Talk2Image: A Multi-Agent System for Multi-Turn Image Generation and Editing

Shichao Ma, Yunhe Guo, Jiahao Su, Qihe Huang, Zhengyang Zhou*, Yang Wang*

University of Science and Technology of China

Abstract

Text-to-image generation tasks have driven remarkable advances in diverse media applications, yet most focus on single-turn scenarios and struggle with iterative, multi-turn creative tasks. Recent dialogue-based systems attempt to bridge this gap, but their single-agent, sequential paradigm often causes intention drift and incoherent edits. To address these limitations, we present **Talk2Image**, a novel multi-agent system for interactive image generation and editing in multi-turn dialogue scenarios. Our approach integrates three key components: intention parsing from dialogue history, task decomposition and collaborative execution across specialized agents, and feedback-driven refinement based on a multi-view evaluation mechanism. Talk2Image enables step-by-step alignment with user intention and consistent image editing. Experiments demonstrate that Talk2Image outperforms existing baselines in controllability, coherence, and user satisfaction across iterative image generation and editing tasks.

Introduction

Recent years have seen remarkable progress in text-to-image (T2I) generative models, especially diffusion-based approaches (Ho, Jain, and Abbeel 2020; Podell et al. 2023), which produce high-quality and diverse images from concise textual prompts. This has driven widespread adoption in art creation, graphic design, advertising, and other domains. Meanwhile, vibrant open-source communities like HuggingFace (Face 2018), Civitai (Civitai 2022), and OpenArt (OpenArt 2021) have accelerated the sharing of models and workflows, broadening the selection of models available to users.

Despite these advances, most T2I models still lack mechanisms for dynamic interactivity and multi-turn control. Users often find it challenging to iteratively refine complex visual intentions through natural language, limiting the progressive articulation of creative goals (Ma et al. 2025). Recent dialogue-based systems, such as GenArtist (Wang et al. 2024a), DialogDraw (Ma et al. 2025), and DialogGen (Huang et al. 2024), integrate large language models (LLMs) to close the loop between user intention and image generation. However, they typically adopt a single-agent, sequential paradigm, which limits modularity collaboration and often leads to *intention drift* (misalignment with cumulative user goals) and *incoherent edits* (visual inconsistencies across iterations) in complex tasks.

Multi-agent systems (MAS) offer a promising alternative to address these limitations through task decomposition and collaboration (Calegari et al. 2021; Cardoso and Ferrando 2021). With the growing availability of specialized generative models on platforms like Civitai, a natural question arises: ***Can we leverage these resources to develop a MAS framework supporting sustained, multi-turn image generation and editing?*** To turn this vision into reality, three key challenges must be addressed: (1) accurately parsing user intention and transforming it into structured, executable prompts; (2) decomposing complex tasks into diverse subtasks (e.g., object addition, style modification) and coordinating their execution via multi-agent collaboration; and (3) refining outputs through iterative feedback to ensure semantic alignment and visual quality.

To tackle these challenges, we propose **Talk2Image**, a novel multi-agent system for multi-turn interactive image generation and editing. The system comprises three core components that address key limitations of existing frameworks: (1) a dynamic intention parsing module that synthesizes structured prompts based on dialogue history, directly mitigating intention drift; (2) a modular agent collaboration mechanism enabling task decomposition and specialized execution, with a Directed Acyclic Graph ensuring consistency and valid execution for coherent edits across iterations; and (3) a multi-view evaluation and refinement loop that maintains step-by-step alignment with user intention.

Our contributions are summarized as follows:

- We introduce **Talk2Image**, the first multi-agent system tailored for image generation and editing in multi-turn dialogues.
- We design a three-stage architecture encompassing intention parsing, multi-agent collaboration, and a closed-loop feedback mechanism.
- We demonstrate that Talk2Image significantly improves controllability, coherence, and user satisfaction across complex iterative image editing tasks.

*Corresponding Author.

Related Work

Image Generation Models and Ecosystems

Diffusion models (Song, Meng, and Ermon 2020) have become the cornerstone of image generation, enabling high-quality samples through iterative noising and denoising. For text-to-image (T2I) tasks, key frameworks include DALL·E 2 (Ramesh et al. 2022) and Stable Diffusion (Rombach et al. 2022a), which use CLIP (Radford et al. 2021) text encoders to condition latent-space generation, and Imagen (Saharia et al. 2022), which leverages T5 (Raffel et al. 2020) with cross-attention for textual guidance. Advances in editing capabilities include ControlNet (Zhang, Rao, and Agrawala 2023) for spatial condition control, Imagic (Kawar et al. 2023) for semantic editing via optimization, and DiffEdit (Couairon et al. 2022) for automatic region inference. Open-source ecosystems such as HuggingFace, which provides pre-trained models and APIs, and community platforms like Openart and Civitai have accelerated adoption by facilitating workflow sharing. However, these tools still lack robust mechanisms for fine-grained control and multi-turn interaction, limiting their applicability to complex iterative tasks.

Dialogue-based Image Generation and Editing

Dialogue-based image generation and editing systems enable intuitive visual content creation through natural language. Many of these systems use multimodal large language models (MLLMs) to handle both dialogue understanding and image generation/editing tasks. For example, DialogGen (Huang et al. 2025) uses a single MLLM, such as Qwen-VL (Bai et al. 2023), for multi-round text-to-image generation, where the MLLM comprehends user input and collaborates with text-to-image models. Similarly, GenArtist (Wang et al. 2024b) integrates multiple models into a unified framework for image generation and editing, with an MLLM orchestrating task decomposition, tool selection, and execution. However, these single-agent systems rely on a strictly sequential processing approach, where one MLLM manages both dialogue and image tasks, lacking modularity and effective collaboration.

