FormalImG: Evaluating Structural Compositional Generalization for T2I Models

Anonymous ECCV 2026 Submission

Paper ID #3409

Abstract. As natural language becomes the primary interface for image generation, evaluating semantic generalization under language instructions is increasingly important. Existing benchmarks emphasize combinations of concepts but rarely examine the internal semantic structure of language. We introduce FormalImG, a first-order-logic-based benchmark for structural compositional generalization. Natural language instructions are formalized as logical formulas and we define structural compositional complexity and ε-structural compositional generalizability to measure how model performance changes with increasing semantic dependency. The benchmark includes two evaluation scenarios and 4,000 instructions across multiple complexity levels, assessed through symbolic verification and model-as-judge. Experiments show that mainstream text-to-image models experience clear performance decline as structural complexity grows, with stable performance mainly at low complexity levels. Further analysis indicates that large language models already handle textual structural reasoning well, while the language-to-vision transformation stage forms the significant bottleneck. Intermediate layout representations can partially mitigate this issue.

Keywords: Text-to-Image · Compositional Generalization · Benchmark

1 Introduction

A long-standing goal of artificial intelligence is to build reliable and general foundation models that assist humans across real-world tasks. Recent advances in large (vision-)language models have accelerated this progress, as natural language instructions now serve as a universal interface for model control. This shift in the control paradigm has profoundly influenced multiple domains. In robotics, control has evolved from manually engineered command sequences [2] to goal-driven natural language instructions [44]. In programming, assistance has progressed from context-based code completion [34] to conversational code generation [9]. In recommender systems, preference modeling has shifted from implicit signals such as clicks [18] to explicit requirement expression via natural language [30]. Within this landscape, text-to-image foundation models have emerged as a representative application with a broad user base, with commercial platforms reporting over 22 billion generated images and assets worldwide [1]. These models enable users to transform visual concepts into images using only

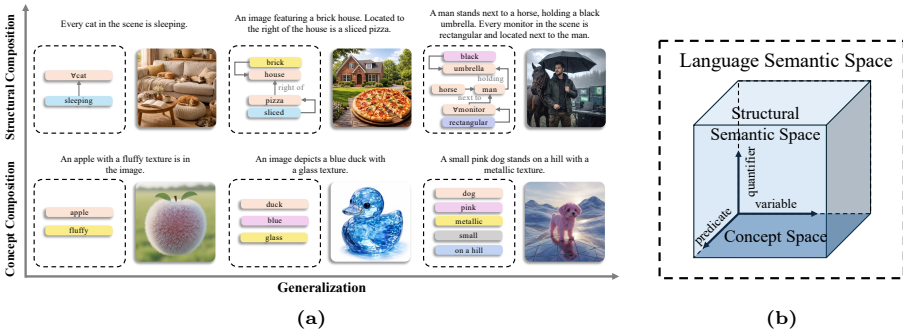


Fig. 1: From concept to structural compositional generalization in language-driven text-to-image models. (a) The evolution of generalization in text-to-image models. (b) The structure of the generalization space. The structural semantic space is spanned by variables, predicates, and quantifiers composed through logical connectives, while the concept space can be viewed as a slice within it.

language instructions, substantially lowering the barrier to content creation. Early models mainly support short or bag-of-words prompts composed of low-level elements such as objects, colors, and textures [27, 33, 35]. As user demands become more complex, recent models can process full language instructions and capture higher-level semantics, including relationships, attribute binding, and logical structure. As text-to-image systems evolve toward stronger language intelligence, understanding their generalization within a semantic space of combinatorial semantic primitives has become a critical research question.

Early text-to-image benchmarks, such as GenEval [10] and T2I-CompBench [20], evaluate basic linguistic capabilities, including rendering a single object, attribute binding, and simple spatial relations. With the rapid progress of text-to-image systems, some modern models have achieved high overall performance on these benchmarks [8, 40], suggesting that fundamental semantic primitives are largely well mastered. More recent work has begun to examine compositional generalization over these primitives. For example, ConceptMix [41] evaluates generalization to unseen combinations of visual concepts by progressively increasing conceptual diversity, such as styles, colors, and object categories. However, ConceptMix mainly emphasizes the diversity and combination of visual concepts, without systematically characterizing the semantic structure underlying language instructions. In particular, existing benchmarks do not explicitly model how objects, attributes, and relations form a structural semantic space through logical structure, nor do they examine how model performance changes as the dependencies among these elements become more complex.

Motivated by this goal, we introduce FormalImG, a benchmark designed to systematically study the structural compositional generalization of text-to-image models within a structural semantic space. The core idea is to explicitly construct a structural semantic space for language instructions and evaluate model performance within this space. To this end, FormalImG adopts first-order logic (FOL)

as a unified semantic representation, formalizing natural language instructions into logical expressions composed of objects, attributes, unary and binary relations, and quantifiers. The use of FOL provides clear expressive boundaries and strong compositional structure, enabling explicit characterization of the semantic complexity of language instructions and analysis of how model performance changes as semantic complexity increases.

Our main contributions are summarized as follows. **First**, we introduce FormalImG, a FOL-based benchmark for studying the semantic generalization of text-to-image models. FormalImG contains 4,000 test samples with progressively increasing complexity, providing a solid testbed for systematic analysis of semantic generalization. **Second**, we formally model the semantic structure of language instructions based on FOL, and define structural compositional complexity and ϵ-structural compositional generalizability, thereby providing a computable tool for analyzing the structural compositional generalization of language-driven generative models. **Third**, to balance evaluation rigor and generality, we construct a dual-scenario evaluation protocol consisting of Knolling and Natural settings. The Knolling scenario provides a style-controlled environment, where generated images are symbolized and substituted into logical formulas for formal verification. The Natural scenario serves as a general text-to-image setting, assessing semantic generalization under more realistic and diverse language instructions. **Fourth**, we reveal a hierarchical bottleneck in the structural generalization of current models. Our experiments show that performance consistently degrades as complexity increases. For advanced models, the primary limitation lies in the transformation from language to visual, whereas weaker models exhibit bottlenecks already at the stage of semantic understanding.

