Chain-of-Image Generation: Toward Monitorable and Controllable Image Generation

Young Kyung Kim [1 2] **Oded Schlesinger** [1 2] **Qiangqiang Wu** [1 3] **J. Matías Di Martino** [2 4] **Guillermo Sapiro** [1 5]

Abstract

While state-of-the-art image generation models achieve remarkable visual quality, their internal generative processes remain a "black box." This opacity limits human observation and intervention, and poses a barrier to ensuring model reliability, safety, and control. Furthermore, their non-human-like workflows make them difficult for human observers to interpret. To address this, we introduce the Chain-of-Image Generation (CoIG) framework, which reframes image generation as a sequential, semantic process analogous to how humans create art. Similar to the advantages in monitorability and performance that Chain-of-Thought (CoT) brought to large language models (LLMs), CoIG can produce equivalent benefits in text-to-image generation. CoIG utilizes an LLM to decompose a complex prompt into a sequence of simple, step-by-step instructions. The image generation model then executes this plan by progressively generating and editing the image. Each step focuses on a single semantic entity, enabling direct monitoring and intervention when errors are detected. We formally assess this property using two novel metrics: CoIG Readability, which evaluates the clarity of each intermediate step via its corresponding output; and Causal Relevance, which quantifies the impact of each procedural step on the final generated image. We further show that our framework mitigates entity collapse by decomposing the complex generation task into simple subproblems, analogous to the procedural reasoning employed by CoT. Our experimental results, validated by a human study, confirm that CoIG substantially enhances monitorability while achieving competitive compositional robustness. The framework is model-agnostic and integrates with any image generation model.

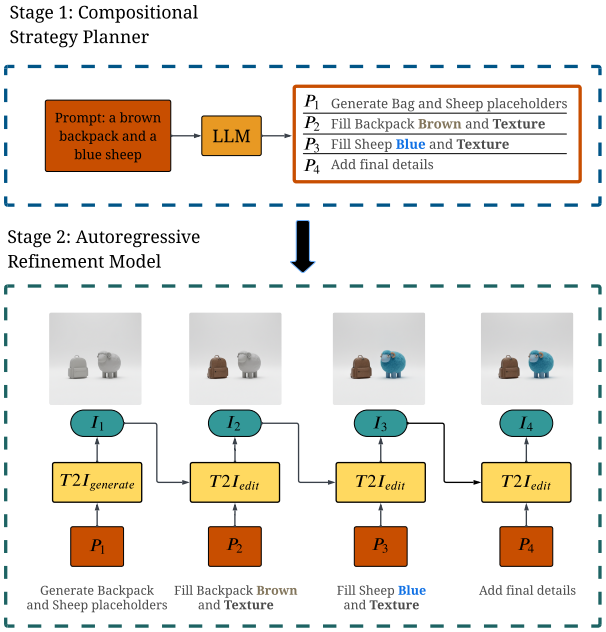


Figure 1. Proposed Chain-of-Image Generation (CoIG) framework. CoIG consists of two principal stages. (1) The Compositional Strategy Planner (CSP), in which an LLM decomposes a single complex prompt into a sequence of simpler sub-prompts $\{P_1, \ldots, P_n\}$ that serves as a human-readable and monitorable plan for iterative editing. (2) The Autoregressive Refinement Model (ARM) executes this plan by first generating an initial image using a text-to-image model (T2I) $I_1 = T2I_{generate}(P_1)$ and then iteratively refining the image via $I_t = T2I_{edit}(I_{t-1}, P_t)$, treating the previous image I_{t-1} as an explicit state that conditions subsequent generations.

1. Introduction

Recent breakthroughs in generative artificial intelligence (AI), primarily driven by diffusion and autoregressive models, have significantly advanced the field of image synthe-

[1]Department of Electrical and Computer Engineering, Princeton University, Princeton, USA [2]Department of Electrical and Computer Engineering, Duke University, Durham, USA [3]Department of Computer Science, City University of Hong Kong, Hong Kong, China [4]Department of Informatics and Computer Science, Universidad Católica del Uruguay, Montevideo, Uruguay [5]Apple, Cupertino, USA. Correspondence to: Young Kyung Kim <yk4491@princeton.edu>, Guillermo Sapiro <guillermos@princeton.edu>.

Published as a paper at the 1st FoGen workshop, ICML 2026, Seoul, South Korea, 2026.

sis (Sohl-Dickstein et al., 2015; Van Den Oord et al., 2016; 2017; Parmar et al., 2018; Ho et al., 2020; Song et al., 2020). The continued scaling of large image-text datasets has dramatically improved the fidelity, versatility, and breadth of generated imagery (Rombach et al., 2022; Ramesh et al., 2021; Saharia et al., 2022; Betker et al., 2023; Podell et al., 2023). Yet, the critical challenge of monitorability remains largely unaddressed despite this rapid progress.

In the broader context of AI, monitorability—defined as the capacity for human (or in some cases, machine) oversight—is fundamental for ensuring accountability, safety, and controllability. This concept is an active area of research within Large Language Model (LLM), where observing Chain-of-Thought (CoT) reasoning pathways is a key area of study (Korbak et al., 2025; Emmons et al., 2025).

Extending the principle of monitorability to image generation, however, introduces unique challenges, as dominant synthesis methodologies are fundamentally different from the human approach to creating images. For instance, diffusion models (Ho et al., 2020; Song et al., 2020) operate by simultaneously refining a complete pixel canvas, while conventional autoregressive models (Li et al., 2024b; Lee et al., 2022) construct an image patch-wise. The human process of image generation, conversely, is inherently both autoregressive and semantic: An artist typically composes a scene by first outlining primary subjects, then gradually adding details to those subjects, and often addressing the background as a final compositional step (Iarussi et al., 2013; Goel, 1995). This semantic progression allows an observer to easily understand the artist's focus at any given stage.

In contrast, the intermediate steps of most generative models are largely non-interpretable to a human observer (or any monitoring system imitating it). The simultaneous refinement of all subjects and background elements in diffusion models obscures clear compositional logic. Similarly, in conventional patch-based autoregressive models, independently determining whether a given patch corresponds to a specific semantic entity or the background is non-trivial.

To address this limitation, we introduce *Chain-of-Image Generation (CoIG)*, a generation framework that makes image generation monitorable by emulating a human-like process analogous to CoT prompting in reasoning tasks. Building on prior work on LLM-based prompt refinement for compositional synthesis (Yang et al., 2024; Hao et al., 2023; Mañas et al., 2024), our framework utilizes an LLM to decompose an input prompt into a sequence of sub-prompts. This sequence is designed to mirror a human artist's workflow; for instance, an initial sub-prompt may define the spatial arrangement of all subjects, while subsequent steps progressively elaborate on specific semantic components, addressing one semantic component per step. The CoIG process operates in an autoregressive manner: An initial image is synthesized from the first sub-prompt, followed by iterative edits based on each succeeding sub-prompt in the sequence. Figure 1 provides an illustrative example of this process.

To quantitatively evaluate monitorability, we introduce two novel metrics: CoIG Readability, which quantifies the interpretability of intermediate steps; and Causal Relevance, which measures the contribution of each intermediate step to the final generated output. Full metric details are provided in Section 4.

Furthermore, as we build CoIG, we identify prevalent failure cases in existing generative models. A notable example occurs when prompts specify multiple, similar entities with different attributes or actions (examples in Appendix B). In such cases, models often exhibit attribute confusion or merge the entities into a single mode, which we define as "entity collapse." We demonstrate that CoIG mitigates this issue by decomposing the complex task into simpler, sequential steps, similar to how CoT prompting breaks down complex reasoning problems for LLM.

Our main contributions are:

- We introduce monitorability as a first-class objective in image generation and propose CoIG, a novel framework that decomposes a complex prompt into a sequence of tractable, semantic steps, each producing a human-readable intermediate image that enables direct monitoring and intervention. To our knowledge, this is the first framework to formally define, measure, and evaluate monitorability in image generation via human-readable intermediate images.
- To quantitatively evaluate monitorability in this and future work, we introduce two novel metrics for image generation: CoIG Readability, which measures the clarity of intermediate steps to a human observer; and Causal Relevance, which assesses the impact of each step on the final output. A human study validates both metrics and confirms that MLLM evaluators serve as reliable proxies.
- We systematically address "entity collapse," a critical form of attribute binding failure in which attributes of similar entities are merged, and demonstrate its mitigation through our proposed CoIG framework.

2. Related Works

A key advancement for LLMs is CoT prompting, which enables step-by-step reasoning before producing an output (Wei et al., 2022; Kojima et al., 2022; Lu et al., 2022; Yao et al., 2023; Zhao et al., 2023), proving effective across arithmetic, symbolic reasoning, and medical question-answering (Wang et al., 2022; Lewkowycz et al., 2022; Yao et al., 2023; Singhal et al., 2023; 2025). The emergence

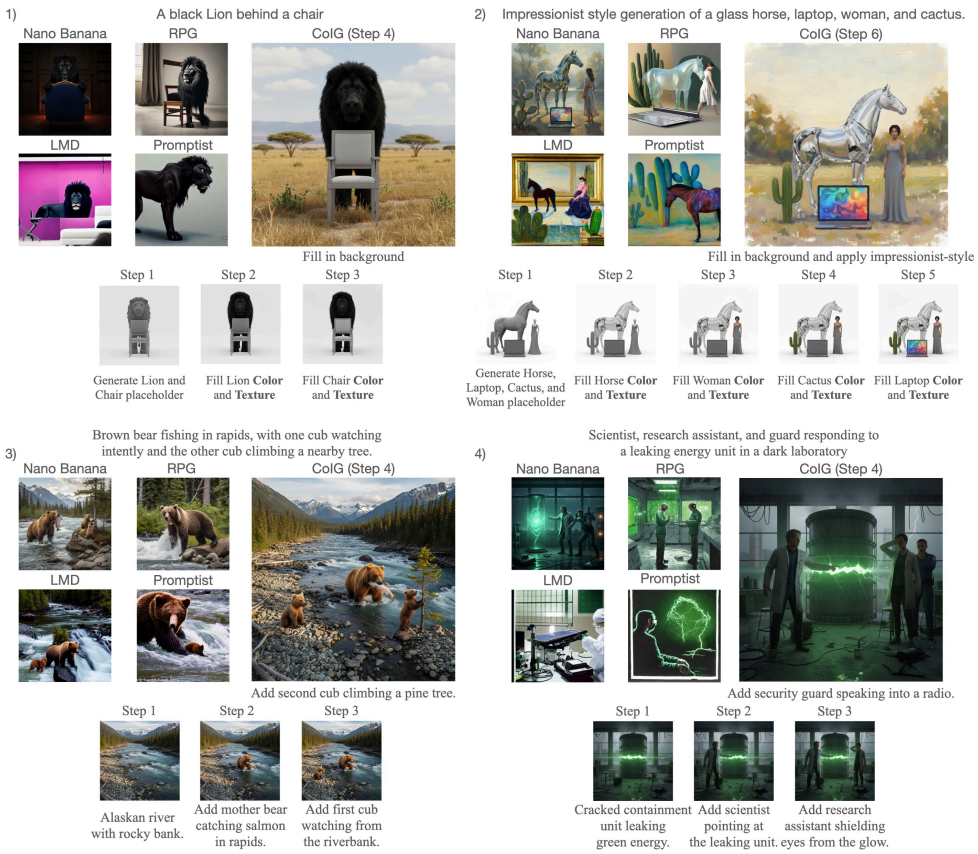


Figure 2. Qualitative comparison between our **CoIG** framework and four baselines—**Nano Banana** (Fortin et al., 2025), **RPG** (Yang et al., 2024), **LMD** (Lian et al., 2023), and **Promptist** (Hao et al., 2023). Across four complex scenarios that require spatial reasoning and multi-entity coordination (1–4), the baselines frequently struggle with entity collapse and incorrect spatial arrangements. In contrast, CoIG decomposes the prompt into explicit, monitorable steps (shown at the bottom of each panel), ensuring precise object placement and correct attribute binding. For additional examples, please see Appendix A.