Multi-Agent Systems

Recent advancements in multi-agent systems (MAS) have enhanced task planning, language collaboration, and multimodal understanding, with frameworks like AutoGPT (Yang, Yue, and He 2023) and CAMEL (Li et al. 2023) now incorporating basic support for image generation and visual inputs. Despite these strides, such frameworks remain underoptimized for the unique demands of visual generation and editing tasks. This gap highlights a critical challenge: developing comprehensive multi-agent systems that integrate dynamic labor division, iterative feedback mechanisms, and multimodal perception capabilities specifically tailored to image generation and editing workflows.

Methodology

Talk2Image is a multi-agent system enabling interactive image generation and editing via natural language dialogue. It addresses three core challenges: (1) tracking user intention across turns to mitigate drift; (2) hierarchically decomposing complex tasks into structured subtasks, coordinated by specialized agents to ensure edit coherence; (3) refining outputs through feedback-aware optimization to maintain semantic alignment. The workflow (Figure 1) and its formalization in Algorithm 1 implement this hierarchical decomposition, multi-level scheduling, and closed-loop refinement.

Dialogue-based Intention Parsing

To parse user intention and mitigate intention drift in multi-turn dialogue, Talk2Image employs a four-stage intention parsing pipeline that systematically preserves, normalizes, and structures cumulative user intention. This design addresses historical information loss, linguistic variance, and unstructured representation, which are key drivers of misalignment across turns.

Dialogue State Memory At each turn t, the system maintains an evolving interaction history:

$$\mathcal{H}_t = \{(u_1, r_1), (u_2, r_2), \ldots, (u_{t-1}, r_{t-1})\} \tag{1}$$

where u_i and r_i denote the user's instruction and the system's response, respectively. The prompt synthesizer f_θ fuses the current user input u_t with history to generate a textual state summary:

$$P_t^{\text{text}} = f_\theta(\mathcal{H}_t, u_t) \tag{2}$$

The resulting prompt captures the cumulative scene description, resolving user intention revisions such as additions, deletions, or substitutions across turns.

Linguistic Normalization To reduce variance in user phrasing, a canonicalization layer $g(\cdot)$ is applied:

$$P_t^{\text{norm}} = g(P_t^{\text{text}}) \tag{3}$$

This rule-based rewriter eliminates redundancy (e.g., duplicated objects), standardizes phrasing, and clarifies spatial relations, serving as a reliable precursor for structured parsing.

Key Information Extraction The normalized prompt is parsed into structured fields by a semantic extractor $h(\cdot)$:

$$I_t = h(P_t^{\text{norm}}) = (\mathbf{s}, \mathbf{a}, \mathbf{b}, \mathbf{t}) \tag{4}$$

where:

- $\mathbf{s} \in$ *SUBJ*: subject entities with optional spatial anchors (e.g., "a cat on the left").
- $\mathbf{a} \in$ *ATTR*: attribute features (e.g., "white", "fluffy").
- $\mathbf{b} \in$ *SCENE*: background scene description (e.g., "grassy field under sunlight").
- $\mathbf{t} \in$ *T*: disallowed or excluded elements (e.g., "right dog", "green color").

This intention structure I_t supports downstream modular execution and backward editing.

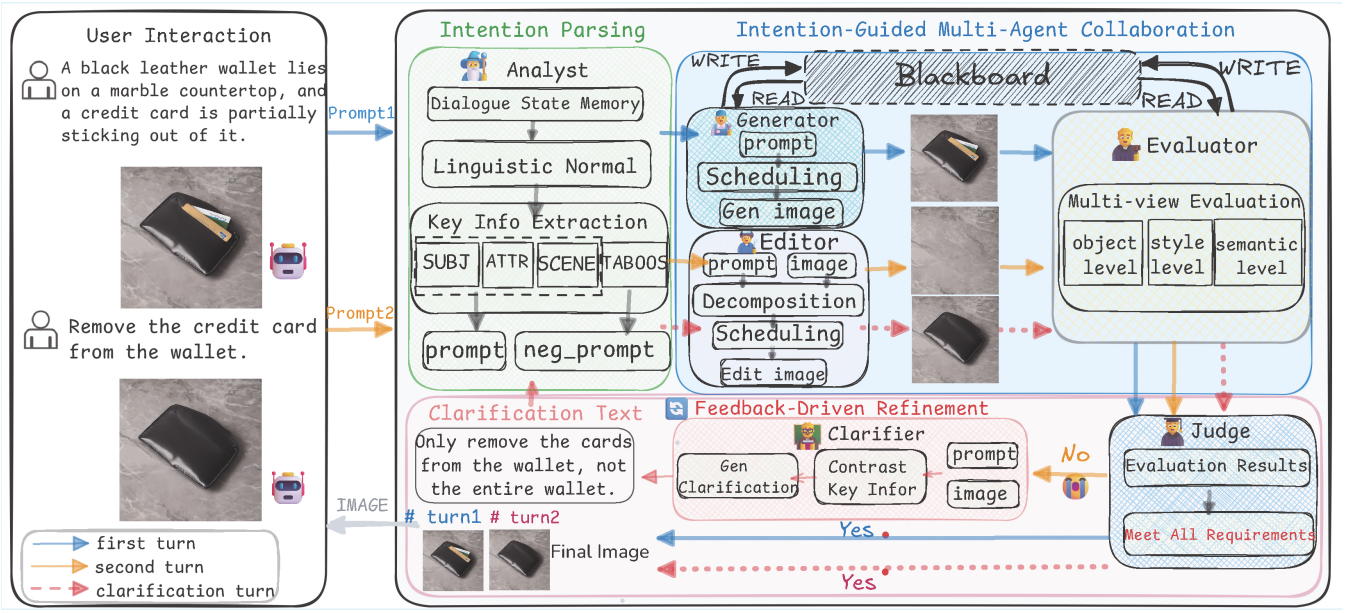

Figure 1: Overview of Talk2Image, depicting its closed-loop pipeline: intention parsing → hierarchical task decomposition → multi-agent scheduling & execution → feedback-driven refinement.

