GENERATE ANY SCENE: SCENE GRAPH DRIVEN DATA SYNTHESIS FOR VISUAL GENERATION TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in text-to-vision generation excel in visual fidelity but struggle with compositional generalization and semantic alignment. Existing datasets are noisy and weakly compositional, limiting models' understanding of complex scenes, while scalable solutions for dense, high-quality annotations remain a challenge. We introduce GENERATE ANY SCENE, a data engine that systematically enumerates scene graphs representing the combinatorial array of possible visual scenes. GENERATE ANY SCENE dynamically constructs scene graphs of varying complexity from a structured taxonomy of objects, attributes, and relations. Given a sampled scene graph, GENERATE ANY SCENE translates it into a caption for text-to-image or text-to-video generation; it also translates it into a set of visual question answers that allow automatic evaluation and reward modeling of semantic alignment. Using GENERATE ANY SCENE, we first design a self-improving framework where models iteratively enhance their performance using generated data. *SDv1.5* achieves an average **4%** improvement over baselines and surpassing fine-tuning on CC3M. Second, we also design a distillation algorithm to transfer specific strengths from proprietary models to their open-source counterparts. Using fewer than 800 synthetic captions, we fine-tune *SDv1.5* and achieve a **10%** increase in TIFA score on compositional and hard concept generation. Third, we create a reward model to align model generation with semantic accuracy at a low cost. Using GRPO algorithm, we fine-tune SimpleAR-0.5B-SFT and surpass CLIP-based methods by **+5%** on DPG-Bench. Finally, we apply these ideas to the downstream task of content moderation where we train models to identify challenging cases by learning from synthetic data.

1 INTRODUCTION

Despite the high-fidelity of modern generative models (text-to-image and text-to-video), we are yet to witness wide-spread adoption ([1; 2; 3; 4; 5]). Controllability remains out of reach ([6]). Generated content appears realistic but often falls short of semantic alignment ([7; 8; 9; 10]). Users prompt models with a specific concept in mind. For example, when prompted to generate a scene of a "A black dog chasing after a rabbit that is eating the grass, in Van Gogh's style, with starlight lightening", some models are likely to generate an image of a dog but might miss the rabbit or get the style incorrect.

We hypothesize that these limitations stem not only from architectural bottlenecks but more fundamentally from the lack of structured, compositionally rich training data ([3]), especially those with uncommon compositions. Popular datasets such as LAION ([11]) and CC3M ([12]) predominantly consist of web-crawled image-caption pairs, which are inherently noisy, weakly compositional, and biased toward single-object, coarse-grained descriptions. Such datasets lack explicit grounding of object-attribute relations and multi-object interactions, restricting models' ability to generalize to complex visual scenes. Efforts to enhance caption quality ([3; 13]) have demonstrated that enhancing the compositional density and semantic richness of captions can significantly improve generative performance. Nevertheless, manual curation of such dense compositional annotations is labor-intensive, while automatic annotation methods (e.g., via MLMs) suffer from hallucination and semantic noise.

Constructing a compositional dataset requires that we first define *the space of the visual content*. Scene graphs are one such representation of the visual space ([14; 15; 16; 17; 18]), grounded in cognitive science ([19]). A scene graph represents objects in a scene as individual nodes in a graph.

Each object is modified by attributes, which describe its properties. For example, attributes can describe the material, color, size, and location of the object in the scene. Finally, relationships are edges that connect the nodes. They define the spatial, functional, social, and interactions between objects (20). For example, in a living room scene, a "table" node might have attributes like "wooden" or "rectangular" and be connected to a "lamp" node through a relation: "on top of". This systematic scene graph structure provides simple yet effective ways to define and model the scene. As such, scene graphs are an ideal foundation for systematically defining the compositional space of visual content in text-to-vision generation.

We introduce GENERATE ANY SCENE, a system capable of efficiently enumerating the space of scene graphs representing a wide range of visual scenes. GENERATE ANY SCENE composes scene graphs of any structure using a rich taxonomy of visual elements, translating each scene graph into an input caption and visual question answers to evaluate the output image or video. In particular, we first construct a rich taxonomy of visual concepts consisting of $28,787$ objects, $1,494$ attributes, $10,492$ relations, $2,193$ scene attributes from various sources. Based on these assets, GENERATE ANY SCENE can synthesize an almost infinite number of scene graphs of varying complexity (21). Besides, GENERATE ANY SCENE allows configurable scene graph generation. For example, evaluators can specify the complexity level of the scene graph to be generated or provide a seed scene graph to be expanded. By automating these steps, our system ensures both scalability and adaptability, providing researchers and developers with diverse, richly detailed scene graphs and corresponding captions tailored to their specific needs. We also conduct comprehensive text-to-vision evaluations using our generated captions, as detailed in Appendix A.

We show that GENERATE ANY SCENE can allow generation models to self-improve. Our diverse captions can facilitate a framework to iteratively improve *Text-to-Vision generation* models using their own generations. Given a model, we generate multiple images, identify the highest-scoring one, and use it as new fine-tuning data to improve the model itself. We fine-tune *SDv1.5* (22) and achieve an average of **4%** performance boost compared with original models, and this method is even better than fine-tuning with the same amount of real images and captions from the Conceptual Captions CC3M over different benchmarks.

We also use GENERATE ANY SCENE to design targeted distillation algorithms. Using our evaluations, we identify limitations in open-sourced models that their proprietary counterparts excel at. Next, we distill these specific capabilities from proprietary models. For example, *DaLL-E 3* (3) excels particularly in generating composite images with multiple parts. We distill this capability into *SDv1.5*, effectively bridging the gap between *DaLL-E 3* and *SDv1.5*. After targeted fine-tuning, *SDv1.5* achieves a **10%** increase in TIFA score (23) for compositional tasks and hard concept generation.

Then we propose a low-cost scene graph-based reward model for RLHF (24) in text-to-image generation. By leveraging synthetic scene graphs generated by GENERATE ANY SCENE, we generate exhaustive question-answer pairs that cover all objects, attributes, and relationships in the caption. Our method enables fine-grained, compositional reward modeling without manual annotation or heavy LLM inference. With GRPO (25), we fine-tune SimpleAR-0.5B-SFT (26) using a scene graph reward model, achieving better compositional alignment than CLIP-based methods (27) (**+5%** on DPG-Bench (28)).

Finally, we apply GENERATE ANY SCENE to the downstream application of content moderation. Content moderation is a vital application, especially as *Text-to-Vision generation* models improve. A key challenge lies in the limited diversity of existing training data. To address this, we leverage GENERATE ANY SCENE to generate diverse and compositional captions, creating synthetic training data that complements existing datasets. By retraining a ViT-T (29) detector with our enriched dataset, we enhance its detection performance, particularly in cross-model and cross-dataset scenarios.

2 GENERATE ANY SCENE

In this section, we present GENERATE ANY SCENE (Figure 1), a data engine that systematically synthesizes diverse scene graphs in terms of both structure and content and translates them into corresponding captions.

