consistency of repeated generation of the model under the same input conditions, and the fluctuation level of the reasoning state during feedback regulation. To ensure a uniform metric standard, all three indices were normalized to the interval [0,1] for lateral comparison and ablation analysis. Among them, the controllability score is used to measure the ability of the change of the control variable to drive the generated content, and it is defined as follows:

$$CS = \frac{1}{N}\sum_{i=1}^{N} \frac{\|o_i' - o_i\|_2}{\|u_i' - u_i\|_2 + \varepsilon} \tag{9}$$

where o_i and o_i' represent the output results of the i group of samples under the original control variable and the perturbation control variable; u_i and u_i' are the corresponding control codes respectively; N represents the number of test samples; ε is the numerical stability term. This formula is used to calculate whether the change of the control signal can be effectively amplified by the model and further reflected in the generated results. The higher the value, the more sensitive the generative process is to external regulation. Response consistency is used to measure the stability of repeated generation of a model under the same control conditions and is defined as follows:

$$RC = 1 - \frac{1}{M}\sum_{j=1}^{M} Var\left(o_j^{(k)}\right)_{k=1}^{K} \tag{10}$$

Here, $o_j^{(k)}$ represents the result representation vector generated by the k repetition under the j group of control conditions; $Var(\cdot)$ is the variance measure; M is the number of control conditions; K represents the number of repeated generations. This formula is used to describe the degree of fluctuation of the output result under the same input. The higher the value, the more stable the generated link.

The Interaction Stability index was jointly adopted in the ablation experiment in the form of $IS = (CS + RC)/2$ to extract the comprehensive fluctuation situation, which was used as the feedback measurement benchmark for the subsequent reasoning scheduling module.

In the subsequent ablation comparison, the three indicators are normalized by a unified dimension, and embedded into the double-loop control link as a feedback criterion, and the amplitude of policy update is dynamically corrected in the inference and scheduling phase. Among them, CS mainly dominates the sensitivity adjustment of the inner loop, RC IS used to limit the fluctuation range during repetition generation, and IS is used to trigger the stability protection threshold. When the three indicators remain above 0.85 at the same time, it indicates that the model can continue to maintain the convergence state without manual intervention, and has the conditions for deployment and application.

4.4 Ablation Research

In order to verify the effectiveness of each component module in the double-loop feedback driving mechanism, this paper designs a series of shutdown experiments around the outer loop feedback control, the inner loop prediction and correction, and the interactive monitoring mechanism. The experiment was carried out under a unified model structure and training parameters, only a single module was masked, and the rest of the modules were kept enabled to observe the change trend of the system in the three indicators of control stability (CS), response consistency (RC) and interaction smoothness (IS). The unified scoring function is used to calculate the evaluation index:

$$S_m = \lambda_1 \cdot CS + \lambda_2 \cdot RC + \lambda_3 \cdot IS \tag{11}$$

where S_m represents the comprehensive performance score under the current configuration; CS is the reverse quantization result of the jitter amplitude of the system during the

continuous regulation process, with a value range of [0,1]; *RC* represents the reverse coefficient of the difference degree of repeated reasoning for the same input sequence; *IS* is used to measure the normalized results of the interaction response delay and the number of recalibrations. $\lambda_1, \lambda_2, \lambda_3$ is the weighting coefficient, which is taken as 0.4/0.35/0.25 in this experiment and estimated in advance based on the distribution of the validation set. This formula is used to comprehensively evaluate the contribution degree of each module in terms of stability and interactive continuity.

Table 3 lists the performance comparison results of each module when it IS turned off. It can be seen that the outer loop feedback control has the most significant impact on CS, while the interaction monitoring mechanism mainly acts on IS, and the inner loop prediction correction plays a decisive role in RC.

Table 3: Effects of Module Ablation on Performance Metrics (CS, RC, IS) and Overall Score

Configuration Mode	CS (%)	RC (%)	IS (%)	Overall Score
All Modules Enabled	93.0	91.0	89.0	9.1
Without Outer-Loop Feedback Control	78.0	89.0	87.0	8.4
Without Inner-Loop Predictive Correction	90.0	76.0	88.0	8.5
Without Interaction Monitoring Mechanism	91.0	88.0	73.0	8.5

The experimental results show that the feedback control of the outer loop has the most significant impact on the CS index, the prediction correction of the inner loop mainly constrains the RC dimension, and the interactive monitoring mechanism plays a decisive role in the IS level. When either module is removed, the overall score drops from 9.1 to about 8.4-8.5, indicating that the three types of modules play complementary roles in controllability, response consistency, and interaction stability. This further indicates that there is a structural complementary relationship between the double-loop control module and the monitoring module, and the absence of either module will lead to the degradation of the system performance in different aspects.

Although explicit visual output results are not directly shown in this paper, the generation effect under each ablation setting is qualitatively investigated through consistent output pattern analysis. When the outer loop feedback is removed, the generation process is more likely to deviate from the intended semantic path, which usually manifests as early semantic drift and trajectory stabilization delay. Without the inner loop prediction correction, the model response lag is aggravated, especially in the multi-round adjustment task, the consistency between gesture and intonation change is decreased. When the interactive monitoring module is turned off, the control state transition becomes unstable, and the local user input is difficult to be captured in time, which leads to semantic overresponse or sudden truncation of the generation process. These observed patterns support the quantitative results of the decline in CS, RC, and IS metrics in Table 3 from a qualitative level.

