Ambiguity-Aware Multi-Round Text-Guided Image Editing via Intent Knowledge Graph Compilation

Yining Liu
School of Computer Science and Technology
Shandong University
Qingdao, China
202300130160@mail.sdu.edu.cn
ORCID: 0009-0004-5517-4734

***Abstract*—Text-driven image editing with diffusion models has achieved significant progress, and most existing methods assume that user instructions are clear and directly actionable. However, in real interactions, users often provide vague or underspecified requests (e.g., "make it more cinematic" or the polysemous word "bank"). To address this problem, we propose DiaKG-Edit, an ambiguity-aware multi-round dialogue image editing framework with two components: Ambiguity-Aware Modeling (AAM) and Graph-to-Token Compilation (G2T). In AAM, we maintain an Editable Knowledge Graph (EKG) as a persistent representation of editing intent, which explicitly records entities, attributes, relations, and their strengths and is incrementally updated across interaction rounds. For each incoming instruction, the system determines whether it is clear or ambiguous, applying conservative updates in clear cases and resolving ambiguity via dialogue-driven refinement, thereby progressively updating the EKG while preserving previously confirmed edits. Based on the stabilized intent representation, G2T maps incremental differences between successive EKG states to executable token-level control strategies for diffusion-based editors, enabling deterministic and stable multi-round execution. Experiments show that DiaKG-Edit outperforms HRV under ambiguous instructions, improving CLIP instruction alignment from 0.77 to 0.79 (+2.6%) and reducing LPIPS from 0.152 to 0.147 (-3.3%).**

***Index Terms*—text-guided image editing, diffusion models, ambiguity resolution, multi-round dialogue, editable knowledge graph, intent graph compilation.**

I. Introduction

Text-driven image editing with diffusion models has achieved rapid progress in recent years, enabling users to modify visual content through natural language instructions [1], [2], [10]–[13]. By manipulating internal representations such as cross-attention maps or latent features, modern diffusion-based editors can perform fine-grained, training-free image modifications while largely preserving the original structure. Despite these advances, most existing methods implicitly assume that user instructions are explicit, unambiguous, and directly executable, which limits their applicability in realistic interactive scenarios.

In practice, user inputs are often vague, subjective, or semantically underspecified. Requests such as "add a bank," or "make this nicer" do not specify concrete visual operations and may admit multiple plausible interpretations. To resolve such ambiguity, users typically engage in multi-round interaction, progressively refining their intent through dialogue. However, ambiguity is not confined to a single instruction: it interacts with the entire multi-round editing process, where earlier decisions must be preserved while later clarifications selectively refine or override uncertain aspects. Directly applying token-level diffusion editing under such conditions frequently leads to unstable results, unintended visual drift, or accidental violation of previously confirmed constraints.

Recent advances in large language models (LLMs) have significantly improved natural language understanding and dialogue-based intent clarification [9] [19]. Nevertheless, treating LLMs merely as prompt rewriters is insufficient for reliable interactive image editing. Two fundamental challenges remain. 1) is how to explicitly model and reliably maintain evolving user intent under ambiguous instructions and multi-round dialogue, while preserving previously confirmed edits without introducing cross-round drift. 2) is how to bridge the gap between high-level, structured semantic intent and stable token-level control required by diffusion-based image editors, especially in multi-round interactive settings.

To address these challenges, we propose DiaKG-Edit, an ambiguity-aware, multi-round dialogue-driven image editing framework built upon explicit intent modeling and semantic-to-token compilation. Our framework consists of two tightly coupled components: Ambiguity-Aware Modeling (AAM) and Graph-to-Token Compilation (G2T).

AAM addresses the first challenge by explicitly representing and maintaining evolving user intent under ambiguous inputs and multi-round interaction using a persistent Editable Knowledge Graph (EKG). By estimating instruction ambiguity and switching between conservative updates and dialogue-based resolution, AAM progressively refines uncertain intent while preserving previously confirmed constraints.

G2T addresses the second challenge by bridging structured semantic intent and executable diffusion-based editing through deterministic mapping from incremental EKG changes to token-level operations. Entity, attribute, and strength updates are translated into token manipulations such as semantic replacement, reweighting, or localized injection, enabling stable, interpretable, and model-agnostic execution.

Extensive experiments demonstrate that DiaKG-Edit significantly improves robustness to ambiguous instructions and cross-round consistency, while remaining competitive un-

