WHEN WORLDS MEET: BENCHMARKING IMPLICIT CAUSAL REASONING WITH IRS-BENCH

Anonymous authors
Paper under double-blind review



Figure 1: **Implicit reaction reasoning in image composition.** Each example combines a background scene with a source object and requires the model to infer the interaction-induced state change or scene effect. Reaction-aware composition realizes the implied reaction while preserving the source object; naive composition may remain visually plausible but leave the interaction absent.

ABSTRACT

An image captures a partial snapshot of a local world. Combining foreground and background snapshots therefore requires more than harmonizing pixels: the merged scene must realize the state changes and scene effects licensed by their interaction. Existing benchmarks for image composition primarily measure perceptual quality, identity preservation, and spatial alignment, leaving this merge-and-reaction problem untested. We formulate it as image-conditioned implicit causal reasoning and introduce **IRS-Bench**. Each instance provides a foreground image F, a background image B, and an instruction q specifying the entities and their placement while withholding the governing relation and its consequence. The source images fix entity identities; post-interaction states and scene effects must be inferred from visual evidence. To accept valid alternatives, IRS-Bench distinguishes deterministic interactions with one admissible state from intentional interactions with an admissible set. The benchmark contains 1,000 pairs across eleven leaf tasks and five causal mechanisms: physical, biological, chemical, institutional, and functional. Each item includes a governing rule, admissible outcomes, and weighted binary visual checks; 172 items also provide photographic post-interaction references. Under the **ME-Ruler** score, the strongest non-ReactionGraph baseline obtains 61.1 and 71.6 under GPT-5.6 SOL and Qwen-3.8-27B, respectively, across 18 systems. A training-free ReactionGraph reference pipeline obtains 77.2 and 88.8 under the same judges. These results show that perceptual coherence and causal faithfulness are distinct capabilities and that models must carry interaction constraints across visual contexts rather than merely produce plausible composites.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

An image captures a partial observation of a local scene state. A multi-image composite merges two such observations, but validity requires more than visual compatibility: an entity entering another context may change state or induce an effect in the surrounding scene. We call this the *world-merging* problem. Image-composition systems have made substantial progress in subject insertion, style harmonization, spatial control, and lighting consistency (Zhang et al., 2023; Song et al., 2023; Zhang et al., 2025; Ahmed et al., 2026), while their evaluations largely measure whether the inserted entity is present, visually compatible, and correctly placed. They rarely test whether the composite depicts what should happen when the two visual contexts meet.

This formulation gives the benchmark a limited world-model perspective without claiming complete world simulation. Here, a "world" is an image-defined context of entities, properties, and relations; the benchmark asks whether a model carries constraints from one context into another. Mechanisms may be physical, biological, chemical, institutional, or functional, so appearance matching or physical-template recall is insufficient for behavioral, material, or rule-governed responses.

We study this capability as *implicit causal reasoning from images*. An instance provides (B, F, q), where B is a background image, F is a foreground image, and q specifies the entities and their placement. The instruction does not name the interaction or its consequence. The pair therefore fixes the identities of the entities, while the post-interaction states and scene effects must be inferred from the visual premises. This setting avoids a central confound in text-conditioned evaluation: if the consequence is written in the prompt, a model can reproduce the description without recovering the relation from the scene.

Correctness also depends on the causal regime. Some interactions, such as those involving mechanical or optical response under fixed conditions, yield a unique target state. Others, such as behavior or institutional response, admit multiple outcomes. We annotate these cases separately as *deterministic* and *intentional*, and score them against a unique target or an admissible set, respectively. This regime-aware definition makes the benchmark reference-independent: a valid alternative is accepted when it satisfies the annotated relation, while an unsupported state remains incorrect.

Based on this formulation, we introduce **IRS-Bench**, a benchmark of 1,000 foreground–background pairs covering eleven leaf tasks across five causal mechanisms: physical, biological, chemical, institutional, and functional. The data combine AI-generated, web or public-dataset, and author-photographed imagery. Each item records the governing relation, admissible outcomes, and an independently reviewed set of weighted binary visual checks that can be judged from the generated composite. A subset of 172 items includes photographic post-interaction references. We additionally release controlled variants that remove or weaken a necessary causal condition and a separate 100-item multi-foreground set for qualitative transfer analysis; neither is included in the primary 1,000-item aggregate score.

We evaluate 18 systems, including heuristic composition, dedicated multi-reference generators, open-weight image generation and editing models, proprietary systems, and three training-free ReactionGraph variants. The benchmark uses two separately specified multimodal judges and reports causal correctness separately from appearance diagnostics. The strongest non-ReactionGraph baseline obtains 61.1 and 71.6 overall under the two judges, while the best ReactionGraph variant obtains 77.2 and 88.8. The gap persists across mechanisms and demonstrates that perceptual coherence is not a reliable proxy for causal faithfulness.

Our contributions are:

- A formal task definition for image-conditioned causal reasoning with an identity lock, state freedom, and separate deterministic and intentional correctness criteria;
- IRS-Bench, a 1,000-item benchmark spanning eleven causal tasks, with governing-rule annotations, admissible outcomes, independently reviewed visual rulers, photographic references for 172 items, and controlled evaluation variants; and
- A reproducible evaluation of 18 systems with two independent multimodal judges, together with ReactionGraph, a training-free diagnostic baseline that makes the inferred interaction explicit before image realization.

Table 1: **Related benchmark comparison.** IRS-Bench jointly supports image-pair inputs, implicit causal reasoning, photographic post-interaction ground truth, and five causality dimension: Physical, biological, chemical, institutional, and functional. Its evaluation covers causal correctness, domain specificity, visual naturalness, spatial reasonableness, and identity preservation. ✓, ○, and ×denote defining, partial, and absent coverage, respectively.

	Dataset			Causality dimension					Evaluation dimension				
Dataset	Img. pair	Impl. reas.	Real GT	Phys.	Bio.	Chem.	Inst.	Func.	Causal corr.	Domain spec.	Visual nat.	Spatial reas.	Identity pres.
I. Text-prompt benchmarks													
GenEval (Ghosh et al., 2023)	×	×	×	×	×	×	×	×	×	×	×	✓	×
DPG-Bench (Hu et al., 2024)	×	×	×	×	×	×	×	×	×	×	×	✓	×
ConceptMix (Wu et al., 2024)	×	×	×	×	×	×	×	×	×	×	×	✓	×
GenAI-Bench (Li et al., 2024)	×	×	×	×	×	×	×	×	×	×	×	✓	×
Commonsense-T2I (Fu et al., 2024)	×	✓	×	✓	✓	×	○	○	✓	✓	×	×	×
GenEval 2 (Kamath et al., 2025)	×	×	×	×	×	×	×	×	×	×	×	✓	×
TIIF-Bench (Wei et al., 2025)	×	×	×	×	×	×	×	×	×	×	×	✓	×
LongBench-T2I (Zhou et al., 2025)	×	×	×	×	×	×	×	×	×	○	×	✓	×
R2I-Bench (Chen et al., 2025)	×	✓	×	○	×	×	○	✓	○	✓	×	○	×
T2I-CompBench++ (Huang et al., 2025b)	×	×	×	×	×	×	×	×	×	×	×	✓	×
UniGenBench++ (Wang et al., 2025)	×	×	×	×	×	×	×	×	×	×	○	✓	×
OneIG-Bench (Chang et al., 2026)	×	×	×	×	×	×	×	×	×	×	○	×	×
WISE (Niu et al., 2025)	×	✓	×	✓	✓	✓	○	○	○	×	✓	×	×
T2I-ReasonBench (Sun et al., 2026)	×	✓	×	✓	✓	✓	✓	×	×	×	✓	×	×
T2I-CoReBench (Li et al., 2026)	×	✓	×	✓	○	○	×	✓	○	×	×	○	×
II. Image-conditioned composition benchmarks													
OmniContext (Wu et al., 2026a)	✓	×	×	×	×	×	×	×	×	×	×	×	✓
XVerseBench (Chen et al., 2026a)	✓	×	×	×	×	×	×	×	×	×	○	×	✓
MICON-Bench (Wu et al., 2026b)	✓	×	×	×	×	×	×	×	✓	×	○	✓	✓
IRS-Bench (ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

2 Related Work

Text-to-Image Image Generation. Text-to-image generation has advanced with models such as Stable Diffusion (Rombach et al., 2022), using latent diffusion to create high-quality images from text prompts. However, these models primarily generate single-prompt images without addressing consistency across multiple visual references. Recent native unified multimodal models extend image generation to multi-reference settings (Li et al., 2025). OmniGen (Xiao et al., 2025) targets unified image generation, whereas OmniGen2 (Wu et al., 2026a) focuses on instruction-aligned multimodal generation. Complementary efforts address challenges in multi-image composition: AIComposer (Li et al., 2025) focuses on cross-domain image composition with stylistic discrepancies between inputs, UNO pursues stable and scalable customized generation, and DreamFuse adapts foregrounds to background environments and interactions for more coherent blending. Further work improves control over entities and spatial structure: XVerse (Chen et al., 2026a) enables fine-grained control over multiple subjects and semantic attributes, including pose, style, and lighting, while LAMIC (Chen et al., 2026c) incorporates spatial-layout awareness to generate coherent and consistent images from multiple references. Despite these advances in composition quality, style integration, and reference control, existing methods do not examine whether composites are causally plausible. Thus, a composite may appear visually natural while remaining logically incorrect. To address this gap, we present the first systematic study of implicit causal reasoning in image generation.

Benchmarks for text-to-image generation. As shown in Table 1, existing T2I benchmarks primarily follow two directions: compositional generation and reasoning-driven generation. Early benchmarks focus on explicit objects, attributes, and spatial relations, while later studies extend evaluation to dense compositions and complex instructions. Reasoning-oriented benchmarks, including Commonsense-T2I (Fu et al., 2024), R2I-Bench (Chen et al., 2025), WISE (Niu et al., 2025), and T2I-CoReBench (Li et al., 2026), further examine commonsense and scientific knowledge. However, text prompts do not fix entity identities, and target outcomes may be implicitly encoded in linguistic cues. Although image-conditioned benchmarks such as OmniContext (Wu et al., 2026a), XVerseBench (Chen et al., 2026a), and MICON-Bench (Wu et al., 2026b) introduce visual inputs, they mainly assess identity preservation, spatial reasonableness, and visual naturalness, without systematically evaluating the implicit causal consequences of visual interactions. IRS-Bench addresses this gap by introducing implicit causal reasoning under fixed visual premises. It covers physical, biological, chemical, institutional, and functional causality, while jointly evaluating causal correctness, domain specificity, visual naturalness, spatial reasonableness, and identity preservation. This design provides a comprehensive evaluation of both causal faithfulness and perceptual quality.

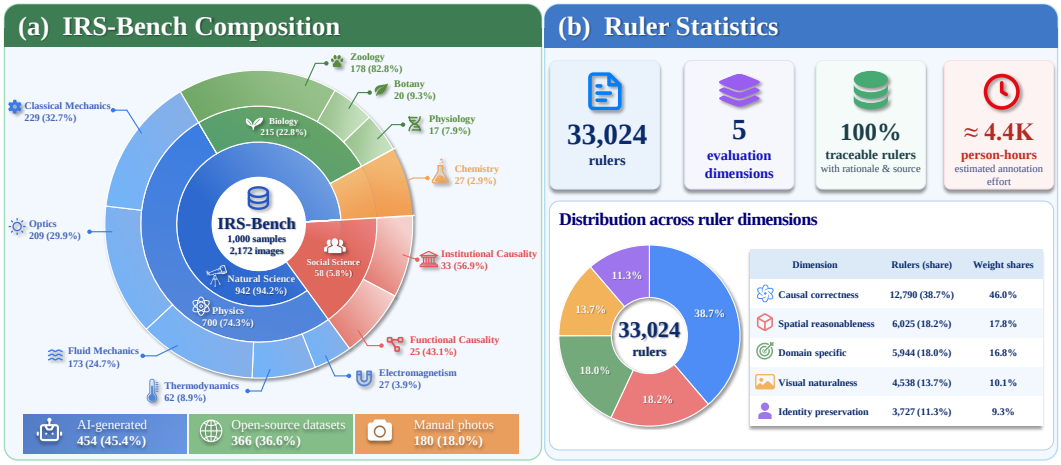


Figure 2: **Composition of IRS-Bench and ME-Ruler statistics.** (a) Hierarchical distribution of the 1,000 benchmark instances across domains, causal mechanisms, leaf tasks, and image provenance. (b) ME-Ruler contains 33,024 traceable binary rules across five evaluation dimensions; the chart reports their distribution, weight shares, and estimated annotation effort.

3 The IRS-Bench Benchmark

3.1 Hierarchical Capability Taxonomy

IRS-Bench tests whether a model can infer and render the reaction implied by merging a foreground with a background while preserving the source identities. The instruction names the entities and their placement but withholds both the relation and its consequence; success therefore requires evidence from the image pair rather than repetition of a textual answer. We annotate each instance along two axes: the mechanism that governs the interaction and the causal regime that defines which outcomes are admissible.