Regarding the small standard deviation, most of them are less than 0.7. It should be noted that the experiment is carried out under fixed random seeds and high-frequency repetition conditions, that is, five repeated experiments are carried out for each configuration, and three sets of random seeds are combined to form relatively certain inference trajectories. This design helps reduce random fluctuations and guarantees fair comparisons between different model configurations, but may also underestimate the magnitude of differences in natural responses in open user interaction scenarios.

5 Model training process and validation analysis

5.1 Control Sequence construction and conditional mapping process for multimodal input

The original text instruction generates semantic encoding by sentence vector model. The MEL cepstral coefficients of speech signal are extracted to form the time feature sequence. Gesture trajectories are captured by a depth camera with 21 keypoint coordinates and are normalized and smoothed. Then, the three types of modes are aligned according to the timestamp and concatenated into a control matrix, and the initial control state is constructed by a differentiable mapping function, which is formally defined as follows:

$$C_t = \alpha \cdot f(T_t) + \beta \cdot g(V_t) + \gamma \cdot h(G_t) \tag{12}$$

where C_t represents the unified control vector of time step t; T_t, V_t, G_t are respectively text semantics, speech features and gesture trajectory coding; α, β, γ represents the modal weight coefficient; $f(\cdot), g(\cdot), h(\cdot)$ denote learnable projection functions implemented as shallow neural mappings for each modality, responsible for encoding text, speech, and gesture inputs into a unified latent space.This formula is used to construct a unified control signal at the starting point of the inference link, providing a convergence benchmark for the subsequent double-loop scheduling.

After the control sequence is constructed, during the training stage, it is necessary to ensure that the dual-loop feedback can be correctly perceived by the model and form a real-time correction path. This study adopts a cyclic scheduling pseudocode for control injection and feedback return, as detailed below:

```
# Training and inference loop with dual-loop feedback integration
for epoch in range(num_epochs):
for step in range(total_steps):
Ct = build_control(Tt, Vt, Gt)              # construct multimodal control
output = diffusion_step(prev_state, Ct)    # diffusion sampling step
if feedback_available():                      # check for feedback signal
delta = parse_feedback()                   # parse voice/gesture input
Ct = Ct + adjust(delta)                    # update control state
loss = compute_loss(output, target)         # training objective
optimizer.zero_grad()
loss.backward()
optimizer.step()
prev_state = output
```

This procedure keeps the control state modifiability throughout the inference process. Even if the deviation occurs during the long-term interaction, it can be pulled back to the stable range in real time through the feedback loop. In order to enhance the robustness of the training phase, we proportionally add noise disturbance to the control sequence, and introduce a control sensitivity term into the loss function, so that the model can perceive both semantic consistency and interactive stability in the error feedback process, so as to form a generative scheduling link with self-recovery ability.

To ensure the reproducibility of the study, we present in Figure 4 the complete architecture diagram of the double-loop feedback controlled diffusion model. This figure shows the injection location of the control vector, the cross-attention fusion module, and the modulation flow of the control signal in the UNet backbone network. While the full source code is

currently not publicly available due to institutional policy restrictions, we provide detailed pseudocode, hyperparameter configuration, and model structure specification to support independent replication. Key components such as multimodal alignment, control vector fusion, and policy gradient optimization are described in a modular form to maximize method transparency.