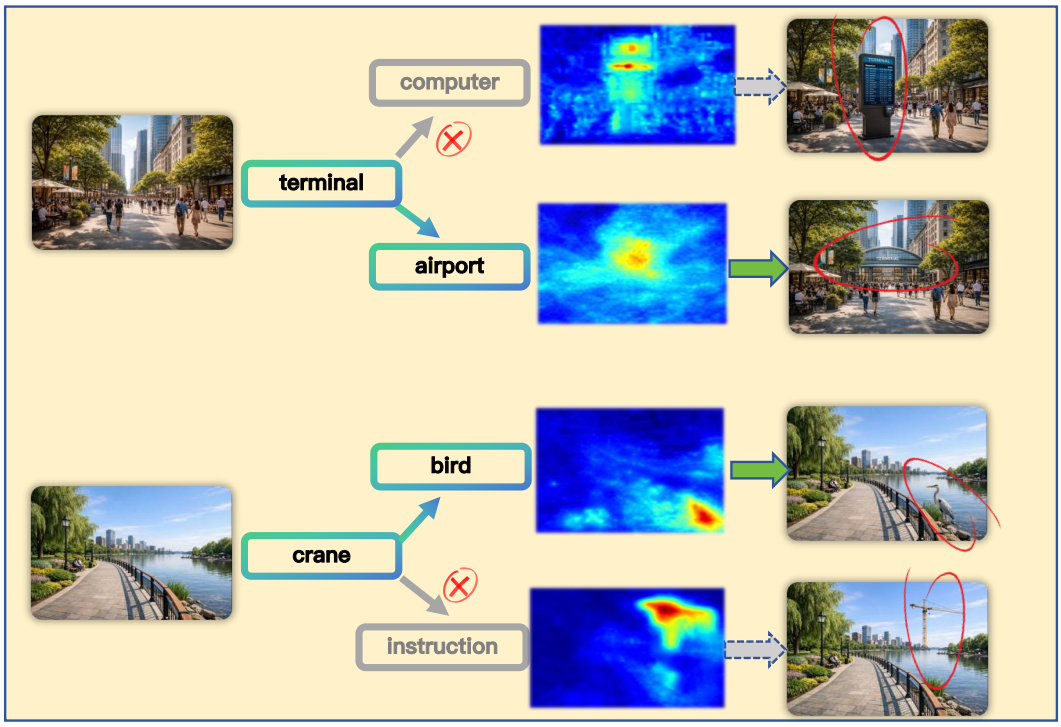

Fig. 1. Teaser of dialogue-driven ambiguity resolution for text-guided image editing. Given an initial image and an ambiguous instruction (e.g., "Add a terminal" or "Add a crane"), direct generation without clarification may lead to underspecified or semantically incorrect edits. Our method actively engages in multi-round dialogue to resolve lexical ambiguity, incrementally updates an editable knowledge graph (EKG) to prune invalid interpretations, and compiles the refined intent into a stable editing process. As a result, the final edits accurately reflect the user's clarified intent while preserving previously confirmed scene structure and layout.

der explicit-instruction settings. Importantly, our method is training-free and can be directly applied to existing diffusion-based image editing frameworks, providing a practical solution for more natural, reliable, and interactive image editing.

Our contributions are summarized as follows:

- We identify two fundamental challenges in realistic multi-round image editing: ambiguity arising from user inputs combined with dialogue-based intent evolution, and the semantic-to-token gap between high-level intent and diffusion-based execution.
- We propose Ambiguity-Aware Modeling (AAM), which explicitly represents and incrementally updates user intent using an Editable Knowledge Graph, enabling stable intent tracking and clarification across dialogue rounds.
- We introduce Graph-to-Token Compilation (G2T), a deterministic framework that maps structured semantic intent updates to token-level diffusion editing operations, supporting stable and interpretable multi-round execution without additional training.

II. Related Work

Research on text-driven interactive image editing primarily focuses on understanding user intent and executing edits within diffusion-based generative models [1], [2]. Existing work can be broadly grouped into two directions: (1) ambiguous intent understanding and multi-round interaction, and (2) diffusion-based image editing with token- or attention-level control [3], [10], [12], [13]. Despite recent progress, most approaches treat intent understanding and edit execution as loosely coupled stages, lacking a unified mechanism to maintain evolving intent and translate it into stable, executable editing operations.

A. Ambiguous Intent Understanding and LLM-based Dialogue Editing

In realistic interaction settings, user editing instructions are often subjective, incomplete, or semantically ambiguous. Prior work addresses such ambiguity through uncertainty detection, clarification strategies, or dialog-based interaction, including recent methods that leverage large language models (LLMs) for high-level reasoning and cross-round coherence [9], [18]–[20]. These approaches improve intent interpretation by generating clarifying questions or refining instructions through dialogue [18], [19].

However, most existing methods assume convergence to a fixed editing goal or encode historical constraints implicitly in natural language. As a result, they lack an explicit and persistent representation of evolving editing intent. Without such a representation, it is difficult to reliably preserve previously confirmed edits, resolve conflicts across interaction rounds, or ensure stable intent refinement when instructions are progressively updated rather than specified upfront.

B. Diffusion-Based Image Editing with Token-Level Control

Recent diffusion-based image editing methods have demonstrated strong controllability by manipulating internal representations such as tokens, attention maps, or latent features [3], [10]–[14], [16], [17]. By operating directly on these low-level components, training-free approaches can achieve fine-grained edits while preserving global image structure under explicit instructions [3], [14], [16], [17].

However, most token- or attention-level control methods implicitly assume that the target editing objective is well-defined and stable. When user instructions are vague, subjective, or evolve across multiple dialogue rounds, repeatedly reinterpreting free-form text often leads to instability, over-editing, or drift in attention and visual content [3], [10], [12], [16]. As a result, while existing diffusion-based editors provide effective low-level execution mechanisms, they lack a principled way to connect evolving high-level semantic intent with stable token-level control, limiting their applicability in interactive and dialogue-driven editing scenarios.