Scene graph. A scene graph is a structured representation of a visual scene, where objects are represented as nodes, their attributes (such as color and shape) are properties of those nodes, and the relationships between objects (such as spatial or semantic connections) are represented as edges. In recent years, scene graphs have played a crucial role in visual understanding tasks, such as those found in Visual Genome ([14]) and GQA ([30]) for visual question answering (VQA). Their utility has expanded to various *Text-to-Vision generation* tasks. For example, the DSG ([31]) and DPG ([10]) benchmarks leverage scene graphs to evaluate how well generated images align with captions.

Taxonomy of visual elements. To construct a scene graph, we use three main metadata types: **objects**, **attributes**, and **relations**. We further introduce **scene attributes** that capture global visual contexts, such as art style, to facilitate comprehensive caption synthesis. The statistics and source of our metadata are shown in Table 1. Additionally, we build a hierarchical taxonomy that categorizes metadata into distinct levels and types, enabling fine-grained analysis. This structure supports precise content synthesis, from broad concepts like "flower" to fine-grained instances such as "daisy."

Table 1: Summary of the quantities and sources of visual elements. Details are in Appendix B.

Metadata Type	**Number**	**Source**
Objects	28,787	WordNet ([32])
Attributes	1,494	Wikipedia ([33]), etc.
Relations	10,492	Synthetic Visual Genome ([34])
Scene Attributes	2,193	Places365 ([35]), etc.

2.1 Generating data with scene graphs

Step 1: Scene graph structure enumeration. Our engine pre-computes a library of directed scene-graph topologies subject to user-specified *structural constraints*: complexity (total number of objects, relations, and attributes) ([36]), average node degree, and number of connected components. We first sample the number of object nodes and then systematically enumerate feasible edge sets and attribute attachments that satisfy these constraints. We provide 3 optional controls: (i) *degree-profile* bounds per-node in/out-degree, (ii) *seed-graph preservation* embeds a user-provided seed graph as a subgraph of each enumerated structure, and (3) *commonsense plausibility filtering* prunes implausible contents while retaining compositional diversity (See Appendix. H.1). All enumerations are performed once per parameter tuple and cached for fast querying.

Step 2: Populate the scene graph structure with metadata. Given a generated scene graph structure, the next step involves populating the graph with metadata. For each object node, attribute node, and relation edge, we sample the corresponding content from our metadata. This process is highly customizable and controllable: users can define the topics and types of metadata to include, for instance, by selecting only commonsense metadata or specifying relationships between particular objects. By determining the scope of metadata sampling, we can precisely control the final content of the captions and easily extend the diversity and richness of scene graphs by adding new metadata.

Step 3: Sample scene attributes. We also include scene attributes that describe aspects such as the art style, viewpoint, time span (for video), and 3D attributes (for 3D content). These scene attributes are sampled directly from our metadata, creating a list that provides contextual details to enrich the description of the visual content.

Step 4: Translate scene graph to caption. We introduce a deterministic and programmatic algorithm that converts scene graphs with scene attributes into captions. It traverses scene graphs by converting objects/attributes/relations into descriptive text in topological order, while tracking each object's references to ensure coherence. Programmatic grammar rules are employed (e.g., disambiguating identical objects with "the first/second" and skipping already mentioned objects) to prevent duplication and misreference, resulting in clear captions. We also provide LLM paraphrasing as an optional step to diversify wording; however, our studies (see Appendix A.3) show that paraphrasing does not materially affect results. We adopt the programmatic caption converter as the default for its speed and low hallucination rate.

Step 5: Convert scene graph to a series of question-answer pairs. Given a synthetic scene graph, Generate Any Scene automatically enumerates exhaustive QA pairs using templates that query object attributes (e.g., What color is the sphere?), spatial relations (e.g., What is to the left of the cube?), and other compositional elements. Each answer maps directly to an object, attribute, or edge, ensuring full coverage of the graph at minimal cost. This enables both VQA-based evaluation of

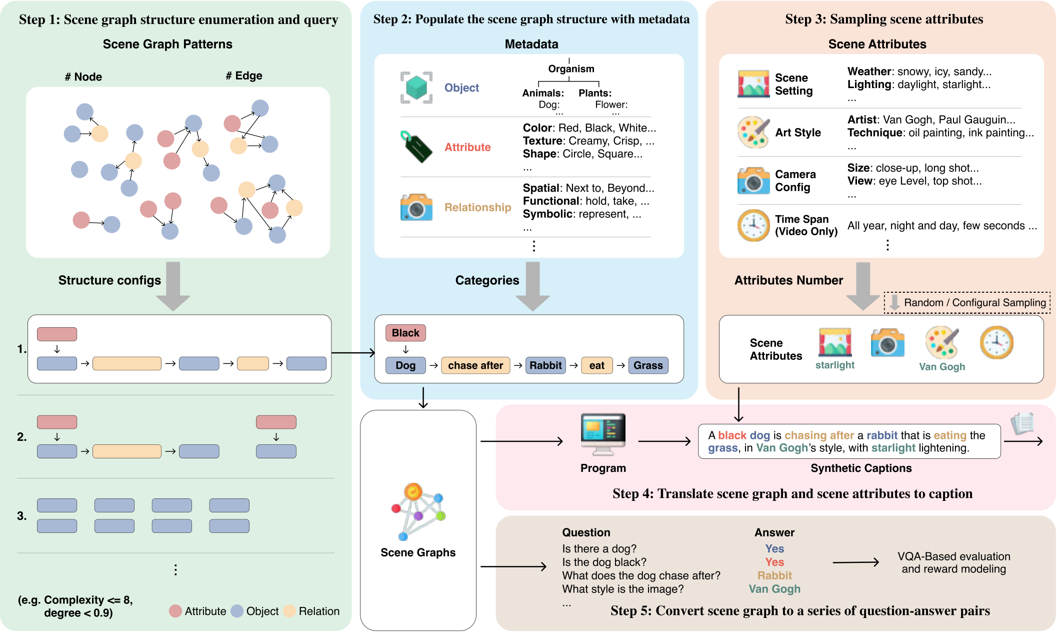

Step 1: Scene graph structure enumeration and query
Scene Graph Patterns
# Node
# Edge
Structure configs
1.
2.
3.
(e.g. Complexity <= 8, degree < 0.9)
● Attribute ● Object ● Relation
Step 2: Populate the scene graph structure with metadata
Metadata
Object
Organism
Animals: Dog: ...
Plants: Flower: ...
Attribute
Color: Red, Black, White...
Texture: Creamy, Crisp, ...
Shape: Circle, Square...
Relationship
Spatial: Next to, Beyond...
Functional: hold, take, ...
Symbolic: represent, ...
Categories
Black
Dog → chase after → Rabbit → eat → Grass
Step 3: Sampling scene attributes
Scene Attributes
Scene Setting
Weather: snowy, icy, sandy...
Lighting: daylight, starlight...
Art Style
Artist: Van Gogh, Paul Gauguin...
Technique: oil painting, ink painting...
Camera Config
Size: close-up, long shot...
View: eye Level, top shot...
Time Span (Video Only)
All year, night and day, few seconds ...
Attributes Number
Random / Configural Sampling
Scene Attributes
starlight
Van Gogh
Scene Graphs
Program
A black dog is chasing after a rabbit that is eating the grass, in Van Gogh's style, with starlight lightening.
Synthetic Captions
Step 4: Translate scene graph and scene attributes to caption
Question
Is there a dog?
Is the dog black?
What does the dog chase after?
What style is the image?
...
Answer
Yes
Yes
Rabbit
Van Gogh
VQA-Based evaluation and reward modeling
Step 5: Convert scene graph to a series of question-answer pairs