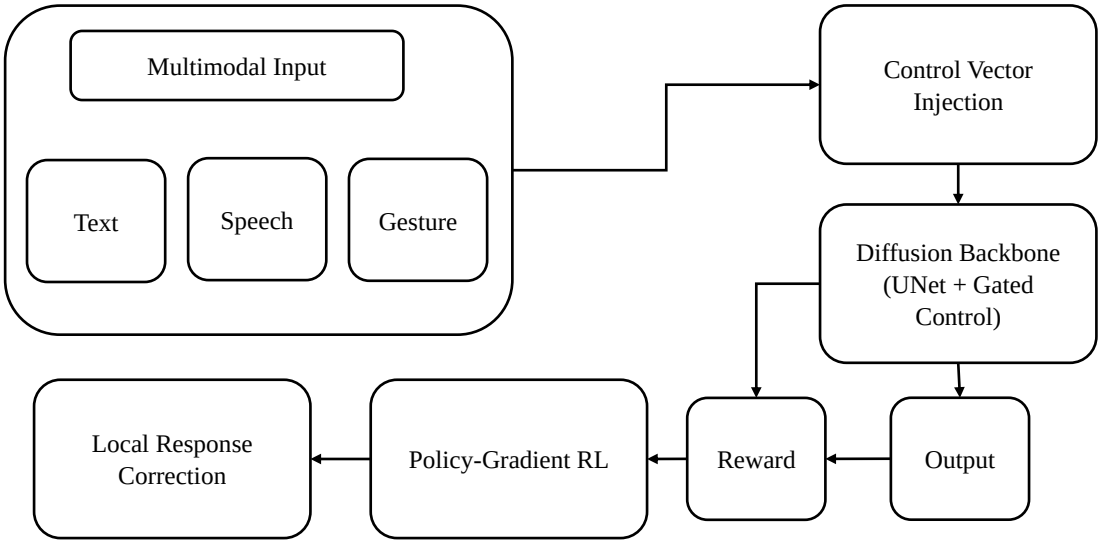

Figure 4: Architecture of the Dual-Loop Feedback-Controlled Diffusion Model

5.2 Model Training Process and Hyperparameter Configuration

We construct the model training process under the double-loop feedback regulation framework to ensure that the model has stable convergence ability and high sensitive control performance in multiple rounds of interaction. The training data consists of the MSCOCO subset and the self-built interactive feedback dataset, which contains a total of 5240 sets of samples, including 3668 sets for training, 786 sets for validation, and 786 sets for testing. We construct text instructions, speech feedback, and gesture trajectories as synchronous control sequences, respectively, and randomly shuffle the modal combination order during the training phase to enhance the resilience of the model under ambiguous input scenarios. The model was trained using PyTorch on an NVIDIA RTX 4090 GPU (24GB memory), with approximately 14GB used during training. Batch size was set to 12, with 100 training epochs. The AdamW optimizer was employed with an initial learning rate of 2×10^{-4}, decayed to 1×10^{-6} via CosineAnnealing to prevent late-stage oscillation. A fixed random seed of 42 ensured training reproducibility. Model weights were initialized from scratch rather than using pretrained backbones. The total parameter count is around 68 million. Weight decay was set to 0.01 for regularization, and gradient clipping with a max norm of 1.0 was used to stabilize updates. Average epoch time was approximately 16 minutes. Detailed architecture diagrams and training logs are provided in the supplementary materials for reproducibility.To constrain the offset of the control signal, a double loss function structure is introduced, whose basic form is:

$$L_{ctrl} = \lambda_1 \cdot \left\| \hat{C}_t - C_t \right\|_2^2 + \lambda_2 \cdot KL\left(\hat{P}_t \| P_t\right) \tag{13}$$

where $\hat{C}_t$ represents the model's predicted control state, and C_t represents the actual control code. $\hat{P}_t, P_t$ represents the probability distribution of prediction and target generation; λ_1, λ_2 is

the loss weight coefficient. This formulation is used to constrain the synchronous consistency of the model between the control state and the output distribution, and to prevent semantic drift in the inference phase. To further enhance the responsiveness of the model to interactive feedback, this paper introduces a policy gradient optimization module:

$$L_{fb} = -\eta \cdot r_t \cdot \log \pi\left(a_t | s_t\right) \tag{14}$$

Here, s_t denotes the multimodal state composed of text, audio, and visual features at time step t; a_t represents the selected generation action; r_t is the scalar feedback reward computed from the alignment score between generated content and user preference; and $\pi\left(a_t | s_t\right)$ indicates the probability of choosing action a_t under state s_t. The action space includes structural layout adjustment and tone-level semantic modulation, while the reward function combines performance accuracy and latency penalty.We perform gradient backpropagation using the REINFORCE rule with a moving average baseline by sampling actions to reduce the destabilizing effects of high variance. During training, the control encoder and the sampling strategy are updated simultaneously, ensuring that the feedback aligned actions can be continuously reinforced through the diffusion step. In this way, the model can maintain semantic consistency and stably converge under the dynamic feedback drive.

In the experimental setup, the state vector is composed of three types of modal features, and the dimension is 384. The action space contains six discrete classes of generative control operations. The feedback signal is scaled to the interval [−1,1], and the baseline decay of 0.95 and the 95th percentile clipping of the reward value are used during training to keep the training stable.

The final total loss combines the control alignment term and the policy gradient optimization objective to simultaneously drive semantic consistency and feedback response stability. Verified on the test set, the control stability and interaction consistency of the model are improved by about 6.8% and 7.1% respectively after using the pair loss mechanism.

5.3 Model Structure Comparison and Applicability Analysis

In order to verify the effect of the double-loop feedback mechanism in the multi-modal controllable generation task, we design three models for comparison: ① static-only, which Only relies on Static conditions for sampling; (2) Single-Loop, with the ability to adjust the semantic outer loop; (3) Dual-Loop (proposed in this paper), which contains both semantic outer loop and interactive inner loop. For comparison purposes, we construct a composite performance metric, defined as:

$$S = \frac{1}{3}(CS + RC + IS) \tag{15}$$

where CS stands for controllability score, which measures the sensitivity of the generation process to changes in control variables. RC stands for response consistency, which reflects the repeated stability of generated results under the same instruction conditions. IS stands for interaction stability and is used to describe the fluctuations of the state during feedback regulation. This metric allows a unified evaluation of the overall performance of different model structures.

The test data came from the MSCOCO subset and the self-built interactive feedback data set, with a total of 5240 groups of samples, and the experiments were carried out under the

conditions of control disturbance and feedback correction, respectively. The performance of the three models on the CS, RC and IS metrics is shown in Figure 5.