As shown in Figure 2, the mechanism hierarchy has two domains, five dimensions, and eleven leaf tasks. Natural Science comprises Physical Causality (Classical Mechanics, Optics, Fluid Dynamics, Thermodynamics, and Electromagnetism), Biological Causality (Zoology, Botany, and Physiology), and Chemical Causality (Chemical Reaction). Social Science comprises Institutional Causality (Institutional Rule Response) and Functional Causality (Functional Interaction). We classify an instance by the relation it tests rather than by the objects it contains, so the categories remain stable as object coverage grows. Appendix B gives task-level definitions and counts.

The causal regime is annotated separately. A *deterministic* instance admits one target state under the depicted conditions, whereas an *intentional* instance admits a set of valid responses because behavior, convention, or purpose does not force a unique realization. IRS-Bench contains 727 deterministic and 273 intentional instances. Evaluation uses the unique state or the admissible set accordingly, preventing a valid alternative from being penalized merely for differing from one reference image.

3.2 Construction of IRS-Bench

IRS-Bench contains 1,000 binary merge instances over the eleven leaf tasks. Instance i provides (F_i, B_i, q_i) and withholds a record containing the active relation, regime, admissible reaction, and scoring rulers. A system returns one composite $\hat{I}_i = G(F_i, B_i, q_i)$. Because q_i specifies only entities and placement, neither the reaction nor its evaluation criteria are available at generation time.

Collection and annotation. The 1,000 pairs comprise 454 AI-generated, 366 web or public-dataset, and 180 author-photographed instances. We retain a pair only when the foreground is identifiable, the background supports a definite interaction, and the entailed reaction is absent from both inputs. Annotators assign the governing leaf task and regime separately, then record *subject property, scene*

property, reaction, and *governing rule*; records whose rule merely restates the reaction are revised. Every item receives an independent second review under the protocol in Appendix C.2.

Ground truth and release. A subset of 172 instances has photographic ground truth captured after the causal process under a fixed camera and illumination. The remaining instances are evaluated from their annotations and rulers because one reference image cannot enumerate every admissible realization. Each instance carries an independently reviewed set of binary rulers linked to its source annotation. The audit release also includes two independently generated causal graphs per instance, a 100-instance multi-foreground subset, and quality-control records; causal graphs and hidden records are excluded from model inputs in the reported evaluation. Appendix C documents the full construction pipeline.

3.3 Ruler-Based Evaluation

GPT-5 proposes candidate rulers from the image pair and annotated admissible outcomes. Four annotators retain only questions that are relevant and decidable from the composite, add missing requirements, and ensure that intentional instances do not penalize legitimate alternatives. Each annotator rates every retained ruler from 1 to 10; ratings are normalized within an instance and averaged across $V = 4$ annotators. For n_i rulers on instance i, the ruler weight and evaluator-specific benchmark score are

$$w_{i,j} = \frac{1}{V}\sum_{v=1}^{V}\frac{100\, s_{i,j}^{(v)}}{\sum_{k=1}^{n_i} s_{i,k}^{(v)}}, \qquad S_m = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n_i} w_{i,j}\,\mathbf{1}\Big[J_m\Big(\hat{I}_i, r_{i,j}\Big) = \text{yes}\Big], \qquad V = 4, \quad (1)$$

where N is the number of evaluated instances, $s_{i,j}^{(v)} \in \{1, \ldots, 10\}$ is annotator v's importance rating, and $J_m(\hat{I}_i, r_{i,j})$ is evaluator m's binary verdict. Because $\sum_j w_{i,j} = 100$ for every instance, $S_m \in [0, 100]$ is the mean satisfied ruler weight. GPT-5.6 SOL and Qwen-3.8-27B judge independently, yielding S_1 and S_2.

The rulers decompose performance into five *assessment axes*, distinct from the mechanism dimensions in Section 3.1: causal correctness, domain specificity, visual naturalness, spatial reasonableness, and identity preservation. Identity covers the inserted entities' appearance and count; pose, action, orientation, position, scale, and lighting remain state-dependent. Appendix D provides the ruler schema and judging protocol.

4 Evaluation Protocol & Baseline Suite

4.1 Evaluated Baselines & Protocols

The suite is built to separate competence at rendering from competence at inference, so it spans systems that perform no generation at all, systems trained or engineered to insert a reference subject into a scene, and general generation–editing foundation models that receive the task as an instruction. We evaluate eighteen existing systems together with one training-free formulation of our own, instantiated on three generator stacks. The systems fall into four groups, defined by the technical route that produces the composite, and these are the row groups of Table 2.

Heuristic Baselines. Copy-paste places the foreground into the background without any generative modification. It encodes no model of the interaction between the two sources, and we record it as a lower bound rather than as a competitor: a system scoring below it is not merely failing to infer the reaction, it is rendering a state less correct than the input already provided.

Text-to-Image Image Generation. OmniGen2 (Wu et al., 2026a), AIComposer (Li et al., 2025), UMO (Cheng et al., 2025), LAMIC (Chen et al., 2026c), DreamFuse (Huang et al., 2025a), SHINE (Lu et al., 2026) and CareCom (Chen et al., 2026b) are dedicated to multi-reference generation and subject customization: they condition on one or more reference images and are trained or engineered to preserve the appearance of the inserted entity across images. This is the route the benchmark is designed to probe, because identity preservation is a constraint the task imposes rather than the objective these systems optimize.

Open-source Image Generation and Editing Model. Skywork UniPic 3.0 (Wei et al., 2026), Qwen-Image-3.0 (Wu et al., 2025), Qwen-Image-Edit-2511 (Wang et al., 2024), Wan2.7-Image (Wan et al., 2025), FireRed-Image-Edit (Team et al., 2026), HunyuanImage 3.0-Instruct (Cao et al., 2025) and FLUX.2-klein-9B (Labs et al., 2025) are open-weight unified models that accept an arbitrary image–text instruction and are not specialized for composition.

Close-source Image Generation and Editing Model. ChatGPT Images 2.0, Nano Banana 2 and Seedream 5.0 lite (Seedream et al., 2025) serve the same role through a public interface. The instruction specifies only entities and placement: the reaction remains encoded in the image pair, preventing a text-conditioned model from receiving the target consequence as a prompt.

Common protocol. Every system receives the same public input (B, F, q), where q specifies entities and placement but not their relation or its consequence. Existing baselines use their official composition interfaces; ReactionGraph additionally applies the fixed inference and compilation prompts of Section 4.2. No method accesses the withheld record, admissible set, knowledge graphs, or scoring rulers. Each method returns one composite, scored only after generation using Section 3.3. We run open-weight systems on eight NVIDIA A100 (80G) GPUs at their recommended settings, call proprietary systems through public interfaces with default configurations, and run judges at temperature zero under the resolution limits in Appendix D.

4.2 A Diagnostic Training-Free Baseline

We propose **ReactionGraph** as a diagnostic, training-free baseline. It adds an explicit inference step to a fixed image compositor, so that we can measure whether making the implied reaction explicit improves generation. The method receives only the public input (B, F, q), where B is the background, F is the foreground, and q specifies the requested entities and their placement.

$$\begin{aligned} M^{(0)} = \mathcal{G}_{\mathrm{comp}}(B, F, q), \mathcal{H} \quad &= \Psi_{\mathrm{infer}}(B, F, M^{(0)}, q), \\ \mathcal{P} = \Gamma_{\mathrm{comp}}(\mathcal{H}, B, F, q), \hat{I} \quad &= \Phi_{\mathrm{real}}(M^{(0)}, B, F, \mathcal{P}). \end{aligned} \tag{2}$$

The compositor first creates a *neutral seed* $M^{(0)}$ with the requested placement, scale, and approximate appearance. The remaining three operations infer the missing reaction, convert it into image-edit requirements, and realize those requirements on the seed. We execute Equation 2 once per item: there is no iterative repair, reranking, or output selection. Conditioning the final edit on $M^{(0)}$, B, and F provides identity references, but identity preservation remains an evaluated constraint rather than an architectural guarantee.

Reaction-graph inference. The input does not state what should happen after the merge. Ψ_{infer} therefore examines the visible properties of B and F and derives the state change and scene effects licensed by their interaction. It may use $M^{(0)}$ to locate entities and contact regions, but the seed is used for localization only; its appearance is not evidence for the reaction. The output is a small provenance-carrying directed acyclic graph

$$\mathcal{H} = \big(\mathcal{V}_O \dot{\cup} \mathcal{V}_C, \mathcal{E}, \tau\big), \quad \mathcal{V}_O = \mathcal{V}_B \dot{\cup} \mathcal{V}_F, \quad \mathcal{V}_C = \mathcal{V}_{\mathrm{state}} \dot{\cup} \mathcal{V}_{\mathrm{effect}}, \qquad \mathcal{E} \subseteq (\mathcal{V}_O \dot{\cup} \mathcal{V}_C) \times \mathcal{V}_C \times \Lambda, \tag{3}$$

where $\mathcal{V}_B$ and $\mathcal{V}_F$ contain observed scene and foreground properties, and $\mathcal{V}_C$ contains the inferred consequences. We divide $\mathcal{V}_C$ into (i) states adopted by the foreground entity and (ii) effects induced in the scene. An edge records the rule or physical relation that supports a consequence; τ orders observed properties before the interaction and consequences after it. The graph is acyclic, every consequence is connected to an observed property, and no entity may be introduced that is absent from B and F. It is therefore a structured derivation and provenance record, not a structural causal model estimated from two images.

Constraint compilation and state realization. A pixel generator cannot directly consume a graph. Γ_{comp} converts each consequence into one or more atomic, typed requirements:

$$c_k = (d_k, e_k, \phi_k, s_k, q_k), \qquad d_k \in \{\textsc{position}, \textsc{state}, \textsc{effect}, \textsc{identity}\}, \tag{4}$$

Here, e_k identifies the entity or region to which a requirement applies, ϕ_k specifies the visible condition to be realized, s_k records the supporting graph node and rule, and q_k, when present, is a post-realization predicate used internally to check the corresponding requirement. Position and State constraints specify the required configuration of an entity, Effect constraints specify the

Table 2: Quantitative comparison of our method with existing methods on the 11 IRS-Bench subtasks, measured by ME-Ruler scores using GPT-5.6 SOL and Qwen-3.8-27B as independent evaluators, respectively. Mech: Classical Mechanics; Opt: Optics; Fluid: Fluid Dynamics; Thermo: Thermodynamics; EM: Electromagnetism; Zool: Zoology; Bot: Botany; Physio: Physiology; Chem: Chemical Reaction; Inst: Institutional Rule Response; Func: Functional Interaction.