III. Method

Based on the above analysis, existing text-driven image editing methods still struggle with two concerns. 1) is how to model and maintain user intent under ambiguous instructions and multi-round interactions, while preserving previously confirmed edits and avoiding cross-round drift. 2) is how to translate high-level, structured semantic intent into concrete and stable token-level control signals that diffusion-based image editors can directly execute.

To address these concerns, we propose DiaKG-Edit, a dialogue-driven ambiguity-aware image editing framework composed of two key modules. To tackle the first concern, we introduce Ambiguity-Aware Modeling (AAM). To tackle the second concern, we introduce Graph-to-Token Compilation (G2T). In the following, we present the problem formulation, overall framework, and detailed design of each module.

A. Problem Definition

An editing session starts with an initial image x_0 and an initial source prompt p_0 (where p_0 can be provided by the user or automatically obtained via an image captioning model). At the k-th interaction round, the user provides an editing instruction u_k, together with the historical context $\mathcal{H}_{k-1} = \{(u_i, x_i, p_i)\}_{i=0}^{k-1}$. The system outputs an edited image x_k and is required to maintain intent consistency across multiple rounds of interaction.

B. Overall Framework

Our method consists of two core modules:

- Ambiguity-Aware Intent Modeling(AAM): We maintain a persistent *Editable Knowledge Graph* (EKG) G_k to represent structured editing intent. The system explicitly estimates instruction ambiguity and performs gated switching between a "direct mapping mode" and an "ambiguity resolution mode."
- Graph-to-Token Compilation (G2T): We compile incremental updates of the EKG into executable token-level editing plans for diffusion-based image editors, including token selection, operation types (e.g., semantic replacement, attribute reweighting, and localized token injection), and corresponding control schedules.

C. Ambiguity-Aware Modeling (AAM)

`Overview`: As discussed in the fist concern, a key difficulty in multi-round editing is keeping track of what the user has already confirmed, so later instructions do not accidentally overwrite earlier decisions [23]. However, many prior methods treat multi-round editing as if the target goal is fixed, so they do not keep an explicit evolving intent state across rounds [3], [10], [11], [23]. To address this, we propose Ambiguity-Aware Modeling (AAM), which explicitly maintains an evolving intent state throughout the interaction. AAM combines an ambiguity-aware gating mechanism (AAG) to decide conservative vs. resolution updates and an *Editable Knowledge Graph* (EKG) with provenance-aware, incremental updates to preserve confirmed intent while refining unclear parts over rounds.

`Ambiguity-Aware Gating(AAG)` **:** Given the current user instruction u_k and the historical interaction context $\mathcal{H}_{k-1}$, we estimate an ambiguity score $\rho_k \in [0, 1]$ that measures how clearly the instruction specifies an actionable editing intent. We model instruction ambiguity as the combined effect of multiple sources of uncertainty in interactive editing. Formally, the ambiguity score is defined as

$$\rho_k = \sum_i \alpha_i \, \phi_i(u_k, \mathcal{H}_{k-1}), \tag{1}$$

where each $\phi_i(\cdot) \in [0, 1]$ measures the degree of a specific ambiguity type and α_i denotes its relative contribution. In practice, different ϕ_i correspond to common ambiguity sources, including `lexical ambiguity`, `missing entities or attributes`, `subjective or qualitative descriptions` (e.g., "nicer" or "more cinematic"), `unclear scope of effect`, and `potential conflicts` with previously confirmed intent. A larger ρ_k indicates that the instruction admits multiple plausible interpretations and requires further resolution before reliable execution.

Based on the ambiguity score, a fixed threshold τ is used to route the instruction into different processing modes:

$$\text{Mode}(u_k) = \begin{cases} \text{Direct Mapping Mode (DMM)}, & \rho_k < \tau, \\ \text{Ambiguity Resolution Mode (ARM)}, & \rho_k \geq \tau. \end{cases} \tag{2}$$

For non-ambiguous instructions handled in Direct Mapping Mode (DMM), the system performs conservative intent updates. Explicit entities and attributes are extracted from u_k, slot filling and entity alignment are applied, and only minimal modifications to the EKG are allowed, avoiding the introduction of inferred attributes not directly specified by the user.

