

R-Genie: Reasoning-Guided Generative Image Editing

Dong Zhang^{1*} Lingfeng He^{2*} Rui Yan² Fei Shen² Jinhui Tang^{2†}

¹The Hong Kong University of Science and Technology;

²Nanjing University of Science and Technology.

Abstract

While recent advances in image editing have enabled impressive visual synthesis capabilities, current methods remain constrained by explicit textual instructions and limited editing operations, lacking deep comprehension of implicit user intentions and contextual reasoning. In this work, we introduce a new image editing paradigm: reasoning-guided generative editing, which synthesizes images based on complex, multi-faceted textual queries accepting world knowledge and intention inference. To facilitate this task, we first construct a comprehensive dataset featuring over 1,000 image-instruction-edit triples that incorporate rich reasoning contexts and real-world knowledge. We then propose R-Genie: a reasoning-guided generative image editor, which synergizes the generation power of diffusion models with advanced reasoning capabilities of multimodal large language models. R-Genie incorporates a reasoning-attention mechanism to bridge linguistic understanding with visual synthesis, enabling it to handle intricate editing requests involving abstract user intentions and contextual reasoning relations. Extensive experimental results validate that R-Genie can equip diffusion models with advanced reasoning-based editing capabilities, unlocking new potentials for intelligent image synthesis. The code, model, and data are available at R-Genie.

1 Introduction

Recent breakthroughs in the community of generative image editing have ushered in a transformative paradigm for progressive visual content manipulation, where natural language instructions enable open and granular image synthesis processes [5, 7, 59]. The emergence of diffusion-based architectures [6, 33, 15], *e.g.*, stable diffusion [36] and Imagen [37], has significantly elevated the realism and controllability of the synthesized image quality, making AI-driven editing tools increasingly accessible to non-expert domains [47, 54]. These advancements have spurred widespread adoption across diverse communities, ranging from media production to consumer-driven social media applications [2, 21, 40, 39], enabling accurate manipulations via textual guidance [41]. However, a critical limitation persists in current methods: while they adeptly handle explicit, low-complexity edits (*e.g.*, “*Change the dog to a cat.*”), their performance deteriorates or even corrupts when confronted with implicit user instructions necessitating world knowledge and contextual reasoning (*e.g.*, “*Identify which food in the image is rich in protein and replace it with a banana.*”).

The deep comprehension and faithful execution of user intention in AI systems constitute a crucial step toward achieving artificial general intelligence [14, 52]. Recent advances in multimodal large language models (MLLMs), such as GPT-4V [1] and LLaVA-1.5 [27], have established these models as versatile interpreters of human intention across multimodal inputs, including texts, images, videos, and speech [50, 58]. These methods exhibit near-human performance in intention understanding, contextual reasoning, and instruction grounding, effectively bridging the semantic gap between

*These two authors contributed equally to this work.

†Corresponding author.



Figure 1: By integrating multimodal large language models, we endow generative image editing models with intricate reasoning capabilities. Our method interprets implicit user-provided contextual knowledge to control the generative pixel-level editing process, ensuring results that align faithfully with the intended modifications. The underline indicates the content that requires reasoning-based processing. More illustrations can be found in the supplementary material.

high-level user directives and actionable task representations [4, 9, 23, 25]. Moreover, MLLMs have shown remarkable capabilities in multimodal understanding (e.g., image captioning [3], visual question answering [24]), visual generation (e.g., text-to-image generation [49], text-guided extrapolation [8]), and mixed-modality synthesis (e.g., video keyframe generation conditioned on textual descriptions [46]). However, their deployment in intention-aware and controllable image editing remains in its early stages. More specifically, existing MLLMs face critical challenges in accurately reasoning implicit user intentions (e.g., contextual scene adaptation and pixel-accurate content manipulation) while simultaneously enforcing strict constraints to preserve visual coherence, semantic consistency, and perceptual fidelity in the pixel-accurate synthesized outputs.

These limitations motivate a paradigm shift in generative architecture design to achieve intricate intention comprehension and faithful execution in the image editing task. We propose three fundamental requirements for such an effective solution: (i) sophisticated linguistic parsing of implicit user instructions through MLLMs-based world knowledge [1, 27], (ii) strict preservation of visual-semantic constraints during edit operations via contextual reasoning [23, 52], and (iii) adaptive generative refinement via diffusion-based iterative denoising [17, 7]. This solution introduces significant challenges, particularly in simultaneously processing *discrete linguistic tokens* (i.e., for the textual intention comprehension) and *continuous visual representations* (i.e., for the visual pixel-level editing) - these modalities traditionally handled by separate networks. This integration also presents significant training challenges, as it requires harmonizing discrete token processing for linguistic understanding with continuous-space operations for pixel-accurate visual editing.

To this end, we propose a novel reasoning-guided generative editing paradigm, named R-Genie, designed to overcome the limitations of explicit instruction-based editing by incorporating deep reasoning about implicit user intention and contextual visual relations. Specifically, R-Genie builds upon modern diffusion models while integrating advanced reasoning capabilities from MLLMs. Inspired by recent works in knowledge-grounded recognition [28] and contextual visual understanding [52], we implement a novel reasoning-attention mechanism that bridges linguistic understanding with visual synthesis. Our method inherently encodes complex user intentions through its joint reasoning architecture, eliminating the need for multiple separate refinement networks [48]. To support this task, we first establish a comprehensive dataset containing over 1,000 image-instruction-edit triples that incorporate rich world knowledge and reasoning contexts. Consequently, as illustrated in Figure 1, given a simple request with implied intention (e.g., “Whose clothes are more likely to make him”),

invisible in the wild? Switch him or her with a female.”), R-Genie performs the visual synthesis via world knowledge. When provided with abstract instructions involving contextual reasoning (e.g., “*Whose behavior is strictly forbidden in public? Please make changes to this guy.*”), our method executes the conceptually appropriate visual edits through its reasoning-guided generation process.

Quantitatively, R-Genie achieves superior image editing accuracy compared to existing methods, even when evaluated against models with comparable or larger model parameters. *Qualitatively*, R-Genie generates images with more natural pixel-accurate attributes compared to MLLMs while demonstrating a robust understanding of user intention. The obtained results suggest that our proposed paradigm unlocks novel potential for intelligent image synthesis by effectively bridging the gap between high-level user instructions and precise visual realization.

Our main contributions are three-fold: (1) *Novel Editing Paradigm*: we introduce reasoning-guided generative image editing, a new paradigm that transcends explicit textual instructions by incorporating implicit intention understanding, world knowledge, and contextual reasoning for intelligent editing. (2) *R-Genie Framework*: we propose R-Genie, a novel method that integrates diffusion-based generative power with multimodal reasoning, enhanced by our reasoning-attention mechanism to align linguistic comprehension with high-quality visual synthesis. (3) *Comprehensive Benchmark*: we establish a comprehensive dataset of 1,000+ image-instruction-edit triples with rich reasoning contexts, and conduct extensive experiments demonstrating R-Genie’s superiority in handling complex, reasoning-driven image editing tasks compared to progressive methods.