System	Natural Science											Social Science		Overall
	Physical Causality						Biological Causality				Chem.	Inst.	Func.	
	Mech.	Opt.	Fluid	Thermo.	EM	*avg*	Zool.	Bot.	Physio.	*avg*				
Heuristic Baselines														
Copy-paste	34.7/47.7	36.5/47.5	38.2/48.7	33.0/41.9	53.5/73.2	39.2/51.8	35.5/43.4	32.4/37.1	29.9/31.8	32.6/37.4	45.0/51.9	38.3/49.6	31.1/44.7	36.4/47.0
Text-to-Image Image Generation														
OmniGen2	29.3/41.9	37.6/58.3	32.9/48.2	31.7/41.6	41.4/61.9	34.6/50.4	28.1/40.2	29.0/39.4	28.5/39.0	28.5/39.5	37.8/49.0	36.7/56.5	29.6/47.0	32.4/47.3
AIComposer	35.8/51.1	36.0/50.9	33.7/50.7	31.5/50.3	46.3/60.8	36.7/52.8	30.2/40.2	35.5/57.5	37.7/46.2	34.5/48.0	38.9/52.8	38.3/49.5	32.1/49.3	34.6/49.3
UMO	29.1/42.2	37.9/58.4	32.7/48.3	31.7/42.5	42.4/60.8	34.8/50.4	28.0/40.2	29.1/36.4	29.2/40.1	28.8/38.9	36.5/49.4	38.6/56.1	33.7/53.9	32.5/47.6
LAMIC	19.2/27.4	22.3/33.9	21.5/32.2	19.1/28.6	26.3/43.9	21.7/33.2	20.8/29.1	26.0/38.9	24.5/31.2	23.8/33.1	28.7/39.6	31.3/37.1	33.5/48.2	22.0/31.9
DreamFuse	48.4/61.3	53.7/69.4	46.3/60.0	37.4/51.4	48.8/66.2	46.9/61.7	42.5/52.2	40.6/51.0	43.2/51.8	42.1/51.6	46.9/55.7	47.6/62.5	45.0/57.0	47.0/60.1
SHINE	36.5/43.7	37.7/52.3	39.4/51.5	29.2/41.4	36.1/53.1	35.8/48.4	34.2/43.6	39.9/50.5	28.6/31.9	34.2/42.0	45.5/58.5	36.0/50.7	35.0/49.2	36.5/47.6
CareCom	36.0/53.7	35.6/55.1	38.3/54.2	32.5/46.4	44.2/68.5	37.3/55.6	36.3/49.9	30.6/46.5	28.1/37.7	31.7/44.7	45.2/55.3	38.2/56.2	32.5/46.3	36.4/52.9
Open-source Image Generation and Editing Model														
Skywork UniPic 3.0	50.8/62.9	54.4/68.2	52.2/66.2	39.9/47.6	51.0/64.1	49.7/61.8	36.8/45.2	32.5/45.0	32.5/36.0	33.9/42.1	51.7/60.9	48.5/62.4	38.7/50.7	47.6/59.4
Qwen-Image-3.0	59.5/68.9	64.8/76.5	68.3/77.8	50.6/62.5	59.1/71.3	60.5/71.4	51.7/59.5	55.9/62.6	51.6/60.0	53.1/60.7	56.4/67.3	59.0/69.7	56.5/67.6	59.8/69.7
Qwen-Image-Edit-2511	53.1/65.1	60.2/76.2	54.6/66.5	43.9/53.6	44.5/59.8	51.3/64.2	44.3/51.7	30.5/43.4	39.2/49.5	38.0/48.2	50.4/59.1	52.9/70.8	50.2/61.1	51.7/63.7
Wan2.7-Image	54.4/66.0	59.7/73.9	59.7/71.0	43.0/53.0	52.9/67.3	53.9/66.2	45.6/55.5	44.4/55.2	39.8/47.5	43.3/52.7	55.4/67.7	51.8/62.4	47.1/57.6	53.4/65.1
FireRed-Image-Edit	54.4/67.3	62.5/77.8	58.4/70.8	43.4/54.4	50.7/61.0	53.9/66.3	45.4/52.9	42.2/50.3	39.7/43.4	42.4/48.9	56.3/70.0	53.9/65.9	47.3/63.6	53.7/65.7
HunyuanImage 3.0-Instruct	55.9/67.5	62.6/79.5	63.0/74.9	47.0/57.8	50.9/67.0	55.9/69.3	48.7/57.8	50.9/65.1	50.7/56.0	50.1/59.6	53.3/66.6	56.0/67.7	56.4/67.0	56.3/68.7
FLUX.2-klein-9B	53.9/65.7	61.7/76.8	58.6/71.1	47.2/56.3	51.4/68.5	54.6/67.7	46.4/55.7	49.8/53.1	43.9/56.3	46.7/55.0	50.8/62.2	52.6/69.6	50.9/63.3	54.1/66.2
Close-source Image Generation and Editing Model														
ChatGPT Images 2.0	56.1/69.6	63.4/80.6	60.6/75.3	49.6/60.8	54.7/60.4	56.9/69.3	48.9/60.5	53.3/67.1	47.1/56.4	49.7/61.3	55.0/67.3	56.8/71.1	52.4/65.5	56.4/70.1
Nano Banana 2	60.9/71.4	66.5/80.8	69.3/79.2	57.1/71.6	65.6/78.2	63.9/76.3	51.2/58.0	51.9/53.2	53.8/58.4	52.3/56.5	61.7/72.6	57.5/64.8	50.8/64.1	61.1/71.6
Seedream 5.0 lite	54.6/64.6	61.0/75.5	60.6/72.5	45.1/55.7	54.4/72.1	55.1/68.1	46.2/55.4	45.9/53.0	43.6/55.5	45.3/54.6	54.3/62.8	53.2/67.5	44.4/58.1	54.2/65.7
ReactionGraph (w/ Image 2.0)	74.8/87.0	74.8/**92.3**	74.7/89.3	76.9/87.1	**80.4/92.1**	76.3/89.6	71.4/82.1	73.5/88.7	69.5/**82.2**	71.5/**84.3**	82.1/92.9	**75.3/91.0**	**74.4/91.2**	74.5/88.1
ReactionGraph (w/ Nano Banana 2)	**77.5/88.0**	**77.6**/92.1	**81.6/92.6**	**80.5/92.7**	77.0/91.0	**78.8/91.3**	**72.5/82.4**	**73.6/90.7**	**70.7**/78.5	**72.3**/83.9	**84.5/93.0**	70.3/81.3	73.9/88.4	**77.2/88.8**
ReactionGraph (w/ Qwen-Image-3.0)	68.6/82.7	69.7/87.3	74.0/86.7	74.2/88.8	74.1/88.5	72.1/86.8	63.2/76.3	66.4/85.1	67.5/80.3	65.7/80.6	70.9/86.1	68.5/79.7	66.7/81.1	69.3/83.7

changes that should appear in the surrounding scene, and IDENTITY constraints specify properties that must be preserved. The compiler generates requirements for every consequence node in H without adding entities or spatial requirements unsupported by the graph. The predicates q_k are internal generation aids and are not used to compute the benchmark's ME-Ruler score.

The realization stage takes $\mathcal{P}$, B, F, and $M^{(0)}$ as input and produces the final image $\hat{I}$ in a single edit. It is instructed to realize the specified position, state, and scene effects while preserving foreground identity and avoiding changes to unrelated regions. We use the same procedure across the three reasoner–generator stacks, with prompts and constraint schemas held fixed.

5 EXPERIMENTS

5.1 QUANTITATIVE COMPARISON

To validate the effectiveness of our ReactionGraph and assess the challenge of the constructed IRS-Bench benchmark, we evaluate and compare representative methods with our ReactionGraph on all 11 IRS-Bench subtasks, as shown in Table 2. We use our ME-Ruler metric for evaluation, where each entry is a/b denote the ME-Ruler scores computed using GPT-5.6 SOL and Qwen-3.8-27B, respectively, as independent the MLLM evaluator. Existing systems span a wide range in the overall column, from 22.0 to 61.1, so the benchmark separates methods that a perceptual metric would rank closely together. ReactionGraph leads on every task under both judges, verifing its effectiveness. Besides, results under the ME-Ruler metric with Claude Sonnet 5 as the MLLM evaluator and under FID, LPIPS and CLIP-I are given in the Table 5 and Table 6 of Appendix F.2.

5.2 QUALITATIVE COMPARISON

Figure 3 compares ReactionGraph with representative baselines. Existing systems miss entities, place them wrongly, blend them unnaturally, or render consequences the contact does not support. ReactionGraph preserves the input entities and produces interactions consistent with the governing mechanism, and where a real composite exists its output is closer to the recorded interaction in appearance, state and consequence. The visual results match the quantitative ones; more are in Figure 4, Figure 5 and Figure 6 of the Appendix F.4.

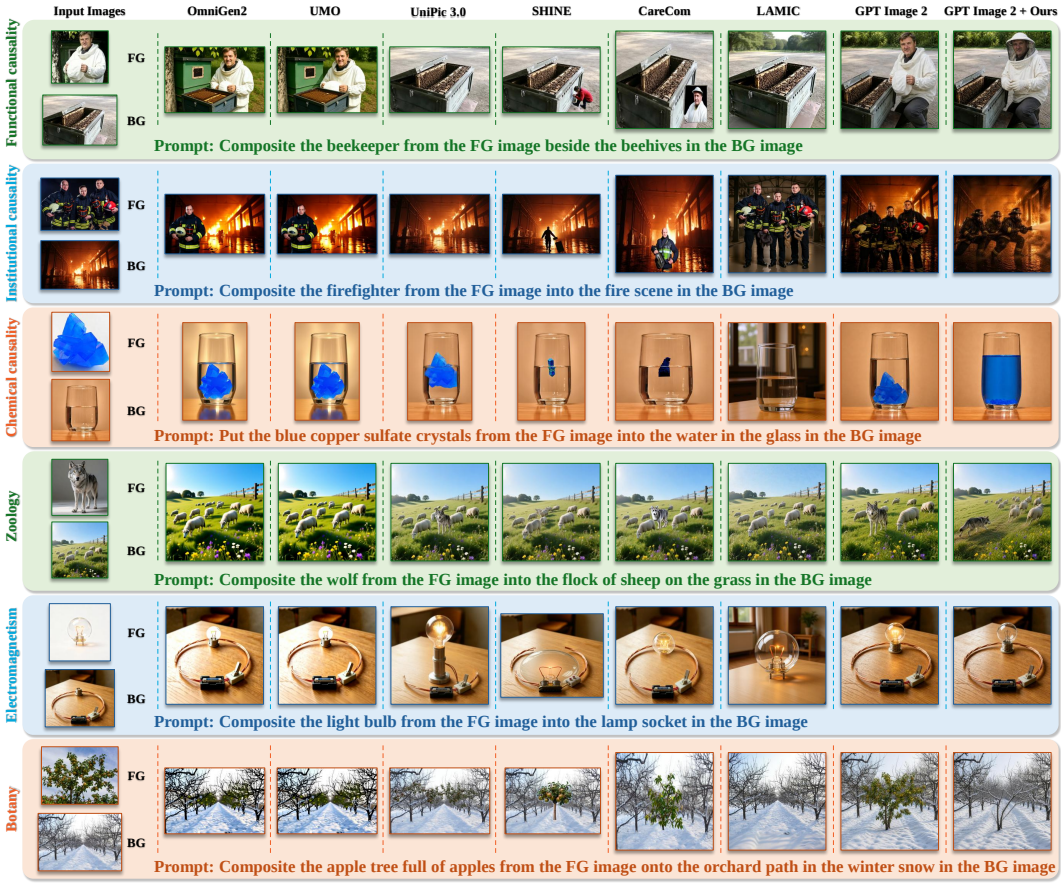


Figure 3: Qualitative comparison of our ReactionGraph with representative methods on our IRS-Bench dataset. For more results, please refer to Figure 4, Figure 5 and Figure 6.

.

5.3 Ablation Study

To evaluate the effectiveness of each component, we construct and assess several variants on 155 IRS-Bench groups, as shown in Table 7 of Appendix F.3. Direct foreground–background composition with ChatGPT Images 2.0 achieves a score of 60.96. Adding reaction-graph inference and constraint compilation with state realization improves the score to 71.16 and 62.40, respectively. Combining all components yields a score of 74.09, demonstrating the contribution of each component and the necessity of the complete framework.

5.4 Further Study

Comparison of different Knowledge Graph methods. To examine the compatibility between Reaction-Graph Inference and Constraint Compilation and State Realization, we conduct an additional ablation study by replacing Reaction-Graph Inference with existing knowledge-graph methods, including Davidsonian Scene Graph (DSG) (Cho et al., 2024) and Knowledge-Bridger (KB) (Ke et al., 2025), while keeping the remaining components unchanged. We evaluate each variant on 155 IRS-Bench groups, with the results reported in Table 8. Our method achieves the best performance across all evaluated variants, demonstrating its effective integration with Constraint Compilation and State Realization.

Causal reasoning beyond memorization. To distinguish mechanism-based reasoning from memorized co-occurrence, we construct negative and benign subsets from 20 subsets selected from IRS-

Bench dataset. Each subset is evaluated with a 40-question ruler covering the five evaluation dimensions described in Section 3.3 and background preservation. Negative subsets remove a condition required by the causal mechanism while preserving the entities and scene, whereas benign subsets retain the reasoning chain but weaken its strength. As shown in Table 9, our method improves over the direct synthesis baseline by 16.43, 6.92, and 25.10 points on the negative, benign, and original scenes, respectively. The larger gain on negative subsets compared with benign subset indicates that the method can detect broken causal conditions and preserve causal consistency, confirming that its improvements mainly arise from mechanism-based reasoning rather than memorized surface co-occurrence. The qualitative results are shown in Figure 7.

5.5 Generalization Study

Three-way merges. To evaluate the generalization ability of ReactionGraph in the multi-image synthesis task, we additionally constructed a triplet subset containing 100 samples. Each sample consists of 1 background image and 2 foreground images. We evaluated ReactionGraph and existing representative methods on this subset, and the qualitative comparison results are shown in Figure 8. The results demonstrate that ReactionGraph can effectively fuse different inputs, preserve coherent spatial relationships and causal connections between entities, and avoid naive entity splicing, which verifies its effectiveness and generalization ability in the three-image synthesis task.

Video. To evaluate the generalization ability of ReactionGraph in collision generation for video worlds, we apply it to the video generation task, as shown in Figure 9, Figure 10, Figure 11 and Figure 12. The results show that without ReactionGraph, the model can generate superficially coherent videos, yet suffers from temporal inconsistency, including drifts in entities and spatial relationships, as well as state transitions that violate causal logic. After introducing ReactionGraph, the generated videos can better preserve the relationships between foreground and background, and produce continuous state changes and outcomes consistent with their interactions. These results demonstrate that ReactionGraph can transfer the learned causal representations from static image synthesis to the video continuation task, so as to effectively guide video generation.