Prompt Construction via Template Emission To convert the structured intention I_t into textual directives suitable for generation and editing, the prompt emitter $\Phi(\cdot)$ produces three complementary prompt components:

$$\Phi(I_t) := \begin{cases} \mathrm{p}_{\mathrm{pos},t} = \Phi^{+}(\mathbf{s}, \mathbf{a}, \mathbf{b}) \in \mathcal{G}_{\mathrm{pos}}, \\ \mathrm{p}_{\mathrm{neg},t} = \Phi^{-}(\mathbf{t}) \in \mathcal{G}_{\mathrm{neg}}, \\ \mathrm{p}_{\mathrm{scene},t} = \Phi^{\mathrm{scene}}(\mathbf{b}) \in \mathcal{G}_{\mathrm{scene}}. \end{cases} \tag{5}$$

Each mapping Φ^{*} instantiates symbolic slots using deterministic prompt templates drawn from predefined grammar domains:

$$\mathcal{G}_{\mathrm{pos}} ::= \texttt{SUBJ with ATTR in SCENE} \tag{6}$$

$$\mathcal{G}_{\mathrm{neg}} ::= \texttt{List(ExcludedItem)} \tag{7}$$

$$\mathcal{G}_{\mathrm{scene}} ::= \texttt{SCENE} \quad \text{(no subject placeholders)} \tag{8}$$

This template-based emission ensures consistent formatting, stable spatial grounding, and modular control over desired content, excluded elements, and background context, enabling interpretable prompts at each dialogue turn.

Intention-Guided Multi-Agent Collaboration

Talk2Image employs a heterogeneous multi-agent architecture to enable compositional visual reasoning, integrating three core components: hierarchical task decomposition, multi-level scheduling, and blackboard-based communication. These components collectively support robust multi-turn, feedback-driven visual generation and editing.

Hierarchical Task Decomposition To transform complex user intentions into executable actions, the structured intention $\mathcal{G}_t$ is decomposed into a hierarchy of subgoals:

$$\mathcal{G}_t = \bigcup_{l=1}^{L} \mathcal{G}_t^{(l)}, \qquad L = 3, \tag{9}$$

where higher levels capture coarse-grained editing intents and lower levels specify atomic operations. Concretely, the hierarchy follows:

levels specify atomic operations. Concretely, the hierarchy follows:

- **Level 1:** Composite intentions (e.g., `Add`, `Replace`, `Remove`, `Move`);
- **Level 2:** Task-specific substeps (e.g., `Add` → `Localize` / `Generate` / `Fuse`);
- **Level 3:** Atomic operations executed by agents.

This structure enables granular control while preserving coherence across the entire editing workflow.

Multi-Level Scheduling Given the three-level task hierarchy, the system schedules execution in a corresponding three-stage manner: (1) enforcing cross-level dependencies, (2) assigning the appropriate agent to each subgoal, and (3) executing atomic operations in valid order.

(1) Task-Level Scheduling. Dependencies among high and mid-level subgoals (e.g., `Remove` before `Add`) are modeled using a directed acyclic graph (DAG):

$$G_t = (V_t, E_t), \tag{10}$$

where $V_t = \{v_g \mid g \in \mathcal{G}_t\}$ and edges E_t encode precedence $g \prec h$. The induced partial order satisfies

$$g \prec h \wedge h \prec k \Rightarrow g \prec k, \qquad \neg(g \prec g),$$

ensuring that higher-level intentions are expanded and scheduled in a coherent order.

(2) Agent-Level Scheduling. Each decomposed subgoal g is matched to an agent capable of executing its required

Algorithm 1: Hierarchical Multi-Agent Execution with Multi-View Feedback

Require: Dialogue history $\mathcal{H}$, user input u_t
Ensure: Final image $\hat{I}_{\text{final}}$

$I_t \leftarrow f_\theta(\mathcal{H}, u_t)$ ▷ Parse into structured intention
$retry \leftarrow 0$
repeat
  $\{\mathcal{G}_t^{(l)}\}_{l=1}^{L} \leftarrow \Pi(I_t)$ ▷ L1–L3 hierarchical goal decomposition
  $\mathcal{G}_t \leftarrow \bigcup_{l=1}^{L} \mathcal{G}_t^{(l)}$
  $G_t = (V_t, E_t) \leftarrow \texttt{BuildDAG}(\mathcal{G}_t)$ ▷ Task-level dependencies
  $\sigma_t \leftarrow \texttt{TopologicalSort}(G_t)$ ▷ Valid execution order
  $\hat{I}_t \leftarrow \texttt{Execute}(\sigma_t, \mathcal{B}_t)$ ▷ Agent-/execution-level scheduling via blackboard
  Compute $\mathcal{J}_{\text{obj}}, \mathcal{J}_{\text{style}}, \mathcal{J}_{\text{match}}$ ▷ Multi-view evaluation
  $s \leftarrow \mathcal{J}(\hat{I}_t, I_t)$ ▷ Aggregate feedback score
  if $s < \tau$ **then**
    $I_t \leftarrow A_{\text{clarify}}(\hat{I}_t, I_t)$ ▷ Refine intention via clarifier
    $retry \leftarrow retry + 1$
  end if
until $s \geq \tau$ **or** $retry \geq N_{\max}$
return $\hat{I}_t$ as $\hat{I}_{\text{final}}$