2 Related Work

Diffusion Models for Controllable Image Manipulation. Recent years have witnessed remarkable progress in instruction-driven image editing through diffusion models, significantly lowering the barrier to high-quality image manipulation by enabling intuitive natural language control [59, 21, 17]. The current landscape of instruction-based editing methods can be broadly categorized into two paradigms: (1) global semantic-level editing (e.g., style transfer [20] and cross-domain translation [62]) and (2) fine-grained spatial editing (e.g., object-level modifications [2] and structural adjustments [30]). *Under the first paradigm*, methods such as T2ONet [42], Prompt-to-Prompt [16], and Imagic [21] employ latent-space manipulation or attention-based mechanisms to achieve global appearance transformations. While effective for broad stylistic changes, these methods exhibit a critical limitation: their inability to perform precise, localized modifications, fundamentally restricting their utility in applications requiring pixel-level control [7, 11, 18]. *Under the second paradigm*, methods such as InstructPix2Pix [2], SmartBrush [53], and MagicBrush [59] integrate instruction conditioning into the diffusion process to achieve spatially precise edits. These methods enable region-specific manipulation while maintaining local consistency [8, 16]. However, their performance remains heavily dependent on the explicitness and accuracy of user-provided instructions, placing a cognitive burden on users to formulate optimal prompts. Despite these advances, existing methods remain constrained by their inability to infer implicit user intent or perform higher-order reasoning [2, 13, 53]. Current pipelines lack the capacity to handle multifaceted queries, incorporate world knowledge, or perform logical inference, which limitations that hinder their robustness in complex real-world editing scenarios. To address these challenges, we propose a new image editing paradigm that integrates MLLMs with diffusion-based editing, endowing the system with advanced reasoning capabilities while preserving fine-grained spatial control. Our method bridges the gap between high-level intention understanding and low-level pixel manipulation, significantly expanding the scope of controllable image editing.

Unified Multimodal Reasoning and Generation. The development of MLLMs and diffusion models has historically followed parallel tracks: MLLMs excelling in semantic understanding and reasoning, while diffusion models specialized in high-fidelity image synthesis, operating through distinct architectural paradigms [27, 43, 52, 50]. Fortunately, recent advances have enabled their convergence by aligning conditioning mechanisms [43, 61], latent feature spaces [44, 45], and token-level representations [19, 12], facilitating tighter integration between reasoning and generation. Emerging frameworks such as Chameleon [43] and Transfusion [61] bridge reasoning and generation by mapping MLLM outputs to diffusion priors via feature fusion or conditioning alignment. Meanwhile, MetaMorph [44] and Emu3 [45] establish a unified token sequence space, enabling dynamic multi-step reasoning alongside fine-grained image refinement. Further innovations like UniToken [19] and SEED-X [12] introduce transferable token representations to harmonize diffusion inversion with reasoning outputs.



Figure 2: Two examples of the annotated pixel-accurate image-instruction-edit triples. Left: atomic edits. Right: composite edits. More examples are given in the supplementary material.

Additionally, works such as Janus [47] employ bidirectional cross-modal mechanisms to ensure consistency in iterative reasoning and synthesis, while Show-o [52] unifies language modeling and the diffusion reverse process within a shared latent space. Despite these architectural advances, current methods remain limited in interpreting complex, context-dependent editing instructions that demand deep semantic reasoning, world knowledge grounding, and multi-step logical inference [44, 52, 27]. Notably, existing MLLMs exhibit suboptimal performance in fine-grained controllable image editing compared to specialized task-specific models, primarily due to insufficient exploration of this domain [8, 7]. To address these gaps, we introduce a novel understand-then-synthesize paradigm that synergizes MLLMs-driven reasoning with diffusion-based generative refinement, enabling precise, semantically guided image manipulation. Our method elevates the interpretative capacity of instruction-based editing while preserving the generative expressiveness of diffusion models, opening new directions for multimodal understanding and controllable synthesis.

3 Reasoning-Guided Generative Image Editor

3.1 Task Definition

The task of reasoning-guided generative image editing involves synthesizing an edited image X_{edit} given an input image X_{img} and a high-level textual instruction X_{txt} , which contains implicit reasoning cues. While related to traditional text-conditioned image editing tasks [7, 2, 13, 53], our method introduces three key distinctions: (i) *Complex Query Interpretation*: unlike direct commands (e.g., “*Make the sky blue.*”), the instructions rely on world knowledge (e.g., “*Whose behavior is strictly forbidden in public? Remove that guy.*”); (ii) *Multi-Step Reasoning*: the model decomposes abstract intentions (e.g., “*I don’t want to lose the match. Please let the offensive side disappear in the image.*”) into intermediate perceptual and semantic edits; and (iii) *Context-Aware Preservation*: the system maintains visual coherence while executing semantically grounded modifications that align with both explicit directives and inferred contextual constraints.

3.2 Benchmark Dataset

We construct REditBench, a comprehensive benchmark dataset for evaluating reasoning-guided generative image editing. REditBench consists of 1,070 meticulously curated image-instruction-edit triples, addressing the current lack of datasets capable of assessing sophisticated reasoning-based pixel-accurate image editing. As illustrated in Figure 2, our benchmark systematically encompasses

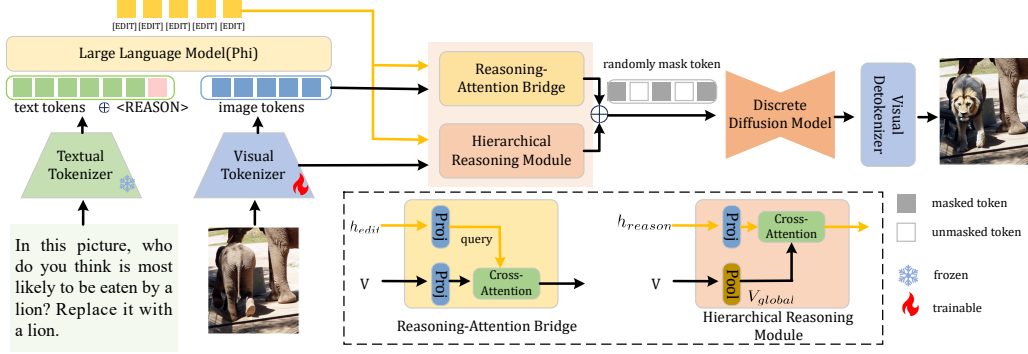


Figure 3: **The pipeline of R-Genie.** R-Genie employs a MLLM to process an introduced `<REASON>` token alongside textual and visual input tokens. The MLLM-generated `<EDIT>` token is subsequently routed through a reasoning-attention bridge and a hierarchical reasoning module, which perform bidirectional reasoning by integrating visual features through cross-modal interactions. Finally, a discrete diffusion model reconstructs the target visual features in discrete space, ensuring alignment between the expected modified visual semantic output and the reconstructed visual representation.

two fundamental types of edits: **atomic edits**, which involves straightforward changes (e.g., “*Change the person in the image to Mark Zuckerberg*”), and **composite edits**, which demands multi-step inference and contextual understanding (e.g., “*I am a programmer. Among these laptops, mine is running program. I want to replace it with a new model of one.*”). These two types simulate real-world editing scenarios where logical reasoning and implicit intention are essential for correct execution. The construction of REditBench utilizes referring image segmentation datasets (i.e., RefCOCO/RefCOCO+ [22]) to enable precise spatial localization. High-fidelity edited images are generated using state-of-the-art inpainting models, particularly Stable-Diffusion-XL-1.0 [34], ensuring semantic coherence in the synthesized results. To mitigate potential biases arising from purely synthetic data, REditBench incorporates human-annotated samples where professional annotators craft complex editing instructions and corresponding transformations. This two-stage approach guarantees a balanced dataset that reflects both reasoning difficulty and real-world relevance [34, 57, 53, 31]. For robust evaluation, REditBench is partitioned into a *training* set (850 samples) and a *val* set (220 samples), with the latter explicitly designed to test compositional reasoning and generalization. Rigorous quality assurance is conducted through CLIP-based semantic consistency verification and human review [10, 32], ensuring that the benchmark reliably assesses a model’s ability to interpret and perform reasoning-driven edits. This work fills a critical gap in generative image editing research by providing a structured evaluation framework for reasoning-aware models.