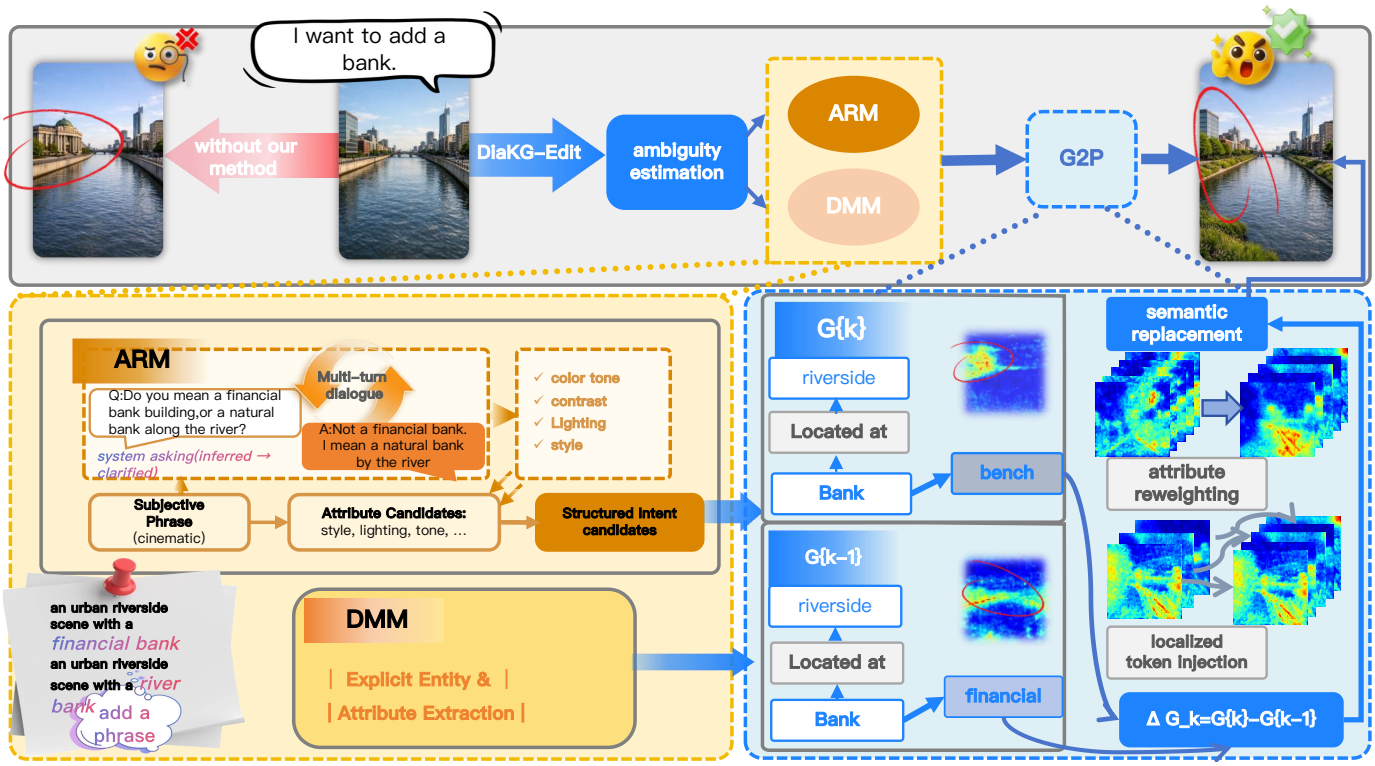

Fig. 2. Overview of the proposed DiaKG-Edit framework. Given an initial image and a sequence of user instructions, the system first applies Ambiguity-Aware Modeling (AAM) to assess instruction ambiguity and incrementally update a persistent Editable Knowledge Graph (EKG) that explicitly represents evolving editing intent across dialogue rounds. Based on the incremental difference between successive EKG states, Graph-to-Token Compilation (G2T) deterministically maps structured intent updates to token-level editing strategies, including semantic replacement, attribute reweighting, and localized token injection. These token-level control signals are then executed by a diffusion-based image editor, enabling stable, interpretable, and consistent multi-round image editing under ambiguous user instructions.

For ambiguous instructions, Ambiguity Resolution Mode (ARM) is activated. First, subjective or high-level descriptions are expanded into structured attribute candidates (e.g., style, lighting, tone, and contrast) with initial strength weights. If critical information is still missing or conflicts are detected, a small set of clarification questions $\{q_k^j\}_{j=1}^m$ is generated, and user responses are incorporated into the intent update. Through this progressive grounding process, the original instruction u_k is resolved into a more explicit intermediate representation $\tilde{u}_k$, which is used exclusively to update the EKG rather than being directly passed to the diffusion-based editor.

`Editable Knowledge Graph with Provenance Annotation(EKG)`: We represent the continuously evolving editing intent as an *Editable Knowledge Graph* (EKG),

$$G_k = (\mathcal{V}_k, \mathcal{E}_k), \tag{3}$$

where the graph structure provides a persistent and structured representation of user intent across interaction rounds. The node set $\mathcal{V}_k$ consists of three types of nodes: entity nodes that correspond to editable objects or regions in the image (e.g., *car*, *sky*, *background*), attribute nodes that describe editable properties of these entities (e.g., *color*, *material*, *lighting*, *style*), and value nodes that specify concrete attribute values (e.g., *red*, *metallic*, *warm*, *cinematic*). The edge set $\mathcal{E}_k$ encodes editing constraints and relations using typed edges such as `has_attr`, `has_value`, `modify`, `spatial_constraint`, and `negation`, enabling explicit modeling of how different intent elements interact.

To support fine-grained control and cross-round stability, each node and edge in the EKG is associated with additional metadata. Specifically, each intent element is assigned a continuous strength weight $w \in [0, 1]$ that reflects its relative importance, an optional scope descriptor that restricts its spatial or semantic extent (e.g., global, background-only, or coarse region tags), and a provenance label $s \in \{\texttt{explicit}, \texttt{inferred}, \texttt{clarified}\}$ indicating how the element was introduced. This provenance-aware design improves interpretability and allows the system to apply conservative updates for non-ambiguous instructions, while enabling controlled inference and progressive confirmation when resolving ambiguous intent.