Figure 1: The generation pipeline of GENERATE ANY SCENE. **Step 1:** Enumerate diverse scene graph structures under user-defined constraints. **Step 2:** Populate structures with sampled objects, attributes, and relations. **Step 3:** Sample scene attributes such as style, perspective, or time span. **Step 4:** Translate scene graph and attributes into coherent captions. **Step 5:** Automatically generate QA pairs covering all elements for evaluation and reward modeling.

generated images and the construction of fine-grained reward models without manual labeling or costly LLM inference.

3 SELF-IMPROVING MODELS WITH SYNTHETIC CAPTIONS

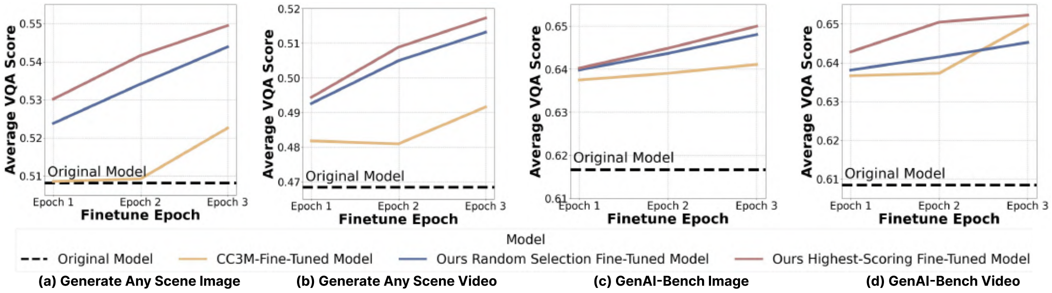


Figure 2: **Results for Self-Improving Models**. Average VQA score of *SDv1.5* fine-tuned on different data across 1K GENERATE ANY SCENE image/video evaluation set and GenAI-Bench image/video benchmark ([37]).

With GENERATE ANY SCENE, we develop a self-improvement framework to improve generative capabilities. By generating scalable compositional captions from scene graphs, GENERATE ANY SCENE expands the textual and visual space, allowing for a diversity of synthetic images that extend beyond real-world scenes. Our goal is to utilize these richly varied synthetic images to further boost model performance.

Iterative self-improving framework. Inspired by DreamSync ([38]), we designed an iterative self-improving framework using GENERATE ANY SCENE with *SDv1.5* as the baseline model. With

Table 2: **Quality and diversity comparison on GenAI-Bench**. Fine-tuning with GENERATE ANY SCENE captions improves global semantic fidelity and perceptual quality without reducing generation diversity.

	SDv1.5	CC3M-FT	GAS-FT
CLIPScore	0.3167	0.3196	0.3206
ImageReward	0.2056	0.3842	0.3927
LPIPS	0.7297	0.7356	0.7329

Table 3: **Generalization to unseen compositions.** On a 400-caption test set containing only unseen combinations of seen elements, the model fine-tuned with GENERATE ANY SCENE achieves the best compositional generalization.

	SDv1.5	CC3M-FT	GAS-FT
VQAScore	0.5823	0.6044	0.6109
CLIPScore	0.2876	0.2927	0.2938
ImageReward	0.4861	0.2602	-0.2497

VQA Score, which shows strong correlation with human evaluations on compositional images (39), we guide the model's improvement throughout the process. Specifically, GENERATE ANY SCENE generates 3 × 10K captions across three epochs. For each caption, *SDv1.5* generates 8 images, and the image with the highest *VQA Score* is selected. From each set of 10K optimal images, we then select the top 25% (2.5K image-caption pairs) as the training data for each epoch. In subsequent epochs, we use the fine-tuned model from the prior iteration to generate new images. We employ LoRA (40) for parameter-efficient fine-tuning.

Baselines. We conduct comparative experiments with the CC3M dataset, which comprises high-quality and diverse real-world image-caption pairs (12). We randomly sample 3 × 10K captions from CC3M, applying the same top-score selection strategy for iterative fine-tuning of *SDv1.5*. Additionally, we include a baseline using random-sample fine-tuning strategy to validate the advantage of our highest-scoring selection-based strategy. We evaluate our self-improving pipeline on *Text-to-Vision generation* benchmarks, including GenAI Bench (37). For the *Text-to-Video generation* task, we use *Text2Video-Zero* as the baseline model, substituting its backbone with the original *SDv1.5* and our fine-tuned *SDv1.5* models.

Fine-tuning with our synthetic captions can surpass high-quality real-world image-caption data. Our results show that fine-tuning with GENERATE ANY SCENE-generated synthetic data consistently outperforms CC3M-based fine-tuning across *Text-to-Vision generation* tasks (Figure 2), achieving the highest gains with our highest-scoring selection strategy. This highlights GENERATE ANY SCENE's scalability and compositional diversity, enabling models to effectively capture complex scene structures. In Table 2, we further evaluate SDv1.5, the CC3M-finetuned model, and the model finetuned with GENERATE ANY SCENE captions on additional metrics from GenAI-Bench. Fine-tuning with GENERATE ANY SCENE yields higher CLIPScore and ImageReward while preserving LPIPS, demonstrating that our method not only strengthens compositional alignment but also improves global semantic fidelity and perceptual quality without reducing generation diversity. In Table 3, we additionally evaluate whether our self-improving framework enhances combinatorial generalization. We extract all objects, attributes, and relations from the CC3M fine-tuning data and retain the metadata sampled by GENERATE ANY SCENE. Using the same element set as in the fine-tuning data, we synthesize 200 CC3M-element-based and 200 GENERATE ANY SCENE-element-based captions while excluding all seen combinations, forming a 400-caption test set of unseen compositions. The model fine-tuned with GENERATE ANY SCENE achieves the highest VQAScore, CLIPScore, and ImageReward, indicating stronger compositional generalization than both SDv1.5 and the CC3M-finetuned baseline. Additional experiment settings and results are in Appendix C.