2 Related Work

Language-Driven Text-to-Image Models. Text-to-image models aim to generate visual content that is semantically consistent with user-provided textual descriptions. Early approaches were primarily based on generative adversarial networks (GANs) [11, 26] or conditional variational autoencoders (CVAEs) [22, 38]. In this stage, the control interface was relatively limited, often relying on category labels or simple conditional signals. With the emergence of diffusion models [17] and the introduction of multimodal encoders such as CLIP [32], modern text-to-image systems, including GLIDE [27], Stable Diffusion [35], and DALL·E 2 [33], began to leverage text embeddings for conditional generation. These models support short or bag-of-words prompts composed of multiple keywords, enabling more flexible semantic control. More recently, the rapid progress of large language models (LLM) and vision–language models (VLM) has led to the adoption of more powerful text encoders in text-to-image frameworks [5, 36, 40], allowing models to process full natural language descriptions. Natural language has thus evolved from a simple prompting signal into a general-purpose control interface, through which users can impose fine-grained constraints on objects, attributes, relationships, and global scene composition. Meanwhile, several

studies have explored unified architectures [8, 25, 42] that jointly model text understanding and image generation within a single framework, further strengthening language-driven generation capabilities. Overall, the evolution of text-to-image models indicates a clear trend: natural language has become the primary interface for controlling visual generation.

Evaluation of Text-to-Image Models. Early evaluation of text-to-image models mainly relied on image quality metrics such as FID [16], IS [37], and text–image similarity measures [15]. These metrics primarily measure visual quality or coarse text–image matching. They provide only limited evidence about whether a model follows specific linguistic constraints. As text-to-image models improved, researchers introduced dedicated benchmarks for more fine-grained evaluation. One line of work examines basic semantic capabilities, such as generating single objects, binding attributes, satisfying simple spatial relations, and handling counting tasks [10, 19, 20]. These approaches break textual meaning into individual semantic elements and evaluate models on these basic units. However, such evaluations usually focus on isolated elements or fixed templates, and do not systematically test how these elements are combined in open-ended natural language. More recent studies investigate compositional generalization in text-to-image generation [21, 23, 41]. For example, ConceptMix [41] evaluates generalization to unseen combinations by controlling the number of visual concepts. Despite this progress, existing benchmarks mainly emphasize specific skills or novel visual combinations, while paying limited attention to the linguistic structure of the instructions. As natural language becomes the primary interface for controlling text-to-image systems, an important question is whether models can consistently satisfy semantic constraints under diverse expressions and structural compositions. Generalization in the language space therefore remains an important yet insufficiently studied problem for text-to-image models.

3 FormalImG

This section introduces the construction of FormalImG. Our core idea is to define a structural semantic space grounded in first-order logic, and to construct a set of language instructions whose semantics can be mapped to a single first-order logic formula in this space. This formal representation is then used for subsequent generalization analysis and evaluation.

3.1 Structural Semantic Space

Natural language instructions could be formulated as structural expressions generated through recursive composition of vocabulary under grammatical constraints. The vocabulary forms a discrete set of symbols, while grammar provides compositional rules that allow sentences to expand through structural combination. At the semantic level, such recursive composition gives rise to representations of entities, their attributes, and relations between entities, together with the logical structure among these elements. From a structural perspective, the

space of language instructions therefore constitutes a structural semantic space. We formally define the structural semantic space as follows.

Definition 1 (Structural Semantic Space). Let $\mathcal{O}$ denote the set of object symbols, $\mathcal{A}$ the set of attribute symbols, $\mathcal{R}$ the set of relation symbols, and $\mathcal{S}$ the set of semantic operators. Define the set of atomic semantic units as $\mathcal{P} = \{a(o) \mid o \in \mathcal{O},\ a \in \mathcal{A}\} \cup \{r(o_1, o_2) \mid o_1, o_2 \in \mathcal{O},\ r \in \mathcal{R}\}$. The structural semantic space is then defined as $\mathcal{E} = \text{Closure}_{\mathcal{S}}(\mathcal{P})$.

The structural semantic space consists of all well-formed structural expressions that are recursively generated from atomic semantic units under the compositional rules induced by the semantic operators.

The mapping from natural language sentences to structural expressions is generally not unique. Vocabulary choices are open-ended, expressions admit diverse formulations, and quantifier scope as well as coreference resolution may yield multiple interpretative paths. This flexibility makes structural semantic analysis and verification challenging in the full natural language space.

To enable formal analysis of semantic structures in language instructions, we conduct our study in an unambiguous semantic subspace. In this subspace, entities, relations, and scopes are explicitly specified, and semantic composition follows deterministic rules. We adopt FOL as a formal surrogate of this structural semantic space. FOL provides explicit variable binding mechanisms, well-defined quantifier scope, and decidable satisfiability, allowing semantic structures of language instructions to be precisely characterized and supporting subsequent executable verification and complexity analysis.

A first-order formula ϕ consists of variables, predicates, quantifiers, and logical connectives. Variables represent objects; unary predicates represent attributes; binary predicates represent relations; quantifiers specify variable scope; and logical connectives determine the overall logical structure. This yields a compositional and unambiguous semantic representation, enabling explicit analysis of model behavior under different semantic structures. In practice, we instantiate first-order logic in the visual domain via a Domain Specific Language (DSL). The DSL defines a finite vocabulary of objects, attributes, and predicates, providing a closed and controllable platform for our study.

3.2 Structural Compositional Generalization

To investigate model generalization within the structural semantic space, we need to examine model performance across different levels of structural complexity. To this end, we introduce a computable measure of structural compositional complexity. Due to the combinatorial explosion inherent in this space, it is infeasible to enumerate all possible structures. Following prior work [41], we therefore conduct analysis along a monotonic and extensible dimension. To capture the reasoning difficulty induced by complex structures, we characterize structural compositional complexity in terms of variable dependency scale. Based on this, we provide the following definition.