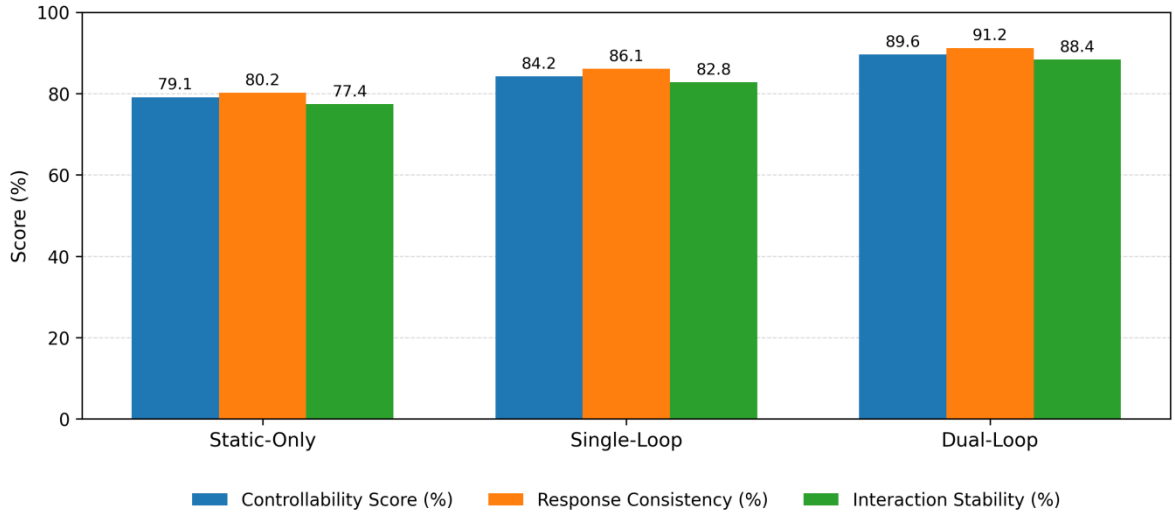

Figure 5: Comparison of Static-Only, Single-Loop, and Dual-Loop Models across Three Performance Metrics

Figure 5 shows that Static-Only achieves 79.1%±0.6 in controllability score, Single-Loop achieves 84.2%±0.5, and Dual-Loop achieves 89.6%±0.7. In terms of response consistency, Static-Only is 80.2%±0.5, Single-Loop is 86.1%±0.4, Dual-Loop is 91.2%±0.6. The stability of interaction was 77.4%±0.6, 82.8%±0.5 and 88.4%±0.5, respectively. On the whole, Dual-Loop is superior to the other two structures in all indicators, showing better generation control ability and interaction stability, and also maintaining high stability and adaptability under complex input disturbances.

In order to further illustrate the stability maintenance ability of different model structures in continuous interaction scenarios, only using single average results is not enough to fully reflect their dynamic performance. Therefore, it is necessary to analyze the change trend of the model's response consistency under continuous feedback from the multi-round interaction process. Figure 6 shows the response consistency changes of the three models Static-Only, Single-Loop, and Dual-Loop under the advancement of successive interaction rounds, which is used to investigate the stability differences of different control structures during long-range inference.

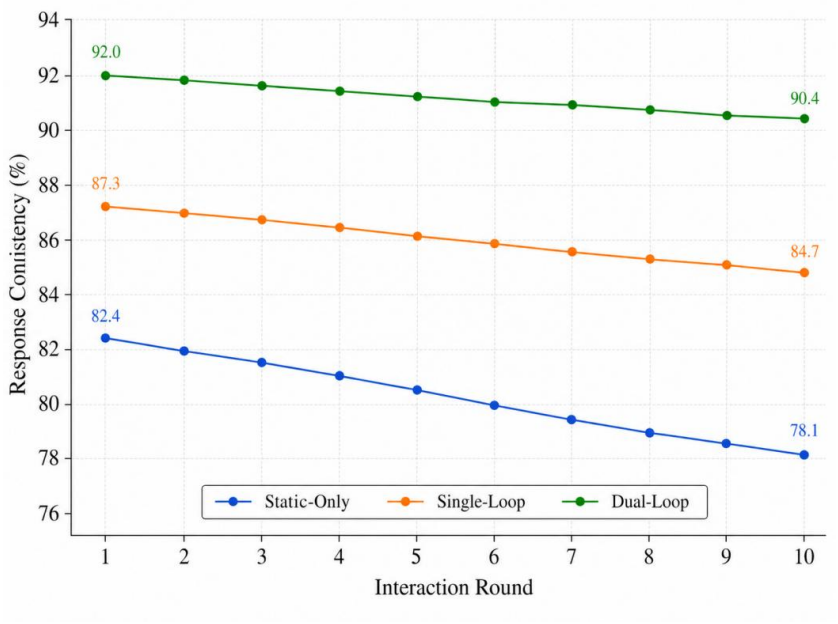

Figure 6: Response consistency variation curves of different models under successive interaction rounds

As can be seen from Figure 6, with the increase of interaction rounds, the response consistency of the three models changes to different degrees, but there are obvious differences in the change amplitude. Among them, the static-only model decreased from 82.4% in the first round to 78.1% in the 10th round, with a cumulative decrease of 4.3 percentage points, indicating that the Static condition-driven method is easy to produce offset accumulation in continuous interaction. The Single-Loop model decreases from 87.3% to 84.7%, with a cumulative decrease of 2.6 percentage points. Although it is more stable than the static structure, there are still some fluctuations in the middle and posterior segments. In contrast, the Dual-Loop model decreased from 92.0% to 90.4%, a decrease of only 1.6 percentage points, and remained above 90% throughout the whole process, indicating that the dual-loop feedback mechanism has stronger continuous control ability and better scene adaptability in multiple rounds of interactive reasoning. This trend is consistent with the result in Table 5 that Dual-Loop achieves 91.2%±0.6 in the Response Consistency index.