4 DISTILLING TARGETED CAPABILITIES

Although self-improving with GENERATE ANY SCENE shows clear advantages over high-quality real-world datasets, its efficiency is inherently limited by the model's own generation capabilities. To address this, we leverage the taxonomy and systematical generation capabilities within GENERATE ANY SCENE to identify specific strengths of proprietary models (*DaLL-E 3*), and distill these capabilities into open-source models. More details are in Appendix D.

We evaluate multiple models using GENERATE ANY SCENE controllably generated captions and observe that *DaLL-E 3* achieves *TIFA Score* **1.5** to **2** times higher than those of other models. As shown in Figure 4a, when comparing *TIFA Score* across captions with varying numbers of elements

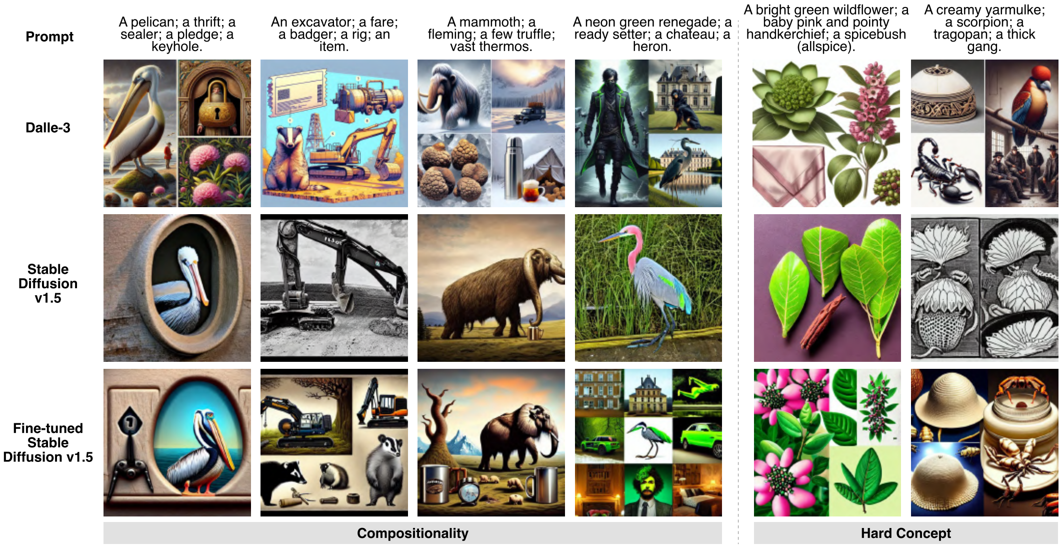

Prompt
A pelican; a thrift; a sealer; a pledge; a keyhole.
An excavator; a fare; a badger; a rig; an item.
A mammoth; a fleming; a few truffle; vast thermos.
A neon green renegade; a ready setter; a chateau; a heron.
A bright green wildflower; a baby pink and pointy handkerchief; a spicebush (allspice).
A creamy yarmulke; a scorpion; a tragopan; a thick gang.
Dalle-3
Stable Diffusion v1.5
Fine-tuned Stable Diffusion v1.5
Compositionality
Hard Concept

Figure 3: **Examples for Distilling Capabilities.** Examples of images generated by *DaLL-E 3*, the original *SDv1.5*, and the fine-tuned versions. The left four captions demonstrate fine-tuning with multi-object captions generated by GENERATE ANY SCENE for better compositionality, while the right two columns focus on understanding hard concepts.

(objects, relations, and attributes), *DaLL-E 3* **counterintuitively** maintains consistent performance regardless of element count. The performance of other models declines as the element count increases, which aligns with expected compositional challenges. We suspect that these differences are primarily due to *DaLL-E 3*'s advanced capabilities in compositionality and **understanding hard concepts**, which ensures high faithfulness across diverse combinations of element types and counts.

Distilling compositionality from DaLL-E 3. When analyzing model outputs from our synthetic captions, we find that *DaLL-E 3* tends to produce straightforward combinations of multiple objects (Figure 3). In contrast, open-source models like *SDv1.5* often omit objects from the captions, despite being capable of generating each one individually. This difference suggests that *DaLL-E 3* may benefit from training data emphasizing multi-object presence, even without detailed layout or object interaction. Such training likely underpins *DaLL-E 3*'s stronger performance on metrics like *TIFA Score* and *VQA Score* that prioritize object inclusion. To effectively distill these compositional abilities into *SDv1.5*, we employ GENERATE ANY SCENE for targeted synthesis of 778 multi-object captions, paired with images generated by *DaLL-E 3*, for finetuning *SDv1.5*.

Distilling hard concepts understanding from DaLL-E 3. Figure 3 shows that *DaLL-E 3* is capable not only of handling multi-object generation but also of understanding and generating rare and hard concepts, such as a specific species of flower. We attribute this to its training with proprietary real-world data. Using the taxonomy of GENERATE ANY SCENE, we evaluate both models on 10K GENERATE ANY SCENE captions that broadly cover the taxonomy. For each concept, we gather all captions in which it appears and average their generation scores to obtain a concept-level score for each model. Comparing these concept-level scores lets us identify the 81 concepts where *SDv1.5* shows the largest gap relative to *DaLL-E 3*; the full list is provided in Appendix D. For distilling, we increase the sampling frequency of these hard concepts and generate 778 captions incorporating these hard concepts with other elements, and use *DaLL-E 3* to produce corresponding images.

Baselines. For the baseline, we randomly synthesize 778 captions using GENERATE ANY SCENE paired with *DaLL-E 3*-generated images to fine-tune the model. To evaluate model improvements, we generate another 1K multi-object captions and 1K hard-concept captions separately.

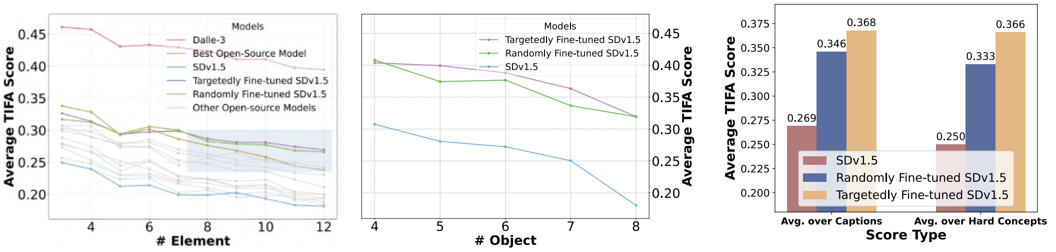
Targeted caption synthesis via GENERATE ANY SCENE enables effective distillation of compositional abilities and hard concept understanding. We analyze images generated by *SDv1.5* before and after fine-tuning on high-complexity captions (Figure 3). Surprisingly, with fewer than 1K LoRA fine-tuning steps, *SDv1.5* effectively learns *DaLL-E 3* 's capability to arrange and compose multiple objects within a single image. Quantitatively, Figure 4b shows a 10% improvement in *TIFA Score* after targeted fine-tuning, surpassing the performance of the randomly fine-tuned model. On a broader set of 10K GENERATE ANY SCENE-generated captions, the targeted fine-tuned model consistently outperforms randomly fine-tuned and original counterparts across complex scenes (Figure 4a). These results confirm not only the effectiveness but also the scalability and efficiency of GENERATE ANY SCENE. Also, the results in Figure 4c show that our targeted fine-tuning with hard concepts leads to improved model performance, reflected in higher average scores across captions and increased scores for each challenging concept.