atomic operation. With capability vectors $\phi(a)$ and requirement vectors $\psi(g)$, the selected agent is:

$$a_g = \arg\max_{a \in \mathcal{A}} \texttt{Compat}(\phi(a), \psi(g)), \tag{11}$$

where the compatibility score is defined as a weighted cosine similarity:

$$\texttt{Compat}(x, y) = \frac{x^\top W y}{\|x\|_2 \, \|y\|_2}. \tag{12}$$

(3) Execution Scheduling. A topological sort of G_t yields an executable sequence $\sigma_t = [g_1, \ldots, g_n]$ that respects all dependencies. A subgoal is executed only when all required outputs are available:

$$\mathcal{S}(g, G_t, \mathcal{B}_t) = \begin{cases} a_g(\mathcal{O}_g^*) & \text{if } \mathcal{O}_g^* \subseteq \mathcal{B}_t, \\ \perp & \text{otherwise,} \end{cases} \tag{13}$$

where $\mathcal{O}_g^* = \{o_h \mid h \in \text{Pred}(g)\}$ and $\mathcal{B}_t$ is the blackboard state.

Together, these stages ensure that complex intentions (L1) are decomposed, assigned, and executed (L2→L3) in a coherent, dependency-aware manner.

Blackboard-based Collaborative Communication To support the execution of atomic operations (L3), agents share intermediate results through a persistent blackboard $\mathcal{B}_t$, represented as a tuple $\langle \mathcal{I}_t^*, \mathcal{O}_t^*, \mathcal{R}_t^* \rangle$ containing intention traces, agent outputs, and feedback records. Communication follows two primitive operations: *READ* ($\texttt{READ}(a, \mathcal{B}_t, \chi)$) for retrieving relevant information, and *WRITE* ($\texttt{WRITE}(a, \mathcal{B}_t, o)$) for appending new results.

The output of agent a on subgoal g is:

$$o_g = a(\texttt{READ}(a, \mathcal{B}_t, \chi_g) \cup \{g\}), \tag{14}$$

where $\chi_g(o) = (\exists h \in \text{Pred}(g) : o = o_h)$ ensures that each atomic operation receives all required predecessor outputs, enabling dependency-aware yet loosely coupled collaboration among agents.

Multi-View Feedback-Driven Refinement

To ensure semantic alignment and visual quality, Talk2Image employs a closed-loop optimization mechanism that evaluates generated images from multiple perspectives and refines them incrementally via feedback. This stage builds on parsed intentions and multi-agent execution, using feedback to correct discrepancies between output and user intention.

Compositive Scoring Mechanism Let $\hat{I}_t$ denote the generated image at turn t, and I_t the parsed instruction. The multi-view feedback function $\mathcal{J}$ combines three normalized scores ($[0, 1]$):

$$\mathcal{J}(\hat{I}_t, I_t) = \lambda_o \cdot \mathcal{J}_{\text{obj}} + \lambda_s \cdot \mathcal{J}_{\text{style}} + \lambda_m \cdot \mathcal{J}_{\text{match}} \tag{15}$$

where $\mathcal{J}$obj is the F1 score measuring consistency between detected objects and the target nouns specified in I_t, $\mathcal{J}$style denotes the cosine similarity between the generated image and pre-defined reference style embeddings, and $\mathcal{J}_{\text{match}}$ represents CLIP-based semantic alignment between $\hat{I}_t$ and the corresponding textual prompt.

Refinement Loop As shown in Algorithm 1, the system iteratively refines I_t and G_t until $\hat{I}_t$ meets the quality threshold. The final image is output when the feedback score $s = \mathcal{J}(\hat{I}_t, I_t)$ stabilizes above τ or the retry limit $N_{\max}$ is reached:

$$\hat{I}_{\text{final}} = \hat{I}_t\big|_{s \geq \tau \vee \text{retry} = N_{\max}} \tag{16}$$

This result integrates cumulative user intent, coherent multi-agent editing operations, and feedback-driven refinements, ensuring consistent alignment with the evolving conversation history while preserving overall visual quality.

System Perspective

Talk2Image can be interpreted as a multi-turn interactive framework integrating three core components: LLM-based intention modeling, task-oriented symbolic control, and weakly supervised multi-agent execution. Its modular design and multi-turn interaction mechanism enable seamless integration of diverse AI models, support collaborative human-AI interaction, and facilitate controllable real-time editing. Meanwhile, the feedback-driven cyclic refinement ensures persistent semantic alignment between user inputs and generated images across iterations. Overall, Talk2Image offers a systematic solution for high-quality image generation and fine-grained editing in conversational scenarios.