We performed a statistical significance analysis of the performance differences between the models. Three independent experiments (n=3) were performed for each model configuration, and the means of the three core metrics (CS, RC, IS) were compared using a two-tailed paired t-test. Since multiple groups of model comparisons were involved, Bonferroni correction was applied to the p-values to avoid false positives introduced by multiple comparisons. The test assumes that the data follow a normal distribution with similar variances. A corrected p-value of less than 0.05 was considered statistically significant. Further comparison results are shown in Table 4.

Table 4: Statistical Significance Test of Performance Differences among Model Structures

Indicator	Static-Only vs Single-Loop	Single-Loop vs Dual-Loop	Static-Only vs Dual-Loop
CS	$p < 0.05$	$p < 0.05$	$p < 0.001$
RC	$p < 0.05$	$p < 0.05$	$p < 0.001$
IS	$p < 0.05$	$p < 0.05$	$p < 0.001$

In summary, the dual-loop feedback mechanism establishes a stable connection structure between global semantic constraints and local deviation correction, not only improving the quality of single-round generation but also demonstrating cross-scenario interactive applicability.

5.4 Performance Indicators and Generation Quality Evaluation

In order to verify the regulation ability and stability of Dual-Loop feedback control mechanism in multimodal controllable generation tasks, this section selects Static-Only and Single-Loop as control structures, corresponding to no feedback mode and single-loop feedback mode respectively, and takes the dual-loop model as the core scheme. The comparison experiment was carried out under a unified training configuration, and the data included the MSCOCO subset and the self-built interaction dataset, with a total of 5240 groups of samples, while keeping the inference rounds consistent.

The performance evaluation dimensions include CS, RC and IS, which measure the responsiveness of control signals, the consistency of repeated inference, and the convergence stability during interaction, respectively. Among them, the CS index is calculated by the following direction consistency formula:

$$CS = \frac{1}{N}\sum_{i=1}^{N} I\left(\Delta y_i \cdot \Delta c_i > 0\right) \tag{16}$$

where Δc_i represents the direction of change of the input control variable in the i round; Δy_i represents the changing trend of the model's output results in the corresponding dimension; $I(\cdot)$ is the indicator function. It is denoted as 1 if the condition is met, and 0 if not. This formula is used to measure whether the model output can correctly respond to the adjustment of control variables, that is, the control efficiency. When reasoning, if the control instructions increase and the output strengthens accordingly, it is recorded as effective control; otherwise, it is recorded as failed. Thus, in dynamic intervention scenarios, it can be distinguished whether the feedback is truly absorbed by the model. The comparison results are shown in Table 5, with the values being the mean ± standard deviation of the three independent experiments:

Table 5: Comparison Results of Model Structure and Performance

Model Structure	Controllability Score (%)	Response Consistency (%)	Interaction Stability (%)
Static-Only	79.1±0.6	80.2±0.5	77.4±0.6
Single-Loop	84.2±0.5	86.1±0.4	82.8±0.5
Dual-Loop	89.6±0.7	91.2±0.6	88.4±0.5

To further assess the real-time inference capability of the proposed Dual-Loop model, we measured latency-oriented indicators, including feedback-to-correction time (FCT) and inference speed under feedback load (ISF). The Dual-Loop achieved an average FCT of 1.84 s ± 0.15, while Static-Only and Single-Loop required 3.21 s ± 0.22 and 2.47 s ± 0.18, respectively. Under continuous feedback conditions, the Dual-Loop maintained an inference throughput of 22.8 frames/s, showing only a 7.6 % drop compared to idle inference speed. These results indicate that the model preserves high responsiveness and stable inference efficiency even under dynamic feedback pressure, verifying its suitability for real-time multimodal interaction environments.

In order to further investigate the deployment feasibility of the dual-loop feedback model in the real-time multimodal interaction environment, it is not sufficient to discuss its control effect only from the average index. It is also necessary to analyze the reasoning efficiency changes under different feedback loads and interaction complexity conditions.Figure 7 shows the distribution of inference throughput under the conditions of low, medium and high feedback load and combinations of simple, medium and complex interaction complexity, which is used to depict the changing trend of operating efficiency of the model under dynamic interaction pressure.