of explicit rationales prompted research into their faithfulness (Jacovi & Goldberg, 2020; Turpin et al., 2023; Arcuschin et al., 2025; Chen et al., 2025) and, more broadly, monitorability: the capacity for a human to observe and intervene in the generative process (Korbak et al., 2025; Emmons et al., 2025). Despite these advances, such principles have not been translated to image generation, primarily because diffusion and autoregressive models operate in a non-semantic manner inherently difficult for a human to monitor. This architectural limitation motivates the need for a new framework capable of producing a human-readable generative process for images.

A central challenge in text-to-image synthesis is compositionality: faithfully rendering multiple entities with specified attributes and relationships (Liu et al., 2022; Huang et al., 2023b; Zheng et al., 2023; Li et al., 2023). Benchmarks such as T2I-CompBench (Huang et al., 2023a), GenEval (Ghosh et al., 2023), and others (Li et al., 2024a; Hu et al., 2024; Wu et al., 2024; Wei et al., 2025; Li et al., 2025b) systematically test attribute binding and spatial relationships. Beyond these known challenges, we investigate "entity collapse," a specific failure in which models incorrectly merge features between semantically similar entities.

A significant body of work addresses compositional fidelity through different mechanisms. Inference-time attention-control methods (Feng et al., 2022; Rassin et al., 2023; Chefer et al., 2023) modify internal cross-attention maps to improve attribute binding, but require architecture access and produce no human-visible intermediates. Training-based approaches (Wang et al., 2024; Zarei et al., 2024) similarly improve binding without exposing the generative process. LLM-based planners decompose prompts into structured text plans (Yang et al., 2024; Hao et al., 2023; Lian et al., 2023; Li et al., 2025a), adding transparency at the planning stage, but the subsequent image generation remains a single black-box call. More recently, GoT (Fang et al.) produces language reasoning chains with spatial coordinates and ImageGen-CoT (Guo et al., 2025) verifies autoregressive tokens with reward models; both introduce reasoning before generation, yet neither produces human-

Figure 3. Visualizing the monitorable advantage of CoIG. Unlike the black-box Baseline, which suffers from attribute leakage (e.g., bleeding "green" onto the cup) or ignores texture constraints (e.g., missing "fluffy"), CoIG's readable steps ensure precise attribute binding. More examples in Appendix C.

readable intermediate images during the generative process itself. The closest paradigm to ours is iterative agentic editing (Kovalev et al., 2025; Liang et al., 2025; Jaiswal et al., 2026; Mohebbi et al., 2025), where a VLM critic repeatedly refines a draft image toward the target. These methods do produce visible intermediate images, but each intermediate is a draft of the full final scene rather than a single-entity, semantically interpretable step—making it difficult for a human observer to isolate what changed and why. Crucially, none of these lines of work formally defines or measures monitorability. CoIG differs by making monitorability a first-class objective: each step is constrained to a single semantic update, and we provide dedicated metrics (Readability, Causal Relevance) and a human study to verify that the generative process is both interpretable and causally faithful.

3. Chain-of-Image Generation Framework

The CoIG framework (Figure 1) is designed to enhance the monitorability of the image generation pipeline. Our approach consists of two core components: a Compositional Strategy Planner (CSP), leveraging an LLM to decompose the prompts; and an Autoregressive Refinement Model (ARM), employing a text-to-image model (T2I) for iterative image synthesis. This two-stage process transforms a single, complex prompt into a monitorable, step-by-step generation process that mimics the human artistic process.

3.1. Compositional Strategy Planner

The CSP transforms conventional one-step generation into a transparent, sequential process. We implement CSP using a pre-trained LLM, prompted with specific decomposition rules (described in Appendix D), to parse the original prompt into a logically ordered sequence of n simplified sub-prompts $P_1, P_2, \ldots, P_n$, each focusing on a single semantic component (e.g., the color or shape of a specific object). This decomposition serves two primary functions:

Ensuring monitorability. The sequence of sub-prompts serves as a human-readable "plan of action." Before image generation, one can inspect this plan and supervise the model's intended trajectory. Such transparency underpins monitorability by enabling verification and intervention, thereby enhancing model reliability and safety.

Enforcing compositionality. By separating a complex scene into discrete tasks, we prevent the model from "blending" or "collapsing" attributes between similar entities, by handling each attribute in a separate step. This is demonstrated in Figure 2.

3.2. Autoregressive Refinement Model

The ARM uses the generated sequence of sub-prompts $\{P_1, P_2, \ldots, P_n\}$ to progressively guide the model toward the intended final image. The initial image I_1 is directly generated from P_1, written as

$$I_1 = T2I_{\text{generate}}(P_1), \tag{1}$$

where $T2I_{\text{generate}}(\cdot)$ denotes the T2I model's generative operation. This initial image establishes the basic composition, such as a placeholder for each subject or the background of the scene.

In each subsequent step, the T2I model refines the previous output I_{t-1} according to the current sub-prompt P_t,

$$I_t = T2I_{\text{edit}}(I_{t-1}, P_t), \tag{2}$$

where $T2I_{\text{edit}}(\cdot)$ denotes the T2I model's editing operation. Each intermediate image I_{t-1} thus functions as an explicit state, carrying the cumulative visual context up to step $t-1$. This mechanism is analogous to state propagation in conventional autoregressive models, but with each state fully observable and monitorable. When necessary, we can edit P_t to guardrail the generative process within our intended trajectory.

4. Monitorability in Image Generation

In generative AI, particularly concerning LLMs, monitorability is predicated on two fundamental principles (Korbak et al., 2025; Emmons et al., 2025): (1) **Readability**, the capacity of a model to produce interpretable reasoning; and (2) **Causal Relevance**, the existence of a causal linkage between this reasoning and the final output. If the reasoning lacks this causal relationship, monitoring it is meaningless, as intervening in or altering intermediate steps does not

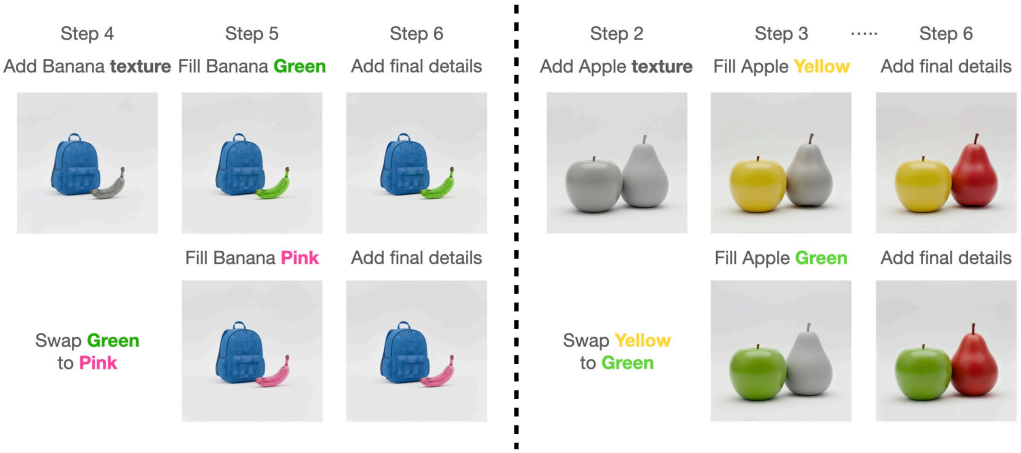


Figure 4. Qualitative demonstration of Causal Relevance. Rows 1 and 3 show the final images from two original CoIG sequences. Rows 2 and 4 show final images resulting from single perturbations applied during intermediate steps, which change the color of the apple (row 2) and the banana (row 4). The persistence of these targeted changes confirms the causal link between intermediate steps and the final output.

guarantee changing the final outcome (Adebayo et al., 2018; Carloni et al., 2025). While a monitorable LLM reasoning mimics human cognition, a monitorable image-generation process should likewise approximate human image creation. Accordingly, we provide formal definitions of **Readability** and **Causal Relevance** in CoIG,

Definition (Readability of CoIG)

Readability of CoIG requires that each intermediate generative step be **visually self-explanatory**. By requiring that each update introduce a single distinct visual concept—whether a specific object, a coherent group of entities, or a targeted attribute—we ensure that users can discern the purpose of each step simply by observing the image. This restriction minimizes ambiguity, allowing errors to be immediately localized to specific instructions rather than obscured within complex, multifaceted updates. Figure 3 illustrates this principle.

Definition (Causal Relevance)

Causal Relevance requires each intermediate generative step to contribute demonstrably and persistently to the final output. Analogous to an artist's brushstroke, once a subject is introduced or a detail refined, its effect should persist through subsequent stages. If an action's effects disappear or are overwritten by the final step, that step lacks causal relevance and becomes meaningless to monitor. Figure 4 illustrates this principle.