5.6 User study

To verify whether our proposed evaluation aligns with human perception, we conduct a single-blind user study covering 22 physical interaction scenarios across 21 models, with the consensus from 4 independent annotators serving as the ground truth (GT). As shown in Table 11, our framework consistently outperforms conventional automatic metrics across all backbones. Specifically, ME-Ruler implemented with GPT-5.6 Sol and Claude Sonnet 5 achieves superior rank correlations of 0.877 and 0.869 in Spearman's ρ, and up to 0.724 in Kendall's τ, while ME-Ruler (Qwen3.8-27B) obtains the lowest rank displacement with a Rank MAE of 2.190. In contrast, standard benchmarks such as CompBench and Share-CoT exhibit severe misalignment with human judgment, showing substantially lower correlations ($\rho \leq 0.598, \tau \leq 0.438$) and large average rank errors exceeding 4.5 positions. These results demonstrate that ME-Ruler captures subtle causal plausibility and composition quality markedly closer to human consensus, exhibiting robust model-agnostic generalization across both proprietary and open-source LLM backbones.

6 Conclusion

We introduced **IRS-Bench**, a benchmark for image-conditioned implicit causal reasoning in multi-image composition. Each foreground–background pair defines local scene contexts whose interaction constrains entity states and environmental effects; the benchmark fixes source identities, distinguishes deterministic from intentional interactions, and evaluates admissible alternatives with weighted binary rulers. Results across 1,000 pairs, eleven leaf tasks, and five causal mechanisms show that perceptual coherence and causal faithfulness are distinct: visually integrated composites can still violate interaction-specific constraints. ReactionGraph, a training-free diagnostic pipeline, improves performance by making the inferred interaction explicit before rendering, supporting the separation of causal inference from image realization. IRS-Bench is a controlled test of interaction-conditioned visual reasoning, rather than a measure of general world-model competence; its operational annotations, uneven coverage, and evaluator-specific scores define its scope.

AI USE STATEMENT

We used generative AI tools to assist with the literature survey and drafting and polishing of this manuscript. As mentioned earlier, some datasets are generated by AI tools, and language models assisted in the annotation pass described in Section 3, under author supervision and subject to the independent review described there. All AI-assisted text was reviewed and verified by the authors against the cited sources, and all claims, evaluations, and artifacts are the authors' own. We take full responsibility for the final content of this work.

ETHICS STATEMENT

IRS-Bench does not collect personal attributes or private communications. Human annotation, where reported, is restricted to judgements about image content and was performed by the authors. A benchmark of causal correctness in composition has no obvious dual use beyond that already available to users of image generators, and its intended effect is to make a currently hidden failure mode measurable. Release terms and any source-specific licensing or attribution constraints will be documented per component; third-party imagery will not be relicensed beyond the permissions of its original source.

REPRODUCIBILITY STATEMENT

We will release the benchmark data, annotation schema, construction and evaluation protocols, judge prompts, and framework code. Exact model identifiers, decoding parameters, prompt text, and graph serialization are required in that release for the reported results to be independently reproducible.

REFERENCES

Waqas Ahmed, Dean Diepeveen, and Ferdous Sohel. Multishadow: Multi-object shadow generation for image compositing via diffusion model. *arXiv preprint arXiv:2603.02743*, 2026.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv e-prints*, pp. arXiv–2509, 2025.

Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. Oneig-bench: Omni-dimensional nuanced evaluation for image generation. *Advances in Neural Information Processing Systems*, 38, 2026.

Bowen Chen, Haomiao Sun, Li Chen, Xu Wang, Daniel Du, Xinglong Wu, et al. Xverse: Consistent multi-subject control of identity and semantic attributes via dit modulation. *Advances in Neural Information Processing Systems*, 38:7437–7459, 2026a.

Jiaxuan Chen, Bo Zhang, Qingdong He, Jinlong Peng, and Li Niu. Carecom: Generative image composition with calibrated reference features. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2877–2885, 2026b.

Kaijie Chen, Zihao Lin, Zhiyang Xu, Ying Shen, Yuguang Yao, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. R2i-bench: Benchmarking reasoning-driven text-to-image generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12606–12641, 2025.

Yuzhuo Chen, Zehua Ma, Jianhua Wang, Kai Kang, Shunyu Yao, and Weiming Zhang. Lamic: Layout-aware multi-image composition via scalability of multimodal diffusion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3174–3182, 2026c.

Yufeng Cheng, Wenxu Wu, Shaojin Wu, Mengqi Huang, Fei Ding, and Qian He. Umo: Scaling multi-identity consistency for image customization via matching reward. *arXiv preprint arXiv:2509.06818*, 2025.

Jaemin Cho, Yushi Hu, Jason Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *International conference on learning representations*, volume 2024, pp. 15625–15645, 2024.

Xingyu Fu, Muyu He, Yujie Lu, William Yang Wang, and Dan Roth. Commonsense-t2i challenge: Can text-to-image generation models understand commonsense? *arXiv preprint arXiv:2406.07546*, 2024.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Junjia Huang, Pengxiang Yan, Jiyang Liu, Jie Wu, Zhao Wang, Yitong Wang, Liang Lin, and Guanbin Li. Dreamfuse: Adaptive image fusion with diffusion transformer. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–10. IEEE, 2025a.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3563–3579, 2025b.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. Geneval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025.

Guanzhou Ke, Shengfeng He, Xiaoli Wang, Bo Wang, Guoqing Chao, Yuanyang Zhang, Yi Xie, and Hexing Su. Knowledge bridger: Towards training-free missing modality completion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25864–25873. IEEE, 2025.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024.

Haowen Li, Zhenfeng Fan, Zhang Wen, Zhengzhou Zhu, and Yunjin Li. Aicomposer: Any style and content image composition via feature integration. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16840–16850. IEEE, 2025.

Ouxiang Li, Yuan Wang, Xinting Hu, Huijuan Huang, Rui Chen, Jiarong Ou, Xin Tao, Pengfei Wan, Xiaojuan Qi, and Fuli Feng. Easier painting than thinking: Can text-to-image models set the stage, but not direct the play? In *International Conference on Learning Representations*, volume 2026, pp. 86729–86758, 2026.

Shilin Lu, Zhuming Lian, Zihan Zhou, Shaocong Zhang, Chen Zhao, and Adams Wai-Kin Kong. Does flux already know how to perform physically plausible image composition? *arXiv preprint arXiv:2509.21278*, 2026.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Fanqing Meng, Kunpeng Ning, et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18310–18319. IEEE, 2023.

Kaiyue Sun, Rongyao Fang, Chengqi Duan, Xian Liu, Aoxue Li, and Xihui Liu. T2i-reasonbench: Benchmarking reasoning-informed text-to-image generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 8919–8944, 2026.

Super Intelligence Team, Changhao Qiao, Chao Hui, Chen Li, Cunzheng Wang, Dejia Song, Jiale Zhang, Jing Li, Qiang Xiang, Runqi Wang, et al. Firered-image-edit-1.0 technical report. *arXiv preprint arXiv:2602.13344*, 2026.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024.

Yibin Wang, Zhimin Li, Yuhang Zang, Jiazi Bu, Yujie Zhou, Yi Xin, Junjun He, Chunyu Wang, Qinglin Lu, Cheng Jin, et al. Unigenbench++: A unified semantic evaluation benchmark for text-to-image generation. *arXiv preprint arXiv:2510.18701*, 2025.

Hongyang Wei, Hongbo Liu, Zidong Wang, Yi Peng, Baixin Xu, Size Wu, Xuying Zhang, Xianglong He, Zexiang Liu, Peiyu Wang, et al. Skywork unipic 3.0: Unified multi-image composition via sequence modeling. *arXiv preprint arXiv:2601.15664*, 2026.

Xinyu Wei, Jinrui Zhang, Zeqing Wang, Hongyang Wei, Zhen Guo, Bairui Li, and Lei Zhang. Tiif-bench: How does your t2i model follow your instructions? *arXiv preprint arXiv:2506.02161*, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Chenyuan Wu, Jiahao Wang, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, et al. Omnigen2: Towards instruction-aligned multimodal generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21964–21975, 2026a.

Mingrui Wu, Hang Liu, Jiayi Ji, Xiaoshuai Sun, and Rongrong Ji. Micon-bench: Benchmarking and enhancing multi-image context image generation in unified multimodal models. *arXiv preprint arXiv:2602.19497*, 2026b.

Xindi Wu, Dingli Yu, Yangsibo Huang, Olga Russakovsky, and Sanjeev Arora. Conceptmix: A compositional image generation benchmark with controllable difficulty. *Advances in Neural Information Processing Systems*, 37:86004–86047, 2024.

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. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13294–13304. IEEE, 2025.

Bo Zhang, Yuxuan Duan, Jun Lan, Yan Hong, Huijia Zhu, Weiqiang Wang, and Li Niu. Controlcom: Controllable image composition using diffusion model. *arXiv preprint arXiv:2308.10040*, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Zhiqiu Zhang, Dongqi Fan, Mingjie Wang, Qiang Tang, Jian Yang, and Zili Yi. Region-to-region: Enhancing generative image harmonization with adaptive regional injection. *arXiv preprint arXiv:2508.09746*, 2025.

Yucheng Zhou, Jiahao Yuan, and Qianning Wang. Draw all your imagine: A holistic benchmark and agent framework for complex instruction-based image generation. *arXiv preprint arXiv:2505.24787*, 2025.

A Formal Task Definition

A world is a pair $W = (E, R)$: entities $E = \{e_1, \ldots, e_n\}$ and causal relations $R \subseteq E \times E$ between them. Each entity carries an identity $id(e) \in \mathcal{I}$, which fixes *what* it is and how it appears, and a state $st(e) \in \Sigma$, the configuration it currently occupies. A merge of a foreground world W_F and a background world W_B yields

$$W_{\mathrm{m}} = W_F \oplus W_B, \qquad \text{s.t.} \quad \begin{array}{l} id\big(\mathrm{ent}_F(W_{\mathrm{m}})\big) = id\big(\mathrm{ent}(W_F)\big) \\ \mathrm{ent}(W_{\mathrm{m}}) \subseteq \mathrm{ent}(W_F) \cup \mathrm{ent}(W_B), \end{array} \tag{5}$$

where $\mathrm{ent}_F(W_{\mathrm{m}})$ descends from the foreground. Equation 5 states the *identity lock*: that entity survives with its identity intact, and no entity enters the merge that neither source contained. Its complement, *state freedom*, leaves only the state free, so the merge may rewrite *how* an entity is but never *which* entity it is. This separates the task from generation from text, where an inconvenient premise can simply be regenerated.

A relation r between the inserted entity and an entity $e \in \mathrm{ent}(W_B)$ induces a set-valued reaction map $\kappa_r : \Sigma \to 2^{\Sigma}$. The benchmark assigns the governing leaf task and the causal regime separately at the instance level. The regime $\rho \in \{\textsc{det}, \textsc{int}\}$ specifies whether the depicted conditions determine a unique admissible target state or permit more than one response. This avoids equating a broad subject area, such as biology, with a fixed notion of admissibility. For the active relations $R(e_F, E_B)$, the jointly admissible foreground states are

$$S^*(e_F, E_B) \;=\; \bigcap_{r \in R(e_F, E_B)} \kappa_r\big(st(e_F)\big), \tag{6}$$

where an empty intersection marks an inconsistent annotation rather than a valid item. A faithful composite need not reproduce one reference image; it must realize a state in S^* while preserving source identity and rendering the environmental effects entailed by the active relations.

Regime and admissibility. The regime fixes whether a reaction map is single-valued:

$$\rho = \textsc{det} \;\Rightarrow\; |S^*| = 1, \qquad \rho = \textsc{int} \;\Rightarrow\; |S^*| \geq 1. \tag{7}$$

Since a relation may also require the merge to modify the background rather than to restate the foreground, we state admissibility as

$$\mathrm{Adm}(C) = \mathbb{1}\big[\, st(\mathrm{ent}_F(C)) \in S^* \,\big] \cdot \mathbb{1}\big[\, \mathrm{eff}(C) \supseteq \textstyle\bigcup_r \kappa_r^{\mathrm{env}} \,\big], \tag{8}$$

where $\mathrm{eff}(C)$ collects the effects observable in composite C and κ_r^{env} denotes the environmental effects entailed by relation r. Equation 8 tests causal satisfaction and effect realization; the identity condition in Equation 5 is checked separately. The intentional condition permits a singleton set because an intentional process may have only one response compatible with the particular depicted constraints, even though the mechanism is not deterministic in general.

B Task Taxonomy and Coverage

The eleven leaf tasks of Section 3.1 fall into five causal dimensions under two domains. A dimension is defined by the mechanism that governs the relation a sample tests rather than by the objects that instantiate it, so the definitions below remain fixed as the inventory grows. Each leaf task states what the merge has to realize.