Model	Task				Quantitative Metrics				
	Generate	Edit	VQA	Chat	L2 ↓	CLIP-I ↑	CLIP-T ↑	DINO ↑	Human ↑
Null Text Inversion	×	✓	×	×	$0.0335_{\uparrow 0.0037}$	$0.8468_{\downarrow 0.0730}$	$0.2710_{\downarrow 0.0447}$	$0.7529_{\downarrow 0.1201}$	$0.7218_{\downarrow 0.1046}$
HIVE	×	✓	✓	×	$0.0557_{\uparrow 0.0259}$	$0.8004_{\downarrow 0.1194}$	$0.2673_{\downarrow 0.0484}$	$0.6463_{\downarrow 0.2267}$	$0.6848_{\downarrow 0.1416}$
InstructPix2Pix	×	✓	×	×	$0.0598_{\uparrow 0.0300}$	$0.7924_{\downarrow 0.1274}$	$0.2726_{\downarrow 0.0431}$	$0.6177_{\downarrow 0.2553}$	$0.7232_{\downarrow 0.1032}$
MagicBrush	×	✓	×	×	$0.0353_{\uparrow 0.0055}$	$0.8924_{\downarrow 0.0274}$	$0.2754_{\downarrow 0.0403}$	$0.8273_{\downarrow 0.0457}$	$0.7864_{\downarrow 0.0400}$
Genartist	✓	✓	×	×	$\mathbf{0.0298}_{=}$	$0.9071_{\downarrow 0.0127}$	$0.3067_{\downarrow 0.0090}$	$0.8492_{\downarrow 0.0238}$	$0.7985_{\downarrow 0.0279}$
Talk2Image(Ours)	✓	✓	✓	✓	**0.0298**	**0.9198**	**0.3157**	**0.8730**	**0.8264**

Table 1: Multi-turn editing results. ✓ indicates support for the feature; × indicates lack of support, and bold indicates the best results. Subscripts show performance differences relative to the best-performing model (arrows: ↑ worse for lower-better metrics, ↓ worse for higher-better metrics).

Experiments

Experimental Setup

Datasets. We evaluate the single-turn generation capability of Talk2Image on the T2I-CompBench benchmark (Huang et al. 2023). For multi-turn editing, we use the MagicBrush benchmark (Zhang et al. 2023), the first large-scale, manually annotated dataset specifically designed for multi-turn image editing.

Baselines. We benchmark Talk2Image against three categories of representative baselines to cover both single-turn generation and multi-turn editing settings:

(1) Single-turn Text-to-Image Generation Models, including Stable Diffusion v2 (Rombach et al. 2022b), DALL-E 2/3 (Ramesh et al. 2022; OpenAI 2023), StructureDiffusion (Feng et al. 2023), GORS (Huang et al. 2023) and SDXL (Podell et al. 2023).

(2) Multi-turn Image Editing Models, such as MagicBrush (Zhang et al. 2023), InstructPix2Pix (Brooks, Holynski, and Efros 2023), Null Text Inversion (Mokady et al. 2023), HIVE (Thusoo et al. 2009) and GenArtist (Wang et al. 2024a).

(3) Dialogue-based Multimodal Systems, including SDXL, Seed-X (Ge et al. 2024), Hunyuan-DiT (Li et al. 2024) and Qwen2.5-VL-32B Image-Edit[1] (Bai et al. 2025).

Model	Spatial relationship		Non-Spatial	
	UniDet ↑	Human ↑	CLIP ↑	Human ↑
Stable v2	0.1342	0.3467	0.3127	0.5827
DALL-E 2	0.1283	0.3640	0.3043	0.5964
Structure v2	0.1386	0.3467	0.3111	0.6745
GORS	0.1815	0.4560	0.3193	0.6853
SDXL	0.2133	0.4425	0.3119	0.7160
DALL-E 3	0.2543	0.4956	0.3003	0.7255
Talk2Image (ours)	**0.2676**	**0.5240**	**0.3269**	**0.7853**

Table 2: Single-turn generation results on spatial and non-spatial relationships.

[1]Qwen2.5-VL does not have built-in image editing capabilities. The Image-Edit tool utilized is a feature under Qwen2.5-VL, sourced from https://chat.qwen.ai/.

Implementation Details. Talk2Image is a zero-shot framework using Qwen2.5-VL-7B as the agent executor. It runs efficiently on a single NVIDIA A100 (80GB) with no training or fine-tuning required.

Metrics. We use both automatic and human evaluation metrics: (1) For single-turn generation, UniDet accuracy (Zhou, Koltun, and Krähenbühl 2022), CLIP score (Radford et al. 2021), and Human score; (2) For multi-turn editing, L2 error, CLIP-I/T, DINO similarity (Zhang et al. 2022), and Human score. Both Human scores follow T2I-CompBench (25 prompts/category, 300 images/model) with 3 annotators rating alignment 1–5, then normalized and averaged.

Supplementary Mechanism and Hyperparameter Validation To ensure the robustness of our compositive scoring mechanism and parameter settings, we conduct additional supplementary experiments, detailed in the *Appendix*. Specifically, these include: (1) comparative experiments on dynamic weight adjustment strategies, contrasting fixed weights, rule-based dynamic weights, and random dynamic weights; (2) sensitivity analysis of key hyperparameters, involving systematic tests on feedback threshold τ and maximum retry count $N_{\max}$. These experiments further validate the rationality of our design choices.

Single-turn Generation Evaluation

We evaluate Talk2Image on single-turn text-to-image generation, focusing on spatial relationships and non-spatial attributes. As shown in Table 2, it outperforms state-of-the-art T2I models across all metrics: its UniDet score is 5.2% higher than DALL-E 3 and 25.5% higher than SDXL in spatial tasks, with human ratings surpassing DALL-E 3 by 5.7%. For non-spatial attributes, its CLIP score is 2.4% above GORS, and human ratings are 8.2% higher than DALL-E 3.