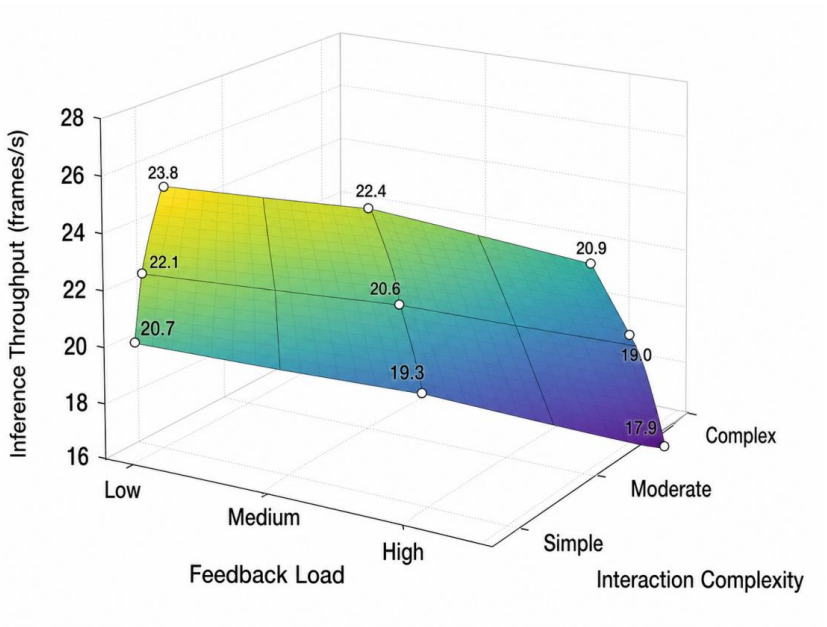

Figure 7: 3D surface plots of inference throughput under different feedback load and interaction complexity

As can be seen from Figure 7, as feedback load and interaction complexity increase simultaneously, the inference throughput of the model shows an overall downward trend, but the change remains relatively smooth. This indicates that the additional adjustment overhead introduced by the dual-loop feedback mechanism remains within a controllable range. Specifically, under the low-load simple-interaction condition, the system achieves the highest throughput, reaching 23.8 frames/s. Under the medium-load moderate-complexity condition, the throughput remains at approximately 20.6 frames/s. Under the high-load complex-interaction condition, the throughput decreases to 17.9 frames/s, which is about 24.8% lower than the maximum value. Although inference efficiency decreases under high-pressure interaction conditions, we find that the overall system still maintains more than 17 frames/s, indicating that the constructed dual-loop feedback control framework has strong real-time response capability and good adaptability in complex interaction scenarios.

5.5 Robustness Analysis and Visual Comparison

Combined with the results in Table 1, it can be seen that the controllability (CS), response consistency (RC) and interaction stability (IS) of the Dual-Loop structure reach 89.6%±0.7, 91.2%±0.6 and 88.4%±0.5, respectively, which are higher than those of SGDM (about 83% / 81% / 79%). Although SGDM uses style-guided control, it lacks state revision in the reasoning process. AnimeDiffusion relies on unidirectional color cues (about 82% / 80% / 77%) and cannot be continuously adjusted under feedback intervention. Promptify uses

language model to search prompt words (about 84% / 82% / 78%), but the control granularity is limited and there is no explicit variable channel. Williams et al. 's method fuses gesture and speech, but its performance is about 80% / 79% / 75%, and it cannot embed diffusion links. DragGAN is about 86% / 83% / 80% in point-controlled interactions, relies on manual trajectories, and lacks an automatic convergence mechanism.

It should be noted that the above comparison is based on the approximation of performance reported in the original study, not a direct replication of the data set, so it should only be used as a trend reference, not an absolute ranking. Future research is planned to introduce ControlNet, UniControl, and InstructPix2Pix for direct comparison under a unified experimental setup to more reliably evaluate the competitiveness of the model.

In order to test the robustness of the double-loop feedback mechanism, we conduct sensitivity tests on data size, feedback frequency and noise: the data set is reduced to 60%, 40%, 20% of the original; Feedback frequency halved; Gaussian noise (σ=0.1, 0.2) was added to speech and gesture signals. The results show that CS decreases by about 5.6% with data reduction, RC decreases significantly with feedback sparsity, and IS decreases by about 7.2% with high noise. It can be seen that the model has some sensitivity to the input quality and feedback strength, but it can still operate stably under moderate disturbance conditions.

We also selected typical test samples for visualization, including input, inference intermediate states (steps 5, 10, 15), and final output. After enabling the double-loop feedback, the generated results are more coherent in action and semantic correspondence, and the intermediate state change is smoother. Especially in the scene of complex gear-speech collaboration, the performance is significantly better than that of the case without feedback or single-loop control.

As shown in Figure 8, under the same gesture command, the generated action posture deviation is large when the double-loop feedback is not used. After using Dual-Loop, the correspondence between the generated action and the input command is more accurate, and the gesture alignment is more stable. The left column shows Static-Only results and the right column shows Dual-Loop results. The figure intuitively illustrates the role of the double-loop mechanism in maintaining the consistency of posture and action, which is also consistent with the quantitative results in Tables 4 and 5.