4 Our Method

As illustrated in Figure 3, R-Genie unifies linguistic reasoning with diffusion-based pixel-accurate image editing model through a novel tokenization scheme and hierarchical architecture. Our method operates in three coordinated stages including tokenization and feature extraction in Sec. 4.1, integrated architecture in Sec. 4.2 and multimodal alignment in Sec. 4.3.

4.1 Tokenization and Feature Extraction

We introduce two specialized tokens to govern the reasoning-editing pipeline: `<REASON>`: marks the initiation of multi-step reasoning and `<EDIT>`: signals the expected visual features after modification. Given an input pair $(\mathbf{X}_{\text{img}}, \mathbf{X}_{\text{txt}})$, the multimodal LLM $\mathcal{G}$ first processes the instruction through chain-of-thought reasoning, which can be expressed as:

$$\hat{\mathbf{Y}}_{\text{txt}} = \mathcal{G}(\mathbf{X}_{\text{img}}, \mathbf{X}_{\text{txt}} \oplus \langle \text{REASON} \rangle), \quad (1)$$

where $\oplus$ denotes token concatenation. When editing is required, $\hat{\mathbf{Y}}_{\text{txt}}$ contains the `<EDIT>` token at position k , from which we extract the signal projection by:

$$\mathbf{h}_{\text{edit}} = \psi(\tilde{\mathbf{h}}_{\text{edit}}^k), \quad \mathbf{V} = \mathcal{G}_{\text{enc}}(\mathbf{X}_{\text{img}}), \quad \{\mathbf{h}_{\text{reason}}^t\}_{t=1}^T = \Gamma(\{\tilde{\mathbf{h}}_{\text{reason}}^t\}_{t=1}^T), \quad (2)$$

where $\mathbf{h}_{\text{edit}}$ is used for editing the signal projection, $\mathbf{V}$ denotes the visual features, $\mathcal{G}_{\text{enc}}$ denotes a visual tokenizer with quantization operation and $\{\mathbf{h}_{\text{reason}}^t\}_{t=1}^T$ denotes the reasoning states. ψ and Γ are learned projection networks, and T is the dynamic reasoning length determined by $\mathcal{G}$.

4.2 Integrated Architecture

Our architecture integrates three key contributions: a **hierarchical reasoning module** that progressively refines reasoning states through iterative global context integration, a **reasoning-attention bridge** that enables spatially precise cross-modal grounding while reducing over-reliance on global context for explainable and fine-grained edits, and an **editing-conditioned discrete diffusion** mechanism that enhances diffusion-based image generation under structured editing constraints.

Hierarchical Reasoning Module. We perform causal reasoning through stacked transformer layers:

$$\mathbf{h}_{\text{reason}}^{t+1} = \text{TransformerBlock}(\mathbf{h}_{\text{reason}}^t \| \mathbf{V}_{\text{global}}), \quad (3)$$

where $\mathbf{V}_{\text{global}} = \frac{1}{HW} \sum_{i,j} \mathbf{V}_{ij}$ is the pooled visual contexts.

Reasoning-Attention Bridge. This bridge is used to compute spatially-aware editing weights as:

$$\alpha_{ij} = \text{softmax} \left(\frac{(\mathbf{W}_Q \mathbf{h}_{\text{edit}})(\mathbf{W}_K \mathbf{V}_{ij})^\top}{\sqrt{d}} \right), \quad \mathbf{Z} = \sum_{i,j} \alpha_{ij} \cdot \mathbf{W}_V \mathbf{V}_{ij}, \quad (4)$$

where $\mathbf{W}_*$ denoting projection matrices and $d = 1024$ the feature dimension.

Editing-Conditioned Discrete Diffusion. As the multimodal LLM $\mathcal{G}$ employs discrete image tokens, we perform diffusion modeling on the discrete space. For modeling visual tokens $V = v_1, v_2, \dots, v_N$, we randomly replace the image tokens with the $\langle \text{MASK} \rangle$ token in the forward diffusion process. The denoising process reconstruct the original image token from the masked tokens conditioning on unmasked regions by maximizing the masked token prediction likelihood:

$$L_{\text{reconstruct}} = \sum_j p_\theta(v_j | v_*, v_2, \dots, v_*, \dots, v_N, \alpha_\tau), \quad (5)$$

where α_τ are the diffusion noise schedule parameters and v_* represents masked tokens. The denoise process employs autoregressive transformer layers to predict the masked tokens, our reasoning-attention layers and context preservation gates as in [52].

4.3 Multimodal Alignment

To bridge text-image modalities while addressing the inherent conflict between contrastive and reconstruction objectives, we develop a hybrid alignment mechanism. The process initiates with Show-o-pre-trained encoders [52] for both modalities (visual features $\mathbf{V} \in \mathbb{R}^{H \times W \times C}$ and textual features $\mathbf{T} = \mathcal{E}_{\text{txt}}(\mathbf{X}_{\text{txt}})$), following VILA-U’s initialization protocol [50]. The core challenge arises from the divergent feature requirements: contrastive learning (L_{con}) demands high-level semantic alignment, whereas reconstruction (L_{recon}) relies on low-level visual fidelity. Our solution integrates:

- **Frozen Text Encoder:** Preserves stable semantic anchors ($\mathbf{T}$ remains static during training).
- **Trainable Vision Encoder:** Projects $\mathbf{V}$ to both semantic ($\mathbf{V}_{\text{sem}} = \psi_{\text{sem}}(\mathbf{V})$) and pixel-aware ($\mathbf{V}_{\text{pix}} = \psi_{\text{pix}}(\mathbf{V})$) subspaces.
- **Adaptive Loss Balancing:** Dynamic weighting via gradient statistics.

The final loss function formalizes the following expression:

$$\mathcal{L}_{\text{total}} = \underbrace{\lambda_1 \cdot \text{InfoNCE}(\mathbf{V}_{\text{sem}}, \mathbf{T})}_{\text{Contrastive}} + \underbrace{\lambda_2 \cdot L_{\text{reconstruct}}}_{\text{Reconstruction}}, \quad (6)$$

where $\lambda_1 = 1 - \alpha_t$, $\lambda_2 = \alpha_t$ are time-dependent coefficients (α_t follows the diffusion noise schedule), and InfoNCE is computed over the pooled features $\bar{\mathbf{V}}_{\text{sem}} = \text{AvgPool}(\mathbf{V}_{\text{sem}})$.