This superiority stems from two strengths. First, our multi-view feedback mechanism enables comprehensive evaluation (e.g., semantic alignment, visual consistency) and iterative refinement to correct subtle text-image mismatches. Second, integrating specialized models as tools leverages both large foundational models (e.g., DALL-E 3) and task-specific optimizations (e.g., GORS), avoiding single-model limitations. This hybrid framework, combined with self-

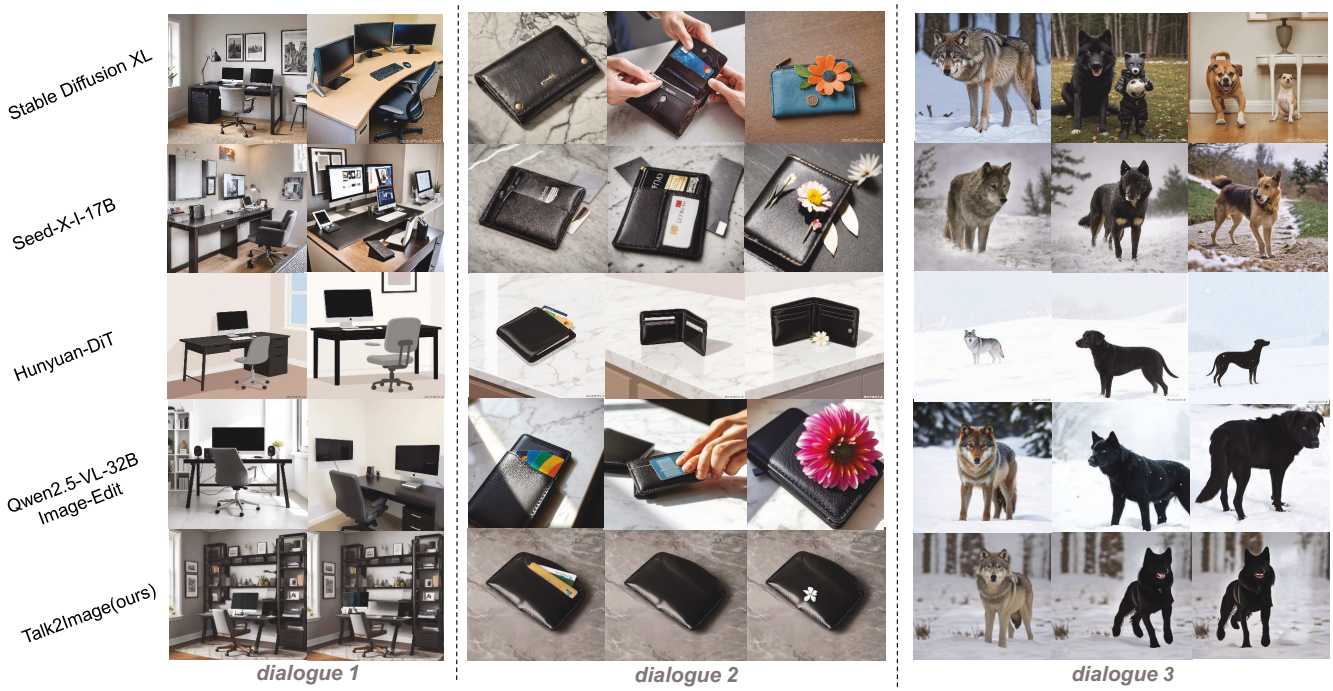

Figure 2: Visualization of Dialog Outputs. The above are three multi-round dialogues. **Dialogue1**: First, a home office with a black desk, gray chair, and monitor on the desk; then, move the monitor left. **Dialogue2**: Initially, a black leather wallet (credit card partially sticking out) on a marble countertop; next, remove the card; finally, place a flower on the wallet. **Dialogue3**: First, a gray wolf in the snow; then, replace it with a black dog; subsequently, move the dog left.

correction, ensures precise text-image control, which is critical for complex prompts with intricate attribute bindings or spatial relationships.

Multi-turn Editing Evaluation

We evaluate Talk2Image on multi-turn image editing, where users iteratively refine images via conversational feedback. Table 1 shows Talk2Image uniquely supports a unified set of capabilities (generation, editing, VQA, chat), while baselines are functionally limited. Quantitatively, it outperforms all baselines, with leads of 1.4% in CLIP-I, 9.0% in CLIP-T, and 2.8% in DINO over Genartist, plus a 3.5% advantage in human evaluations.

The core reasons lie in its modular design. The multi-turn intention parser tracks evolving user intention across dialogue turns, effectively mitigating intention drift by maintaining context consistency, ensuring each edit aligns with both current requests and prior interactions. Additionally, multi-agent collaboration enables fine-grained task division: specialized agents handle distinct editing needs (e.g., style adjustments, object additions), while a coordination module integrates their outputs to avoid incoherent edits. Unlike single-model baselines, which struggle with diverse editing demands, Talk2Image leverages the strengths of multiple models through this collaborative framework, adapting flexibly to complex, multi-step refinement tasks.

Qualitative Evaluation

To assess multi-turn coherence and visual controllability, we compare representative outputs from Talk2Image and baselines on a shared set of dialogue prompts. As shown in Figure 2, Talk2Image preserves semantic consistency across turns, accurately accumulates user intention, and maintains visual quality. In contrast, single-agent or one-shot models often suffer from layout drift, or incomplete edits.