To reduce drift in multi-round editing, we introduce an anchor-based stabilization mechanism. An initial anchor graph G_0 is retained to capture relatively stable scene structure and early confirmed intent, and is fused with the most recent graph state as

$$\bar{G}_{k-1} = \textsc{Fuse}(G_0, G_{k-1}), \tag{4}$$

where FUSE preserves immutable or high-confidence con-

straints from G_0 while incorporating editable user-specified intent from G_{k-1}. This anchoring strategy helps prevent accumulated errors and unintended overwriting of previously confirmed edits.

The EKG is updated incrementally at each interaction round. Given the intent updates $\mathcal{U}_k$ produced by the ambiguity-aware gating process (including optional clarification results), the graph is updated as

$$G_k = \textsc{Update}(G_{k-1}, \mathcal{U}_k). \tag{5}$$

The update procedure is implemented through a set of atomic graph-editing operations, including node and edge addition, strength adjustment, deactivation of invalid elements, and explicit conflict resolution. When conflicts arise, they are resolved by prioritizing intent elements with higher provenance confidence (`explicit` > `clarified` > `inferred`), followed by recency and user-specified priority when available.

D. Graph-to-Token Compilation (G2T)

As discussed in the second concern, the key difficulty is translating high-level and sometimes ambiguous user intent into concrete token-level control signals that diffusion-based image editors can reliably execute [3], [14], [16], [17]. To address this, G2T establishes a deterministic mapping from structured intent updates to token-level editing strategies by compiling incremental changes in the Editable Knowledge Graph (EKG).

Instead of reinterpreting the full intent in every round, we compute the incremental graph difference

$$\Delta G_k = \textsc{Diff}(G_{k-1}, G_k), \tag{6}$$

which localizes modified entities, attributes, values, and their associated strengths.

G2T compiles ΔG_k into a token-level editing plan

$$\Pi_k = \{(\mathrm{op}_i, \mathcal{T}_i, \gamma_i)\}_{i=1}^{n_k}, \tag{7}$$

where op_i denotes the operation type, including `semantic replacement`,`attribute reweighting`, and`localized token injection`, denotes the operation type, $\mathcal{T}_i$ the target tokens, and γ_i the corresponding strength and scheduling parameters derived from EKG weights.

The operation type is selected deterministically based on the change pattern in ΔG_k: discrete changes in entities or mutually exclusive attribute values trigger `semantic replacement`; continuous adjustments of attribute strength trigger `attribute reweighting`; and localized additions or constraint refinement trigger `localized token injection`. This design ensures a consistent and reproducible mapping from semantic intent updates to token-level execution.

Target tokens are obtained by aligning EKG elements with their corresponding token spans, accounting for tokenizer behavior. EKG strength weights are further mapped to diffusion-step control schedules, e.g.,

$$\alpha_t = \sigma(a \cdot w + b) \cdot \mathbb{I}[t \in [t_s, t_e]], \tag{8}$$

where earlier steps emphasize global structure preservation and later steps focus on local refinement.

IV. Experiments

A. Experimental Settings

We use a linear noise schedule with the number of diffusion steps set to $T = 1000$, where the noise variance increases linearly from $\beta_{\text{start}} = 10^{-4}$ to $\beta_{\text{end}} = 0.02$. Unless otherwise specified, all experiments are conducted under identical random seeds and input image conditions to eliminate performance variance introduced by sampling randomness or parameter differences.

B. Dataset Construction

To evaluate multi-round, dialog-driven image editing, we construct a benchmark of 600 interaction episodes, each consisting of an initial image and a complete dialog trajectory. Each episode is represented as $\{U_1, F_1, U_2, F_2, \ldots, U_K\}$, where U_k denotes the user instruction at round k and F_k the corresponding feedback, confirmation, or clarification. An edited image is generated after each round, enabling us to evaluate not only single-round instruction alignment but also cross-round consistency under evolving intent. The average number of rounds is $\bar{K} = 3.2$, with episode lengths ranging from 2 to 5 turns.

The initial images are sampled from public natural-image sources and cover diverse editing scenarios, including object replacement, attribute modification, background adjustment, and style refinement. To support fine-grained analysis of ambiguity handling, each episode is annotated by the dominant ambiguity type in the first-round instruction, including lexical ambiguity, missing attributes, subjective style requests, unclear scope, and cross-round conflicts.

The dialog scripts are manually designed to simulate realistic closed-loop interactions. In most cases, the first-round instruction is intentionally underspecified or ambiguous, while subsequent turns progressively provide confirmations, corrections, or missing constraints. During evaluation, the same initial images and dialog scripts are applied to all compared methods, ensuring a controlled and fair comparison across different editing frameworks.

C. Quantitative Results for Multi-Round Dialogue-Based Image Editing

Table I reports the quantitative comparison of different methods on the multi-round dialogue-based image editing task, covering image–text semantic alignment, visual stability across rounds, intent consistency, and subjective user evaluation.