Definition 2 (Structural Compositional Complexity). Let ϕ be a closed FOL formula in the semantic subspace. According to variable-sharing relations, ϕ is decomposed into connected components with pairwise disjoint variable sets, $\phi = \bigwedge_{i=1}^{N} \phi_i$. Let $\mathrm{Var}(\phi_i)$ denote the set of quantified variables appearing in component ϕ_i. The structural compositional complexity of ϕ is defined as

$$K(\phi) = \max_i |\mathrm{Var}(\phi_i)|. \tag{1}$$

This measure captures the maximal dependency scale among variables in the formula. If multiple atomic clauses share quantified variables, their truth values must be jointly evaluated within the same variable assignment space, thereby forming a semantically coupled component. When $K(\phi) = 1$, substructures associated with different variables are independent, and reasoning decomposes into single-entity judgments. When $K(\phi) > 1$, variables are connected through binary predicates, and the model must perform multi-entity joint reasoning. As complexity increases, the number of variables that must be coordinated simultaneously grows, and the difficulty of reasoning increases accordingly.

During dataset construction, we further constrain each instruction to contain only a single connected component after decomposition. This design ensures that all quantified variables participate in a unified dependency structure, so that $K(\phi)$ faithfully reflects the depth of coupled relational reasoning rather than a simple aggregation of independent subtasks.

Structural compositional complexity characterizes the scale of joint reasoning required by a formula. However, it does not by itself describe model behavior. To study model performance in the structured semantic space, we examine how accuracy varies with structural complexity. An ideal compositional reasoning model should maintain stable performance as structural complexity increases. We therefore ask whether the model can sustain acceptable performance when $K(\phi)$ grows to higher levels. We refer to this cross-level capability as structural compositional generalization.

Definition 3 (ε-Structural Compositional Generalizability). Let Φ denote the set of all FOL instruction formulas. For any integer $k \in \mathbb{N}$, define the complexity level $\Phi_k = \{\phi \in \Phi \mid K(\phi) = k\}$. Let x_ϕ denote the natural language instruction corresponding to formula ϕ. Let $S(y, \phi) \in \{0, 1\}$ be a semantic satisfaction function indicating whether a generated output y satisfies ϕ. For a model f, let $SR_k(f) = \mathbb{E}_{\phi \sim \mathcal{U}(\Phi_k), y \sim f(\cdot|x_\phi)}[S(y, \phi)]$ denote the average satisfaction rate of f at complexity level k. Given a performance threshold $\varepsilon \in (0, 1)$, the ε-structural compositional generalizability of f is defined as

$$G_\varepsilon(f) = \max\{K \mid SR_k(f) \geq \varepsilon,\ \forall k \leq K\}. \tag{2}$$

The ε-structural compositional generalizability measures the range of structural compositional complexity levels over which a model can stably maintain acceptable performance. The condition $SR_k(f) \geq \varepsilon$ for all $k \leq K$ requires that the model achieve a satisfaction rate above the threshold at every complexity level up to K. Consequently, $G_\varepsilon(f)$ represents the largest continuous structural compositional complexity level that the model can reliably generalize.

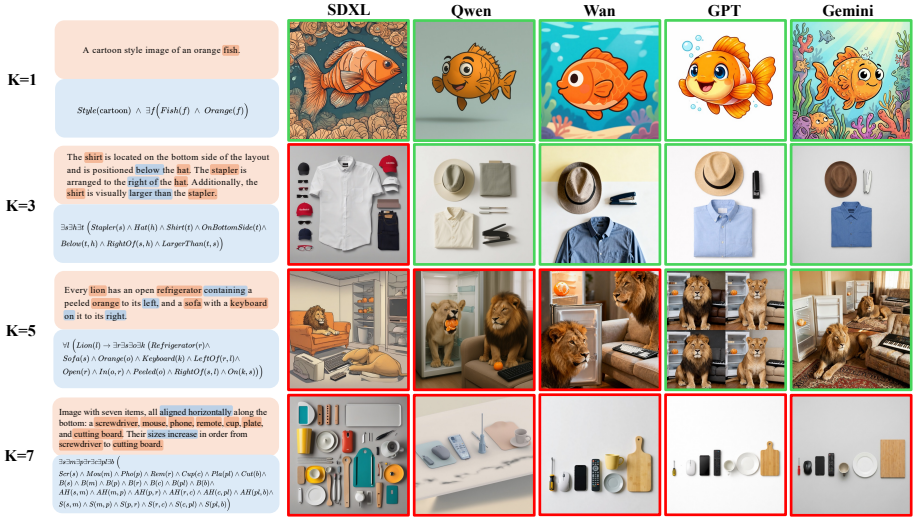


Fig. 2: Generation examples of models under instructions with increasing complexity. Green borders denote images that satisfy the instruction, while red denote failures.

3.3 Benchmark Construction

Inspired by [41], we generate data through a controlled LLM-based synthesis pipeline with both syntax and semantic checking.

Instead of directly sampling logical forms, we first sample a generation configuration in DSL domain that defines the target properties of each instruction, such as the desired K, object categories, quantifier structure and predicates. This configuration serves as a structural prior of an instruction.

Conditioned on this configuration, an LLM jointly generates a natural language instruction and its corresponding DSL expression. Generating both representations within a single semantic context encourages globally coherent descriptions and avoids the unnatural or contradictory structures that often arise from programmatic enumeration of logical forms.

The generated data are then subjected to formal verification implemented in Python. The DSL is checked for syntactic validity, variable scope consistency, domain compliance, and correct complexity assignment. An additional LLM-based semantic check detects potential semantic conflicts and ensures alignment between the natural language instructions and their corresponding DSL representations. Samples failing any verification step are discarded and regenerated.

To ensure both experimental rigor and broader applicability, we construct two complementary scenarios: Knolling and Natural. The Knolling scenario features structured object arrangements with visual grounding, facilitating symbolic abstraction and formal verification. The Natural scenario reflects general text-to-image generation, allowing us to examine structural compositional generalization under more open semantic conditions. Examples are shown in Fig. 2.

3.4 Evaluation

We adopt scenario-specific evaluation strategies. The Knolling setting is evaluated via formal symbolic verification, whereas the Natural setting relies on model-as-judge for assessment.