5 REINFORCEMENT LEARNING WITH A SYNTHETIC REWARD FUNCTION

Reinforcement Learning with Human Feedback (RLHF) has become an increasingly popular fine-tuning strategy in text-to-image generation ([41]; [42]; [26]). However, defining an effective reward model that accurately captures semantic alignment for text-to-image generation remains an open challenge. Existing reward models like CLIP offer only coarse-grained image-text similarity signals, which fall short in assessing compositional correctness and lack interpretability. Alternative approaches have explored using visual question answering (VQA) as a proxy for evaluating semantic alignment, aiming for finer-grained assessments, yet require either labor-intensive datasets with dense annotations or large volumes of contextually relevant questions via advanced LLMs. Leveraging its structured scene graph synthesis capabilities, GENERATE ANY SCENE offers a scalable alternative by producing exhaustive semantic queries with negligible overhead, enabling low-cost, compositional reward modeling (Sec 2.1).

Experiment setup. Building on this scene graph-based reward modeling strategy, we adopt Group Relative Policy Optimization (GRPO) as our reinforcement learning algorithm. We fine-tune the SimpleAR-0.5B-SFT model for one epoch using 10K captions generated by GENERATE ANY SCENE, each paired with their scene graph-derived QA sets. For reward evaluation, we use Qwen2.5-VL-3B, a lightweight open-source vision-language model, to answer these QA pairs given the model-generated images. The reward is computed as the accuracy across all questions. This fine-grained, scene graph-aligned reward provides precise feedback on compositional faithfulness. As a baseline, we compare against SimpleAR-0.5B-RL, trained with CLIP-based rewards on 11K captions from real world datasets for one epoch. We evaluate our scene graph-based reward model on three benchmarks: DPG-Bench ([10]), GenEval ([9]), and GenAI-Bench ([37]). More details are in Appendix E.

GENERATE ANY SCENE rewards outperform CLIP. As shown in Table 4, our method outperforms both SFT and CLIP-RL models and achieves a significant improvement, demonstrating superior compositional faithfulness driven by explicit scene graph rewards. Importantly, this performance gain is directly enabled by the GENERATE ANY SCENE engine, which constructs explicit scene graphs to generate compositional captions. GENERATE ANY SCENE provides a structured and cognitively



(a) **Distilling compositionality from DaLL-E 3**: Model results on TIFA vs. total element numbers in captions in 10K general GENERATE ANY SCENE captions. ("Best Open-Source Model" refers to Flux.1-schnell)

(b) **Distilling compositionality from DaLL-E 3**: Model results on TIFA vs. total element numbers in captions in 1K multi-object GENERATE ANY SCENE captions.

(c) **Distilling hard concepts understanding from DALL-E 3**: Models' average *TIFA Score* performance over captions and hard concepts in 1K hard concepts GENERATE ANY SCENE captions.

Figure 4: **Results for Distilling Capabilities**. The left two figures show the results for **Distilling compositionality**, while the rightmost figure shows the results for **Distilling hard concepts understanding from DALL-E 3**.

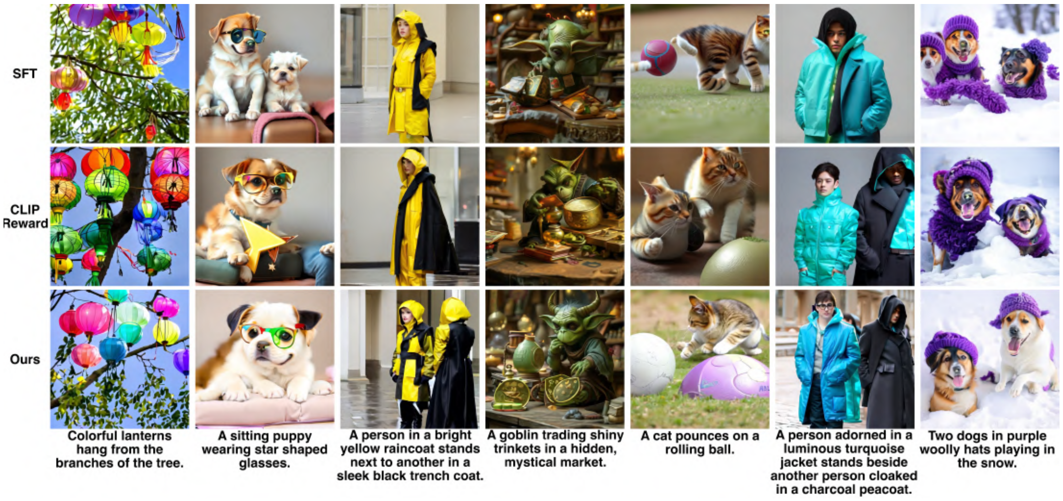

SFT
CLIP Reward
Ours
Colorful lanterns hang from the branches of the tree.
A sitting puppy wearing star shaped glasses.
A person in a bright yellow raincoat stands next to another in a sleek black trench coat.
A goblin trading shiny trinkets in a hidden, mystical market.
A cat pounces on a rolling ball.
A person adorned in a luminous turquoise jacket stands beside another person cloaked in a charcoal peacoat.
Two dogs in purple woolly hats playing in the snow.

Figure 5: **Comparison of generated images.** Our reward model enables image generation with better semantic alignment, realism, and visual quality than baselines.

aligned visual representation, from which we derive exhaustive QA pairs with minimal additional cost. Combined with lightweight VLM judge, this approach offers a scalable, low-cost solution for semantic-level reward modeling.

Table 4: Evaluation on the DPG, GenEval and GenAI benchmark. GRPO training with our reward model outperforms both SFT baseline and CLIP-RL models. TO: two objects, P: position, CA: color attribute.