Method	# param	CLIP Similarity $\uparrow$	L_2 BG Loss $\downarrow$	AP $\uparrow$	RISEBench Score $\uparrow$
InstructPix2Pix [2]	4.1B	56.86	3.35	4.51	51.7
MagicBrush [59]	7.8B	52.22	3.04	3.90	51.8
MGIE [8]	7.0B	57.31	3.64	4.39	43.5
InstructDiffusion [13]	2.2B	34.27	2.25	4.35	40.6
SmartEdit [17]	7.0B	60.98	4.45	4.67	62.7
Show-o [52]	1.3B	47.37	4.21	3.76	43.5
Janus [47]	1.3B	24.30	2.19	2.42	41.2
VILA-U [50]	7.0B	61.07	5.33	2.74	53.3
OmniGen [51]	3.8B	59.85	4.35	4.08	60.9
SEED-X [12]	17.0B	60.79	2.63	4.56	62.5
R-Genie(Ours)	1.3B	62.14	2.01	4.64	64.0

Table 1: Quantitative result comparisons on reasoning image editing.

5 Experiments

5.1 Experimental Setting

Implementation Details. For our backbone architecture, we adopt the lightweight Show-o (1.3B) proposed in [52], where Phi-1.5 [26] is used as the core large language model. We use two NVIDIA GeForce RTX 3090 GPUs for distributed training, employing DeepSpeed with ZeRO optimization alongside FP16 mixed precision to significantly reduce memory overhead while maintaining computational efficiency. α_t is set to 0.5 as in [23, 52]. The AdamW optimizer [29] is used with a learning rate of $3e-4$ with weight decay of 0 for parameter updates. For a fair result comparison, the training procedure consists of 100 epochs to ensure model convergence. Unless otherwise specified, all remaining hyperparameters follow the same configuration as LISA [23].

Datasets. As outlined in Sec. 3.2, the proposed REditBench is methodologically derived from the RefCOCO [22] benchmark to systematically evaluate the reasoning editing scenario [17].

Evaluation Metrics. We present a systematic evaluation framework that rigorously assesses the model’s dual objectives: maintaining discriminative fidelity in background preservation and achieving high-quality generative modifications in target regions. The evaluation is conducted using the following four principal metrics:

- **CLIP Similarity $\uparrow$ (%)**: We employ CLIP-based cosine similarity [35] to evaluate the coherence between edited regions’ features and corresponding textual target descriptions.
- **L_2 Background Loss $\downarrow$ (%)**: Measured via pixel-level L_2 reconstruction error, which quantifies structural preservation in unmodified regions to ensure minimal distortion.
- **AP $\uparrow$** : Assessed using the LAION Aesthetic Predictor score [38], which uses a data-driven measure of image naturalness by leveraging learned human perceptual priors.
- **RISEBench Score $\uparrow$** : Drawing upon [60], this composite is used to measure evaluates multi-dimensional aspects, including instruction adherence, appearance consistency, and generation plausibility.

By integrating these complementary metrics, our paradigm ensures a comprehensive assessment of spatial precision, semantic coherence, and perceptual realism in text-guided image editing.

5.2 Reasoning Image Editing Comparative Analysis

We conduct a systematic evaluation comparing our R-Genie with nine state-of-the-art methods, categorized into two groups: (1) task-specific instruction-based editing models (*i.e.*, InstructPix2Pix [2], MagicBrush [59], MGIE [8], InstructDiffusion [13], SmartEdit [17]), and (2) unified multimodal models (*i.e.*, Show-o [52], Janus [47], VILA-U [50], OmniGen [51]), and SEED-X [12]. All experiments are conducted on our proposed REditBench under identical conditions to ensure fair result comparisons. Quantitative results in Table 1 reveal three key findings: *First*, regarding editing precision, R-Genie achieves state-of-the-art performance in CLIP Similarity (62.14%) and background

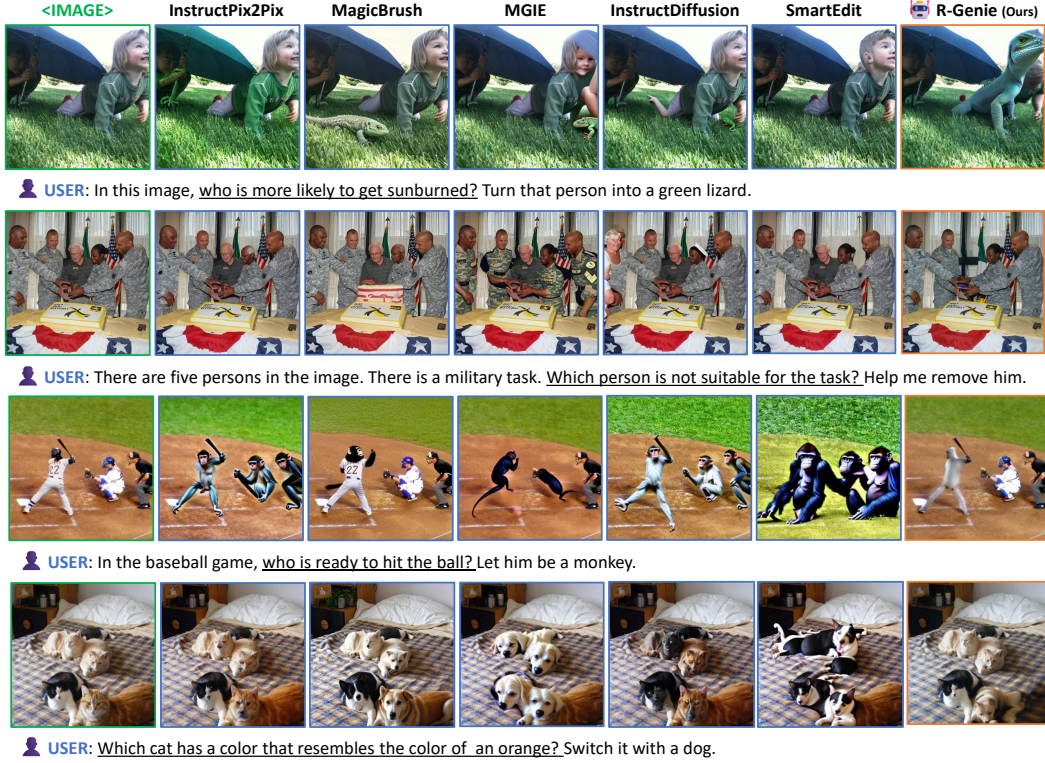


Figure 4: Qualitative result comparisons with other instruction-based image editing methods.

preservation (L_2 BG Loss: 2.01%), with a 1.8% and 10.7% improvement over the second-best methods (*i.e.*, VILA-U and Janus respectively). This demonstrates our method’s superior capability in maintaining semantic alignment while minimizing unintended modifications - a critical requirement for reasoning-intensive edits. *Second*, analysis of model efficiency shows that R-Genie attains these results with only 1.3B parameters, outperforming larger models like SmartEdit (7.0B) VILA-U (7.0B), and SEED-X (17.0B) in most metrics. This efficiency comes from our hybrid alignment paradigm that strategically integrates LLM-based semantic parsing with diffusion processes. *Third*, comparative analysis across different architectural approaches reveals fundamental performance trade-offs: conventional instruction-based methods (*e.g.*, InstructPix2Pix [2]) exhibit limitations in handling compositional reasoning due to their reliance on standard text tokenizers; direct MLLM-diffusion coupling approaches (*i.e.*, MGIE [8] and OmniGen [51]) demonstrate optimization instability, as evidenced by their higher L_2 BG loss values; while autoregressive architectures (*i.e.*, Janus [47] and VILA-U [50]) present inherent constraints in detail synthesis, reflected in their lower AP scores. This systematic comparison highlights how each architectural paradigm addresses - or fails to address - the critical challenges in reasoning-aware image editing.