Knolling. Let y denote a generated image and ϕ the target DSL instruction. A symbolic scene representation $S(y)$ is constructed through a hybrid parsing process: object categories and attributes are recognized using Qwen3-VL-8B [3], a VLM with strong open-vocabulary object grounding capability, while geometric properties such as positions, sizes, and spatial relations are computed from detected object layouts. The correctness of the generation is evaluated by logical satisfaction, defined as $\text{Score}(y, \phi) = 1$ if $S(y) \models \phi$, and 0 otherwise. This evaluation provides an executable and fine-grained measure of whether the generated scene satisfies the required semantic constraints.

Natural. For natural image evaluation, explicit symbolic parsing is not available. We therefore employ Gemini-3-Pro [14] as an automated judge. The model takes the generated image together with the natural language instruction and the corresponding DSL instruction, and produces a binary decision indicating whether the instruction is satisfied. The DSL instruction is provided as a reference to reduce ambiguity in the natural description.

4 Experiments and Key Observations

4.1 Experimental Setup

Benchmark Composition. FormalImG consists of two evaluation scenarios: Knolling and Natural. Each scenario contains 2,000 test samples, covering complexity levels from $K = 1$ to $K = 10$. Each complexity level includes 200 samples, resulting in a balanced distribution across different values of K. We further group complexity levels into three subsets: Easy ($K = 1$ to 3), Medium ($K = 4$ to 6), and Hard ($K = 7$ to 10).

Model Selection. We evaluate multiple text-to-image models, including both advanced close-source models (e.g., GPT-Image-1.5 [29] and Gemini-3-Pro-Image [13]) and well-validated open-source models (e.g., SDXL [31] and Qwen-Image [40]). For the Knolling setting, some models with limited context capacity may not reliably follow long style descriptions. For these models, we use a simplified style prompt to ensure stable generation.

Metric. Following the evaluation method described in Sec. 3.4, we compute the average instruction satisfaction rate. Results are reported separately for each complexity level K, along with the overall average success rate. In addition, under the threshold $\varepsilon = 0.7$, we compute and report the ε-structural compositional generalizability $G_{0.7}$ for each model.

Table 1: Instruction satisfaction rates of text to image models on FormalImG across two evaluation scenarios and K. E, M and H denote the average performance on Easy, Medium and Hard subset. Overall denote the average performance across all levels.

Model	K										E	M	H	O	$G_{0.7}$
	1	2	3	4	5	6	7	8	9	10					
Natural															
Gemini-3-Pro-Image [13]	92.5	78.5	**76.0**	**79.5**	**78.5**	**74.0**	**68.0**	**76.0**	**71.5**	**65.5**	82.3	**77.3**	**70.2**	**76.0**	**6**
Gemini-2.5-Flash-Image [12]	90.0	63.5	54.0	54.5	61.5	49.0	38.0	39.5	42.0	28.5	69.2	55.0	37.0	52.1	1
GPT-Image-1 [28]	95.0	81.0	71.0	76.5	57.0	59.0	61.5	54.5	51.0	47.5	82.3	64.2	53.6	65.4	4
GPT-Image-1.5 [29]	**96.0**	**84.0**	73.5	73.5	63.5	63.5	62.0	59.5	52.5	53.0	**84.5**	66.8	56.8	68.1	4
Seedream-4.5 [7]	95.5	83.5	55.0	67.0	58.0	48.5	41.5	44.5	35.5	30.0	78.0	57.8	37.9	55.9	2
Wan-2.6-T2I [39]	92.5	72.0	58.5	54.0	45.0	45.0	36.5	41.0	30.0	34.0	74.3	48.0	35.4	50.9	2
SD1.5 [35]	55.5	10.0	2.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	22.5	0.2	0.0	6.8	0
SDXL [31]	72.5	26.5	7.5	2.0	1.0	0.5	0.0	0.0	0.0	0.0	35.5	1.2	0.0	11.0	1
Flux1.dev [5]	73.0	47.5	29.5	23.5	18.5	13.5	10.5	9.0	4.0	1.5	50.0	18.5	6.2	23.1	1
Qwen-Image [40]	90.0	66.5	43.0	38.0	35.5	26.5	17.0	22.0	13.0	11.0	66.5	33.3	15.8	36.3	1
Hunyuan-3.0 [6]	77.5	58.5	40.5	35.0	32.0	23.5	15.0	18.0	14.0	9.5	58.8	30.2	14.1	32.4	1
BAGEL [8]	85.0	49.0	30.0	22.0	26.5	14.5	6.5	9.0	7.0	1.5	54.7	21.0	6.0	25.1	1
Knolling															
Gemini-3-Pro-Image [13]	93.0	**90.0**	**80.0**	**70.0**	**66.5**	**63.5**	**58.5**	**49.0**	**53.0**	**47.0**	**87.7**	**66.7**	**51.9**	**67.1**	**4**
Gemini-2.5-Flash-Image [12]	**94.5**	73.0	57.5	53.5	36.5	35.0	23.5	18.5	18.5	12.5	75.0	41.7	18.2	42.3	2
GPT-Image-1 [28]	94.0	71.0	59.5	49.5	45.5	37.5	36.5	24.0	20.5	16.0	74.8	44.2	24.2	45.4	2
GPT-Image-1.5 [29]	92.0	75.5	62.5	52.0	51.5	48.0	42.0	30.0	28.0	20.0	76.7	50.5	30.0	50.2	2
Seedream-4.5 [7]	85.0	65.0	48.5	34.0	33.5	26.5	21.0	17.0	12.0	12.5	66.2	31.3	15.6	35.5	1
Wan-2.6-T2I [39]	63.0	53.0	39.5	30.0	30.0	28.0	24.0	14.5	18.0	12.0	51.8	29.3	17.1	31.2	0
SD1.5 [35]	43.5	17.0	4.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	21.5	0.2	0.0	6.5	0
SDXL [31]	50.0	31.5	7.5	1.0	0.5	0.0	0.0	0.0	0.0	0.0	29.7	0.5	0.0	9.1	0
Flux1.dev [5]	56.5	35.0	16.0	9.0	2.5	3.5	1.5	0.0	0.5	0.0	35.8	5.0	0.5	12.5	0
Qwen-Image [40]	65.0	39.0	27.0	23.0	13.0	7.5	8.0	4.0	4.5	1.5	43.7	14.5	4.5	19.3	0
Hunyuan-3.0 [6]	49.5	35.0	20.0	12.5	5.0	3.5	3.5	3.0	3.5	0.0	34.8	7.0	2.5	13.6	0
BAGEL [8]	38.0	32.0	5.0	1.5	1.0	0.0	0.5	0.0	0.0	0.0	25.0	0.8	0.1	7.8	0