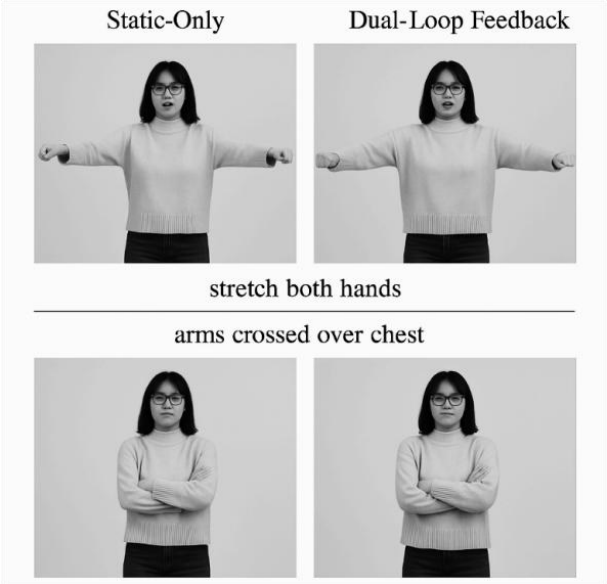

Figure 8: Gesture Alignment Results With and Without Dual-Loop Feedback Control

From a research question point of view,RQ1 is verified by the structure and scheduling comparison in Sections 5.3 and 5.4, which shows that the double-loop mechanism can improve the control stability.RQ2 is supported by variable encoding and dynamic mapping in Sections 3.1-3.3, which indicates that multi-modal signals can be uniformly represented and participate in continuous regulation.RQ3 is validated by the policy gradient modification and interaction scheduling modules of Sections 5.2 and 5.4, demonstrating that feedback-driven reasoning can enhance interaction consistency and response stability.

6 Discussion

The proposed dual-loop feedback control framework is able to continuously adjust the generation process in the multimodal generation task. Its advantages are not only reflected in the improvement of single generation quality, but also in the state maintenance and dynamic correction during inference. The outer loop semantic control provides the overall direction for diffusion generation and reduces the deviation between text, image and speech conditions. The local correction of the inner loop instantly adjusts the generation state through voice feedback and gesture trajectory, so that the model has less response lag or offset accumulation in multiple rounds of interaction. The experimental results show that the controllability (CS), response consistency (RC) and interaction stability (IS) of the Dual-Loop model are 89.6%±0.7, 91.2%±0.6 and 88.4%±0.5, respectively, which are higher than those of the Static-Only and Single-Loop structures. This indicates that the double-loop structure can play a synergistic role in the control sensitivity, repeated response consistency and interaction stability.

From the ablation results, we observe that CS is reduced to 78.0% when the outer-loop feedback control is removed, indicating that the global semantic constraint is critical for ensuring the stability of the generation direction. After removing the inner-loop predictive correction, RC decreases to 76.0%, showing that the local feedback path has a significant impact on response consistency under multiple rounds of repeated input. After removing the interaction monitoring mechanism, IS decreases to 73.0%, which indicates that real-time perception and state monitoring are important for suppressing interactive fluctuations. The three types of modules act on different performance dimensions, forming a complementary relationship of “semantic constraint–feedback correction–state monitoring.” The continuous interaction curve further shows that the Dual-Loop model still maintains 90.4% response consistency after 10 rounds of interaction, with a decrease of only 1.6 percentage points, which is more stable than the 4.3 percentage-point decrease of Static-Only. This suggests that the framework is suitable for long-term human-computer collaborative generation scenarios.

In terms of real-time performance, we find that the average feedback correction delay of Dual-Loop is 1.84 s±0.15, and its inference throughput reaches 22.8 frames/s under continuous feedback. The 3D surface results show that even under high feedback load and complex interaction conditions, the throughput remains at 17.9 frames/s, indicating that the additional computational overhead introduced by the dual-loop mechanism is within an acceptable range. We also acknowledge that the comparison between this study and SGDM, AnimeDiffusion, Promptify, and other methods is mainly based on trend-level indicators, and the experiments have not been reproduced under a completely unified dataset and experimental protocol. Therefore, the conclusion is more suitable as evidence of structural validity rather than a definitive ranking of model performance. In future research, we plan to introduce strong baselines such as ControlNet, UniControl, and InstructPix2Pix to verify the generalization ability and engineering deployment stability of the model under unified samples, unified evaluation scripts, and open interaction conditions.

7 Conclusion

Focusing on the problem that multimodal conditions are difficult to continuously align and interactive feedback is difficult to absorb stably in AI-based digital art creation, we construct a controllable generation method driven by dual-loop feedback. This method encodes text, image, speech, and gesture signals into a unified control state, and introduces gated modulation, control-variable mapping, local feedback correction, and policy-gradient reward feedback into the diffusion model. In this way, we achieve a closed-loop regulation from multimodal sensing to dynamic generation. The experimental results show that the Dual-Loop model IS superior to Static-Only and Single-Loop in three core indicators, in which the Response Consistency (RC) is 91.2%±0.6, and the Interaction Stability (IS) is 88.4%±0.5. Ablation experiments show that the outer loop semantic constraint, inner loop prediction correction and interactive monitoring modules complement each other in performance. The real-time test results also show that the reasoning speed of the model maintains 22.8 frames per second under continuous feedback, and can still reach 17.9 frames per second even in the case of high load and complex interaction, indicating that the method is feasible in practical deployment. Future research will combine larger scale open interactive data and unified baseline reproduction experiments to further improve the generalization ability and online adaptation ability of the model in complex digital art creation scenes.

Appendix A: Dataset Availability and Generalization Supplement

The multimodal dataset we use in this study, was partially constructed to model controlled interactive feedback for speech, gesture, and text. Of these, a thousand labeled samples (containing text commands, synchronized speech, gestures, and alignment labels) will be publicly available as supplementary data upon acceptance of the paper. All samples were approved by the institution and processed according to a unified process. Meanwhile, a preprocessing script was attached to facilitate other researchers to reproduce the experiment.