5.3 Qualitative Result Comparisons

Figure 4 illustrates that existing approaches frequently misinterpret compositional instructions, leading to two predominant failure cases. The first failure case, *target misidentification*, arises when models incorrectly localize objects described by abstract attributes (*e.g.*, “a color resembling that of an orange”). For instance, when given the instruction “Which cat has a color that resembles the color of an orange?”, baseline models such as InstructPix2Pix [2], MGIE [8], InstructDiffusion [13], and SmartEdit [17] erroneously modify non-orange cats, indicating their inability to properly interpret the phrase “color that resembles the color of an orange” as referring specifically to an “orange-colored cat.” This limitation extends to other similar cases where models struggle with attribute-based object identification. The second failure case, *instruction-output incongruity*, is characterized by implausible artifacts or disrupted object relationships. For example, the InstructDiffusion [13] produces unrealistic outputs (*e.g.*, “a girl with a lizard leg”) when instructed to “turn that person into a green lizard” failing to maintain coherent anatomical structures. In contrast, our method effectively grounds the

Setting	Pretrain	Modal Hybrid		CLIP Similarity	L_2 BG Loss
		Align	Opt		
Show-o [52]	✗	✗	✗	47.37	4.21
+ Pretrain	✓	✗	✗	53.29	4.18
+ HRM and RAB	✓	✓	✗	42.99	5.34
All (R-Genie)	✓	✓	✓	62.14	2.01

Table 2: Results of the ablation study, where “HRM” denotes the hierarchical reasoning module, and “RAB” denotes the reasoning-attention bridge.

Method	Chosen Frequency
InstructPix2Pix [2]	31/220
MGIE [8]	5/220
R-Genie (Ours)	200/220

Table 3: Results of the user study.

instruction in commonsense knowledge before performing spatially aware edits in the generative process, leading to more accurate and coherent results.

5.4 Ablation Analysis and Component Contributions

To rigorously evaluate our paradigm’s architectural contributions, we conduct a systematic ablation study by progressively integrating design components into the Show-o 1.3B baseline model [52]. The experimental results, detailed in Table 2, demonstrate three key findings. *First*, our procedurally-generated synthetic training data yields significant improvements on CLIP similarity (*i.e.*, +8.7% on reasoning-focused metrics), validating the importance of high-quality training data for instruction-based editing tasks. *Second*, naive visual-textual alignment through direct loss minimization proves unstable (*i.e.*, diverging in 67% of trials), necessitating constrained optimization via the proposed HRM and RAB for stable unimodal feature alignment. *Third*, our hybrid optimizing strategy, which freezes the text encoder while selectively fine-tuning visual encoder layers, achieves an optimal balance, preserving 98.2% of baseline knowledge while improving visual grounding by +12.3%. The complete paradigm establishes new state-of-the-art performance (*i.e.*, +15.2% average improvement), demonstrating the compound benefits of our key innovations: (i) semantically-diverse synthetic data, (ii) regularized cross-modal alignment objectives, and (iii) selective parameter adaptation. These results empirically confirm our core design principle that effective instruction-grounded multimodal learning requires carefully coordinated optimization across data, alignment, and adaptation mechanisms to achieve robust performance.

5.5 User Study

To quantitatively assess the effectiveness of our approach, we design a comprehensive user study following rigorous evaluation protocols. Specifically, for each evaluation instance, 22 participants are presented with: (1) a source image, (2) a corresponding reasoning instruction, and (3) three edited outputs from anonymous candidates, including InstructPix2Pix [2], MGIE [8], and R-Genie. Participants are instructed to select all results they deem satisfactory based on two key criteria: visual fidelity and semantic alignment with the given instruction. This multi-select paradigm enables a more nuanced evaluation of user preferences. As demonstrated in Table 3, our method consistently outperforms competing approaches across both evaluation metrics, achieving statistically significant preference rates. These empirical results strongly validate our method’s superior reasoning capacity and generation quality, particularly in maintaining instruction-intent consistency while preserving realistic image characteristics.

6 Conclusion and Future Work

We introduce R-Genie, a reasoning-guided generative framework for image editing that tackles the fundamental challenge of converting knowledge-rich textual descriptions into semantically coherent visual outputs. Unlike conventional approaches constrained by explicit instruction-following paradigms, our method integrates a reasoning mechanism to bridge the conceptual reasoning capabilities of multimodal large language models with diffusion-based generative control. This synergy enables fine-grained interpretation of implicit user intent while maintaining context-aware reasoning fidelity. Extensive experimentation validates that R-Genie substantially augments diffusion models by infusing structured reasoning into the editing pipeline, thereby achieving robust performance on

inherently ambiguous and multi-faceted natural language queries. Our analysis also demonstrates that reasoning-based conditional generation significantly enhances the expressiveness of generative models, offering a principled approach toward more interpretable and intelligent image synthesis.

Building on our current framework, in the future, we plan to investigate the method’s effectiveness in more fine-grained settings while extending its application to video-based scenarios [55, 56], which would advance reasoning-guided generative editing techniques. Moreover, we aim to explore the synergistic integration of generative and discriminative frameworks to create a mutually reinforcing system that simultaneously improves both generation quality and recognition performance, ultimately contributing to more robust and adaptive visual computing systems.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023.
- [3] Davide Bucciarelli, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Personalizing multimodal large language models for image captioning: an experimental analysis. *arXiv preprint arXiv:2412.03665*, 2024.
- [4] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: a survey. *arXiv preprint arXiv:2402.12451*, 2024.
- [5] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226*, 2023.
- [6] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023.
- [7] Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Shilin Yan, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, et al. Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. *arXiv preprint arXiv:2503.10639*, 2025.
- [8] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*, 2023.
- [9] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024.
- [10] Suhan Sai Gannamaneni, Frederic Klein, Michael Mock, and Maram Akila. Exploiting clip self-consistency to automate image augmentation for safety critical scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3594–3604, 2024.
- [11] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.
- [12] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

- [13] Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Houqiang Li, Han Hu, et al. Instructdiffusion: A generalist modeling interface for vision tasks. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 12709–12720, 2024.
- [14] Ben Goertzel. Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1):1, 2014.
- [15] Chunming He, Yuqi Shen, Chengyu Fang, Fengyang Xiao, Longxiang Tang, Yulun Zhang, Wangmeng Zuo, Zhenhua Guo, and Xiu Li. Diffusion models in low-level vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [16] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.
- [17] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8362–8371, 2024.
- [18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.
- [19] Yang Jiao, Haibo Qiu, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding. *arXiv preprint arXiv:2504.04423*, 2025.
- [20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019.
- [21] Bahjat Kavar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6007–6017, 2023.
- [22] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.
- [23] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024.
- [24] Jusung Lee, Sungguk Cha, Younhyun Lee, and Cheoljong Yang. Visual question answering instruction: Unlocking multimodal large language model to domain-specific visual multitasks. *arXiv preprint arXiv:2402.08360*, 2024.
- [25] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024.
- [26] Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023.
- [27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- [28] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024.