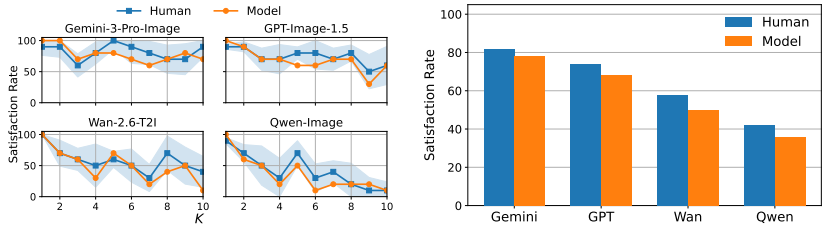
4.2 Limited Structural Compositional Generalization

Tab. 1 shows that in both the Natural and Knolling evaluation scenarios, the instruction satisfaction rate consistently decreases as K increases.

As complexity grows, a clear performance gap gradually emerges between close-source and open-source models. Although some open-source models remain competitive at low K, their performance declines rapidly as K increases. In the high-K regime, several open-source models exhibit near-zero satisfaction rates. In contrast, close-source models demonstrate a more gradual degradation trend and maintain non-trivial performance across a broader range of complexity levels.

Within the close-source model group, performance differences also become more pronounced as K increases. At low K, models perform comparably; however, the gap widens significantly in the high-K regime. Gemini-3-Pro-Image maintains a higher instruction satisfaction rate and exhibits a relatively smoother decline, yet its performance still decreases steadily as complexity increases.

From the perspective of structural compositional generalizability $G_{0.7}$, most models achieve stable generalization only at low structural complexity levels. Among close-source models, the majority can reliably generalize only up to $K = 2$. Even the strongest model, Gemini-3-Pro-Image, generalizes only up to $K = 6$ in the Natural scenario and $K = 4$ in the Knolling scenario. For open-source



(a) Satisfaction rates across K. **(b)** Ranking consistency across models.

Fig. 3: Comparison between model-based judgments and human evaluation.

models, none satisfies the generalization threshold at $K = 2$. Some models fail to generalize even in single-object generation tasks involving global constraints and multiple attribute bindings.

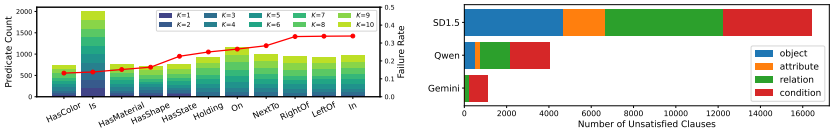
These findings indicate that current text-to-image models still lack robust structural compositional generalization when handling language instructions, highlighting a substantial gap that remains to be addressed.

4.3 Human Evaluation and Reliability of Model-as-Judge

To assess the reliability of automatic evaluation in the Natural scenario, we conduct a human evaluation on a stratified subset of the benchmark, following the scale adopted in prior work [41]. We randomly sample 100 instructions (10 per complexity level $K \in [1, 10]$) and collect generated images from four text-to-image models, yielding 400 image–instruction pairs. Forty volunteers participate in the evaluation, each rating 50 samples, resulting in 2,000 ratings in total. Each image is evaluated by five annotators on a five-point Likert scale for semantic compliance, and the final score is the average rating. a sample is considered to satisfy the language instruction if its mean score is at least 4.

Human ratings show high internal consistency (Cronbach's $\alpha = 0.882$). Comparing binarized human judgments with the model-as-judge outputs yields an accuracy of 0.805 and an F1 score of 0.840. Ranking consistency is measured by the Spearman correlation between human and automatic model rankings across K, yielding a mean correlation of 0.844. Fig. 3 shows that automatic satisfaction rates closely follow human trends across K. In most cases, the automatic results fall within the variability caused by inter-annotator disagreement. Overall model rankings from automatic evaluation exactly match those from human evaluation, although automatic scores are slightly more conservative.

These findings demonstrate that semantic compliance in FormalImG can be consistently assessed by human evaluators, and that the model-as-judge framework provides a reliable and scalable alternative to human evaluation.



(a) Predicates count and failure rate. **(b)** Unsatisfied count across models.

Fig. 4: Failure analysis across predicates and clause types.

4.4 Clause-Level Analysis of Compositional Failures

To further understand the causes of failure in structural compositional generalization for text-to-image models, we conduct fine-grained statistical analysis in the Natural scenario, which contains richer vocabulary of objects, attributes, and relations. We first convert each DSL expression into conjunctive normal form, representing the semantic specification as a conjunction of multiple clauses. During the model-as-judge evaluation process, the judge model is required to return the list of clauses it considers satisfied. We then categorize these clauses into four semantic groups according to their functional roles: object, attribute, relation, and condition. Based on this categorization, we compute the frequency of major predicates across different K levels, the failure rates associated with each predicate, and the proportion of failures across clause categories for different models. The results are shown in Fig. 4.