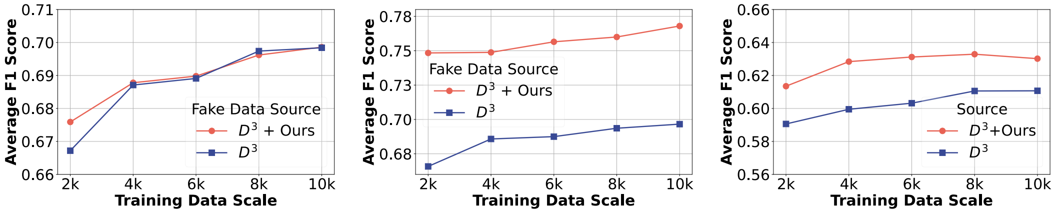
Method	DPG-Bench			GenEval				GenAI-Bench		
	Global	Relation	Overall	TO	P	CA	Overall	Basic	Advanced	All
SimpleAR-0.5B-SFT	85.02	86.59	78.48	0.73	0.22	0.23	0.53	0.74	0.60	0.66
SimpleAR-0.5B-RL (Clip)	86.64	88.51	79.66	**0.82**	0.26	**0.38**	0.59	**0.75**	0.60	0.67
SimpleAR-0.5B-RL (Ours)	**88.46**	**90.13**	**80.50**	0.81	**0.31**	**0.38**	**0.61**	**0.75**	**0.61**	**0.68**

6 IMPROVING GENERATED-CONTENT DETECTION

Advances in *Text-to-Vision generation* underscore the need for effective content moderation (43). Major challenges include the lack of high-quality and diverse datasets and the difficulty of generalizing detection across models *Text-to-Vision generation* (44; 45). GENERATE ANY SCENE addresses these issues by enabling scalable, systematical generation of compositional captions, increasing the diversity and volume of synthetic data. This approach enhances existing datasets by compensating for their limited scope-from realistic to imaginative-and variability.

Experiment setup. To demonstrate GENERATE ANY SCENE's effectiveness in training generated content detectors, we used the D^3 dataset (46) as a baseline. We sampled 5K captioned real and SDv1.4-generated image pairs from D^3 and generated 5K additional images with GENERATE ANY SCENE captions. We trained a ViT–T (47) model with a single-layer linear classifier, and compared models trained with samples solely from D^3 against those trained with samples GENERATE ANY SCENE and D^3.

GENERATE ANY SCENE improves generated content detectors. We evaluate the detector's generalization on the GenImage (48) validation set and images generated using GENERATE ANY SCENE captions. Figure 6 demonstrates that combining GENERATE ANY SCENE-generated images with real-world captioned images consistently enhances detection performance, particularly across cross-model scenarios and diverse visual scenes. More details are in Appendix F.



(a) **In-domain testing (Same Model - SD v1.4)**: Detection results on images generated by SD v1.4 using the GenImage dataset.

(b) **In domain testing (cross-model)**:Average detection results on images generated by multiple models using our captions.

(c) **Out of domain**: Average detection results on images generated by multiple models using captions from the GenImage dataset.

Figure 6: **Results for Application 4: Generated content detector**. Comparison of detection performance across different data scales using D^3 alone versus the combined D^3 + GENERATE ANY SCENE training set in cross-model and cross-dataset scenarios.

7 COMPREHENSIVE EVALUATION WITH GENERATE ANY SCENE

Beyond showcasing GENERATE ANY SCENE in model training, we also show that GENERATE ANY SCENE is a valuable resource for comprehensive and compositional evaluation. Specifically, we synthesize 10K captions for text-to-image, 10K for text-to-video, and 1K for text-to-3D, covering diverse scene structures and content topics. We evaluate 12 text-to-image, 9 text-to-video, and 5 text-to-3D models. Evaluations combine GENERATE ANY SCENE synthetic scene graphs with existing metrics (e.g., CLIP Score (49), VQA Score (39), TIFA Score (23; 31)) to assess semantic similarity, faithfulness, and human preference alignment. Our key findings include: (1) DiT-backbone text-to-image models align more closely with input captions than UNet-backbone models. (2) Text-to-video models struggle with balancing dynamics and consistency, while both text-to-video and text-to-3D models show notable gaps in human preference alignment. Except for aggregating quantitative results, we also leverage GENERATE ANY SCENE 's controllable captioning to evaluate models on fine-grained factors: perplexity, scene complexity, commonsense reasoning, and content category variation for case study.

Overall, GENERATE ANY SCENE yields stable, human-aligned rankings across T2I/T2V/T2-3D. Through broad, controllable coverage of objects, attributes, relations, and categories, it serves as a compositional stress test that reliably exposes plausibility gaps, category brittleness, and long-tail concept failures in current models (see Appendix A).

8 RELATED WORK

***Text-to-Vision generation* models.** *Text-to-Image generation* advances are driven by diffusion models and LLMs. Some open-source models (22; 50; 51; 52; 53; 54) use UNet backbones to refine images iteratively. In parallel, Diffusion Transformers (DiTs) architectures(55; 56; 57; 58) have emerged as a better alternative in capturing long-range dependencies and improving coherence. Proprietary models like DALL-E 3 (3) and Imagen 3 (59) still set the state-of-the-art. Based on *Text-to-Image generation* method, *Text-to-Video generation* models typically utilize time-aware architectures to ensure temporal coherence across frames (60; 61; 62; 63; 64; 65; 66; 67). In *Text-to-3D generation*, recent proposed models (4; 68; 69; 70; 71) integrate the diffusion models with Neural Radiance Fields (NeRF) rendering to generate diverse 3D objects. Recent studies (26; 42; 72; 73) have also explored the integration of image generation into a unified multimodal language model (MLM) framework based on auto-regressive transformer architectures, demonstrating promising improvements in both performance and efficiency.

Synthetic captions for *Text-to-Vision generation*. Captions for *Text-to-Vision generation* models vary greatly in diversity, complexity, and compositionality. This variation makes it challenging and costly to collect large-scale and diverse captions written by humans. Consequently, synthetic captions have been widely used for both training (74; 38; 75; 76; 8; 77; 78; 79) and evaluation purposes (7). For example, training methods like LLM-Grounded Diffusion (74) leverage LLM-generated captions to enhance the model's understanding and alignment with human instruction. For evaluation, benchmarks such as T2I-CompBench (7) and T2V-CompBench (8) utilize benchmarks generated by LLMs. However, LLMs are hard to control and may introduce exhibit systematic bias. In this work, we propose a programmatic scene graph-based data engine that can generate infinitely diverse captions for improving *Text-to-Vision generation* models.

Finetuning techniques for *Text-to-Vision generation*. To accommodate the diverse applications and personalization needs in text-to-vision models, numerous fine-tuning techniques have been developed. LoRA (40) reduces fine-tuning costs via low-rank weight updates, while Textual Inversion (80; 81) introduces new word embeddings for novel concepts without altering core parameters. DreamBooth (82) adapts models to specific subjects or styles using a few personalized images, and DreamSync (38) enables models to self-improve by learning from their own high-quality outputs. Recently, RLHF (26; 41; 42) in *Text-to-Vision generation* has shown promise as an efficient fine-tuning strategy. In this work, we use several fine-tuning techniques with GENERATE ANY SCENE to improve *Text-to-Vision generation* models.

9 CONCLUSION

We present GENERATE ANY SCENE, a system leveraging scene graph programming to generate diverse and compositional synthetic captions for *Text-to-Vision generation* tasks. It extends beyond existing real-world caption datasets to include comprehensive scenes and even implausible scenarios. To demonstrate the effectiveness of GENERATE ANY SCENE, we explore four applications: (1) self-improvement by iteratively optimizing models, (2) distillation of proprietary model strengths into open-source models, (3) a scene-graph-based efficient reward model within the GRPO, and (4) robust content moderation with diverse synthetic data. GENERATE ANY SCENE highlights the importance of synthetic data in improving *Text-to-Vision generation*, and addresses the need to systematically define and scalably produce the space of visual scenes.