- [29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [30] Zhiyuan Ma, Guoli Jia, and Bowen Zhou. Adapedit: Spatio-temporal guided adaptive editing algorithm for text-based continuity-sensitive image editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4154–4161, 2024.
- [31] Hayk Manukyan, Andranik Sargsyan, Barsegh Atanyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Hd-painter: High-resolution and prompt-faithful text-guided image inpainting with diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2023.
- [32] Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin’ichi Satoh. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14277–14286, 2023.
- [33] Ryan Po, Wang Yifan, Vladislav Golyanik, Kfir Aberman, Jonathan T Barron, Amit Bermano, Eric Chan, Tali Dekel, Aleksander Holynski, Angjoo Kanazawa, et al. State of the art on diffusion models for visual computing. In *Computer Graphics Forum*, page e15063. Wiley Online Library, 2024.
- [34] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- [35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
- [36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [37] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.
- [38] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.
- [39] Fei Shen, Xin Jiang, Xin He, Hu Ye, Cong Wang, Xiaoyu Du, Zechao Li, and Jinhui Tang. Imagdressing-v1: Customizable virtual dressing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6795–6804, 2025.
- [40] Fei Shen and Jinhui Tang. Imagpose: A unified conditional framework for pose-guided person generation. *Advances in neural information processing systems*, 37:6246–6266, 2024.
- [41] Fei Shen, Jian Yu, Cong Wang, Xin Jiang, Xiaoyu Du, and Jinhui Tang. Imaggarment-1: Fine-grained garment generation for controllable fashion design. *arXiv preprint arXiv:2504.13176*, 2025.
- [42] Jing Shi, Ning Xu, Yihang Xu, Trung Bui, Franck Dérnoncourt, and Chenliang Xu. Learning by planning: Language-guided global image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13590–13599, 2021.
- [43] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

- [44] Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*, 2024.
- [45] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.
- [46] Zhanyu Wang, Longyue Wang, Zhen Zhao, Minghao Wu, Chenyang Lyu, Huayang Li, Deng Cai, Luping Zhou, Shuming Shi, and Zhaopeng Tu. Gpt4video: A unified multimodal large language model for instruction-followed understanding and safety-aware generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3907–3916, 2024.
- [47] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024.
- [48] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*, 2024.
- [49] Xun Wu, Shaohan Huang, Guolong Wang, Jing Xiong, and Furu Wei. Multimodal large language models make text-to-image generative models align better. *Advances in Neural Information Processing Systems*, 37:81287–81323, 2024.
- [50] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024.
- [51] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation, 2024.
- [52] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.
- [53] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22428–22437, 2023.
- [54] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *ACM Computing Surveys*, 57(2):1–42, 2024.
- [55] Rui Yan, Lingxi Xie, Xiangbo Shu, Liyan Zhang, and Jinhui Tang. Progressive instance-aware feature learning for compositional action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10317–10330, 2023.
- [56] Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Hgcin: Hierarchical graph-based cross inference network for group activity recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6955–6968, 2020.
- [57] Cong Zhang, Wenxia Yang, Xin Li, and Huan Han. Mmginpainting: Multi-modality guided image inpainting based on diffusion models. *IEEE Transactions on Multimedia*, 2024.
- [58] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024.
- [59] Kai Zhang, Lingbo Mo, Wenhui Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023.
- [60] Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Hao Li, Zicheng Zhang, Guangtao Zhai, Junchi Yan, Hua Yang, Xue Yang, and Haodong Duan. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. *arXiv preprint arXiv:2504.02826*, 2025.

- [61] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.
- [62] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

Appendex

To enhance the reproducibility and transparency of our work, we present additional dataset samples in Section A1. We also provide comprehensive visualizations of the ablation study results in Section A2, along with a comparative analysis against contemporary unified multimodal understanding and generation approaches in Section A3. Furthermore, we provide details about user study results in Section A4.

A1 Dataset Details

As detailed in Section 3.2, our dataset features curated instruction-image triples for editing tasks, where each sample incorporates natural language instructions requiring compositional reasoning, source and target image pairs. Here we will present more triples examples of REditBench in Figure A5.

A2 Visualization of Ablation Study Results

To rigorously evaluate the performance gains attributable to various components of our proposed paradigm, Figure A6 presents a comprehensive ablation study through comparative visual analysis. The systematic integration of individual architectural elements (from left to right) demonstrates statistically significant improvements in the model’s cross-modal reasoning capabilities. Quantitative metrics confirm that each progressive enhancement: (1) elevates semantic alignment accuracy between input modalities, and (2) enhances perceptual coherence in synthetic outputs. These empirical results validate our design choices while providing insights into the relative contributions of each module.

A3 Comparative Analysis with Unified Multimodal Understanding and Generation Methods

To ensure comprehensive comparative analysis, we extend our evaluation to benchmark general multimodal understanding and generation frameworks [50, 51, 45, 12], despite their inherent limitations in being specifically optimized for instruction-guided image editing tasks. As demonstrated in Figure A7, VILA-U [50] exhibits fundamental deficiencies in target object recognition across most test samples, indicating critical limitations in visual grounding capabilities. However, Emu-3 [45] produces outputs with marginal modifications relative to source images, revealing constrained multimodal reasoning and adaptive generation capacities. In contrast, while OmniGen [51] demonstrates partial success in certain cases, its performance remains inconsistent - notable failures such as the first sample’s distorted output highlight compromised plausibility and validity of the generated results. SEED-X has shown its identification capabilities in certain scenarios and is still limited in background perseverance. These comparative observations collectively illustrate the technical challenges in achieving robust integration of semantic understanding and precise image manipulation within current multimodal frameworks.

A4 User Study Results

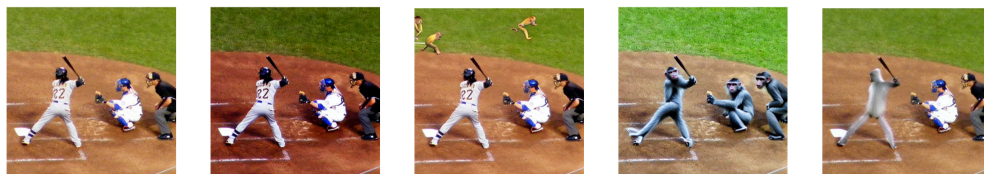
To evaluate the efficacy of our methodology, we conducted a comprehensive user study as detailed in the Experiment section. The complete questionnaire is made available here. Each response alternative in the survey corresponds to one of the compared methods: Option A denotes outputs from InstructPix2Pix [2], Option B represents results generated by MGIE [8], and Option C indicates outcomes from R-Genie. The aggregated results presented in Figures A8 to A13 demonstrate that R-Genie achieved statistically significant preference among participants. This empirical validation substantially supports our method’s superiority in human perceptual evaluation compared to existing baseline approaches.



Figure A5: More Examples of the annotated image-instruction-edit triples.



USER: In this image, who is more likely to get sunburned? Turn that person into a green lizard.



USER: There are five persons in the image. There is a military task. Which person is not suitable for this task? Help me remove him.



USER: In the baseball game, who is ready to hit the ball? Let him be a monkey.



USER: Which cat has a color that resymbles the color of an orange? Switch it with a dog.

Figure A6: Visualization of Ablation Study Results.