Our method consistently achieves the best overall performance across all metrics. In particular, it obtains the highest CLIP score of 0.79, outperforming Stable Flow (0.77) and HRV-P2P (0.75), indicating more accurate alignment with the target intent at the current interaction round [21]. Meanwhile, DiaKG-Edit achieves the lowest LPIPS score of 0.147, compared to 0.152 and 0.148 of the competing methods,

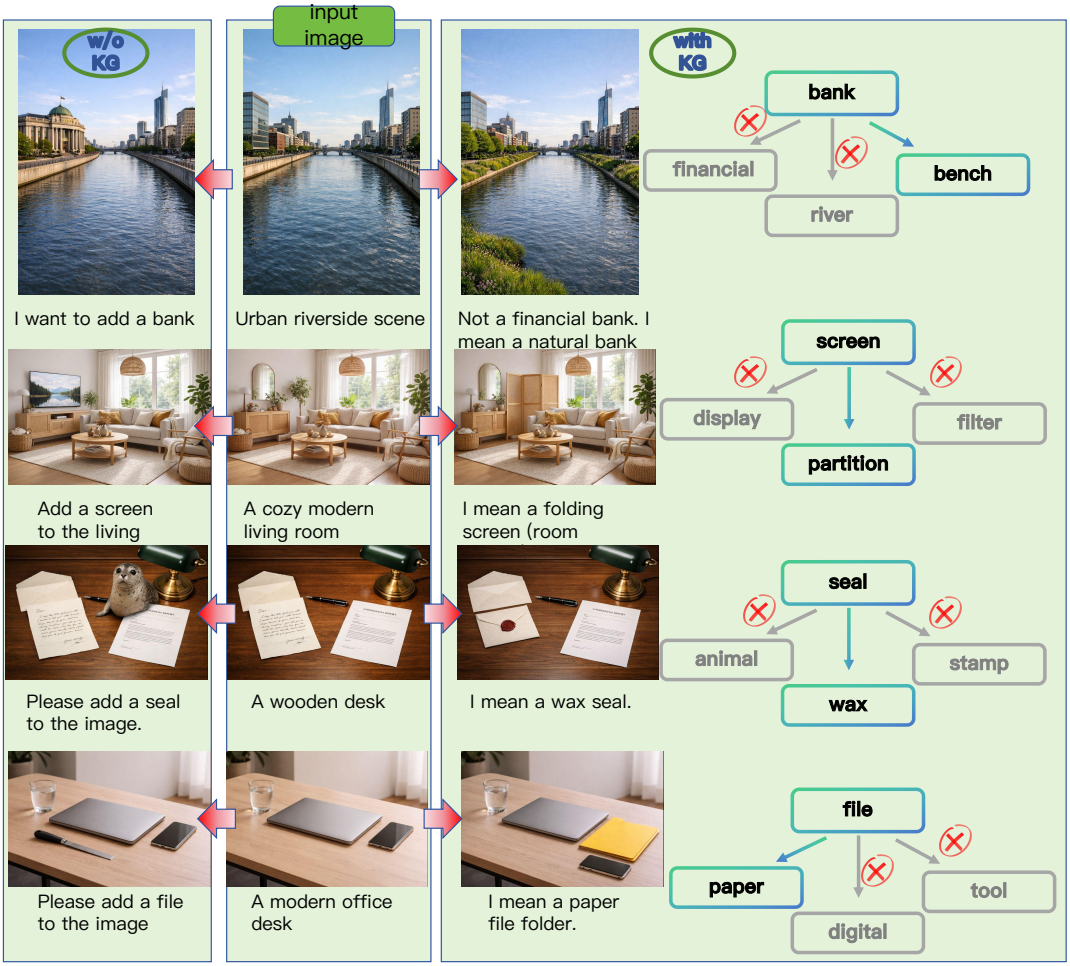

Fig. 3. Qualitative example of multi-round dialogue-driven image editing, illustrating how an initially ambiguous instruction is progressively disambiguated and refined across dialogue rounds, while preserving previously confirmed constraints and edits.

TABLE I

QUANTITATIVE COMPARISON OF MULTI-ROUND DIALOGUE-BASED IMAGE EDITING. **CONSISTENCY** IS COMPUTED AS THE PROPORTION OF PREVIOUSLY CONFIRMED CONSTRAINTS THAT REMAIN SATISFIED IN THE FINAL OUTPUT, **VIOLATION** IS COMPUTED AS THE PROPORTION OF CONFIRMED CONSTRAINTS THAT ARE VIOLATED BY UNINTENDED EDITS, AND **USER SCORE** IS COMPUTED AS THE MEAN HUMAN RATING (1–5) OVER ALL DIALOGUE SESSIONS.

Metric	P2P [3]	HRV [7]	Stable Flow [8]	Ours
CLIP [21] ↑	0.65	0.75	0.77	**0.79**
LPIPS [22] ↓	0.162	0.148	0.152	**0.147**
Consistency ↑	0.61	0.69	0.68	**0.70**
Violation ↓	0.27	0.21	0.23	**0.21**
User Score ↑	3.50	3.98	4.05	**4.42**

suggesting fewer unnecessary visual changes and reduced cumulative drift during progressive refinement [22].