Fig. 4a shows that the distribution of predicates across K is overall relatively balanced. Apart from the `Is` predicate, which appears in all instructions, the remaining predicates also occur at comparable frequencies. In terms of failure rates, models perform relatively stably on unary predicates that describe objects and attributes, whereas binary predicates that express relations exhibit noticeably higher failure rates. Fig. 4b further illustrates the differences in failure patterns across models. For the weaker model SD1.5, failures occur frequently in both object and relation clauses, indicating limited capability in handling basic semantic atoms. For the stronger model Gemini-3-Pro-Image, the failure proportions for object, attribute, and relation clauses decrease substantially, with errors primarily concentrated in condition clauses. The mid-tier model Qwen-Image exhibits behavior between these two extremes. These observations suggest that current advanced text-to-image models have become relatively reliable in attribute-level semantic control, yet still face clear bottlenecks when handling complex structural compositions involving multiple interacting constraints.

4.5 Modality-Level Diagnosis of Structural Generalization Failure

In Sec. 4.2, we observe substantial performance degradation in text-to-image models as structural compositional complexity increases. A key question is if this degradation arises from insufficient semantic understanding or from failures during language-to-vision generation. To localize the bottleneck, we conduct diagnostic analysis at both the language level and the image generation level.

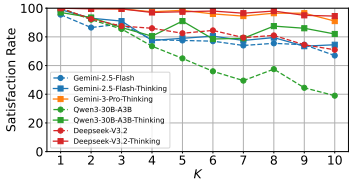


Fig. 5: Performance of LLMs on the Knolling text task across K.

Table 2: Performance summary of LLMs on the Knolling text task.

Model	E	M	H	O	$G_{0.7}$
Gemini-2.5-Flash [12]	90.3	77.5	72.8	79.5	9
Gemini-2.5-Flash-Thinking [12]	94.3	79.0	76.2	82.5	**10**
Gemini-3-Pro-Thinking [14]	**99.8**	97.3	94.6	97.0	**10**
Qwen3-30B-A3B [43]	92.5	64.8	47.6	66.3	4
Qwen3-30B-A3B-Thinking [43]	92.3	83.3	83.6	86.2	**10**
Deepseek-V3.2 [24]	93.3	84.3	76.5	83.9	**10**
Deepseek-V3.2-Thinking [24]	99.7	**97.5**	**96.0**	**97.6**	**10**

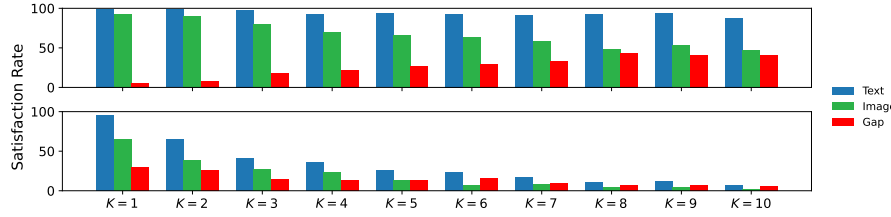


Fig. 6: Comparison of text and image setting across K. The upper plot shows **Gemini-3-Pro-Image**, and the lower plot shows **Qwen-Image**.

We first evaluate LLMs in a pure text setting under the Knolling scenario. Given a natural language instruction, the model outputs a symbolic layout specification including object categories, attributes, and bounding boxes, which is verified using the same DSL interpreter as in the image-based evaluation. As shown in Fig. 5 and Tab. 2, modern language models maintain high satisfaction rates across all complexity levels. Gemini-3-Pro-Thinking and DeepSeek-V3.2-Thinking achieve overall scores of 97.0% and 97.6%, respectively, while sustaining over 90% accuracy even at high K, with both reaching $G_{0.7} = 10$. Reasoning variants consistently outperform their non-reasoning counterparts, particularly at higher complexity. For example, Qwen3-30B-A3B improves from $G_{0.7} = 4$ to $G_{0.7} = 10$ when reasoning is enabled. These results indicate that structural compositional generalization is largely preserved in the language domain.

In contrast, the best text-to-image model achieves only 67.1% overall performance and degrades rapidly with increasing K. To identify the primary source of error, we conduct controlled comparisons on Gemini-3-Pro-Image and Qwen-Image under both text-only and full generation settings. For Gemini-3-Pro-Image we use its text-output mode, while for Qwen-Image we evaluate its language backbone Qwen2.5-VL-7B [4]. As shown in Fig. 6, Gemini-3-Pro-Image maintains nearly 90% performance in text mode, whereas its image generation performance declines sharply with complexity, and the gap between the two modes widens as K increases. This indicates that instruction understanding remains robust while the main limitation arises during language-to-vision generation. In contrast, Qwen-Image exhibits a large gap between text and image outputs even

Table 3: Effect of layout representation and query setting across K. Text and image layouts are compared under both *With Query* and *Without Query* conditions.