Physical Causality covers the five tasks whose outcomes physical law forces. *Classical Mechanics* (229) asks for the mechanical consequence of contact: the deformation, displacement, support failure or stability that weight, load and impact impose. *Optics* (209) asks for the optical consequence of an arrangement: shadow, reflection, refraction, occlusion and illumination, which light, surface and geometry force. *Fluid Dynamics* (173) asks for the hydrodynamic consequence: spreading, flowing, absorption, buoyancy or wetting, which a liquid and the surface or vessel it meets impose. *Thermodynamics* (62) asks for the thermal consequence: heating, ignition, combustion or temperature-driven change, which contact with a heat source imposes. *Electromagnetism* (27) asks for the electromagnetic consequence: attraction, induced magnetism or current-driven force, which a magnetic or energized body imposes on what is placed beside it.

Biological Causality covers the three tasks whose outcomes are responses of a living organism rather than states of an inert body. *Zoology* (178) asks for an animal's behavioral response to the entities the merge introduces, whether predatory, defensive, parental or social. *Botany* (20) asks for a plant's response to the setting it enters, whether stress, growth or dispersal. *Physiology* (17) asks for an organism-level reflex or autonomic response to a stimulus the merge introduces.

Chemical Causality covers *Chemical Reaction* (27), which asks for a transformation of substance: dissolution, reaction, gas evolution or color change, licensed by the material properties of what meets what.

Institutional Causality covers *Institutional Rule Response* (33), which asks for the response that a role, a rule or an authority present in the scene requires of the inserted entity.

Functional Causality covers *Functional Interaction* (25), which asks what the setting does to the function of the inserted object: its use enabled, obstructed or redirected by the space it enters.

The first three dimensions make up the natural-science domain and account for 942 instances; the last two make up the social-science domain and account for 58. The two dimensions of the social domain are deliberately small: their outcomes are licensed rather than forced, so an item demands a stricter notion of correctness and a more careful review.

C Benchmark Construction and Quality Control

Four principles constrained construction. **Identity lock, state freedom** rejects any pair that can be satisfied by replacing the foreground entity. **Regime-appropriate correctness** uses a unique target state for deterministic instances and an admissible set for intentional instances. **Shortcut resistance** uses controlled variants that preserve much of the visible content while removing or weakening a necessary causal condition (Section 5.4). **Reference-independent validity** treats a ground-truth image as one realization rather than the definition of correctness: any composite satisfying Equation 8 and the identity lock is admissible. The separate multi-foreground collection probes qualitative transfer to three-way composition; it is not included in the 1,000-instance benchmark score.

Construction proceeded in four recorded stages as follows.

In the first we drew candidate images from three sources, to control for the systematic bias of any one: existing composition and segmentation datasets, web imagery of real scenes, and photographs taken by the authors. The author-photographed subset is the only one captured under a single illumination and camera configuration, which lets us verify that a photographic ground truth is physically consistent rather than merely plausible. We retained a pair only if the foreground entity is unambiguously identifiable from its image alone, the background contains an entity standing in a definite causal relation to it, and the required reaction is not visible in either input, so that the item cannot be answered by copying evidence. The third condition failed most often, and enforcing it discarded many otherwise attractive pairs.

In the second stage we assigned each retained pair to one of the eleven leaf tasks by the relation governing it and annotated its regime independently. Annotators wrote the relation as $\langle$property of e_F, property of the scene, reaction, rule$\rangle$, which forces the reaction to be derived from stated properties rather than asserted. We returned for revision any record whose rule field merely restated the reaction, because such a record does not establish a causal connection between the premises and the target state.

In the third stage we built the ground truth. For the author-photographed subset it is the photograph of the scene after the causal process has been allowed to occur, which yields 172 pairs with photographic ground truth; for the remainder we generated the post-reaction composite from the two images and the annotated relation with a video generation model, then selected the frame that best realizes the annotated mode, either by sampling temporally even keyframes and having a vision-language model pick the one conforming best to the constraint set, or by having the model plan the generation so that the terminal frame corresponds to the moment the reaction is realized. We ran both routes in parallel and reviewed the selection, because a generative ground truth can exemplify the annotated mode but must not redefine it.

In the fourth stage we compiled each annotation into atomic constraints of four types, position, state, effect, and identity, and rewrote each as a yes-or-no verification question whose expected answer the annotation determines. Identity constraints are phrased negatively, so that a composite satisfying causality by replacing the entity is penalized rather than rewarded, and questions carry importance weights reflecting how diagnostic they are: one whose answer separates the causal mode from its alternatives outweighs one that merely confirms the entity is present.

C.1 Annotation Schema

Each item is stored as one record. Table 3 lists its fields.

Table 3: Fields of one IRS-Bench record.

Field	Contents
`id`	Identifier, unique across the release.
`source`	Provenance of the pair: AI-generated, web-sourced, or author-photographed.
`task`	One of the eleven leaf tasks in Section B.
`regime`	$\rho \in \{\text{DET}, \text{INT}\}$.
`images`	File references for the foreground, the background, and the photographic ground truth where one exists.
`entities`	The subject with its identity features and count, and the background entities.
`relation`	The tuple $\langle$subject property, scene property, reaction, governing rule$\rangle$.
`admissible`	The admissible reactions S^*.
`effects`	The environmental effects κ^{env} the merge must produce.
`ruler`	The questions of Section 3.3, each with its dimension, importance, rationale and source.
`control_type`	Positive, causal-condition-removed, or causal-condition-weakened for items in the controlled analysis of Section 5.4.

The released schema should be accompanied by at least one fully populated example record and a machine-readable field specification. Because those artifacts are not embedded in the manuscript, this table documents the contract required to reproduce benchmark instances without implying that unreleased examples have already been verified.

C.2 Quality Control Protocol

Negative-item construction. Negative items are generated by applying one of two transformations to a positive instance. The *occlusion* transformation keeps both entities, their relative position, and the scene's appearance, but introduces a third element that prevents the causal relation from holding, so that the surface statistics a co-occurrence model conditions on are unchanged while the mechanism is removed. The *property-invalidation* transformation keeps both entities and their relation but perturbs a scene property on which the reaction depends, so that the relation no longer licenses the original effect. In both cases the annotated correct response is the null reaction, and the inserted entity retains the state it held in F. A system that renders the positive reaction on a negative item has responded to association rather than to mechanism.

Annotation and review. Annotation was performed by the authors with language-model assistance. Every item then passed an independent review by a second annotator who had not seen the original decision; disagreements were recorded and resolved by discussion, and the resolution is stored alongside the record. The review checked three things: that the stated rule licenses the stated reaction without restating it; that the admissible set is neither empty nor trivially permissive; and that the constraint set can be jointly satisfied by at least one composite.

Coverage checks. The review stage of ruler construction removes duplicates and questions that a faithful composite would pass regardless, and checks coverage of the five assessment axes where each axis is applicable. It also checks answer polarity and whether all necessary identity, state, and effect constraints are represented. A ruler that fails these checks is returned for revision before the item enters the release.

Automated rejection rules. An item is rejected automatically if its constraint set is empty; if two verification questions demand mutually incompatible states of the same entity; if any identity constraint is unsatisfiable by construction; or if the importance weights of its questions sum to zero.

D EVALUATION PROTOCOL

Question schema. Each ruler record contains an identifier, one of the five assessment axes in Section 3.3, the yes–no question, its expected answer, an importance rating in $\{1, \ldots, 10\}$, a rationale, and a pointer to the source annotation from which it was derived. The source pointer links the ruler to a subject property, scene property, causal rule, or domain-specific checkpoint. These fields make every judgement traceable to the construct it measures.

Judging protocol. The judge receives the foreground, background, and candidate composite with fixed roles. It first locates and counts the inserted subject and records a short appearance summary, then answers each ruler question with a binary verdict. Identity questions compare the composite with the foreground and do not treat state-dependent pose, position, scale, or behavior as identity. Questions about scene preservation compare against the background while excluding changes entailed by the annotated reaction. Judges run at temperature zero.

Auxiliary background analysis. The controlled analysis in Section 5.4 additionally uses a separate 0–10 background-fidelity rating. A second call compares only the background and candidate, inventories unintended scene changes, and excludes necessary local effects of the inserted subject. This rating contributes only to the 155-group ablation protocol defined in Section 5.4; it is not a sixth axis of the benchmark score in Equation 1. Anchors are 10 for preservation apart from necessary effects, 7–9 for small style or texture deviations, 4–6 for visible changes to scene objects or composition, 1–3 for substantial repainting with lost objects, and 0 for wholesale background replacement.

Judge–human calibration values are not reported because the human study described in Section 5.6 is still in progress. The main results should therefore be read as evaluator-specific scores rather than estimates of human accuracy.

E REACTIONGRAPH IMPLEMENTATION

ReactionGraph is training-free, so its behavior after construction of the neutral seed is fixed by three prompts and the schema they exchange. The release must include their verbatim text, model versions, decoding settings, and graph serialization; Table 4 summarizes the role of each prompt.

Table 4: The prompts that define ReactionGraph.

Prompt	Contents
Inference	Receives $(B, F, M^{(0)}, q)$, uses $M^{(0)}$ only for localization, and returns the graph of Equation 3. It specifies the node and edge schemas, temporal ranks, provenance conditions, and the prohibition on entities absent from the public inputs.
Compilation	Converts graph consequences into the typed constraints of Equation 4, retaining a pointer to each supporting node and rule and adding identity and count preservation for every foreground entity.
Specification	Presents the generator with the neutral seed, source references, enumerated requirements, and result-only self-checks. It requests a single edited output and does not expose benchmark rulers or annotations.

The exact prompt text, model identifiers, decoding parameters, and serialization schema are required for independent reproduction and must accompany the artifact release. They are not reproduced here because the current manuscript source does not contain verified versions of those materials.

Table 5: ME-Ruler scores on the 11 IRS-Bench leaf tasks using Claude Sonnet 5 as a third evaluator. Mech: Classical Mechanics; Opt: Optics; Fluid: Fluid Dynamics; Thermo: Thermodynamics; EM: Electromagnetism; Zool: Zoology; Bot: Botany; Physio: Physiology; Chem: Chemical Reaction; Inst: Institutional Rule Response; Func: Functional Interaction.

System	Natural Science											Social Science		Overall
	Physical Causality						Biological Causality				Chem.	Inst.	Func.	
	Mech.	Opt.	Fluid	Thermo.	EM	*avg*	Zool.	Bot.	Physio.	*avg*				
Heuristic Baselines														
Copy-paste	31.6	31.4	34.5	29.6	50.3	35.5	32.1	26.7	25.6	28.1	41.8	38.3	31.1	32.8
Text-to-Image Image Generation														
OmniGen2	26.1	36.9	30.3	26.8	44.3	32.9	23.3	24.7	23.7	23.9	37.6	40.3	32.3	30.0
AIComposer	28.2	27.3	27.1	25.9	31.9	28.1	22.0	27.7	27.0	25.6	35.1	33.3	31.7	27.1
UMO	26.1	36.2	29.3	28.1	40.5	32.1	23.5	24.2	26.7	24.8	36.5	39.8	34.9	29.7
LAMIC	19.0	21.0	18.9	18.9	25.1	20.6	19.2	26.3	22.3	22.6	29.8	30.0	31.2	20.8
DreamFuse	43.1	47.2	44.6	37.0	43.4	43.1	35.7	38.1	39.5	37.7	44.1	49.0	45.4	42.7
SHINE	28.2	30.7	32.3	25.3	33.2	29.9	26.6	27.6	26.2	26.8	40.7	36.9	39.3	29.9
CareCom	30.6	31.9	34.9	27.4	36.7	32.3	29.7	30.8	20.5	27.0	43.6	29.4	34.3	31.7
Open-source Image Generation and Editing Model														
Skywork UniPic 3.0	45.5	48.4	47.2	36.1	47.7	45.0	32.8	27.2	24.6	28.2	50.5	44.3	39.1	42.8
Qwen-Image-3.0	53.7	55.5	60.3	46.9	57.0	54.7	47.3	45.6	45.6	46.2	54.6	59.1	54.6	53.7
Qwen-Image-Edit-2511	47.8	55.0	50.1	41.8	44.4	47.9	40.4	30.4	36.0	35.6	48.6	52.8	48.2	47.6
Wan2.7-Image	50.0	53.2	53.6	39.9	50.7	49.5	42.7	41.4	40.8	41.6	52.6	49.9	43.7	49.0
FireRed-Image-Edit	48.6	53.1	51.4	41.9	44.7	47.9	40.6	36.9	32.3	36.6	55.7	52.3	47.7	47.9
HunyuanImage 3.0-Instruct	49.1	56.3	52.3	45.6	49.7	50.6	44.1	48.9	38.0	43.6	54.8	56.7	56.9	50.5
FLUX.2-klein-9B	48.2	52.6	51.8	42.2	50.0	49.0	41.6	44.3	43.9	43.3	48.9	49.3	48.5	48.2
Close-source Image Generation and Editing Model														
ChatGPT Images 2.0	53.0	57.8	58.3	48.6	50.9	53.7	47.7	49.1	41.8	46.2	55.2	60.1	53.4	53.7
Nano Banana 2	52.7	58.5	58.5	50.7	58.5	55.8	45.4	44.7	41.9	44.0	55.9	56.1	50.2	53.5
Seedream 5.0 lite	49.6	54.9	54.0	43.1	55.6	51.4	42.8	42.8	43.1	42.9	51.3	51.3	42.7	49.7
ReactionGraph (w/ Image 2.0)	65.3	67.7	67.2	66.6	74.9	68.3	65.6	62.2	63.0	63.6	78.5	75.0	71.9	67.3
ReactionGraph (w/ Nano Banana 2)	66.5	67.3	72.2	74.1	68.8	69.8	65.5	60.0	55.5	60.3	82.2	62.9	70.5	68.1
ReactionGraph (w/ Qwen-Image-3.0)	60.5	64.9	65.1	65.5	74.8	66.2	59.4	64.0	52.1	58.5	68.0	65.1	62.9	63.1

F Supplementary Experiments

Implementation details. ReactionGraph executes the four stages in Equation 2 once per item. We use no iterative repair, external verification, reranking, or output selection. Open-weight models run on eight NVIDIA A100 (80G) GPUs at their officially recommended settings.