Our proposed framework implements these two properties through *compositional lock*, which ensures that content generated in prior steps is locked and remains unaltered unless explicitly targeted by the current instruction (Hertz et al., 2022; Simsar et al., 2025). Concretely, each sub-prompt P_t names a single target entity to update and explicitly instructs the editor to preserve all previously generated entities unchanged (position, shape, appearance). This is enforced via a fixed per-step prompt template within $T2I_{\text{edit}}$; full template details are provided in Appendix D. Readability is secured by altering only the prompted subject while keeping the rest fixed, making the visual change clearly tied to the task. Causal relevance is ensured by preserving all prior edits, which yields a visual analog of faithfulness—whether the model's explicit reasoning causally produces the output rather than a post hoc justification after the output is generated (Turpin et al., 2023; Chen et al., 2025). This makes the process meaningful for human monitoring. The compositional lock is enforced via prompt instructions rather than explicit spatial constraints (e.g., masks or bounding boxes). This is a deliberate design choice: prompt-level enforcement preserves model-agnostic compatibility, requiring only a text-guided editing interface without access to internal architecture, attention maps, or segmentation masks. Integrating spatial constraints (e.g., ControlNet (Zhang et al., 2023)) or inference-time attention-control methods could further strengthen preservation in dense scenes and is a promising direction for future work.

We quantitatively evaluate the Readability and Causal Relevance of CoIG via the following methods:

1. **Evaluation of CoIG Readability.** We use a Multimodal Large Language Model (MLLM) as an automated proxy for human judgment of Readability (described in detail in Appendix D). For any intermediate image I_t generated after applying sub-prompt P_t targeting a semantic entity e_t, the MLLM verifies the presence and attributes of e_t in I_t. High fidelity between P_t and the MLLM's verification indicates high

Table 1. Comparison of CoIG and the baseline across three benchmarks, showing competitive compositional performance.

(a) GenEval

Metric	Baseline	CoIG
Single Obj	83.75	98.75
Two Obj	90.91	98.99
Counting	92.50	86.25
Colors	88.30	92.55
Position	92.50	99.00
Color Attri	91.50	93.50
Overall	89.91	94.84

(b) T2I-CompBench

Metric	Baseline	CoIG
Color	90.43	93.31
Shape	82.92	84.02
Texture	83.57	83.85
Spatial	91.10	94.83
Non-Spatial	86.60	84.77
Complex	70.63	69.69
Overall	84.20	85.08

(c) ConceptMix

Metric	Baseline	CoIG
$k = 1$	91.76	91.04
$k = 2$	89.26	86.91
$k = 3$	88.02	84.19
$k = 4$	84.70	78.58
$k = 5$	83.42	81.17
$k = 6$	83.64	79.55
$k = 7$	84.11	78.49
Overall	86.43	82.86

readability.

2. **Evaluation of CoIG Causal Relevance.** We evaluate causal relevance using controlled perturbations and ablations of sub-prompts. First, we flip the target attribute at step t (e.g., changing "red bowl" to "blue bowl") and assess (i) whether the edit appears in the intermediate image I_t, and (ii) whether it persists in subsequent images, especially the final output I_n. An edit that both manifests at step t and persists through to step n indicates a direct and durable causal relevance on the final image. Second, we ablate step t by removing it; if this removal substantially degrades the final output, then step t is causally necessary rather than a mere intermediate visualization.

5. Compositional Robustness and Entity Collapse

5.1. Defining and Analyzing Entity Collapse

Definition (Entity collapse)

Entity collapse occurs when a prompt specifies n semantically similar entities with distinct attributes $\{A_i\}_{i\in\{1,\dots,n\}}$, but the generated image (i) depicts fewer than n instances (merge), (ii) misassigns attributes (swap/leak), or (iii) applies one entity's attributes to all (homogenization).

We attribute this failure to the excessive compositional burden of single-pass generation (Zarei et al., 2024; Campbell et al., 2024): generating the entire image from a complex prompt in one inference overloads the model's capacity to differentiate semantically related entities. Examples are shown in Figure 7 (Appendix B); additional cases appear in Appendix E.

5.2. Mitigation of Entity Collapse via Chain-of-Image Generation

Our proposed CoIG framework mitigates this by reframing the complex task as a sequence of tractable sub-tasks, a strategy analogous to CoT reasoning. CoIG generates a sequential plan: first, it creates placeholders for each entity while considering their interactions. Subsequent steps progressively fill in details for specific attributes and interactions. The final step then fills in a contextually coherent background. This sequential execution, combined with the *compositional lock* (see Section 4), intentionally constrains the model's computational focus. At each step, the model processes only a single entity or interaction. This drastically reduces the compositional load, preventing the attribute-binding failures characteristic of entity collapse.

5.3. The Entity Collapse (EC) Benchmark

To systematically evaluate robustness to entity collapse, we developed the *EC Benchmark* (to be released with this paper), a new 300-prompt benchmark specifically designed to induce this failure mode. Each prompt specifies four semantically similar entities with distinct attributes and pairwise interactions; an MLLM evaluator scores entity count, attribute binding, and interaction correctness per image. Full details on the procedural prompt generation and vocabularies, together with quantitative results, qualitative examples, and the full evaluation protocol, are in Appendix B.

6. Experiments

Our primary baseline for comparative evaluations is Google Nano Banana (Fortin et al., 2025), used in a single pass in contrast to the multi-step CoIG methodology. We implement CoIG using Gemini 2.5 Flash for the CSP and Google Nano Banana for the ARM; both $T2I_{\text{generate}}$ and $T2I_{\text{edit}}$ use Nano Banana's native text-guided editing capability without masks or specialized editing architectures. CoIG requires n sequential model calls (n depends on the complexity of the prompt), increasing inference latency roughly linearly with the number of steps relative to single-pass generation; for our default 6-step configuration, this corresponds to approximately $6\times$ the wall-clock time of a single-pass call. Full details on our experimental setup (evaluation protocols and benchmark descriptions) are in Appendix D, and a qualitative ablation across several distinct LLM planners and image generation backbones—demonstrating that CoIG is

Table 2. Monitorability evaluation on T2I-CompBench. (a) Readability: MLLM accuracy before/after the target step t. (b) Causal relevance (Color): accuracy at step t and final step under the original pipeline, step removal, and attribute flipping. (c) Intervention reliability (Color): detection, recovery, preservation rates, and net accuracy gain. Together, these results confirm that each intermediate step is individually readable, causally necessary for the final output, and actionable for error correction.

(a) Readability

	Before	After
Color	13.1%	88.7%
Shape	59.7%	81.1%
Texture	8.7%	74.7%

(b) Causal relevance (Color)

	Step t	Final
Original	81.0%	94.5%
Remove t	13.9%	15.9%
Flip (orig.)	4.2%	4.0%
Flip (flip.)	79.2%	89.6%

(c) Intervention (Color)

	Accuracy
Detect	84.6%
Recover	90.4%
Preserve	91.9%
Net Δ	44.0%

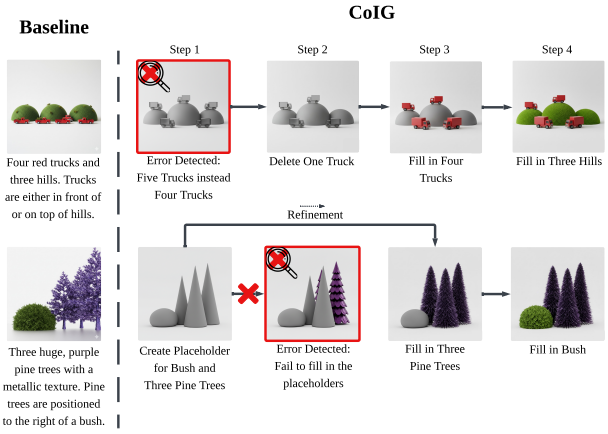


Figure 5. Leveraging monitorability for error correction. CoIG's stepwise structure exposes intermediate failures, allowing a monitor to intervene. Top: A counting error (five trucks instead of four) is corrected via a targeted "Delete One Truck" instruction. Bottom: A placeholder failure is corrected to render "purple pine trees."

not dependent on any single proprietary backbone—is in Appendix F.

6.1. Evaluation on Compositionality

Figure 2 presents a qualitative comparison against four baselines: **Nano Banana** (Fortin et al., 2025), **RPG** (Yang et al., 2024), **LMD** (Lian et al., 2023), and **Promptist** (Hao et al., 2023). The results highlight that the baselines struggle with *entity collapse* in complex prompts—for example, failing to distinguish the specific actions of two bear cubs (Panel 3) or blending the distinct roles of laboratory personnel (Panel 4). In contrast, CoIG's stepwise decomposition ensures high-fidelity adherence to spatial and semantic constraints, while preserving both monitorability and overall generative performance. To demonstrate the general applicability of our framework, we provide an additional qualitative ablation study across several distinct LLM and image generation backbones in Appendix F.

We evaluate CoIG on the GenEval, T2I-CompBench, and ConceptMix benchmarks (Huang et al., 2023a; Ghosh et al., 2023; Wu et al., 2024) to establish its general-purpose effectiveness. Full details on those datasets are in Appendix D. For our evaluation protocol, we use an MLLM evaluator adapted from (Li et al., 2025b). This evaluator is prompted to verify if the prompt's compositional instructions are accurately rendered in the generated image, providing binary (yes/no) answers.

The quantitative results for all three benchmarks are presented in Table 1. We first present the GenEval results in Table 1a. CoIG achieves a better overall score (94.84 vs. 89.91) and outperforms the baseline in the majority of categories, including Single Object, Two Objects, Position, and Color Attribute. Notably, it underperforms the baseline in the "Counting" category.

The evaluation on T2I-CompBench, detailed in Table 1b, shows a more nuanced outcome. While CoIG demonstrates improved performance in the Color and Spatial categories, it lags behind the baseline in Non-Spatial and Complex compositions. The results for Shape and Texture are comparable, leading to a slightly higher overall score (85.08 vs. 84.20), as further supported by Figure 3.

On ConceptMix (Table 1c), CoIG underperforms the baseline. CoIG's transparency lets us diagnose the cause: the zero-shot layout generator produces ambiguous placeholders for dense compositions, and the ARM struggles to fill them without attribute leaking. We quantify the ability to act on such failures in Section 6.2.4.

6.2. Evaluation of Monitorability

6.2.1. Readability of the Chain-of-Image Generation.

We evaluate readability on T2I-CompBench (Section 4). Table 2a shows that accuracy surges sharply after each target step from 13.1% to 88.7% for Color and 8.7% to 74.7% for Texture, confirming that each edit is distinct and attributable. Shape begins higher (59.7%→81.1%) since placeholders already encode structural form. Figure 3 qualitatively illustrates this monitorable progression.