For example, in the second turn of the first dialogue (where the monitor is instructed to move to the left side of the desk), baseline models introduce unintended layout changes compared to their initial outputs, while Talk2Image updates only the specified region. Similar inconsistencies are observed in the other two dialogues. This discrepancy stems from the fact that most baselines rely on a single-model pipeline that rewrites prompts and regenerates entire images at each turn, leading to cross-turn misalignment. In contrast, Talk2Image leverages multi-agent collaboration to decompose tasks, precisely localize changes, generate masks, and apply fine-grained edits on the original image, thus ensuring global consistency throughout the dialogue.

Ablation Study

We conduct an ablation study to evaluate the contribution of the three core components in TALK2IMAGE: the *Multi-Turn Intention Parser*, the *Multi-Agent Collaboration* module, and the *Multi-View Feedback Refinement* mechanism. As shown in Table 3, removing any component leads to

Model	Quantitative Metrics				
	L2 ↓	CLIP-I ↑	CLIP-T ↑	DINO ↑	Human ↑
Talk2Image w/o Multi-turn Intention Parser	$0.0310_{\uparrow 0.0012}$	$0.8988_{\downarrow 0.0210}$	$0.3022_{\downarrow 0.0135}$	$0.8542_{\downarrow 0.0188}$	$0.7968_{\downarrow 0.0296}$
Talk2Image w/o Multi-Agent Collaboration	$0.0315_{\uparrow 0.0017}$	$0.8962_{\downarrow 0.0236}$	$0.3015_{\downarrow 0.0142}$	$0.8356_{\downarrow 0.0374}$	$0.7862_{\downarrow 0.0402}$
Talk2Image w/o Feedback Refinement	$0.0308_{\uparrow 0.0010}$	$0.9011_{\downarrow 0.0187}$	$0.2948_{\downarrow 0.0209}$	$0.8412_{\downarrow 0.0318}$	$0.8031_{\downarrow 0.0233}$
Talk2Image (Ours)	**0.0298**	**0.9198**	**0.3157**	**0.8730**	**0.8264**

Table 3: Ablation study on the impact of three core components: Multi-turn Intention Parser, Multi-Agent Collaboration, and Feedback Optimization.

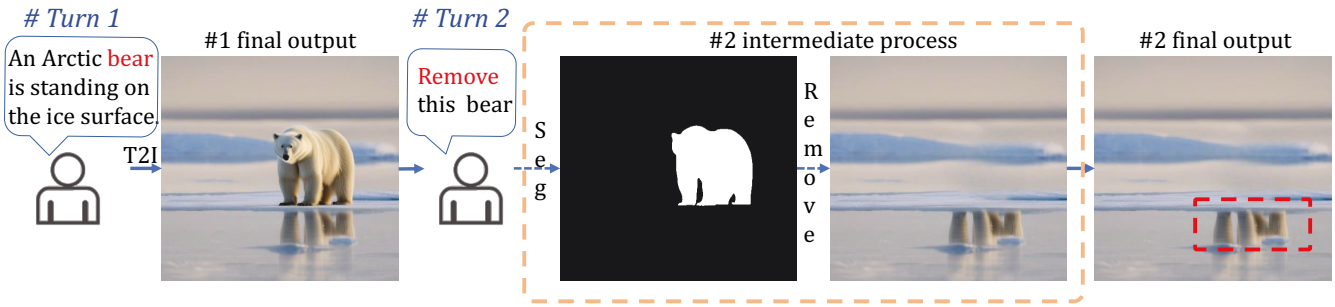

Figure 3: Error Case Example: Incomplete Removal of Bear and Its Reflection.

consistent drops across all quantitative and human-evaluated metrics, demonstrating that the full system benefits from the complementary strengths of all three modules.

Multi-Turn Intent Parser. This module maintains a coherent representation of the user's evolving intent by aggregating dialogue history and normalizing updates across turns. When removed, the system loses the ability to track cumulative constraints, resulting in pronounced *intention drift*. This leads to a 4.3% decrease in CLIP-T alignment and a 2.96-point reduction in human ratings *(Line 1 vs. Line 4)*, reflecting failures in correctly interpreting multi-step instructions.

Multi-Agent Collaboration. This module distributes visual subgoals to specialized agents. Removing it results in monolithic generation unable to support localized edits, causing a 4.3% drop in DINO similarity and 4.9% lower human preference *(Line 2 vs. Line 4)* due to visual incoherence.

Multi-View Feedback Refinement. This closed-loop mechanism refines images via multi-view evaluation. Without it, the model cannot correct errors, leading to a 2.0% lower CLIP-I score and 3.4% higher L2 error *(Line 3 vs. Line 4)*, confirming feedback's role in enhancing alignment and consistency.

Limitation and Error Case Analysis

To better understand our system's boundaries, we examine a representative failure case in Talk2Image (Figure 3). Though the system effectively decomposes intention and schedules tool execution, performance is bounded by underlying models' capabilities. For example, in a removal task ("remove the bear"), the segmentation module successfully detects the animal's body but fails to capture its reflection beneath the ice. As a result, the generated image retains a visible shadow even after removal, deviating from the expected outcome.

This case underscores both a limitation and a strength of Talk2Image. It reveals the system's reliance on third-party tools for fine-grained visual reasoning. However, it also highlights its adaptability: built upon open-source communities, Talk2Image naturally benefits from continual model improvements. More importantly, as a no-training framework, Talk2Image allows seamless integration of updated models without retraining, reflecting its modularity and adaptability in rapidly evolving tool communities.