F.1 Complementary Appearance Metrics

Table 6 reports FID (Heusel et al., 2017), LPIPS (Zhang et al., 2018), and CLIP-I as complementary appearance diagnostics. FID measures a distribution-level feature discrepancy, LPIPS a learned perceptual discrepancy between paired images, and CLIP-I an image-embedding similarity; the arrows in the table indicate their directions. None tests whether the interaction's entailed reaction is present. The release must specify the reference pool, preprocessing, feature extractors, and aggregation used for these values, because otherwise the numbers are not independently reproducible. Causal faithfulness is evaluated separately by ME-Ruler as defined in Section 3.3.

Table 6 reports FID, LPIPS and CLIP-I on the same merges. The 18 systems occupy a narrow band on all three (FID 113.9–147.9, LPIPS 0.50–0.69, CLIP-I 0.80–0.88), so a suite that ME-Ruler spreads over 39 points is nearly flat here; copy-paste even records the lowest LPIPS and a near-maximal CLIP-I, because both metrics reward composites that stay close to their inputs. Measured against its own base generator, which fixes everything but the specification, the added inference raises ME-Ruler by 9.5 to 18.1 overall while moving FID by at most 8.9, LPIPS by at most 0.06 and CLIP-I by at most 0.01, with none of the nine entries improving. The composites are neither more attractive nor visibly degraded: what changes is the state they depict, not how they look.

F.2 Third-Evaluator Results

Table 5 repeats the causal evaluation with Claude Sonnet 5, separate from the two primary evaluators in Table 2. The strongest non-ReactionGraph system has an overall score of 53.7. The three ReactionGraph stacks score 63.1–68.1 overall, improvements of 9.4–14.4 points over that baseline. They

Table 6: Quantitative comparison of our method with existing methods on the 11 IRS-Bench subtasks, measured by FID, LPIPS and CLIP-I. Lower is better for FID and LPIPS and higher for CLIP-I.

System	FID↓	LPIPS↓	CLIP-I↑
Heuristic Baselines			
Copy-paste	128.50	0.50	0.86
Text-to-Image Image Generation			
OmniGen2	137.89	0.69	0.82
AIComposer	140.32	0.60	0.83
UMO	138.00	0.69	0.82
LAMIC	147.87	0.69	0.80
DreamFuse	128.70	0.61	0.85
SHINE	132.70	0.60	0.82
CareCom	128.78	0.62	0.85
Open-source Image Generation and Editing Model			
Skywork UniPic 3.0	123.15	0.63	0.86
Qwen-Image-3.0	116.69	0.55	0.87
Qwen-Image-Edit-2511	121.32	0.61	0.87
Wan2.7-Image	120.15	0.61	0.87
FireRed-Image-Edit	118.33	0.60	0.87
HunyuanImage 3.0-Instruct	113.89	0.61	0.88
FLUX.2-klein-9B	119.14	0.62	0.87
Close-source Image Generation and Editing Model			
ChatGPT Images 2.0	116.34	0.64	0.87
Nano Banana 2	114.23	0.60	0.88
Seedream 5.0 lite	115.94	0.59	0.87
Proposed Method			
ReactionGraph (w/ Image 2.0)	119.20	0.65	0.87
ReactionGraph (w/ Nano Banana 2)	116.75	0.62	0.87
ReactionGraph (w/ Qwen-Image-3.0)	125.56	0.61	0.86

do not uniformly improve FID, LPIPS, or CLIP-I in Table 6, consistent with the intended distinction between appearance similarity and causal faithfulness. These are evaluator-specific measurements; agreement with human judgements remains to be established by the study in Section 5.6.

F.3 Detailed Ablation Study

To evaluate the effectiveness of the proposed components, we construct and assess several variants on IRS-Bench, as shown in Table 7. We use ChatGPT Images 2.0 as the baseline, applying it directly to foreground–background composition; its reproduced results are reported in the first row. We first add reaction-graph inference alone, passing the graph to the generator as free text, so that the causal description reaches it. The resulting gain of 10.2 shows that the description carries information the two images did not already make explicit. We then add constraint compilation and state realization, which factorizes the graph into atomic constraints and their verification questions before realization. The further gain of 1.44 isolates the contribution of structure from that of content. The full framework reaches 74.09 with the gain of 13.13, demonstrating the contribution of each component and the necessity of the complete framework.

Table 7: Ablation study of different component combinations of our ReactionGraph on the 155 groups selected from our IRS-Bench, scored by ME-Ruler scores using GPT-5.6 SOL as the independent evaluator.

Reaction-graph inference	Constraint compilation and state realization	ME-Ruler	Gain
		60.96	0.0
✓		71.16	10.2
	✓	62.40	1.44
✓	✓	74.09	13.13

Table 8: Comparison of different knowledge graph strategies on the 155 groups selected from our IRS-Bench, scored by ME-Ruler scores using GPT-5.6 SOL as the independent evaluator.

Method	ME-Ruler
Davidsonian Scene Graph	62.40
Knowledge-Bridger	53.33
Ours	74.09

Table 9: Comparison of our Samples with the negative and Benign samples

Set	Ours	Naive	Δ
Negative (causal chain cut)	84.00	67.57	+16.43
Benign (causal chain weakened)	75.07	68.15	+6.92
Original Samples	76.70	51.60	+25.10

F.4 More Qualitative Results

Figures 4, 5, and 6 present additional qualitative comparisons between ReactionGraph and representative baseline methods on IRS-Bench. In scenes with dense objects, complex interactions, and subtle visual evidence, existing methods often omit key entities, misidentify relationships. In contrast, ReactionGraph more reliably identifies event-related entities and establishes interactions consistent with the visual evidence, yielding structurally complete and semantically coherent reasoning results. These findings demonstrate the effectiveness of our method.

G Limitations and Dataset Documentation

Coverage. IRS-Bench is deliberately broad but not balanced. Classical mechanics, optics, and zoology account for most instances, whereas electromagnetism, chemistry, botany, physiology, and the two social-science tasks have much smaller samples. Aggregate scores are therefore dominated by the larger tasks, and results for the smaller tasks have higher sampling uncertainty. The benchmark supports task-wise reporting but does not claim exhaustive coverage of any scientific domain.

Construct validity. The deterministic–intentional regime is assigned per instance from the depicted conditions. This distinction simplifies a continuum: physical outcomes can be uncertain under incomplete observation, and a behavior normally treated as intentional can be tightly constrained in a particular scene. The admissible set records the curators' operational judgement rather than a complete model of every possible outcome. The controlled variants reduce reliance on surface co-occurrence, but they do not prove that a system recovered the true causal mechanism.

Evaluation validity. 172 instances have photographic post-interaction references. For the remaining instances, correctness is determined from annotations and rulers rather than a unique target image. Scores are conditional on the selected multimodal evaluator, as shown by reporting multiple judges; the human comparison in Section 5.6 is still in progress.

Generalization. The 100 multi-foreground items and the video examples are qualitative transfer studies. They are not included in the reported 1,000-instance benchmark score and currently have no calibrated rulers. Conclusions about three-way composition and temporal generation should therefore be limited to the examples shown.

Table 10: Four tiers in the single-blind sorting task of Section 5.6. A rater assigns every composite for a scene to exactly one tier; tiers are ordered from 1 (best) to 4 (worst) when computing system ranks.

Column	Tier	Criterion
1	Best	Task completed and the composite is causally correct
2	Partial	Task completed but the composite is not fully causally correct
3	Poor	Task completed but the result is poor
4	Failure	Task not completed

Table 11: Consistency comparison between different evaluation methods and human ranking (GT).

Evaluation Method	Spearman's ρ (SRCC) ↑	Kendall's τ (KRCC) ↑	**Rank MAE ↓**
ME-Ruler(Claude Sonnet 5)	0.869	**0.724**	2.476
ME-Ruler(GPT-5.6 Sol)	**0.877**	0.714	2.286
ME-Ruler(Qwen3.8-27B)	0.858	0.695	**2.190**
CompBench (LLM 1–5)	0.598	0.438	4.571
Share-CoT (VQA 0–100)	0.442	0.333	4.857

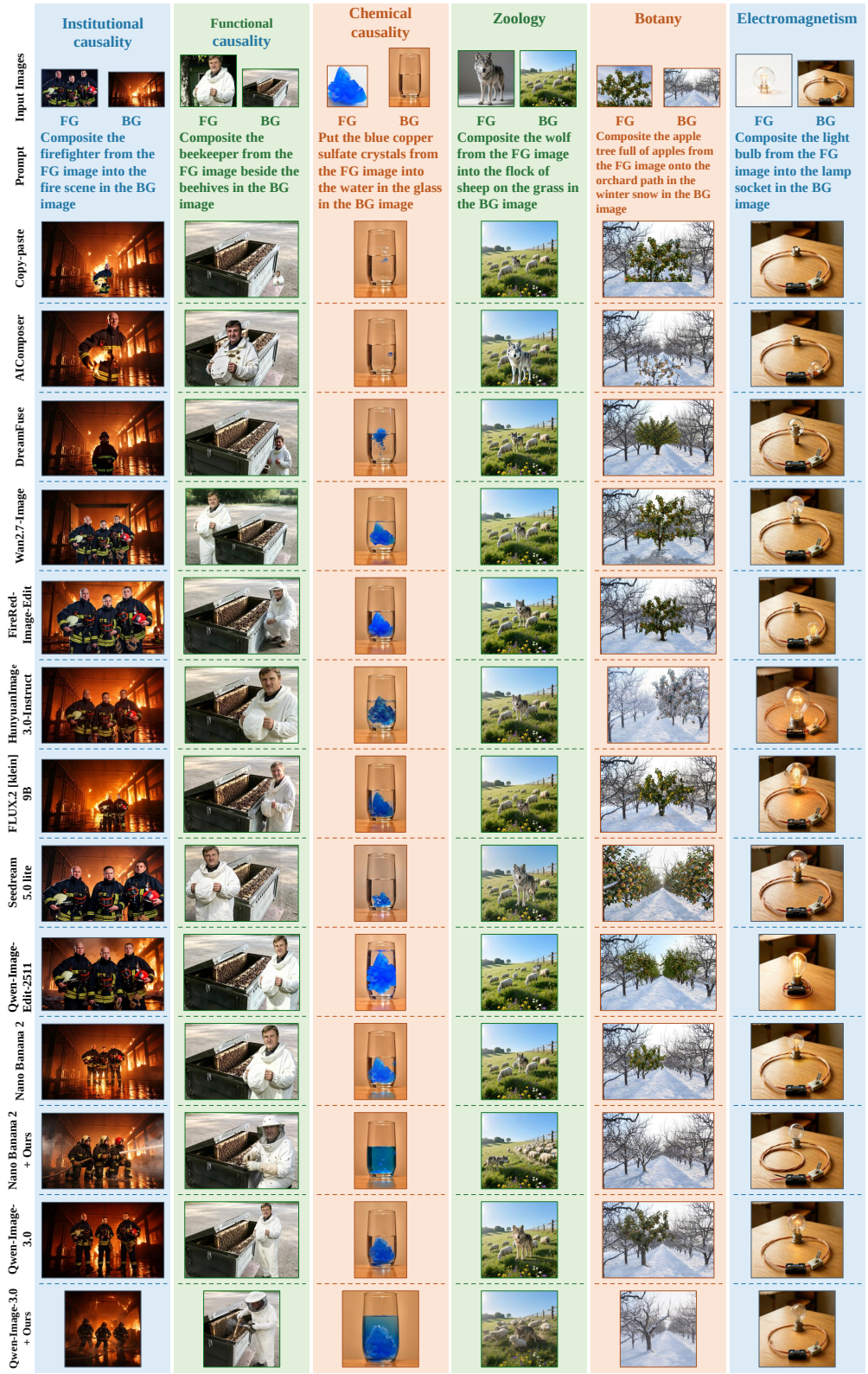


Figure 4: More qualitative comparison of our ReactionGraph with representative methods on our IRS-Bench dataset