Table 3. Human evaluation on the Color dataset (n=12 annotators). (a) Accuracy on readability (step identification), causal relevance (persistence), and error detection (step localization). (b) Cohen's κ between each MLLM and human annotators (mean over all pairs). Inter-human reports agreement among annotators. Human annotators achieve high accuracy across all three tasks, confirming that the intermediate steps are genuinely readable, causally relevant, and sufficient for error detection from a human perspective. The strong agreement between human and MLLM scores further validates the use of MLLMs as evaluation proxies throughout our experiments.

(a) Evaluation accuracy

Evaluator	Readability	Causal Relevance	Error Detection
GPT-4o	82%	98%	74%
Gemini 2.5 Flash	76%	98%	48%
Claude Sonnet	73%	100%	38%
Human (n=12)	74%	92%	72%

(b) Cohen's κ (MLLM vs. human)

Evaluator	Readability	Error Detection
GPT-4o	0.984	0.538
Gemini 2.5 Flash	0.886	0.359
Claude Sonnet	0.842	0.228
Inter-human	0.910	0.427

6.2.2. Causal Relevance.

We evaluate causal relevance on T2I-CompBench using the perturbation method described in Section 4, where step t is the step that introduces the target attribute (e.g., assigning a color to an object). We avoid gray substitutions since our model uses gray for placeholders. In the Original row of Table 2b, accuracy at step t (81.0%) improves further at the final step (94.5%), confirming a durable causal effect. The ablation rows further isolate this finding: removing step t collapses final accuracy from 94.5% to 15.9%, demonstrating that step t is causally necessary for the target attribute to appear in the final image. Flipping the target attribute at step t redirects the final output to match the new target (89.6%) rather than the original (4.0%), confirming that modifications at intermediate steps predictably and persistently alter the final output. Together, these experiments establish the two properties required for meaningful monitoring: each step is individually necessary, and intervening at any step produces a controlled, lasting change in the final image. We focus causal analysis on Color as it offers clean binary verification; extending to shape and spatial attributes is an important direction for future work. Qualitative examples are shown in Figure 4 and Appendix G.

6.2.3. Intervention Reliability.

Readability and causal relevance establish that the intermediate steps of CoIG are interpretable and meaningful; intervention reliability tests whether a monitor can act on this transparency to correct errors. Using the same attribute-flipping setup as the causal relevance evaluation, we inject an error at step t and task an MLLM monitor with detecting the violation and issuing a corrective instruction. Table 2c reports the results. The monitor flags 84.6% of injected violations (Detect). Among the detected cases, the repair succeeds 90.4% of the time (Recover) while preserving unrelated attributes in 91.9% of cases (Preserve). In general, the intervention increases the accuracy by 44.0% relative to the corrupted image (Net Δ). Figure 5 shows qualitative examples of successful corrections.

6.2.4. Human Evaluation.

To validate that CoIG's intermediate steps are interpretable to humans, twelve annotators independently evaluated 50 samples from the Color subset. For each sample, annotators viewed the full 6-step CoIG sequence and performed three tasks: identifying when the target attribute first appears (readability), confirming its persistence through to the final image (causal relevance), and localizing which step contains an injected error (error detection)—a harder task than the binary detection in Table 2c, as it requires identifying the specific step rather than simply flagging a mismatch. Full protocol details are provided in Appendix H.

Results are shown in Table 3a. Human annotators confirm attribute persistence in 92% of cases and correctly identify attribute emergence in 74%, validating that CoIG's steps are both readable and causally relevant. Crucially, humans localize injected errors at 72%, demonstrating that CoIG's transparency is sufficient for effective error monitoring in practice. Among the MLLMs, all three achieve above 73% on readability, confirming they serve as reasonable evaluation proxies. GPT-4o stands out on error detection (74%), surpassing human performance, with Cohen's κ of 0.538 exceeding the inter-human baseline of 0.427 (Table 3b), supporting its use as a scalable substitute for human monitors in the intervention pipeline. Extending this evaluation to additional attribute categories and larger sample sizes is future work.

7. Conclusion

This work addresses monitorability in image generation, a property fundamental to reliability and safety that remains largely unaddressed by black-box generative architectures. We propose CoIG, which reframes generation as a sequential, human-like semantic process with inherently transparent intermediate steps. Our evaluation confirms three findings: CoIG achieves competitive compositionality without sacrificing monitorability; our novel Readability and Causal Relevance metrics, validated by a human study, confirm that intermediate steps are both interpretable and causally

faithful; and CoIG effectively mitigates entity collapse on our newly introduced EC Benchmark. By demonstrating that monitorability and compositional robustness are deeply intertwined, this work lays the foundation for a new class of more reliable generative models.

Acknowledgments

Work partially supported by NSF, ONR, and the Simons Foundation.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. *Advances in Neural Information Processing Systems*, 31, 2018.

Anthropic. Introducing claude sonnet 4.5. *Blog Post, September*, 2025.

Arcuschin, I., Janiak, J., Krzyzanowski, R., Rajamanoharan, S., Nanda, N., and Conmy, A. Chain-of-thought reasoning in the wild is not always faithful. *URL https://arxiv.org/abs/2503.08679*, 2025.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3): 8, 2023.

Campbell, D., Rane, S., Giallanza, T., De Sabbata, C. N., Ghods, K., Joshi, A., Ku, A., Frankland, S., Griffiths, T., Cohen, J. D., et al. Understanding the limits of vision language models through the lens of the binding problem. *Advances in Neural Information Processing Systems*, 37: 113436–113460, 2024.

Carloni, G., Berti, A., and Colantonio, S. The role of causality in explainable artificial intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 15(2):e70015, 2025.

Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., and Cohen-Or, D. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023.

Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., et al. Reasoning models don't always say what they think. *ArXiv Preprint arXiv:2505.05410*, 2025.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *ArXiv Preprint arXiv:2507.06261*, 2025.

Emmons, S., Jenner, E., Elson, D. K., Saurous, R. A., Rajamanoharan, S., Chen, H., Shafkat, I., and Shah, R. When chain of thought is necessary, language models struggle to evade monitors. *ArXiv Preprint arXiv:2507.05246*, 2025.

Fang, R., Duan, C., Wang, K., Huang, L., Li, H., Tian, H., Yan, S., Yu, W., Zeng, X., Dai, J., et al. Got: Unleashing reasoning capability of mllm for visual generation and editing. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Feng, W., He, X., Fu, T.-J., Jampani, V., Akula, A., Narayana, P., Basu, S., Wang, X. E., and Wang, W. Y. Training-free structured diffusion guidance for compositional text-to-image synthesis. *ArXiv Preprint arXiv:2212.05032*, 2022.

Fortin, A., Vernade, G., Kampf, K., and Reshi, A. Introducing gemini 2.5 flash image, our state-of-the-art image model. *Blog Post, August*, 2025.

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

Goel, V. *Sketches of thought*. MIT press, 1995.

Guo, Z. et al. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025.

Hao, Y., Chi, Z., Dong, L., and Wei, F. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:66923–66939, 2023.

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. Prompt-to-prompt image editing with cross attention control. *ArXiv Preprint arXiv:2208.01626*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Hu, X., Wang, R., Fang, Y., Fu, B., Cheng, P., and Yu, G. Ella: Equip diffusion models with llm for enhanced semantic alignment. *ArXiv Preprint arXiv:2403.05135*, 2024.

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

Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., and Zhou, J. Composer: Creative and controllable image synthesis with composable conditions. *ArXiv Preprint arXiv:2302.09778*, 2023b.

Iarussi, E., Bousseau, A., and Tsandilas, T. The drawing assistant: Automated drawing guidance and feedback from photographs. In *ACM Symposium on User Interface Software and Technology (UIST)*. ACM, 2013.

Jacovi, A. and Goldberg, Y. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *ArXiv Preprint arXiv:2004.03685*, 2020.

Jaiswal, S., Prabhudesai, M., Bhardwaj, N., Qin, Z., Zadeh, A., Li, C., Fragkiadaki, K., and Pathak, D. Iterative refinement improves compositional image generation. *ArXiv Preprint arXiv:2601.15286*, 2026.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213, 2022.

Korbak, T., Balesni, M., Barnes, E., Bengio, Y., Benton, J., Bloom, J., Chen, M., Cooney, A., Dafoe, A., Dragan, A., et al. Chain of thought monitorability: A new and fragile opportunity for ai safety. *ArXiv Preprint arXiv:2507.11473*, 2025.

Kovalev, V., Kuvshinov, A., Buzovkin, A., Pokidov, D., and Timonin, D. Craft: Continuous reasoning and agentic feedback tuning for multimodal text-to-image generation. *ArXiv Preprint arXiv:2512.20362*, 2025.

Lee, D., Kim, C., Kim, S., Cho, M., and Han, W.-S. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11523–11532, 2022.

Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., et al. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857, 2022.

Li, B., Lin, Z., Pathak, D., Li, J., Fei, Y., Wu, K., Ling, T., Xia, X., Zhang, P., Neubig, G., et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *ArXiv Preprint arXiv:2406.13743*, 2024a.

Li, M., Hou, X., Liu, Z., Yang, D., Qian, Z., Chen, J., Wei, J., Jiang, Y., Xu, Q., and Zhang, L. Mccd: Multi-agent collaboration-based compositional diffusion for complex text-to-image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13263–13272, 2025a.

Li, O., Wang, Y., Hu, X., Huang, H., Chen, R., Ou, J., Tao, X., Wan, P., Qi, X., and Feng, F. Easier painting than thinking: Can text-to-image models set the stage, but not direct the play? *ArXiv Preprint arXiv:2509.03516*, 2025b.

Li, T., Tian, Y., Li, H., Deng, M., and He, K. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37:56424–56445, 2024b.

Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., and Lee, Y. J. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22511–22521, 2023.

Lian, L., Li, B., Yala, A., and Darrell, T. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*, 2023.

Liang, Z., Sun, J., and Ma, H. An llm-lvlm driven agent for iterative and fine-grained image editing. *ArXiv Preprint arXiv:2508.17435*, 2025.

Liu, N., Li, S., Du, Y., Torralba, A., and Tenenbaum, J. B. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pp. 423–439. Springer, 2022.

Lu, P., Qiu, L., Chang, K.-W., Wu, Y. N., Zhu, S.-C., Rajpurohit, T., Clark, P., and Kalyan, A. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *ArXiv Preprint arXiv:2209.14610*, 2022.

Mañas, O., Astolfi, P., Hall, M., Ross, C., Urbanek, J., Williams, A., Agrawal, A., Romero-Soriano, A., and Drozdzal, M. Improving text-to-image consistency via automatic prompt optimization. *ArXiv Preprint arXiv:2403.17804*, 2024.