Conclusion

We present Talk2Image, a multi-agent system for interactive image generation and editing in multi-turn dialogues. By decomposing user intention into structured prompts and leveraging specialized agents for collaborative execution and iterative refinement, it addresses key limitations of prior single-agent approaches. Experiments confirm its superiority: single-turn generation outperforms DALL-E 3 by 5.2% (UniDet) and SDXL by 25.5%; multi-turn editing leads baselines by 1.4-9.0% in automated metrics, with stronger controllability and consistency. These results validate the potential of modular agent-based systems for intelligent visual dialogue interactions. Future work will explore heterogeneous topology-based collaboration theories to coordinate agents with diverse functions and roles via heterogeneous edges, enhancing flexibility for dynamic role assignments in complex visual dialogue scenarios.

Acknowledgements

This paper is partially supported by the National Natural Science Foundation of China (No.62502488, No.12227901), Natural Science Foundation of Jiangsu Province (BK.20240460), the grant from State Key Laboratory of Resources and Environmental Information System. The AI-driven experiments, simulations and model training are performed on the robotic AI-Scientist platform of Chinese Academy of Science.

References

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18392–18402.

Calegari, R.; Ciatto, G.; Mascardi, V.; and Omicini, A. 2021. Logic-based technologies for multi-agent systems: a systematic literature review. *Autonomous Agents and Multi-Agent Systems*, 35(1): 1.

Cardoso, R. C.; and Ferrando, A. 2021. A review of agent-based programming for multi-agent systems. *Computers*, 10(2): 16.

Civitai. 2022. Civitai. https://civitai.com/.

Couairon, G.; Verbeek, J.; Schwenk, H.; and Cord, M. 2022. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*.

Face, H. 2018. Hugging Face. https://huggingface.co.

Feng, W.; He, X.; Fu, T.-J.; Jampani, V.; Akula, A. R.; Narayana, P.; Basu, S.; Wang, X. E.; and Wang, W. Y. 2023. Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis. In *The Eleventh International Conference on Learning Representations*.

Ge, Y.; Zhao, S.; Zhu, J.; Ge, Y.; Yi, K.; Song, L.; Li, C.; Ding, X.; and Shan, Y. 2024. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Huang, K.; Sun, K.; Xie, E.; Li, Z.; and Liu, X. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 78723–78747.

Huang, M.; Long, Y.; Deng, X.; Chu, R.; Xiong, J.; Liang, X.; Cheng, H.; Lu, Q.; and Liu, W. 2024. Dialoggen: Multi-modal interactive dialogue system for multi-turn text-to-image generation. *arXiv preprint arXiv:2403.08857*.

Huang, M.; Long, Y.; Deng, X.; Chu, R.; Xiong, J.; Liang, X.; Cheng, H.; Lu, Q.; and Liu, W. 2025. DialogGen: Multi-modal Interactive Dialogue System for Multi-turn Text-to-Image Generation. arXiv:2403.08857.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6007–6017.

Li, G.; Hammoud, H. A. A. K.; Itani, H.; Khizbullin, D.; and Ghanem, B. 2023. CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. arXiv:2303.17760.

Li, Z.; Zhang, J.; Lin, Q.; Xiong, J.; Long, Y.; Deng, X.; Zhang, Y.; Liu, X.; Huang, M.; Xiao, Z.; Chen, D.; He, J.; Li, J.; Li, W.; Zhang, C.; Quan, R.; Lu, J.; Huang, J.; Yuan, X.; Zheng, X.; Li, Y.; Zhang, J.; Zhang, C.; Chen, M.; Liu, J.; Fang, Z.; Wang, W.; Xue, J.; Tao, Y.; Zhu, J.; Liu, K.; Lin, S.; Sun, Y.; Li, Y.; Wang, D.; Chen, M.; Hu, Z.; Xiao, X.; Chen, Y.; Liu, Y.; Liu, W.; Wang, D.; Yang, Y.; Jiang, J.; and Lu, Q. 2024. Hunyuan-DiT: A Powerful Multi-Resolution Diffusion Transformer with Fine-Grained Chinese Understanding. arXiv:2405.08748.

Ma, S.; Zhang, X.; Zhao, Z.; Liu, B.; Fan, C.; and Hu, Z. 2025. DialogDraw: Image Generation and Editing System Based on Multi-Turn Dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 24795–24803.

Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6038–6047.

OpenAI. 2023. DALL·E 3 System Card. https://openai.com/dall-e-3. Accessed: 2025-07-28.

OpenArt. 2021. OpenArt. https://openart.ai/.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022a. High-resolution image synthesis with latent

diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022b. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Thusoo, A.; Sarma, J. S.; Jain, N.; Shao, Z.; Chakka, P.; Anthony, S.; Liu, H.; Wyckoff, P.; and Murthy, R. 2009. Hive: a warehousing solution over a map-reduce framework. 2(2): 1626–1629.

Wang, Z.; Li, A.; Li, Z.; and Liu, X. 2024a. Genartist: Multimodal llm as an agent for unified image generation and editing. *Advances in Neural Information Processing Systems*, 37: 128374–128395.

Wang, Z.; Li, A.; Li, Z.; and Liu, X. 2024b. GenArtist: Multimodal LLM as an Agent for Unified Image Generation and Editing. arXiv:2407.05600.

Yang, H.; Yue, S.; and He, Y. 2023. Auto-GPT for Online Decision Making: Benchmarks and Additional Opinions. arXiv:2306.02224.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. arXiv:2203.03605.

Zhang, K.; Mo, L.; Chen, W.; Sun, H.; and Su, Y. 2023. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36: 31428–31449.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhou, X.; Koltun, V.; and Krähenbühl, P. 2022. Simple multi-dataset detection. In *CVPR*.