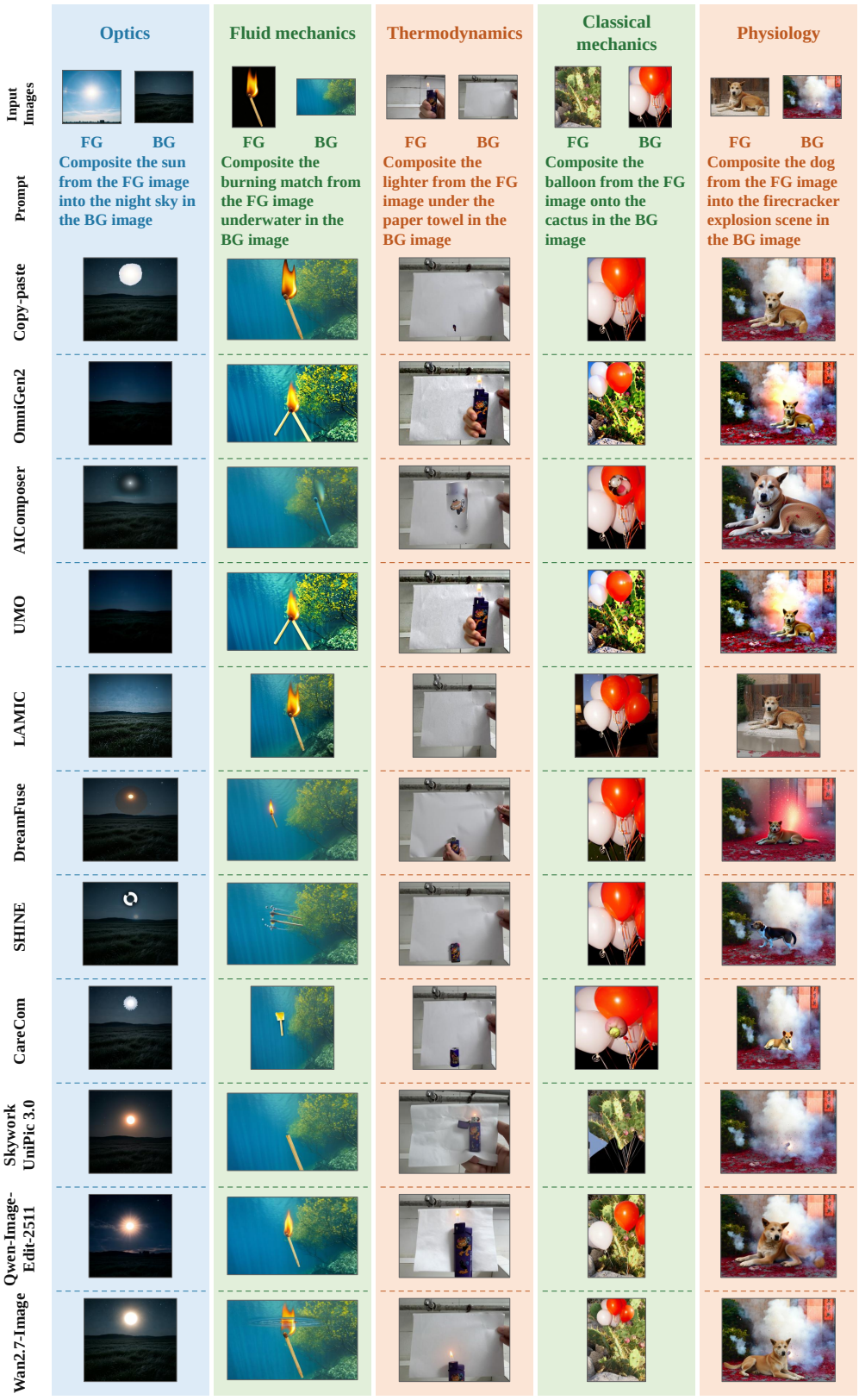


Figure 5: More qualitative comparison of our ReactionGraph with representative methods on our IRS-Bench dataset

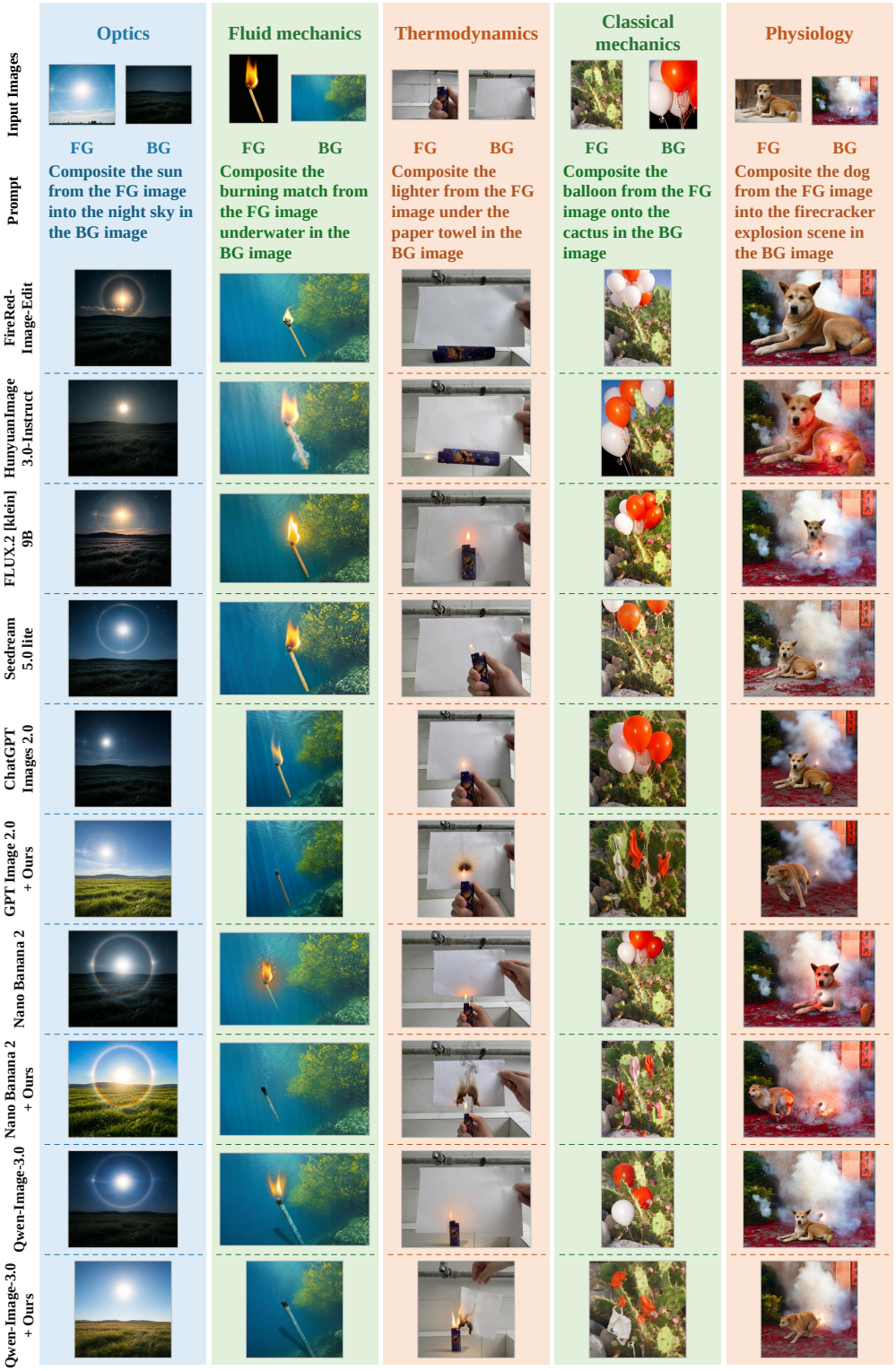


Figure 6: More qualitative comparison of our ReactionGraph with representative methods on our IRS-Bench dataset

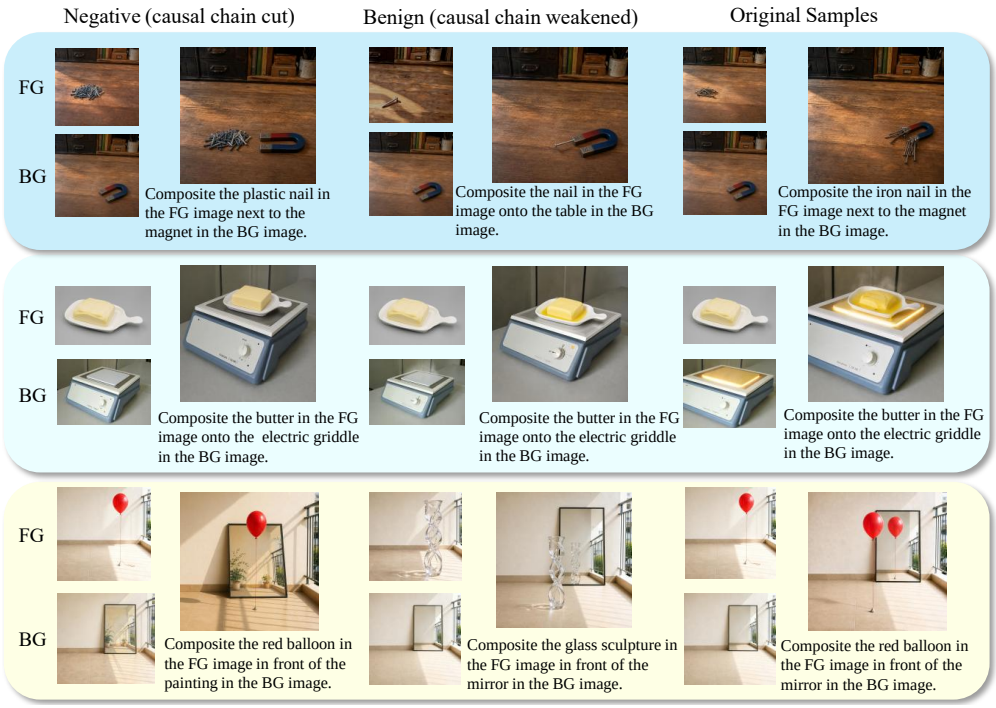


Figure 7: Qualitative results of Negative, Benign and Original samples.

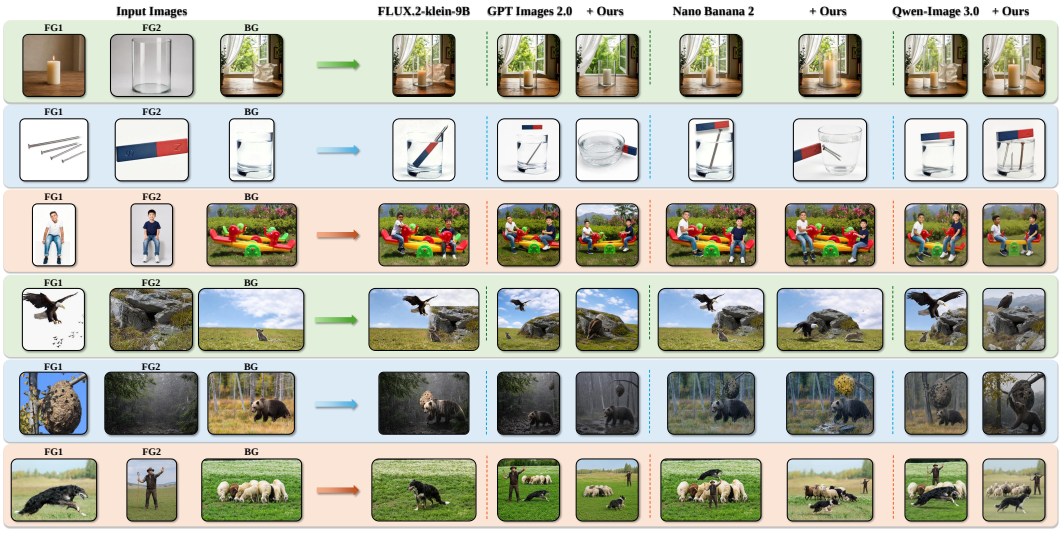


Figure 8: Qualitative comparison of our ReactionGraph with representative methods on our contrusted triplet subset.