Mohebbi, H., Abdulrahman, M., Miao, Y., Poupart, P., and Kothawade, S. Image-poser: Reflective rl for multi-expert image generation and editing. *ArXiv Preprint arXiv:2511.11780*, 2025.

OpenAI. Introducing gpt-5. *Blog Post, August*, 2025a.

OpenAI. Introducing our latest image generation model in the api. *Blog Post, April*, 2025b.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., and Tran, D. Image transformer. In *International Conference on Machine Learning*, pp. 4055–4064. PMLR, 2018.

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

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Rassin, R., Hirsch, E., Glickman, D., Ravfogel, S., Goldberg, Y., and Chechik, G. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, 36:3536–3559, 2023.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

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

Simsar, E., Tonioni, A., Xian, Y., Hofmann, T., and Tombari, F. Lime: localized image editing via attention regularization in diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 222–231. IEEE, 2025.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S. R., Cole-Lewis, H., et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *ArXiv Preprint arXiv:2011.13456*, 2020.

Turpin, M., Michael, J., Perez, E., and Bowman, S. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36: 74952–74965, 2023.

Van Den Oord, A., Kalchbrenner, N., and Kavukcuoglu, K. Pixel recurrent neural networks. In *International Conference on Machine Learning*, pp. 1747–1756. PMLR, 2016.

Van Den Oord, A., Vinyals, O., et al. Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Wang, R., Chen, Z., Chen, C., Ma, J., Lu, H., and Lin, X. Compositional text-to-image synthesis with attention map control of diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5544–5552, 2024.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. *ArXiv Preprint arXiv:2203.11171*, 2022.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837, 2022.

Wei, X., Zhang, J., Wang, Z., Wei, H., Guo, Z., and Zhang, L. Tiif-bench: How does your t2i model follow your instructions? *ArXiv Preprint arXiv:2506.02161*, 2025.

Wu, X., Yu, D., Huang, Y., Russakovsky, O., and Arora, S. Conceptmix: A compositional image generation benchmark with controllable difficulty. *Advances in Neural Information Processing Systems*, 37:86004–86047, 2024.

Yang, L., Yu, Z., Meng, C., Xu, M., Ermon, S., and Cui, B. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *Forty-first International Conference on Machine Learning*, 2024.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36:11809–11822, 2023.

Zarei, A., Rezaei, K., Basu, S., Saberi, M., Moayeri, M., Kattakinda, P., and Feizi, S. Understanding and mitigating compositional issues in text-to-image generative models. *ArXiv E-prints*, pp. arXiv–2406, 2024.

Zhang, L., Rao, A., and Agrawala, M. Adding conditional control to text-to-image diffusion models. In *Proceedings*

of the IEEE/CVF International Conference on Computer Vision, pp. 3836–3847, 2023.

Zhao, R., Li, X., Joty, S., Qin, C., and Bing, L. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. *ArXiv Preprint arXiv:2305.03268*, 2023.

Zheng, G., Zhou, X., Li, X., Qi, Z., Shan, Y., and Li, X. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22490–22499, 2023.

A. Additional Qualitative Results

Additional qualitative results are shown in Figure 6.

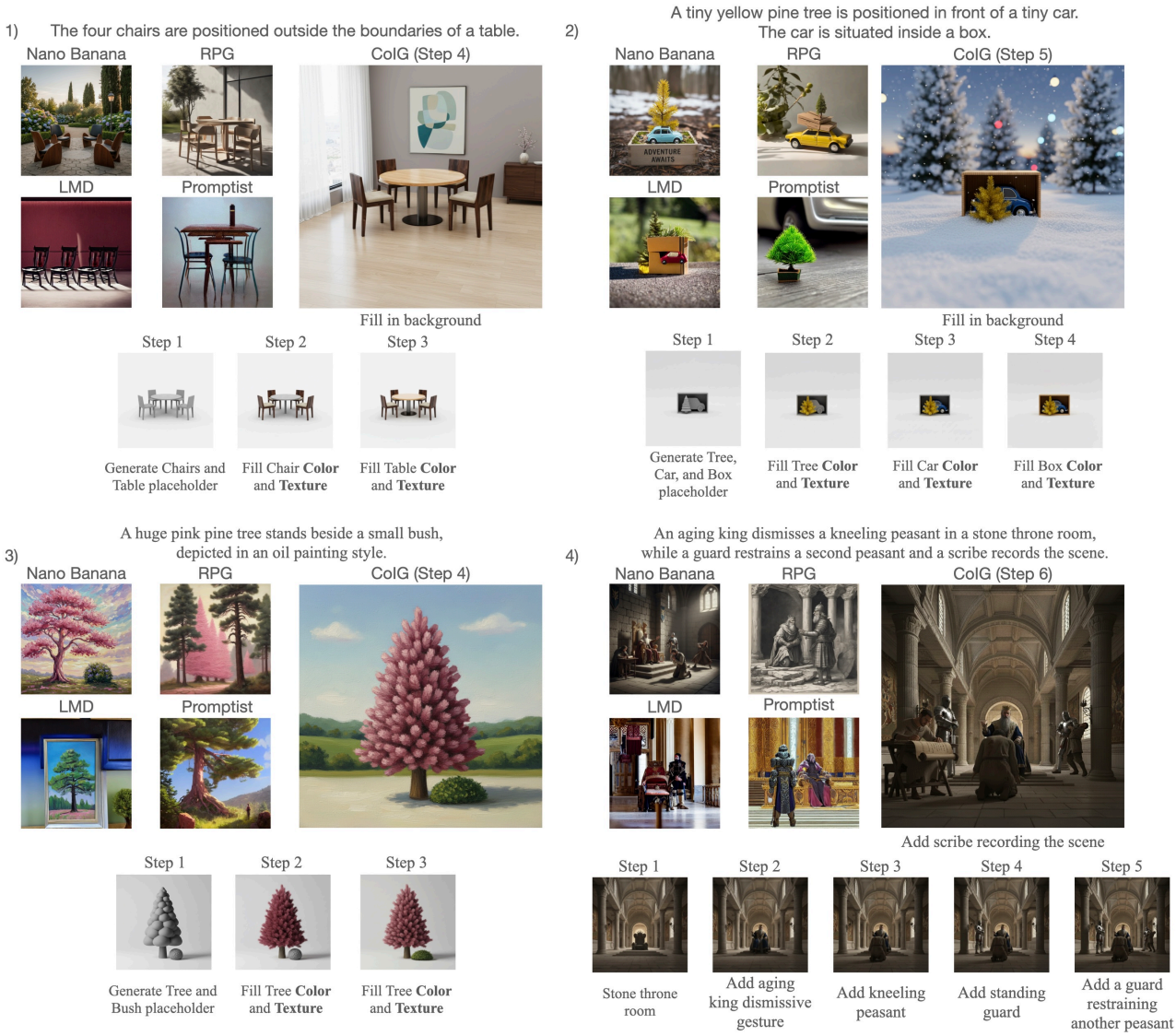


Figure 6. Qualitative comparison between our **CoIG** framework and four baselines—**Nano Banana** (Fortin et al., 2025), **RPG** (Yang et al., 2024), **LMD** (Lian et al., 2023), and **Promptist** (Hao et al., 2023). Across four complex scenarios that require spatial reasoning and multi-entity coordination (1–4), the baselines frequently struggle with entity collapse and incorrect spatial arrangements. In contrast, CoIG decomposes the prompt into explicit, monitorable steps (shown at the bottom of each panel), ensuring precise object placement and correct attribute binding.

B. The Entity Collapse (EC) Benchmark Detail

To systematically evaluate robustness to entity collapse, we developed the *EC Benchmark* (to be released with this paper), which includes 300 prompts procedurally generated by sampling from predefined categorical vocabularies representing `Jobs` (e.g., "airman"), `Attributes` (e.g., "white hair"), and `Interactions` (e.g., "glaring at"). Each prompt is constructed by sampling one `Job` category (to enforce inter-entity similarity), four distinct `Attributes`, and two `Interactions`. This targeted prompt structure (e.g., "Four airmen are in a room. The first is smiling and talking to the second, who is holding a bottle...") is specifically designed to induce and measure the "entity collapse" failure mode. This benchmark is employed to conduct a quantitative comparison of our proposed CoIG framework against baseline "black box" generative

models.

B.1. Evaluation on the Entity Collapse (EC) Benchmark

We test whether CoIG's sequential generation mitigates "entity collapse" by evaluating it against the baseline on our 300-prompt EC Benchmark (Table 4). CoIG achieves a higher total score (5.449 vs. 5.017), driven by a 21.1% gain in Entity Count and 31.8% in Interaction, indicating stronger robustness to entity merging and relationship rendering. The baseline is slightly stronger on Attribute Binding, likely because it can attach several attributes to the few entities it does generate. CoIG's placeholder-based generation explicitly allocates one slot per entity, more reliably preventing distinct entities from collapsing into one, as illustrated in Figure 7.

Table 4. Quantitative comparison on the Entity Collapse (EC) Benchmark. Our model shows significant improvements in entity count and interaction scores, leading to a higher overall score.

Metric	Baseline	CoIG (ours)
Entity Count (out of 1)	0.724 (72.4%)	0.877 (87.7%)
Attribute Binding (out of 4)	3.312 (82.8%)	3.279 (82.0%)
Interaction (out of 2)	0.980 (49.0%)	1.292 (64.6%)
Total Score (out of 7)	**5.017 (71.7%)**	**5.449 (77.8%)**

C. Additional Qualitative Results on Readability

Additional qualitative results are shown in Figure 8.

D. Experimental Setup

Evaluation Protocol: For all experiments requiring semantic verification (GenEval, T2I-CompBench, ConceptMix, CoIG Monitorability, and the EC Benchmark), our evaluation protocol uses an MLLM evaluator adapted from (Li et al., 2025b). We closely adopt the methodology from that paper for the T2I-CompBench and monitorability analyses. These tasks, which involve verifying compositional correctness and evaluating reasoning transparency, utilize Gemini 2.5 Flash (Comanici et al., 2025) as an efficient and scalable MLLM evaluator. Regarding the specific query formulation, we adapt our strategy to the dataset structure. For **GenEval** and **T2I-CompBench**, we employ a targeted QA template (e.g., "Is the {object} present? Is it {count} in count and {color} in color?"), where the attribute slots are dynamically substituted with *shape* or *texture* depending on the specific prompt constraints. In contrast, for **ConceptMix**, we adhere to the benchmark's standardized protocol by utilizing the specific question set provided by the dataset authors. For the EC Benchmark analysis, which requires more nuanced reasoning to specifically detect complex attribute-binding failures like entity collapse, we apply a modified version of their method and utilize the more powerful Gemini 2.5 Pro (Comanici et al., 2025).