More importantly, our approach shows clear advantages in multi-round consistency. It achieves the highest Consistency score (0.70) while simultaneously reducing the Violation rate to 0.21, demonstrating its effectiveness in preserving previously confirmed constraints and avoiding unintended edits across dialogue rounds. This improvement is further reflected in the User Score, where our method reaches 4.42, significantly higher than all baselines, indicating better perceived editing quality and user satisfaction.

Overall, these results validate the effectiveness of the proposed framework in achieving superior semantic alignment, cross-round stability, and reliable intent preservation in multi-round dialogue-based image editing.

D. Qualitative Analysis of Multi-round Dialogues

To illustrate the behavior of the proposed method in realistic interactive scenarios, we present qualitative results of multi-round dialog-driven image editing under ambiguous initial instructions. The analysis focuses on whether the model can progressively resolve ambiguity while preserving structural and attribute constraints confirmed in earlier rounds.

As shown in Fig. 3 , users often begin with vague or underspecified editing requests. Instead of performing aggressive edits, the system incorporates candidate attributes into a structured editing state and incrementally refines them

Fig. 4. Qualitative examples of multi-round dialogue-driven editing. Each row corresponds to an editing session (e.g., cup, person, backpack), and each column shows the progressively edited result at a dialogue round. The results demonstrate that the model incrementally aligns with user intent while preserving previously confirmed structures and attributes, effectively avoiding cumulative drift.

through subsequent dialog rounds based on user feedback. This ambiguity-aware strategy enables gradual disambiguation without introducing unnecessary visual perturbations.

As the dialog progresses, the model responds to newly clarified or refined intent while consistently preserving previously confirmed editing constraints. This explicit maintenance of the editing state prevents repeated modification of stable content and allows the image to converge smoothly toward the user's evolving intent.

Fig. 4 further presents multiple independent editing sessions, covering object attribute changes, human state adjustments, and scene content expansion. Across all cases, the model maintains global structural stability and key semantic relationships, even when the editing target changes substantially, without exhibiting noticeable cumulative drift or semantic conflicts.

Overall, these qualitative results demonstrate that the proposed framework effectively supports progressive disambiguation, accurate intent alignment, and stable multi-round editing, providing intuitive evidence for its consistency and controllability in real-world interactive settings.

E. Ambiguity-Stratified Evaluation

To further analyze model performance under different types of ambiguous instructions, we conduct an ambiguity-stratified evaluation by grouping test samples according to the source of ambiguity to examine the model's ability to understand and align with user intent across diverse forms of ambiguity.

Specifically, based on the primary cause of ambiguity, the test samples are categorized into five groups: polysemy, referential ambiguity, missing attributes, constraint conflicts, and relational ambiguity. Each sample is assigned to a single dominant ambiguity category to avoid interference caused by category overlap.

Fig. 5 presents the success rate comparison across different ambiguity categories, where the success rate indicates whether the final output produced after multi-round interaction matches the user's true intent. It can be observed that, compared with the baseline method, our approach achieves higher success rates across all ambiguity categories, demonstrating stronger robustness in handling various forms of ambiguous instructions.

Furthermore, in more complex scenarios such as constraint conflicts and relational ambiguity, the performance degradation of the baseline is particularly pronounced, whereas our method maintains a stable performance advantage. These results indicate that, through intent clarification and structured state maintenance across multi-round dialogs, the model can effectively resolve conflicts or uncertain relations in instructions and progressively translate them into executable editing constraints.

In summary, the ambiguity-stratified evaluation validates the consistent effectiveness of the proposed method across different ambiguity sources, further supporting its applicability

TABLE II
ABLATION STUDY OF DIAKG-EDIT. WE REMOVE ONE COMPONENT AT A TIME TO EXAMINE ITS IMPACT ON INTENT ALIGNMENT, CROSS-ROUND STABILITY, AND CONSTRAINT SATISFACTION.

Setting	CLIP ↑	LPIPS ↓	Consistency ↑	Violation ↓	Avg. Rounds ↓
Full	**0.79**	**0.147**	**0.70**	**0.21**	2.5
w/o provenance	0.77	0.149	0.66	0.25	2.5
w/o AAG	0.75	0.156	0.62	0.29	1.6
w/o Anchor	0.77	0.152	0.64	0.26	2.5
w/o Graph Diff	0.78	0.148	0.69	0.22	2.5
w/o G2P	0.76	0.153	0.63	0.27	2.3
Single-round	0.74	0.158	0.60	0.31	**1.0**

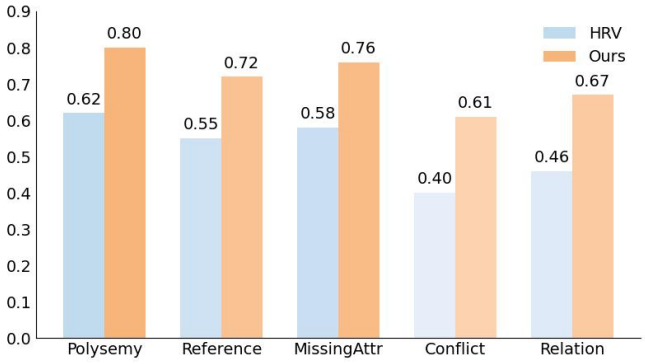

Fig. 5. Ambiguity-stratified success rate comparison

in realistic, open-ended user interaction scenarios.