References

[1] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. 2024. *URL https://openai.com/research/video-generation-models-as-world-simulators*, 3, 2024.

[2] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[4] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024.

[5] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James T. Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *ArXiv*, abs/2310.00426, 2023.

[6] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024.

[7] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *ArXiv*, abs/2307.06350, 2023.

[8] Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. *ArXiv*, abs/2407.14505, 2024.

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

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

[11] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

[12] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[13] Zejian Li, Chenye Meng, Yize Li, Ling Yang, Shengyuan Zhang, Jiarui Ma, Jiayi Li, Guang Yang, Changyuan Yang, Zhiyuan Yang, et al. Laion-sg: An enhanced large-scale dataset for training complex image-text models with structural annotations. *arXiv preprint arXiv:2412.08580*, 2024.

[14] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[15] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015.

[16] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2856–2865, 2021.

[17] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018.

[18] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247, 2020.

[19] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987.

[20] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 852–869. Springer, 2016.

[21] Jieyu Zhang, Weikai Huang, Zixian Ma, Oscar Michel, Dong He, Tanmay Gupta, Wei-Chiu Ma, Ali Farhadi, Aniruddha Kembhavi, and Ranjay Krishna. Task me anything. In *Advances in neural information processing systems*, 2024.

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[23] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering, 2023.

[24] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155, 2022.

[25] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

[26] Junke Wang, Zhi Tian, Xun Wang, Xinyu Zhang, Weilin Huang, Zuxuan Wu, and Yu-Gang Jiang. Simplear: Pushing the frontier of autoregressive visual generation through pretraining, sft, and rl. *arXiv preprint arXiv:2504.11455*, 2025.

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[28] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment, 2024.

[29] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[30] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[31] Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. *ArXiv*, abs/2310.18235, 2023.

[32] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[33] Wikipedia Contributors. Lists of colors. https://en.wikipedia.org/wiki/Lists_of_colors, 2024. Accessed: 2024-11-09.

[34] Jae Sung Park, Zixian Ma, Linjie Li, Chenhao Zheng, Cheng-Yu Hsieh, Ximing Lu, Khyathi Chandu, Quan Kong, Norimasa Kobori, Ali Farhadi, Yejin Choi, and Ranjay Krishna. Synthetic visual genome. In *CVPR*, 2025.

[35] Alejandro López-Cifuentes, Marcos Escudero-Vinolo, Jesús Bescós, and Álvaro García-Martín. Semantic-aware scene recognition. *Pattern Recognition*, 102:107256, 2020.

[36] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11287–11297, 2021.

[37] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Emily Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Genai-bench: A holistic benchmark for compositional text-to-visual generation. In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*, 2024.

[38] Jiao Sun, Deqing Fu, Yushi Hu, Su Wang, Royi Rassin, Da-Cheng Juan, Dana Alon, Charles Herrmann, Sjoerd van Steenkiste, Ranjay Krishna, and Cyrus Rashtchian. Dreamsync: Aligning text-to-image generation with image understanding feedback. *ArXiv*, abs/2311.17946, 2023.

[39] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *ArXiv*, abs/2404.01291, 2024.

[40] J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021.

[41] Lixue Gong, Xiaoxia Hou, Fanshi Li, Liang Li, Xiaochen Lian, Fei Liu, Liyang Liu, Wei Liu, Wei Lu, Yichun Shi, et al. Seedream 2.0: A native chinese-english bilingual image generation foundation model. *arXiv preprint arXiv:2503.07703*, 2025.

[42] Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *arXiv preprint arXiv:2505.00703*, 2025.

[43] Gan Pei, Jiangning Zhang, Menghan Hu, Zhenyu Zhang, Chengjie Wang, Yunsheng Wu, Guangtao Zhai, Jian Yang, Chunhua Shen, and Dacheng Tao. Deepfake generation and detection: A benchmark and survey. *arXiv preprint arXiv:2403.17881*, 2024.

[44] Tianyi Wang, Xin Liao, Kam Pui Chow, Xiaodong Lin, and Yinglong Wang. Deepfake detection: A comprehensive survey from the reliability perspective. *ACM Computing Surveys*, 2024.

[45] Achhardeep Kaur, Azadeh Noori Hoshyar, Vidya Saikrishna, Selena Firmin, and Feng Xia. Deepfake video detection: challenges and opportunities. *Artificial Intelligence Review*, 57(6):1–47, 2024.

[46] Lorenzo Baraldi, Federico Cocchi, Marcella Cornia, Alessandro Nicolosi, and Rita Cucchiara. Contrasting deepfakes diffusion via contrastive learning and global-local similarities. *arXiv preprint arXiv:2407.20337*, 2024.

[47] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers. In *European conference on computer vision*, pages 68–85. Springer, 2022.

[48] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in Neural Information Processing Systems*, 36, 2024.

[49] Tuhin Chakrabarty, Kanishk Singh, Arkadiy Saakyan, and Smaranda Muresan. Learning to follow object-centric image editing instructions faithfully. *ArXiv*, abs/2310.19145, 2023.

[50] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[51] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024.

[52] Pablo Pernias, Dominic Rampas, Mats L Richter, Christopher J Pal, and Marc Aubreville. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. *arXiv preprint arXiv:2306.00637*, 2023.

[53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[54] DeepFloyd Lab at StabilityAI. DeepFloyd IF: a novel state-of-the-art open-source text-to-image model with a high degree of photorealism and language understanding. https://www.deepfloyd.ai/deepfloyd-if, 2023. Retrieved on 2023-11-08.

[55] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

[56] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-\alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

[57] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-\sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692*, 2024.

[58] Black Forest Labs. Flux.1: Advanced text-to-image models, 2024. Accessed: 2024-11-10.

[59] Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, Zach Eaton-Rosen, et al. Imagen 3. *arXiv preprint arXiv:2408.07009*, 2024.

[60] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[61] Fu-Yun Wang, Zhaoyang Huang, Xiaoyu Shi, Weikang Bian, Guanglu Song, Yu Liu, and Hongsheng Li. Animatelcm: Accelerating the animation of personalized diffusion models and adapters with decoupled consistency learning. *arXiv preprint arXiv:2402.00769*, 2024.

[62] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023.

[63] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

[64] Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. Freeinit: Bridging initialization gap in video diffusion models. In *European Conference on Computer Vision*, pages 378–394. Springer, 2025.

[65] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024.

[66] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[67] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, March 2024.

[68] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[69] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023.

[70] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12663–12673, 2023.

[71] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023.

[72] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024.

[73] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

[74] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *Trans. Mach. Learn. Res.*, 2024, 2023.