To specifically quantify *Entity Collapse*, we employ a specialized "visual census" prompting strategy. Unlike standard boolean verification, this protocol requires the MLLM evaluator to act as a rigorous "quality auditor." The model is instructed to scan the image and strictly enumerate every visible entity, assigning a unique identifier (e.g., P_1, P_2) to each. Crucially, the prompt enforces a **visual evidence constraint**: the evaluator must list only attributes and interactions that are unambiguously visible, ignoring any details mentioned in the text prompt that are not rendered in the image. The model outputs a structured JSON object mapping specific attributes and interactions to these unique IDs. This granular extraction allows us to algorithmically detect collapse by verifying if the number of distinct, identified entities matches the requested count ($N_{detected} = N_{requested}$) and ensuring that interaction pairs are correctly bound to separate individuals.

Benchmarks. We conduct our evaluations across four primary benchmarks.

GenEval, T2I-CompBench, ConceptMix (Huang et al., 2023a): We utilize this standard benchmark for two purposes. First, we use them for our main compositionality comparison against the Gemini baseline. These three datasets are widely used to evaluate a model's compositional abilities, each with a unique focus: GenEval provides an object-focused evaluation of properties like object color, position, and count; T2I-CompBench similarly offers structured prompts categorized into groups like attribute binding, object relationships, and complex compositions; and ConceptMix features a scalable method that automatically generates prompts by combining an increasing number of visual concepts, such as style, attribute binding, and size. Visualizations of example prompts from each dataset are shown in Figure 9. We use the T2I-CompBench benchmark's

Figure 7. A qualitative comparison illustrating how CoIG mitigates **entity collapse**. The baseline model (center) exhibits the "merge" failure from Definition 5.1 in Rows 1 and 2, generating only **three** entities instead of the prompted **four**. In contrast, our proposed CoIG framework (right) successfully decomposes the task in all rows, generating all four distinct entities and faithfully rendering their complex, paired interactions. More examples are in Appendix E.

prompts as the basis for our monitorability evaluations; we assess CoIG Readability using the generated intermediate images, and we evaluate Causal Relevance by perturbing the step prompts by swapping an object's color to verify that the change is reflected in the final image.

Entity Collapse (EC) Benchmark. We use our novel 300-prompt benchmark to perform a targeted evaluation of the "entity collapse" failure mode, comparing CoIG directly against the baseline model.

Prompts for Compositional Strategy Planner. To implement the Compositional Strategy Planner (CSP), we utilize a fixed system prompt that directs an off-the-shelf LLM to function as a "systematic prompt architect." This prompt transforms the user's raw input into a strictly ordered sequence of semantically isolated generation steps. The decomposition process is governed by the following core algorithmic rules:

1. **Foundational Anchoring (Step 1):** The generation process begins by establishing the global scene context. Depending on the complexity of the prompt, the CSP defines the first step as either a *foundational sketch* (to lock the spatial layout

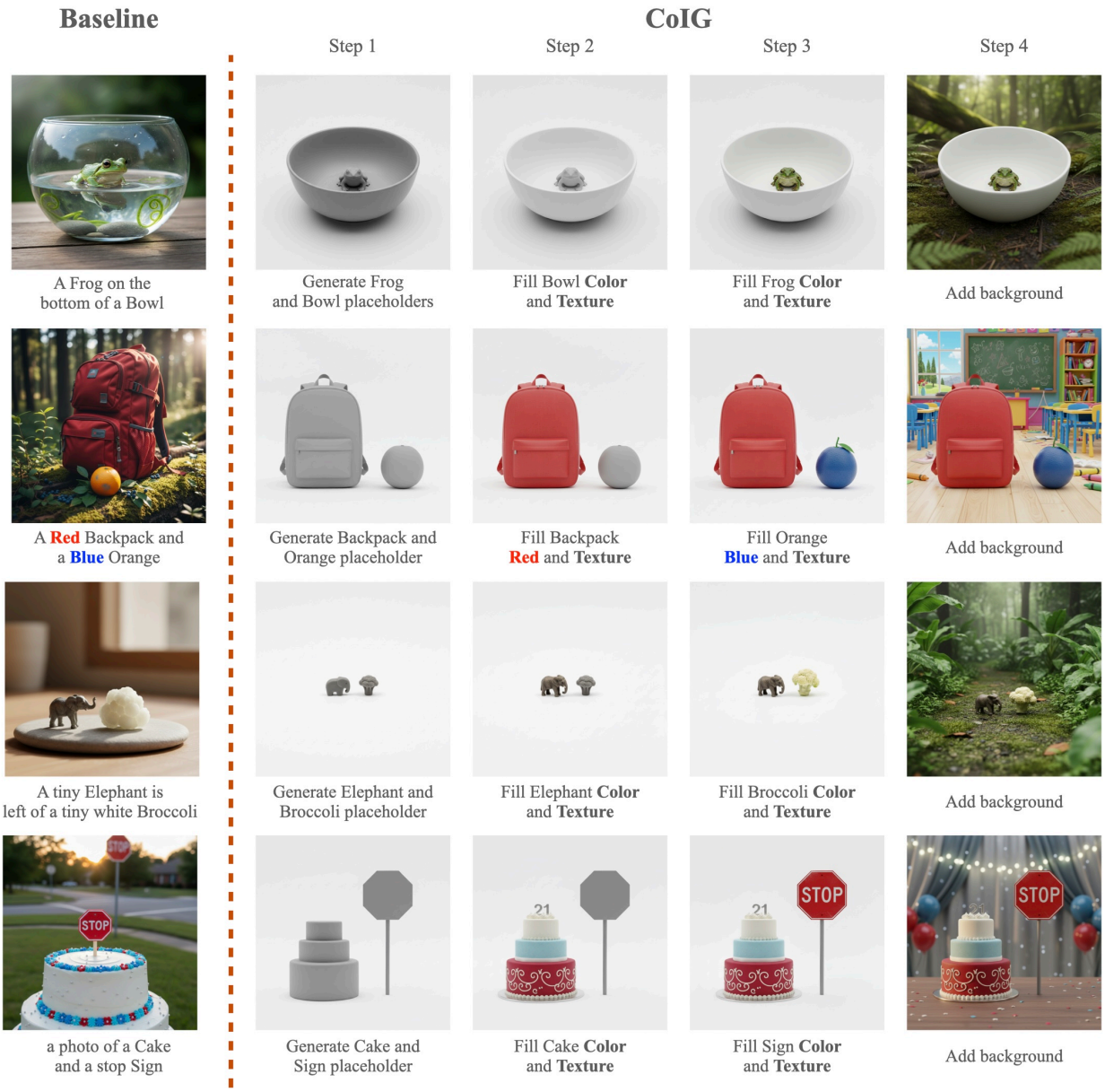


Figure 8. **Visualizing the monitorable advantage of CoIG.** In contrast to the black box Baseline, which often suffers from **attribute leakage** or **semantic bias** (e.g., failing to bind "Blue" to the orange in row 2, or defaulting to a standard "orange" color), CoIG's stepwise process ensures precise attribute binding. By visually separating geometry generation (Step 1), specific object coloring (Steps 2–3), and background synthesis (Step 4), our framework strictly adheres to prompt constraints and allows users to transparently monitor the generation trajectory.

and perspective) or a *background generation* step. This ensures that the compositional structure or environmental context is fixed immediately, providing a stable canvas for subsequent additions.

2. **Sequential Semantic Isolation:** Following the foundational step, the CSP decomposes the remaining prompt into a sequence where **exactly one semantic entity is processed per step**. By isolating each object, the framework ensures that the generative model focuses its attention on a single target at a time. This granular approach prevents attribute leakage and semantic confusion, allowing complex scenes to be constructed piece-by-piece.
3. **Entity Persistence and Immutable Locking:** To ensure consistency and prevent "entity collapse," the system prompt enforces a strict *non-destructive editing policy*. Once an object has been generated or refined in a previous step, it is designated as "locked." The prompt explicitly instructs the model to preserve these previously established regions,

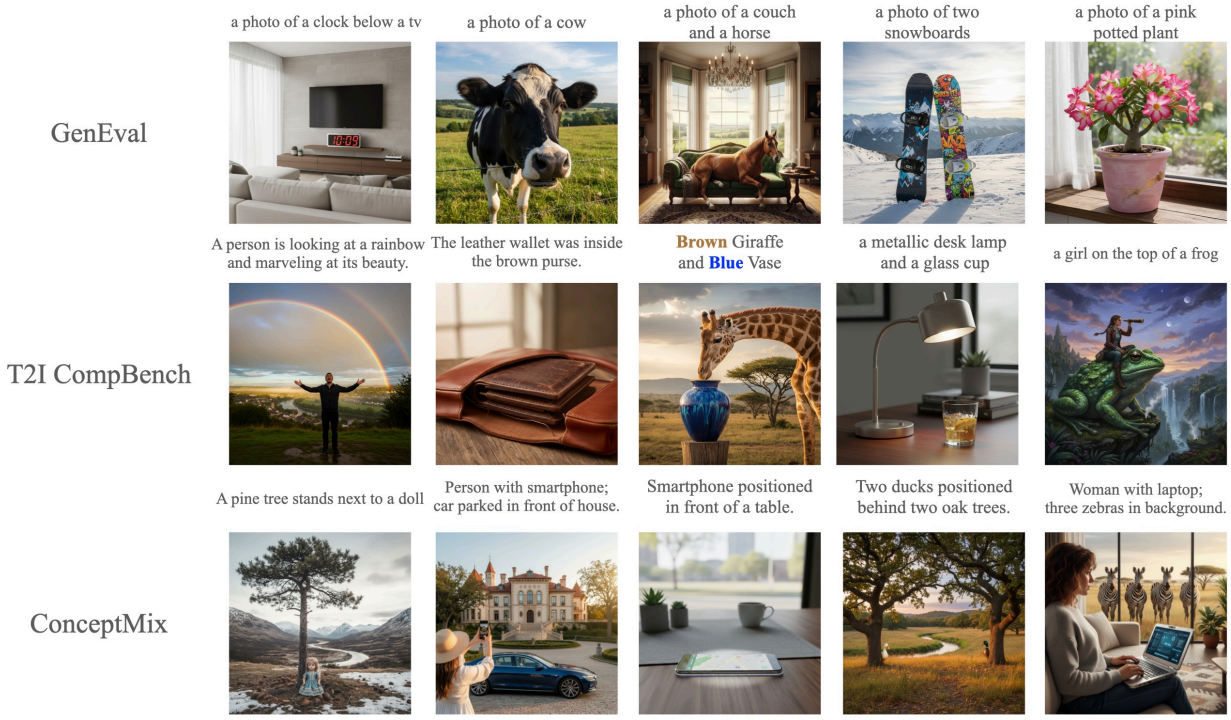


Figure 9. **Visualizations of representative prompts from the evaluation benchmarks.** We display example prompts from **GenEval** (top), **T2I-CompBench** (middle), and **ConceptMix** (bottom). To illustrate the visual content and complexity described by these datasets, we generate the corresponding images using **Nano Banana** (Fortin et al., 2025).

strictly forbidding any alterations to their shape, position, or appearance while new objects are being integrated. This guarantees that progress is cumulative and that earlier generations are not degraded by later steps.