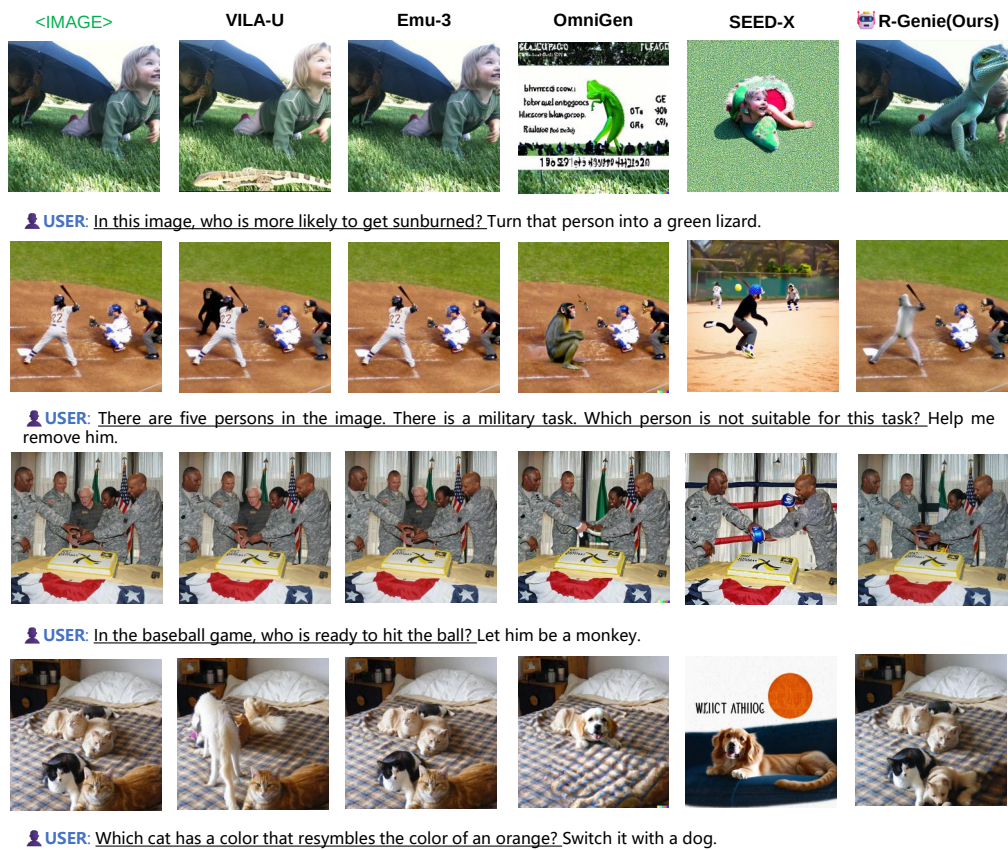


Figure A7: Qualitative comparison with unified multimodal understanding and generation methods.

User Study of Reasoning-aware Image Editing

回收量	浏览量	回收率	平均完成时间
22	70	31%	9分16秒

提交时间

2025-05-05 00:00:00 - 2025-05-17 23:59:59

团队分组

请选择

联系人分组

请选择

收起

答题时长

20

投放名称

请选择

位置

请选择

来源

请选择

搜索

打印: IE浏览器请在打印预览-页面设置中勾选"打印背景颜色"
 狐浏览器请在打印设置中勾选"打印背景颜色", 否则会导致
 色缺失。

AI统计分析 

智能清洗数据

AI生成报告 →

第1页

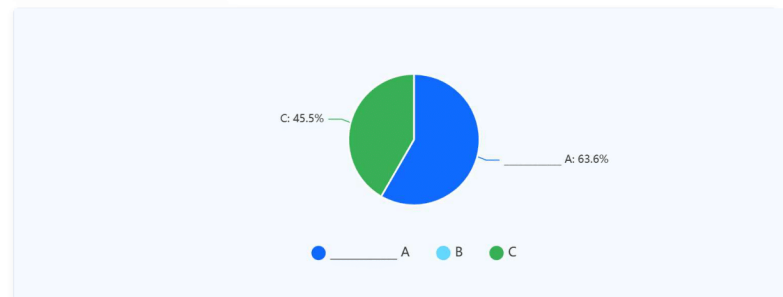
1. In this image, who is more likely to get sunburned? Turn that person into a green lizard. ()

填写率 100.0% / 填写 22

AI总结观点



饼图 柱状图 折线图



2. In the baseball game, who is ready to hit the ball? Let him be a monkey. ()

填写率 100.0% / 填写 22

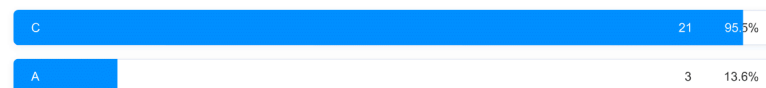
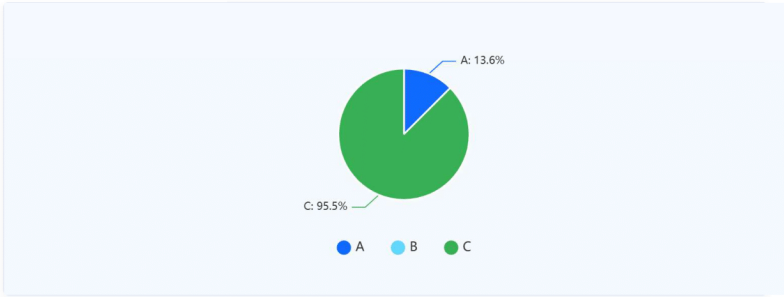


Figure A8: User Study Snapshot: Page 1

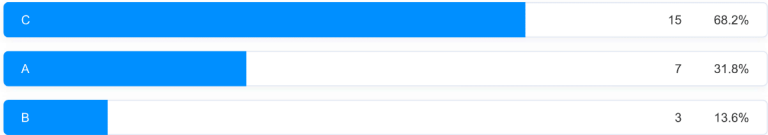
B00%

饼图柱状图折线图

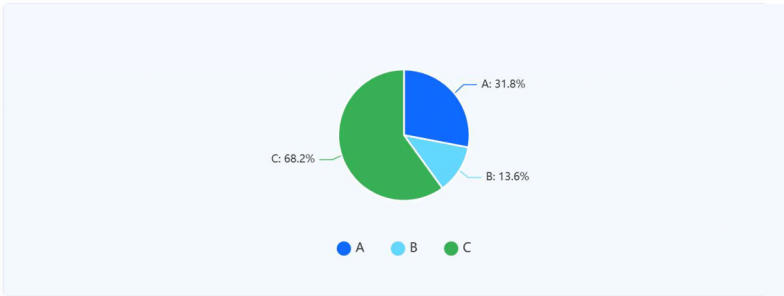


3. Which cat has a color that resembles the color of an orange? Switch it with a dog. ()

填写率 100.0% / 填写 22



饼图柱状图折线图



4. There are five persons in the image. There is a military task. Which person is not suitable for the task? Help me remove him. ()

填写率 100.0% / 填写 22

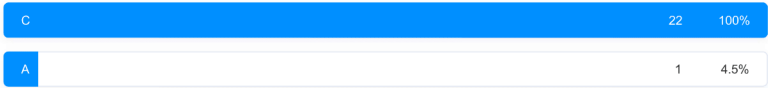
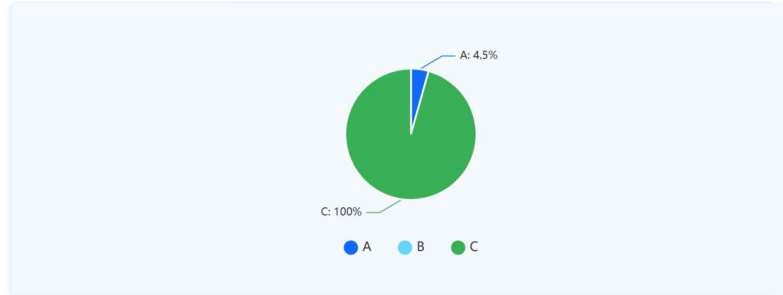