Layout	Query	Model	K										O	$G_{0.7}$
			1	2	3	4	5	6	7	8	9	10		
Text	With	Gemini-3-Pro-Image [13]	91.0↓2.0	89.5↓0.5	83.5↑3.5	76.5↑6.5	71.5↑5.0	70.0↑6.5	70.0↑11.5	62.5↑13.5	59.5↑6.5	51.0↑4.0	72.5↑5.4	7↑3
		GPT-Image-1.5 [29]	95.5↑3.5	82.0↑6.5	67.5↑5.0	55.0↑3.0	58.5↑7.0	48.5↑0.5	39.0↓3.0	35.5↑5.5	33.0↑5.0	21.5↑1.5	53.6↑3.4	2(0)
		Seedream-4.5 [7]	77.5↓7.5	70.0↑5.0	51.0↑2.5	38.0↑4.0	33.0↓0.5	22.5↓4.0	16.5↓4.5	11.0↓6.0	9.5↓2.5	4.0↓8.5	33.3↓2.2	2↑1
	Without	Gemini-3-Pro-Image [13]	64.0↓29.0	83.5↓6.5	74.0↓6.0	61.5↓8.5	61.5↓5.0	54.5↓9.0	46.5↓12.0	43.5↓5.5	42.0↓11.0	38.5↓8.5	57.0↓10.1	0↓4
		GPT-Image-1.5 [29]	58.5↓33.5	78.5↑3.0	64.5↑2.0	52.0(0.0)	41.5↓10.0	37.0↓11.0	26.5↓15.5	21.0↓9.0	17.5↓10.5	12.5↓7.5	41.0↓9.2	0↓2
		Seedream-4.5 [7]	43.5↓41.5	40.0↓25.0	10.5↓38.0	5.0↓29.0	6.0↓27.5	8.0↓18.5	6.5↓14.5	1.0↓16.0	7.5↓4.5	1.5↓11.0	13.0↓22.5	0↓1
Image	With	Gemini-3-Pro-Image [13]	94.0↑1.0	93.5↑3.5	88.0↑8.0	86.0↑16.0	78.0↑11.5	71.5↑8.0	71.5↑13.0	65.0↑16.0	68.5↑15.5	59.0↑12.0	77.5↑10.4	7↑3
		GPT-Image-1.5 [29]	95.5↑3.5	84.0↑8.5	74.0↑11.5	65.0↑13.0	59.0↑7.5	57.5↑9.5	45.5↑3.5	36.0↑6.0	33.5↑5.5	23.5↑3.5	57.4↑7.2	3↑1
		Seedream-4.5 [7]	85.5↑0.5	74.5↑9.5	61.5↑13.0	49.5↑15.5	42.5↑9.0	39.0↑12.5	42.5↑21.5	25.5↑8.5	28.5↑16.5	22.5↑10.0	47.2↑11.7	2↑1
	Without	Gemini-3-Pro-Image [13]	66.5↓26.5	87.0↓3.0	79.0↓1.0	78.0↑8.0	70.5↑4.0	66.0↑2.5	68.0↑9.5	59.0↑10.0	55.5↑2.5	52.5↑5.5	68.2↑1.1	0↓4
		GPT-Image-1.5 [29]	64.0↓28.0	81.0↑5.5	61.5↓1.0	53.0↑1.0	47.5↓4.0	47.0↓1.0	37.0↓5.0	30.0(0.0)	26.5↓1.5	22.0↑2.0	47.0↓3.2	0↓2
		Seedream-4.5 [7]	51.0↓34.0	69.0↑4.0	41.5↓7.0	38.0↑4.0	37.0↑3.5	35.0↑8.5	24.5↑3.5	25.0↑8.0	24.0↑12.0	23.5↑11.0	36.9↑1.4	0↓1

at low complexity. As K increases, its text performance deteriorates rapidly, narrowing the gap and suggesting that language understanding becomes the dominant bottleneck at high complexity.

Overall, these findings reveal a hierarchical bottleneck in structural compositional generalization of text-to-image models. Stronger models are primarily limited by language-to-visual generation, whereas weaker models are constrained by both semantic understanding and visual generation.

4.6 A Promising Solution: Intermediate Layout Representations

Sec. 4.5 shows that failures in structural compositional generalization for text-to-image models largely arise from the language-to-vision transformation stage. A language instruction defines a set of feasible scenes satisfying semantic constraints, whereas image generation requires instantiating this abstract structure into a concrete visual configuration. As structural complexity increases, dependencies among objects tighten and constraints become more strongly coupled, making local generation errors more likely to violate constraints. Providing an explicit instance as an intermediate reference may therefore reduce instantiation ambiguity and stabilize the generation process.

Motivated by this hypothesis, we conduct experiments in the Knolling scenario using intermediate layout representations to guide generation. In the pure text setting, DeepSeek-V3.2-Thinking achieves a 97.6% satisfaction rate. We treat its predicted object categories, attributes, and bounding box coordinates as an approximately correct layout plan and use it as an intermediate condition for image generation. Two layout representations are considered: a textual layout containing coordinates, attributes, and category labels, and a visual layout image. For each representation, models generate images either with the original instruction preserved or conditioned solely on the layout. We evaluate three text-to-image models from different performance tiers that support both text and image inputs. The results in Tab. 3 reveal several consistent trends.

Intermediate layouts stabilize the generation process and improve structural compositional generalization. When the original instruction is preserved, incorporating intermediate layouts yields performance gains at nearly all complexity

levels for stronger models, with more pronounced improvements under the visual layout condition. For Gemini-3-Pro-Image, both layout representations increase structural compositional generalizability to $G_{0.7} = 7$.

The original instruction remains crucial for conveying logical semantic constraints. Although the layout satisfies DSL constraints and forms a valid structural solution, it corresponds to a specific satisfying instance rather than the underlying logical constraints themselves. When the original instruction is removed and generation relies solely on layout conditions, textual layouts lead to substantial performance drops for all models. Visual layouts maintain performance closer to the original setting, yet structural compositional generalizability decreases noticeably in both cases. At $K = 1$, removing the original instruction results in significant degradation for both textual and visual layouts. Further analysis shows that many $K = 1$ instructions correspond to layouts containing only a single object. While the layout provides correct positional coordinates, in the absence of explicit semantic spatial descriptions, models tend to place the object at the center of the image, thereby violating positional constraints.

The form of intermediate representation plays a critical role, and its effectiveness depends strongly on modality alignment. Comparing textual and visual layouts, visual layouts produce more stable and larger overall gains. One possible explanation is that current models receive limited exposure to coordinate-based textual representations during training. In addition, textual layouts must still be processed by the language module, which increases interpretive burden. For weaker models, such as Seedream-4.5, introducing textual layouts can even lead to slight performance degradation.

Overall, these results suggest that intermediate representations can partially mitigate failures in language-to-vision generation by reducing structural ambiguity. However, their effectiveness depends strongly on alignment between the representation format and the model's conditioning modality. How to design more effective intermediate representations for improving structural compositional generalization remains an open question for future work.

5 Conclusion

This paper introduces FormalImG, an FOL-based benchmark for structural compositional generalization, and defines structural compositional complexity together with ε-structural compositional generalizability to characterize the stable performance of text-to-image models under varying scales of semantic dependency. Experimental results show that existing models exhibit significant performance degradation as structural complexity increases, with stable generalization remaining confined to low-complexity regimes. Further analysis reveals a hierarchical bottleneck phenomenon, and shows that incorporating intermediate layout representations can partially mitigate performance degradation.