Figure 9: Qualitative comparison of representative video continuations generated from the same source images, with and without ReactionGraph. (MiniMax H3)

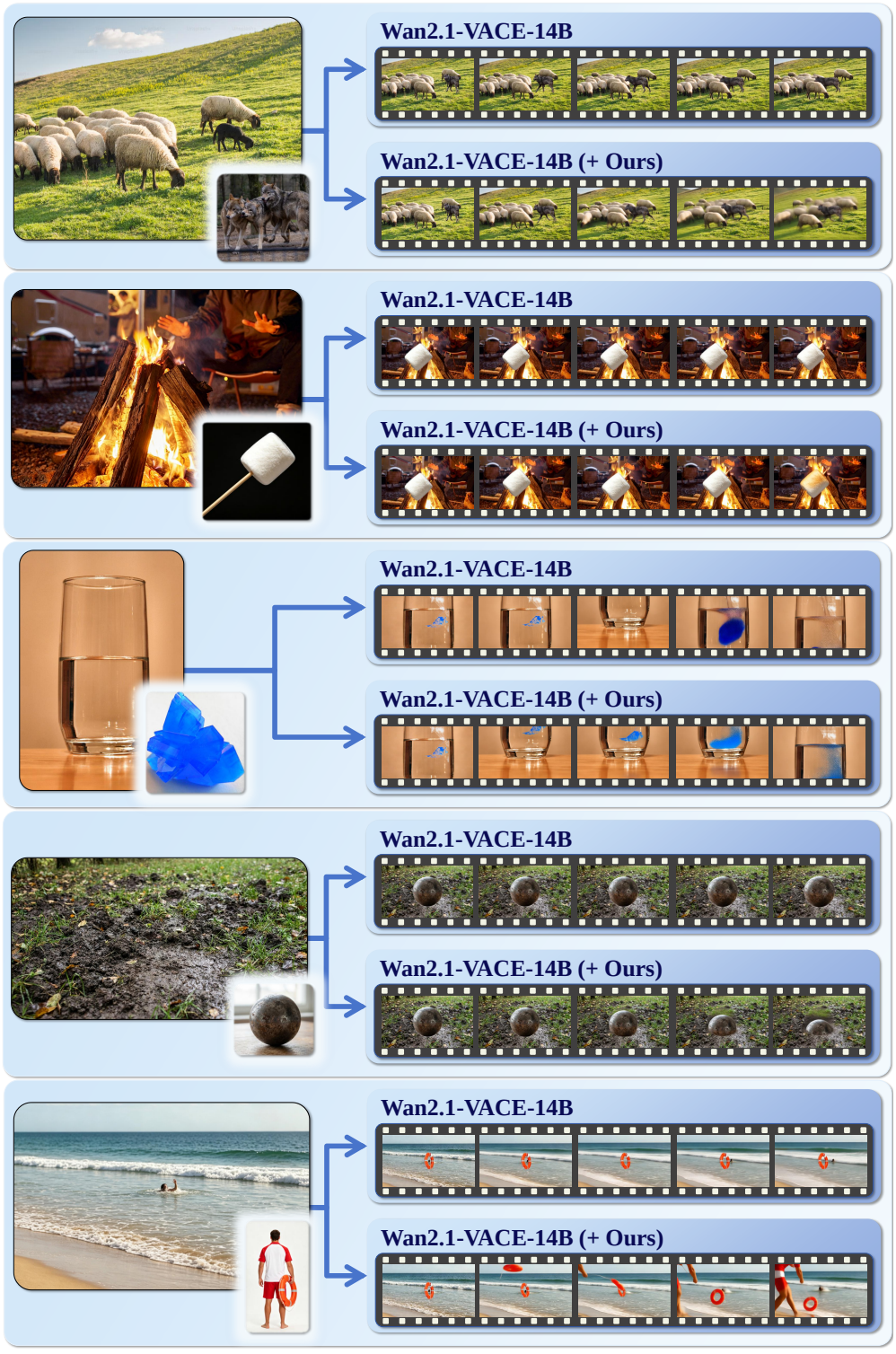


Figure 10: Qualitative comparison of representative video continuations generated from the same source images, with and without ReactionGraph. (Wan2.1-VACE-14B)



Figure 11: Qualitative comparison of representative video continuations generated from the same source images, with and without ReactionGraph. (Bernini-R)



Figure 12: Qualitative comparison of representative video continuations generated from the same source images, with and without ReactionGraph. (SkyReels-A2)

G.1 DATASHEETS

This datasheet follows the reporting goals of Datasheets for Datasets (Gebru et al., 2021). The answers below describe the released benchmark, its intended use, and the conditions that remain dependent on source-specific permissions.

G.1.1 MOTIVATION

This section poses questions primarily to encourage dataset creators to clearly explain why they create the dataset and to improve transparency regarding funding interests. Such transparency may be especially relevant to datasets developed for research.

1. For what purpose was the dataset created?

A: Existing benchmarks for multi-image composition mainly evaluate visual fidelity, semantic alignment, or cross-image consistency. They do not determine whether a generated composite realizes the implicit causal consequence licensed by the foreground–background interaction. IRS-Bench was created to evaluate cross-image interaction understanding, implicit causal reasoning, and causally faithful image synthesis across eleven causal tasks.

2. Who created the dataset (e.g., which team, research group) and on behalf of which entity?

A: IRS-Bench was assembled by the authors of this anonymous submission with support from their affiliated institutions. Author identities and contact details will follow the venue's anonymity policy.

3. Who funded the creation of the dataset?

A: The creation of the dataset was supported by the authors' affiliated institutions. No external funding source is claimed in this manuscript.

G.1.2 COMPOSITION

Most questions in this section aim to give dataset users the information needed to make informed choices about using the dataset for their intended tasks. Some questions seek information on compliance with the General Data Protection Regulation (GDPR) of the EU or comparable regulations in other jurisdictions. Questions specific to datasets relating to people are collected at the end of this section. We recommend interpreting broadly what it means for a dataset to relate to people. For example, any dataset that includes text authored by people falls within this category.

1. What do the instances that comprise our datasets represent (e.g., documents, photos, people, countries)?

A: Each primary instance is a foreground–background image pair together with an annotation of the entities, their causal relation, the governing leaf task, an instance-level causal regime, admissible reactions, environmental effects, and weighted binary verification questions. The images are AI-generated, web-sourced or drawn from public datasets, or photographed by the authors.

2. How many instances are there in total (of each type, if appropriate)?

A: IRS-Bench contains 1,000 foreground–background pairs: 454 use AI-generated imagery, 366 use web imagery or public datasets, and 180 use author-photographed imagery. A subset of 172 pairs has a photographic post-interaction reference. A separate 100-instance multi-foreground collection and the controlled sets in Section 5.4 are supplementary and are not included in the primary 1,000-instance aggregate score.

3. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

A: It is a curated, criterion-based sample rather than an exhaustive or probabilistic enumeration. Candidates were selected to cover diverse causal mechanisms and interaction patterns while satisfying the identity, relation, and absent-reaction criteria described in Section C.

4. Is there a label or target associated with each instance?

A: Yes. Every instance has an annotated causal rule, admissible outcome set, environmental effects, and ruler questions with expected answers. Not every instance has a unique target image: only the

172 photographed cases provide a post-interaction photographic reference, because a single image cannot enumerate all admissible realizations.

5. Is any information missing from individual instances?

A: The released record is intended to be complete for evaluation. A photographic reference is intentionally absent for most instances and is not treated as missing annotation; those cases are evaluated from their admissible outcomes and rulers.

6. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?

A: The benchmark does not define a social or temporal graph between separate instances. Within each record, the foreground–background relation, entity identities, causal rule, and provenance links are explicit.

7. Are there recommended data splits (e.g., training, development/validation, testing)?

A: IRS-Bench is a held-out evaluation benchmark rather than a training resource. We recommend using the complete version as a test set and reporting scores per leaf task and evaluator. Any development or prompt tuning should use a separately documented split or external data.

8. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

A: No. It depends partly on web imagery and public datasets. The release must include a per-item provenance and license manifest; assets that cannot be redistributed should be represented by metadata or source references rather than repackaged.

9. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?

A: The benchmark was not designed to collect private communications, confidential records, or personal attributes. Public or web-sourced images may contain incidental depictions of people, and the release manifest should document any filtering and provide a channel for privacy or removal concerns.

10. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

A: The benchmark is not curated for offensive or sensitive content, but ordinary web imagery can contain incidental people or contextual material. Users should inspect the manifest and apply their own institutional content policies before deployment.

G.1.3 Collection Process

Beyond the objectives described in the preceding section, this section includes questions intended to provide researchers and practitioners with information useful for developing alternative datasets with comparable characteristics. Questions specific to datasets involving people are likewise grouped at the end of the section.

1. How was the data associated with each instance acquired?

A: Candidates were drawn from AI-generated imagery, web imagery and public datasets of real scenes, and photographs taken by the authors. A pair was retained only when the foreground entity was identifiable, the background supported a definite causal interaction, and the entailed reaction was absent from both inputs.

2. What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?

A: Construction includes source filtering, task and regime annotation, post-interaction ground-truth capture for 172 physically realizable cases, and compilation of the annotation into atomic position, state, effect, and identity constraints. Language models assisted candidate generation and anno-

tation; authors and an independent second annotator reviewed each retained record. The review protocol is specified in Appendix C.2.

3. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

A: Sampling was criterion-based, not random. Candidates were filtered by the identity, relation, absent-reaction, and coverage criteria, then assigned exactly one of eleven leaf tasks and one instance-level regime.

G.1.4 Preprocessing, Cleaning, and Labeling

The questions in this section provide dataset users with the information needed to determine whether the processing applied to the raw data is compatible with their intended tasks. For example, text converted into a bag-of-words representation is unsuitable for tasks that depend on word order.

1. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

A: Yes. Each retained pair was annotated with provenance, entities, the governing task and regime, the relation tuple, admissible reactions, environmental effects, and a weighted ruler. Automated checks reject empty constraint sets, incompatible state requirements, unsatisfiable identity constraints, and zero total weight. A second annotator independently reviews the record before disagreements are resolved and stored.

2. Was the 'raw' data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

A: Yes, raw data is accessible.

3. Is the software that was used to preprocess/clean/label the data available?

A: The release should include the construction and evaluation scripts, annotation schema, ruler prompts, and exact serialization format needed to reproduce the benchmark processing. Model-specific generation services or restricted source assets may require users to provide their own access.

G.1.5 Uses

The questions in this section prompt dataset creators to consider which tasks the dataset is and is not appropriate for. Explicitly identifying these intended and unintended uses helps dataset users make informed decisions and avoid potential risks or harms.

1. Has the dataset been used for any tasks already?

A: Yes. This paper uses IRS-Bench to compare image-composition systems, evaluate ReactionGraph, analyze controlled variants, and report evaluator-specific causal scores.

2. Is there a repository that links to any or all papers or systems that use the dataset?

A: An anonymous artifact repository will be provided during review when permitted by the venue, followed by a public repository after publication and source-permission review.

3. What (other) tasks could the dataset be used for?

A: It can support research on multi-image composition, causal scene understanding, evaluator design, controlled shortcut tests, and qualitative transfer to multi-foreground or video continuation settings.

4. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

A: No.

5. Are there tasks for which the dataset should not be used?

A: N/A.

G.1.6 DISTRIBUTION

Dataset creators should answer these questions before distributing the dataset, regardless of whether the recipients are internal members of the institution that created it or external third parties.

1. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

A: No. Our IRS-Bench dataset will be made publicly accessible to the research community.

2. How will the dataset be distributed (e.g., tarball on website, API, GitHub)?

A: We will provide IRS-Bench in the GitHub and the Hugging face repository.

3. When will the dataset be distributed?

A: We plan to release the dataset through a public repository upon official publication of the paper, while maintaining compliance with anonymity policies.

4. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

A: Yes, the dataset will be released under the Creative Commons Attribution NonCommercial-ShareAlike 4.0 International License.

5. Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

A: No.

6. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

A: No.

G.1.7 MAINTENANCE

As with the questions in the preceding section, dataset creators should provide answers before distributing the dataset. The questions in this section are intended to encourage creators to plan dataset maintenance and communicate the maintenance plan to dataset users.

1. Who will be supporting/hosting/maintaining the dataset?

A: The authors will support, host, and maintain the artifact, subject to the permissions of its component sources.

2. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

A: Upon acceptance of the paper, the authors' email addresses will be provided in the paper or on the project website, through which the dataset curators can be contacted.

3. Is there an erratum?

A: No formal erratum is currently available. Corrections and clarifications will be recorded in a versioned errata file.

4. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

A: Any future updates will be announced and documented on the dataset website.

5. Will older versions of the dataset continue to be supported/hosted/maintained?

A: Yes. This initial release will be updated overtime, with earlier versions superseded by newer releases.

6. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

A: Yes. We will provide clear and detailed guidelines to support future dataset extensions.