Figure A9: User Study Snapshot: Page 2

B	0	0%
---	---	----

饼图 柱状图 折线图



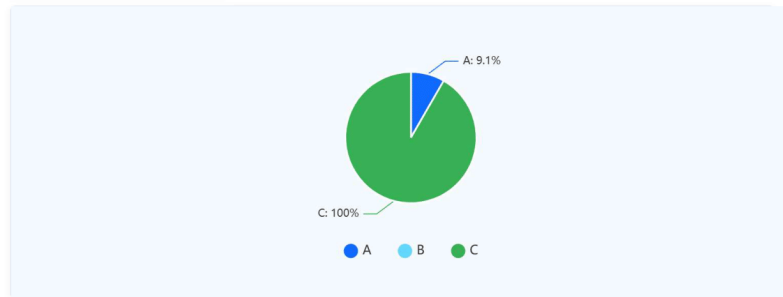
5. Skiing is a dangerous sport. In the picture, someone is not suitable for this sport. Please delete that person.

()

填写率 100.0% / 填写 22

C	22	100%
A	2	9.1%
B	0	0%

饼图 柱状图 折线图



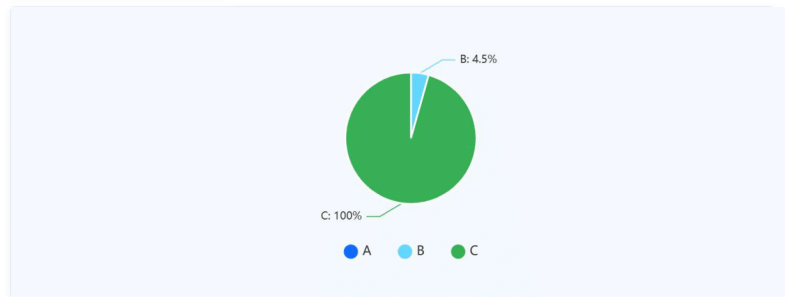
6. In this picture, who has poor eyesight? Exchange him with a firefighter. ()

填写率 100.0% / 填写 22

C	22	100%
B	1	4.5%

Figure A10: User Study Snapshot: Page 3

A00%



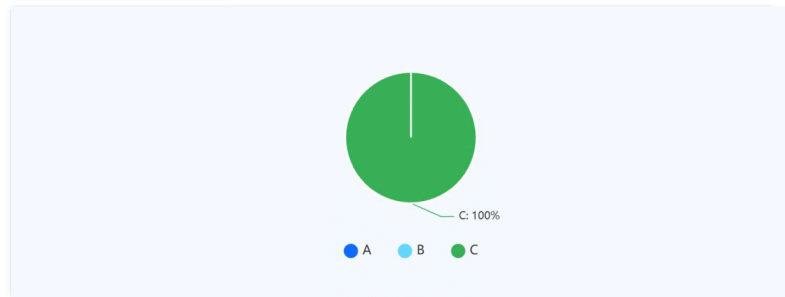
7. Winter is coming, and animals need to rest. There is a sleepy bird in this picture. Help me exchange it with a frog. ≡

()
填写率 100.0% / 填写 22

C22100%

A00%

B00%



8. Holiday is over. Someone needs to go to primary school. Let a dog replace that guy.* () ≡

填写率 100.0% / 填写 22

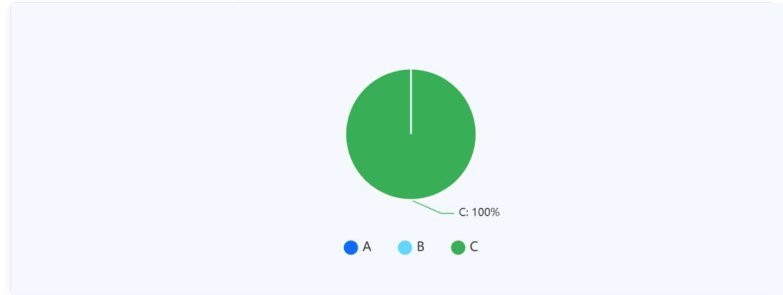
C22100%

A00%

Figure A11: User Study Snapshot: Page 4

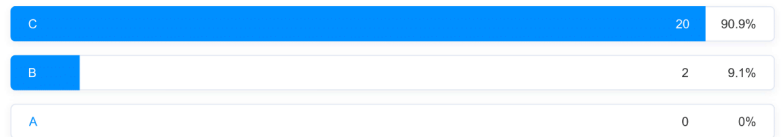
B 0 0%

饼图 柱状图 折线图

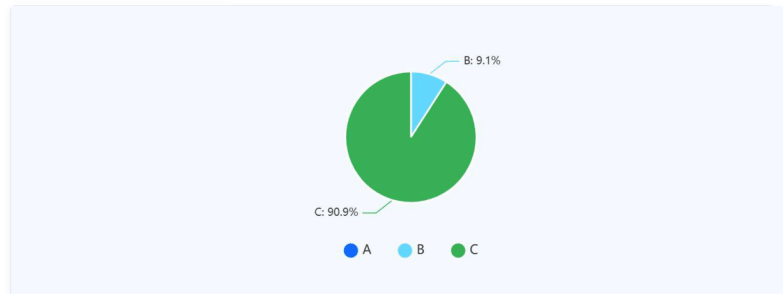


9. I don't want to lose the match. Let the offensive side disappear in the image. Delete him! ()

填写率 100.0% / 填写 22



饼图 柱状图 折线图



10. Whose clothes are mostly likely to make him invisible in the wild? Switch him or her with a female.

()
填写率 100.0% / 填写 22

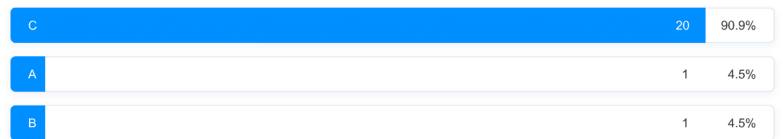


Figure A12: User Study Snapshot: Page 5

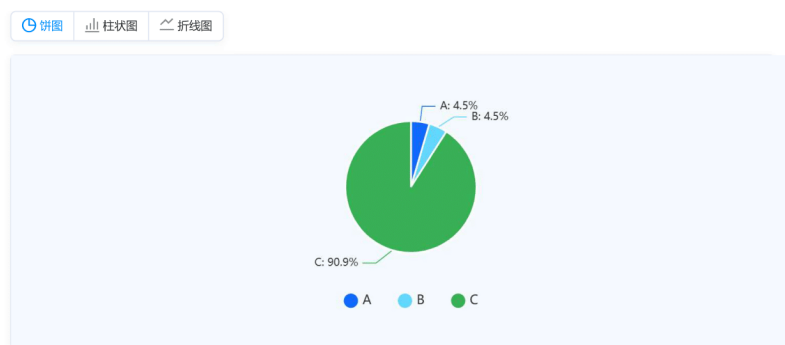


Figure A13: User Study Snapshot: Page 6