References

1. Adobe: Adobe revolutionizes ai-assisted creativity with firefly, the all-in-one home for ai content creation, with new partner and firefly models. https://news.adobe.com/news/2025/04/adobe-revolutionizes-ai-assisted-creativity-firefly (2025) 1
2. Ajaykumar, G., Steele, M., Huang, C.: A survey on end-user robot programming. ACM Computing Surveys **54**(8), 164:1–164:36 (2021) 1
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025) 8
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 12
5. Black Forest Labs: FLUX. https://github.com/black-forest-labs/flux (2024) 3, 9
6. Cao, S., Chen, H., Chen, P., Cheng, Y., Cui, Y., Deng, X., Dong, Y., Gong, K., Gu, T., Gu, X., et al.: Hunyuanimage 3.0 technical report. arXiv preprint arXiv:2509.23951 (2026) 9
7. Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, Y., Jian, X., et al.: Seedream 4.0: Toward next-generation multimodal image generation. arXiv preprint arXiv:2509.20427 (2025) 9, 13
8. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., Guang, S., Fan, H.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025) 2, 4, 9
9. Dong, Y., Jiang, X., Qian, J., Wang, T., Zhang, K., Jin, Z., Li, G.: A survey on code generation with llm-based agents. arXiv preprint arXiv:2508.00083 (2025) 1
10. Ghosh, D., Hajishirzi, H., Schmidt, L.: GenEval: An object-focused framework for evaluating text-to-image alignment. In: Advances in Neural Information Processing Systems. vol. 36, pp. 52132–52152 (2023) 2, 4
11. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems. vol. 27, pp. 2672–2680 (2014) 3
12. Google: Gemini 2.5 flash and gemini 2.5 flash image model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Flash-Model-Card.pdf (2025) 9, 12
13. Google: Gemini 3 pro image model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Image-Model-Card.pdf (2025) 8, 9, 13
14. Google DeepMind: Gemini 3 pro model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf (2025) 8, 12
15. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: CLIPScore: A reference-free evaluation metric for image captioning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 7514–7528 (2021) 4
16. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems. vol. 30, pp. 6626–6637 (2017) 4
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851 (2020) 3

18. Hu, Y., Koren, Y., Volinsky, C.: Collaborative filtering for implicit feedback datasets. In: Proceedings of the IEEE International Conference on Data Mining. pp. 263–272 (2008) 1
19. Huang, K., Duan, C., Sun, K., Xie, E., Li, Z., Liu, X.: 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 (2025) 4
20. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In: Advances in Neural Information Processing Systems. vol. 36, pp. 78723–78747 (2023) 2, 4
21. Kamath, A., Chang, K., Krishna, R., Zettlemoyer, L., Hu, Y., Ghazvininejad, M.: GenEval 2: Addressing benchmark drift in text-to-image evaluation. arXiv preprint arXiv:2512.16853 (2025) 4
22. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: International Conference on Learning Representations (2014) 3
23. Li, B., Lin, Z., Pathak, D., Li, J., Fei, Y., Wu, K., Ling, T., Xia, X., Zhang, P., Neubig, G., Ramanan, D.: GenAI-Bench: Evaluating and improving compositional text-to-visual generation. arXiv preprint arXiv:2406.13743 (2024) 4
24. Liu, A., Mei, A., Lin, B., Xue, B., Wang, B., Xu, B., Wu, B., Zhang, B., Lin, C., Dong, C., et al.: Deepseek-v3.2: Pushing the frontier of open large language models. arXiv preprint arXiv:2512.02556 (2025) 12
25. Lu, J., Clark, C., Zellers, R., Mottaghi, R., Kembhavi, A.: UNIFIED-IO: A unified model for vision, language, and multi-modal tasks. In: International Conference on Learning Representations (2023) 4
26. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014) 3
27. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In: Proceedings of the International Conference on Machine Learning. pp. 16784–16804 (2022) 2, 3
28. OpenAI: Gpt image 1 model. https://platform.openai.com/docs/models/gpt-image-1 (2025) 9
29. OpenAI: New chatgpt images is here. https://openai.com/index/new-chatgpt-images-is-here/ (2025) 8, 9, 13
30. Peng, Q., Liu, H., Huang, H., Yang, J., Yang, Q., Shao, M.: A survey on llm-powered agents for recommender systems. In: Findings of the Association for Computational Linguistics: EMNLP 2025. pp. 11574–11583 (2025) 1
31. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: improving latent diffusion models for high-resolution image synthesis. In: International Conference on Learning Representations (2024) 8, 9
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning. vol. 139, pp. 8748–8763 (2021) 3
33. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125 (2022) 2, 3
34. Raychev, V., Vechev, M.T., Yahav, E.: Code completion with statistical language models. In: Proceedings of the ACM SIGPLAN Conference on Programming Language Design and Implementation. pp. 419–428 (2014) 1

35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10674–10685 (2022) 2, 3, 9
36. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, S.K.S., Lopes, R.G., Ayan, B.K., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: Advances in Neural Information Processing Systems. vol. 35, pp. 36479–36494 (2022) 3
37. Salimans, T., Goodfellow, I.J., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: Advances in Neural Information Processing Systems. vol. 29, pp. 2226–2234 (2016) 4
38. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: Advances in Neural Information Processing Systems. vol. 28, pp. 3483–3491 (2015) 3
39. Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025) 9
40. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025) 2, 3, 8, 9
41. Wu, X., Yu, D., Huang, Y., Russakovsky, O., Arora, S.: ConceptMix: A compositional image generation benchmark with controllable difficulty. In: Advances in Neural Information Processing Systems. vol. 37, pp. 86004–86047 (2024) 2, 4, 5, 7, 10
42. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. In: International Conference on Learning Representations (2025) 4
43. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025) 12
44. Yao, X., Zhou, H., Mees, O., Meng, Y., Xiao, T., Bisk, Y., Oh, J., Johns, E., Shridhar, M., Shah, D., Thomason, J., Huang, K., Chai, J., Bing, Z., Knoll, A.: Bridging language and action: A survey of language-conditioned robot manipulation. arXiv preprint arXiv:2312.10807 (2025) 1