F. Ablation Study

We evaluate the contribution of each component in DiaKG-Edit by removing one module at a time under identical input images, dialogue scripts, and diffusion settings. Table II summarizes the results in terms of instruction alignment (CLIP), structural deviation (LPIPS), and cross-round constraint satisfaction (Consistency/Violation).

Removing provenance labels (*w/o provenance*) degrades constraint preservation, with Consistency dropping from 0.70 to 0.66 and Violation increasing from 0.21 to 0.25, indicating the importance of explicit intent provenance. Disabling ambiguity-aware gating (*w/o AAG*) causes the largest performance drop, reducing CLIP from 0.79 to 0.75 and Consistency from 0.70 to 0.62 while increasing Violation to 0.29, demonstrating that ambiguity-aware decision making is critical under underspecified inputs. Without anchor fusion, LPIPS increases to 0.152 and Consistency decreases to 0.64, showing that anchoring helps suppress cumulative drift across rounds.

Removing graph difference computation or G2T compilation further degrades stability, while the single-round setting yields the lowest Consistency (0.60) and highest Violation (0.31), highlighting the necessity of multi-round intent modeling. Overall, these results confirm that the proposed components are complementary and jointly enable stable and accurate multi-round dialogue-based image editing.

V. Conclusion

In this paper, we presented DiaKG-Edit, an ambiguity-aware multi-round dialogue framework for text-guided image editing with diffusion models. To address the challenges posed by vague instructions and evolving user intent, we introduced Ambiguity-Aware Modeling (AAM), which explicitly maintains a persistent Editable Knowledge Graph to track, refine, and preserve editing intent across interaction rounds. We further proposed Graph-to-Token Compilation (G2T), which deterministically maps incremental intent updates to token-level control strategies, enabling stable and interpretable diffusion-based execution without additional training. Extensive quantitative and qualitative experiments demonstrate that DiaKG-Edit significantly improves robustness to ambiguous instructions, cross-round consistency, and user satisfaction, while remaining competitive under explicit-instruction settings. We believe this work provides a practical and general framework for bridging high-level intent understanding and low-level diffusion control, and opens up new directions for reliable and interactive image editing systems.

References

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[3] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross-attention control," *arXiv preprint arXiv:2208.01626*, 2022. [Online]. Available: https://arxiv.org/abs/2208.01626

[4] Y. He, Y. Zhang, Z. Li, *et al.*, "Towards understanding cross and self-attention in stable diffusion for text-guided image editing," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[5] Z. Wu, H. Wang, J. Yang, *et al.*, "LOCATEdit: Graph Laplacian optimized cross attention for localized text-guided image editing," *arXiv preprint*, 2025. [Online]. Available: https://arxiv.org/abs/2503.21541

[6] C. Li, R. Wang, Y. Zhang, *et al.*, "CPAM: Context-preserving adaptive manipulation for zero-shot real image editing," *arXiv preprint*, 2025. [Online]. Available: https://arxiv.org/abs/2506.18438

[7] Z. Wu, H. Wang, J. Yang, *et al.*, "Cross-attention head position patterns can align with human visual concepts in text-to-image generative models," in *International Conference on Learning Representations (ICLR)*, 2025.

[8] Y. He, Y. Zhang, Z. Li, *et al.*, "Stable Flow: Vital layers for training-free image editing," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[9] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[10] T. Brooks, A. Holynski, and A. A. Efros, "InstructPix2Pix: Learning to follow image editing instructions," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[11] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, "NULL-text inversion for editing real images using guided diffusion models," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[12] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, "Imagic: Text-based real image editing with diffusion models," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[13] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "SDEdit: Guided image synthesis and editing with stochastic differential equations," in *International Conference on Learning Representations (ICLR)*, 2022.

[14] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel, "Plug-and-play diffusion features for text-driven image-to-image translation," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[15] G. Parmar, K. K. Singh, R. Zhang, Y. Li, J. Lu, and J.-Y. Zhu, "Zero-shot image-to-image translation," in *ACM SIGGRAPH*, 2023.

[16] M. Cao, X. Wang, Y. Zhang, *et al.*, "MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing," in *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2023.

[17] H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf, and D. Cohen-Or, "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models," in *ACM SIGGRAPH*, 2023.

[18] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "ReAct: Synergizing reasoning and acting in language models," in *International Conference on Learning Representations (ICLR)*, 2023.

[19] A. Madaan, N. Tandon, P. Gupta, *et al.*, "Self-refine: Iterative refinement with self-feedback," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[20] S. Yao, D. Yu, J. Zhao, *et al.*, "Tree of thoughts: Deliberate problem solving with large language models," *arXiv preprint arXiv:2305.10601*, 2023. [Online]. Available: https://arxiv.org/abs/2305.10601

[21] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. International Conference on Machine Learning (ICML)*, 2021.

[22] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[23] M. Huang, *et al.*, "ParallelEdits: Efficient Multi-Aspect Text-Driven Image Editing," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.