4. **Dual-Context Prompt Structure:** Each generated step is structured into two distinct components to maintain alignment between the global objective and local modifications:
   - **Final Goal (Context):** The complete, original user caption is repeated at every step to serve as a global conditioning anchor.
   - **This Step's Action:** A specific, imperative instruction focused solely on the current operation (e.g., "Fill in the car with red color and metallic texture"), ensuring the generation trajectory remains precise.

E. Additional Qualitative Results on Entity Collapse

Additional qualitative results are shown in Figure 10.

F. Generalizability Across Different LLM and Image Backbones

A key strength of our CoIG framework is its model-agnostic architecture. The framework is designed as a versatile pipeline that can be seamlessly integrated with various Large Language Models (LLMs) and image generation architectures. To demonstrate this flexibility, we apply CoIG using three distinct LLMs (Gemini 2.5 Flash (Comanici et al., 2025), Chat GPT 5 (OpenAI, 2025a), and Claude Sonnet 4.5 (Anthropic, 2025)) combined with two different image models (Nano Banana (Fortin et al., 2025) and GPT Image 1 (OpenAI, 2025b)). As illustrated in Figure 11, the framework consistently produces high-quality final images across all combinations. This indicates that CoIG is robust to the choice of underlying backbone models, maintaining performance stability regardless of the specific components used.

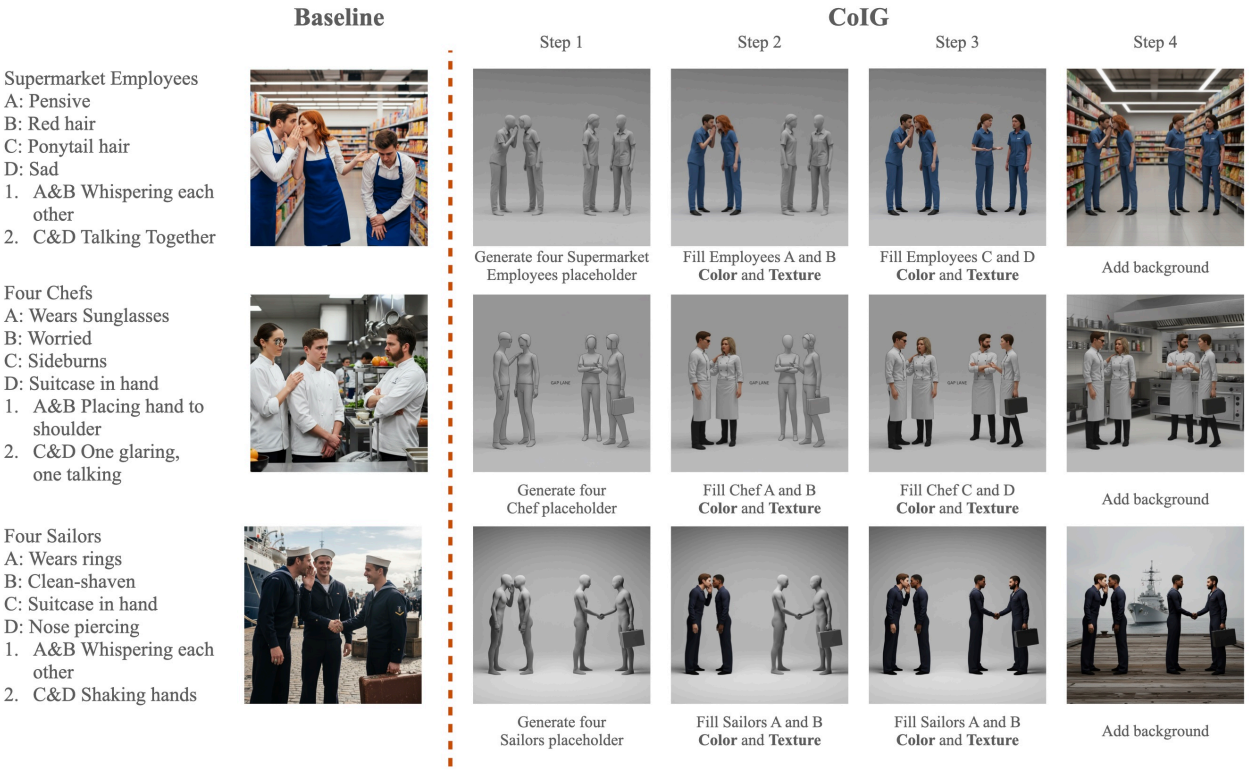


Figure 10. A qualitative comparison illustrating how CoIG mitigates **entity collapse**. The baseline model (center) exhibits severe counting and spatial failures across all examples. In Rows 1, 2, and 3, despite the prompt explicitly requesting **four** distinct individuals (Supermarket Employees, Chefs, and Sailors), the baseline consistently generates only **three** entities, effectively merging two of the subjects. Consequently, the requested paired interactions (e.g., "Whispering" vs. "Shaking hands" in Row 3) become conflated or physically impossible. In contrast, our proposed CoIG framework (right) successfully decomposes the task, utilizing a step-by-step generation process to strictly enforce the creation of all four distinct entities and faithfully render their specific, paired interactions.

G. Additional Qualitative Results on Causal Relevance

Additional qualitative results are shown in Figure 12.

H. Human Evaluation Protocol

Overview. We developed a web-based annotation tool to collect human judgments on the monitorability of CoIG generation sequences. Annotators independently evaluated 50 image sequences from the Color subset of T2I-CompBench across two task blocks, answering a total of three questions per sequence. Each annotator's screens were randomly shuffled using a seeded permutation based on their identifier, ensuring different presentation orders while maintaining reproducibility.

Annotators. We recruited twelve annotators with varying academic and professional backgrounds. None of the annotators are authors of this work, and none were informed of the expected outcomes or the purpose of the study beyond the task instructions.

Stimuli. Each stimulus consists of a 6-step image generation sequence produced by the CoIG pipeline, accompanied by the source text prompt (e.g., "a yellow apple and red bananas"). The first step begins with a gray placeholder; visual attributes such as color are progressively added at subsequent steps. We selected 50 prompts, each containing at least two explicitly colored objects. For the error detection block, we generated perturbed variants by swapping one object's color at a specific step (e.g., replacing "red" with "green" from step 4 onward), producing sequences with a known ground-truth error location.

Block 1: Readability and Causal Relevance (50 screens). For each screen, annotators were shown the text prompt and all six step images side by side. They answered two questions:

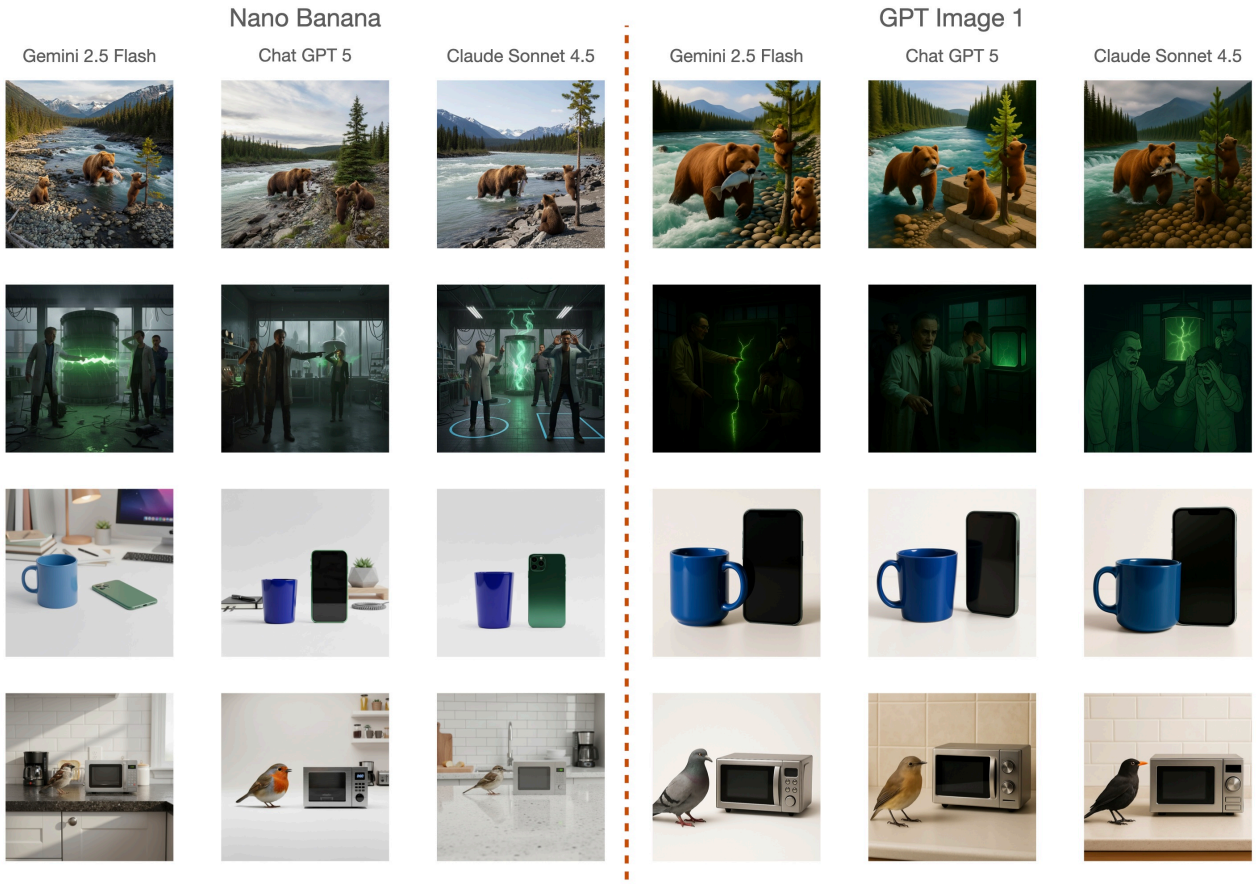


Figure 11. **Qualitative demonstration of architectural versatility.** We present final generation outputs from our CoIG framework using different backbone combinations. The figure is organized into two main sections: the **first three columns** utilize the Nano Banana image model (Fortin et al., 2025), while the **last three columns** utilize the GPT Image 1 model (OpenAI, 2025b). Within each section, we compare the performance of three driving LLMs: Gemini 2.5 Flash (Comanici et al., 2025), Chat GPT 5 (OpenAI, 2025a), and Claude Sonnet 4.5 (Anthropic, 2025). Each row represents a unique text prompt. The consistent high-quality results across this grid demonstrate that our method effectively generalizes across different model architectures.