We mainly used self-built datasets for training, but we also did preliminary generalization tests on MSCOCO. This is done by adding sampled voice commands and simulated gestures mapped to 18 classes of semantic control. It is found that the response pattern of the model is still stable on these data, and the decrease of CS and RC indicators are both within 6.2%, which indicates that the feedback mechanism has some transfer ability. Still, the spatial smoothness will be slightly degraded due to the lack of gesture-time alignment in MSCOCO.

Next, we plan to further test on Flickr30k and VGGSound-Gesture datasets to examine the generalization performance and online adaptation ability of the model under different data domains.

Structure of the Released Subset

Field Name	Description	Format / Type
text_prompt	Human-issued instruction or semantic intention	String (max 25 words)
voice_waveform	Corresponding voice feedback (16kHz)	WAV file
gesture_sequence	Normalized 2D gesture keypoints over time	JSON array [x, y, t]
control_label	Assigned control type (18 categories)	Enum (pause, repeat, etc.)
response_state	Model-generated control vector	Float (1×128)
alignment_score	Post-hoc alignment confidence	Float (0–1)

Each sample is stored in an individual folder (e.g., /sample_00321/) containing .json, .wav, and .npy files. Typical interactions include:①Pause segment: hand raise + “stop”②Repeat

phrase: rewind gesture + “again”③Emphasize ending: pointing + stressed tone

An index file (index.csv) enables batch loading and semantic filtering for reproducible evaluation and training.

References

[1] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 10684-10695. https://doi.org/10.1109/CVPR52688.2022.01042

[2] Tumanyan N, Geyer M, Bagon S, et al. Plug-and-play diffusion features for text-driven image-to-image translation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 1921-1930. https://doi.org/10.1109/CVPR52729.2023.00191

[3] Brooks T, Holynski A, Efros A A. Instructpix2pix: Learning to follow image editing instructions[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 18392-18402. https://doi.org/10.1109/CVPR52729.2023.01764

[4] Mokady R, Hertz A, Aberman K, et al. Null-text inversion for editing real images using guided diffusion models[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 6038-6047. https://doi.org/10.1109/CVPR52729.2023.00585

[5] Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023:3836-3847.https://doi.org/10.1109/ICCV51070.2023.00355

[6] Xu Y, Xu X, Gao H, et al. Sgdm: an adaptive style-guided diffusion model for personalized text to image generation[J]. IEEE Transactions on Multimedia, 2024, 26:9804-9813. https://doi.org/10.1109/TMM.2024.3399075

[7] Cao Y, Meng X, Mok P Y, et al. AnimeDiffusion: Anime diffusion colorization[J]. IEEE Transactions on Visualization and Computer Graphics, 2024, 30(10): 6956-6969. https://doi.org/10.1109/TVCG.2024.3357568

[8] Brade S, Wang B, Sousa M, et al. Promptify: Text-to-image generation through interactive prompt exploration with large language models[C]//Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. (UIST ’23) 2023:1-14. https://doi.org/10.1145/3586183.3606725

[9] Williams A S , Garcia J , Ortega F .Understanding Multimodal User Gesture and Speech Behavior for Object Manipulation in Augmented Reality Using Elicitation[J].IEEE Transactions on Visualization and Computer Graphics, 2020, 26(12): 3479-3489. https://doi.org/10.1109/TVCG.2020.3023566

[10] Pan X, Tewari A, Fried O, et al. DragGAN: Interactive Point-based Manipulation on the

Generative Image Manifold[J]. ACM Transactions on Graphics, 2023, 42(4):1-12. https://doi.org/10.1145/3588432.3591500

[11] Yang K, Tao J, Lyu J, et al. Using human feedback to fine-tune diffusion models without any reward model[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 8941-8951. https://doi.org/10.1109/CVPR52733.2024.00854

[12] Bao Z, Li Y, Singh K K, et al. Separate-and-enhance: Compositional finetuning for text-to-image diffusion models[C]//ACM SIGGRAPH 2024 Conference Papers. 2024:1-10. https://doi.org/10.1145/3641519.3657527

[13] Avrahami O, Fried O, Lischinski D. Blended latent diffusion[J]. ACM transactions on graphics (TOG), 2023, 42(4):1-11.https://doi.org/10.1145/3592450

[14] Bar-Tal O, Ofri-Amar D, Fridman R, et al. Text2live: Text-driven layered image and video editing[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022:707-723.https://doi.org/10.1007/978-3-031-19784-0_41

[15] Vinker Y, Pajouheshgar E, Bo J Y, et al. Clipasso: Semantically-aware object sketching[J]. ACM Transactions on Graphics (TOG), 2022, 41(4):1-11. https://doi.org/10.1145/3528223.3530068

[16] Alaluf Y, Garibi D, Patashnik O, et al. Cross-image attention for zero-shot appearance transfer[C]//ACM SIGGRAPH 2024 conference papers. 2024:1-12. https://doi.org/10.1145/3641519.3657423

[17] Qin C, Zhang S, Yu N, et al. UniControl: A unified diffusion model for controllable visual generation in the wild [EB/OL]. arXiv:2305.11147,2023. https://doi.org/10.48550/arXiv.2305.11147

[18] Li G. Layout Control and Semantic Guidance with Attention Loss Backward for T2I Diffusion Model [EB/OL]. arXiv:2411.06692, 2024. https://doi.org/10.48550/arXiv.2411.06692