- **Q1 (Readability):** "At which step does the [object] first get its color [color]?" — answered by clicking one of the six step images.
- **Q2 (Causal Relevance):** "Does [color] persist to the final image?" — answered Yes or No.

Q1 measures whether a human observer can identify when a specific color attribute first becomes visually readable in the generation chain. Q2 measures causal relevance: whether the color, once introduced, persists through subsequent steps to the final output, confirming that the intermediate step causally influences the result rather than being overwritten.

Block 2: Error Detection (50 screens). Each screen presented a perturbed sequence where one object's color was swapped to an incorrect color starting at a known step. Annotators were instructed that early steps may contain gray placeholders without color and to focus only on steps where color is visible. They answered one question:

- **Q3 (Error Detection):** "At which step does the [object] first appear in the wrong color?" — answered by clicking one of the six step images.

This measures whether a human can localize where an error was introduced in the generation chain, a prerequisite for any intervention or correction mechanism.

Practice Screens. Before each block, annotators completed practice screens (one for Block 1, two for Block 2) that were clearly marked as non-scoring to familiarize them with the interface and question format.

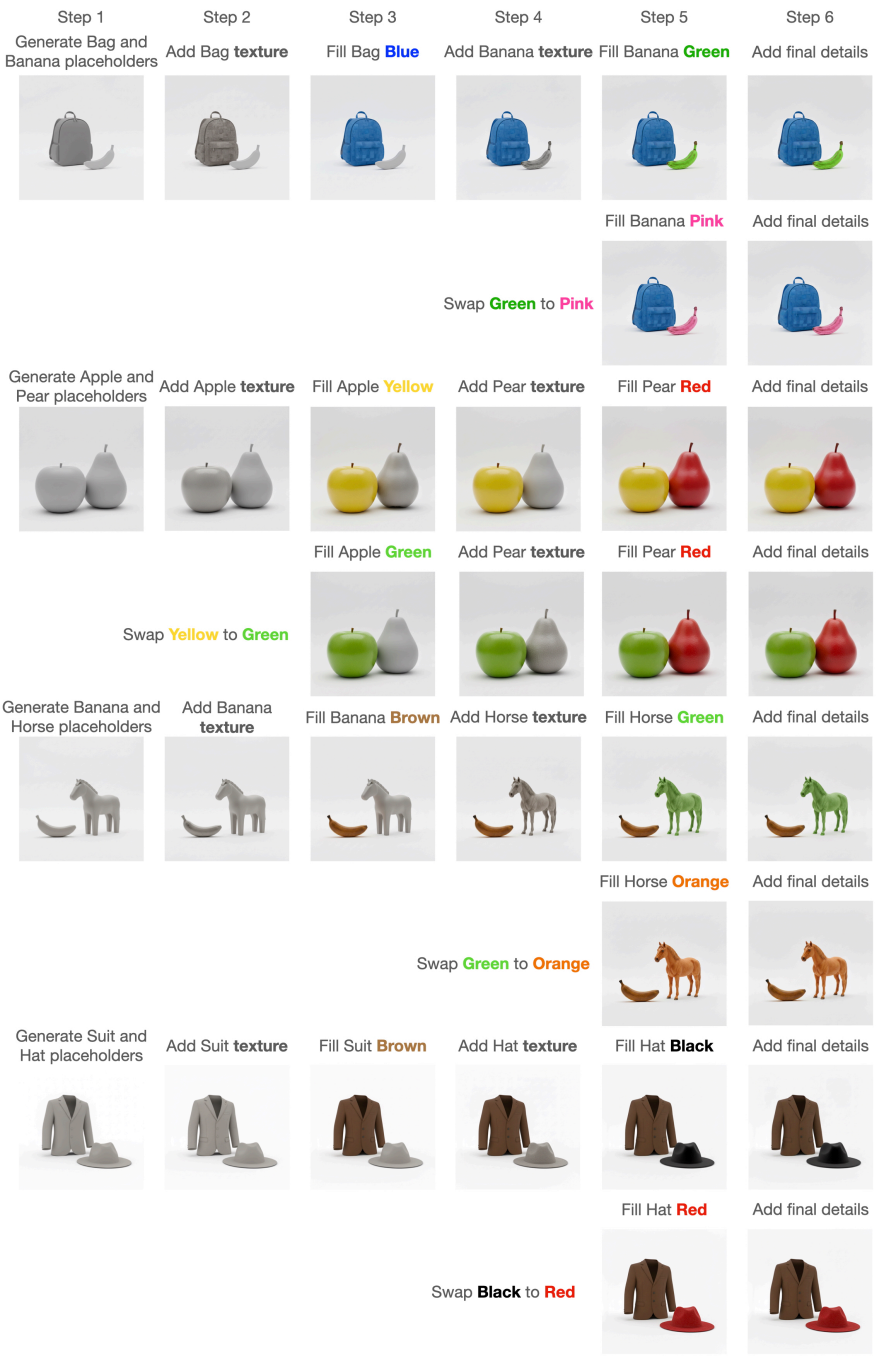
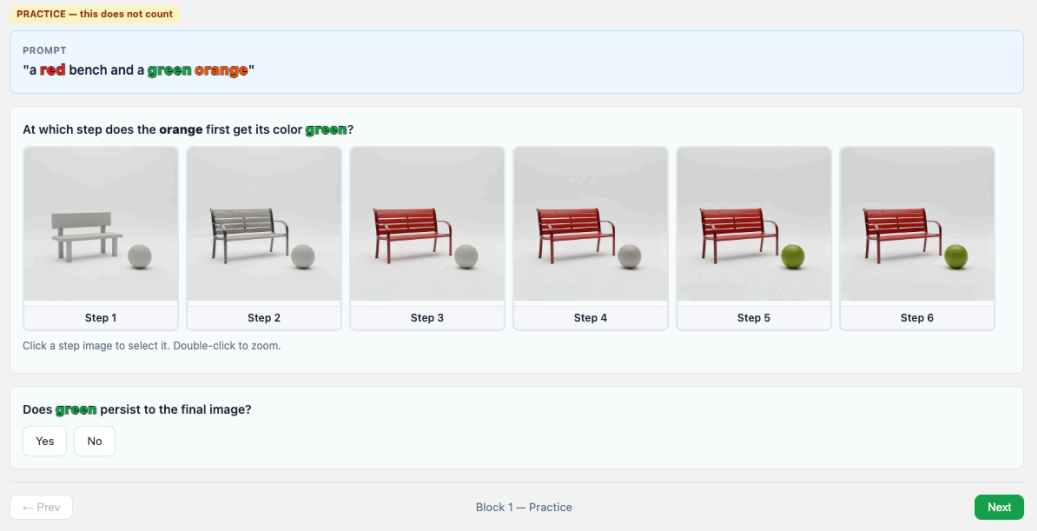


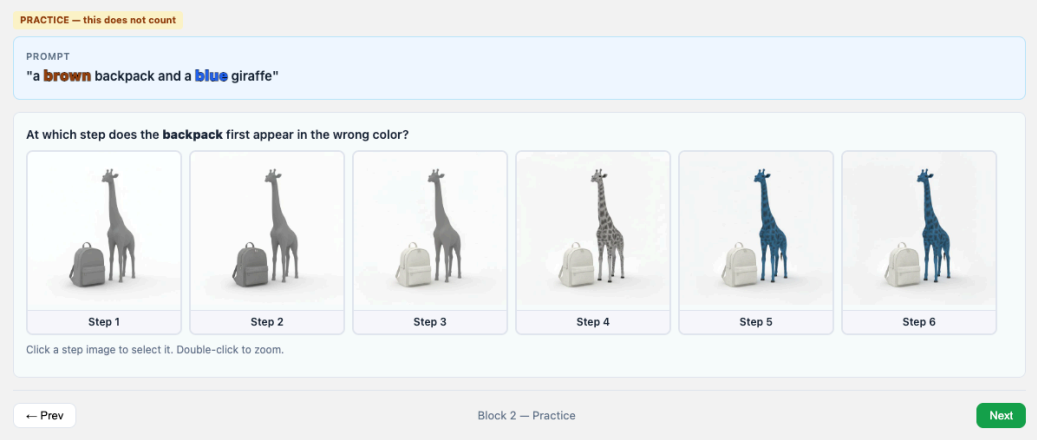
Figure 12. Qualitative demonstration of Causal Relevance. The odd rows (1, 3, 5, and 7) illustrate the final images generated from four distinct original CoIG sequences. The even rows (2, 4, 6, and 8) display the final images resulting from single perturbations applied during intermediate steps within those sequences. For instance, Row 2 shows the banana color changed to pink, Row 4 shows the apple changed to green, Row 6 shows the horse changed to orange, and Row 8 shows the hat changed to red. The consistent preservation of these targeted changes confirms the strong causal link between intermediate steps and the final output.

Interface Details. The annotation tool displayed all six step images in a horizontal row (Figure 13). Annotators selected a step by clicking its image; double-clicking allowed zooming for closer inspection. A progress bar tracked completion within each block. Upon completion of both blocks, annotators exported their responses as a JSON file.

Quality Control. Presentation order was randomized per annotator to mitigate ordering effects. Each annotator worked



(a) Block 1: Readability and Causal Relevance. Annotators identify when a target color first appears and whether it persists to the final image.



(b) Block 2: Error Detection. Annotators localize the step at which an injected color error first appears.

Figure 13. Screenshots of the annotation interface. Six step images are displayed horizontally; annotators respond by clicking the relevant step.

independently without access to others' responses. We computed inter-annotator agreement using Cohen's κ across all annotator pairs to verify annotation reliability.