[75] Jialu Li, Jaemin Cho, Yi-Lin Sung, Jaehong Yoon, and Mohit Bansal. Selma: Learning and merging skill-specific text-to-image experts with auto-generated data. *ArXiv*, abs/2403.06952, 2024.

[76] Rui Zhao, Hangjie Yuan, Yujie Wei, Shiwei Zhang, Yuchao Gu, Lin Hao Ran, Xiang Wang, Zhangjie Wu, Junhao Zhang, Yingya Zhang, and Mike Zheng Shou. Evolvedirector: Approaching advanced text-to-image generation with large vision-language models, 2024.

[77] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In *NeurIPS Datasets and Benchmarks*, 2021.

[78] Song Wen, Guian Fang, Renrui Zhang, Peng Gao, Hao Dong, and Dimitris Metaxas. Improving compositional text-to-image generation with large vision-language models. *ArXiv*, abs/2310.06311, 2023.

[79] Xindi Wu, Dingli Yu, Yangsibo Huang, Olga Russakovsky, and Sanjeev Arora. Conceptmix: A compositional image generation benchmark with controllable difficulty. *ArXiv*, abs/2408.14339, 2024.

[80] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6038–6047, 2022.

[81] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *ArXiv*, abs/2208.01618, 2022.

[82] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, 2022.

[83] Spencer Sterling. zeroscope_v2_576w, 2023. Accessed: 2024-11-10.

[84] Y.C. Guo, Y.T. Liu, R. Shao, C. Laforte, V. Voleti, G. Luo, C.H. Chen, Z.X. Zou, C. Wang, Y.P. Cao, and S.H. Zhang. threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio, 2023.

[85] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation, 2023.

[86] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation, 2023.

[87] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

[88] Kling AI. Kling ai text-to-video. https://klingai.com/text-to-video/new, 2025. Accessed May 23, 2025.

[89] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[90] Meshy AI. Meshy ai – text-to-3d, image-to-3d, and text-to-texture 3d model generator. https://www.meshy.ai, 2025. Accessed May 23, 2025.

[91] Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. Vera: A general-purpose plausibility estimation model for commonsense statements, 2023.

[92] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

[93] Giuseppe Vecchio and Valentin Deschaintre. Matsynth: A modern pbr materials dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22109–22118, 2024.

[94] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3479–3487, 2015.

[95] Jia Xue, Hang Zhang, Kristin Dana, and Ko Nishino. Differential angular imaging for material recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 764–773, 2017.

[96] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[97] Zhe Xu, Dacheng Tao, Ya Zhang, Junjie Wu, and Ah Chung Tsoi. Architectural style classification using multinomial latent logistic regression. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 600–615. Springer, 2014.

[98] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[99] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015.

[100] Colby Crawford. 1000 cameras dataset. `https://www.kaggle.com/datasets/crawford/1000-cameras-dataset`, 2018. Accessed: 2024-11-09.

[101] Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv:2210.14896 [cs]*, 2022.

[102] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.

[103] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[104] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *ICCV*, 2023.

A EVALUATING *Text-to-Vision generation* MODELS WITH GENERATE ANY SCENE

A.1 EXPERIMENT SETTINGS

Models. We conduct experiments on 12 *Text-to-image* models (54; 50; 22; 51; 52; 55; 56; 57; 58; 3), 9 *Text-to-Video* models (63; 83; 62; 60; 61; 64; 67; 66; 65), and 5 *Text-to-3D* models (68; 71; 69; 4; 70).

- For *Text-to-Image generation*, we select a range of open-source models, including those utilizing UNet backbones, such as *DeepFloyd IF* (54), *SDv2.1* (22), *SDXL* (50), *Playground v2.5* (51), and *Wuerstchen v2* (52), as well as models with DiT backbones, including *SD3 Medium* (55), *PixArt-α* (56), *PixArt-Σ* (57), *FLUX.1-schnell* (58), *FLUX.1-dev* (58), and FLUX 1. Closed-source models, such as *DaLL-E 3* (3) and *FLUX1.1 PRO* (58), are also assessed to ensure a comprehensive comparison. All models are evaluated at a resolution of 1024 × 1024 pixels.
- For *Text-to-Video generation*, we select nine open-source models: *ModelScope* (63), *ZeroScope* (83), *Text2Video-Zero* (62), *CogVideoX-2B* (66), *VideoCrafter2* (65), *AnimateLCM* (61), *AnimateDiff* (60), *FreeInit* (64), and *Open-Sora 1.2* (67). We standardize the frame length to 16 across all video models for fair comparisons.
- For *Text-to-3D generation*, we evaluate five recently proposed models: *SJC* (69), *DreamFusion* (68), *Magic3D* (71), *Latent-NeRF* (70), and *ProlificDreamer* (4). We employ the implementation and configurations provided by ThreeStudio (84) and generate videos by rendering from 120 viewpoints. To accelerate inference, we omit the refinement stage. For *Magic3D* and *DreamFusion*, we respectively use *DeepFloyd IF* and *SDv2.1* as their 2D backbones.

Metrics. Across all *Text-to-Vision generation* tasks, we use *Clip Score* (49) (semantic similarity), *VQA Score* (39) (faithfulness), *TIFA Score* (23; 31) (faithfulness), *Pick Score* (85) (human preference), and *ImageReward Score* (86) (human preference) as general metrics:

- *Clip Score*: Assesses semantic similarity between images and text.
- *VQA Score* and *TIFA Score*: Evaluate faithfulness by generating question-answer pairs and measuring answer accuracy from images.
- *Pick Score* and *ImageReward Score*: Capture human preference tendencies.

We also use metrics in VBench (87) to evaluate *Text-to-Video generation* models on fine-grained dimensions, such as consistency and dynamics, providing detailed insights into video performance.

For *Text-to-Video generation* and *Text-to-3D generation* tasks:

- We calculate *Clip Score*, *Pick Score*, and *ImageReward Score* on each frame, then average these scores across all frames to obtain an overall video score.
- For *VQA Score* and *TIFA Score*, we handle *Text-to-Video generation* and *Text-to-3D generation* tasks differently:
  - ○ In *Text-to-Video generation* tasks, we uniformly sample four frames from the 16-frame sequence and arrange them in a 2 × 2 grid image.
  - ○ For *Text-to-3D generation* tasks, we render images at 45-degree intervals from nine different viewpoints and arrange them in a 3 × 3 grid.

This sampling approach optimizes inference speed without affecting score accuracy (39).

Synthetic captions. We evaluate our *Text-to-Image generation* and *Text-to-Video generation* models on 10K randomly generated captions, with scene graph complexity ranging from 3 to 12 and scene attributes from 0 to 5, using unrestricted metadata. The captions exhibit an average graph degree of 1.15, with values spanning from 0.0 to 0.8. The mean number of connected components per scene graph is 3.51, ranging from 1 to 11. For *Text-to-3D generation* models, due to their limitations in handling complex captions and time-intensive generation, we restrict scene graph complexity to 1-3, scene attributes to 0-2, and evaluate on 1K captions.

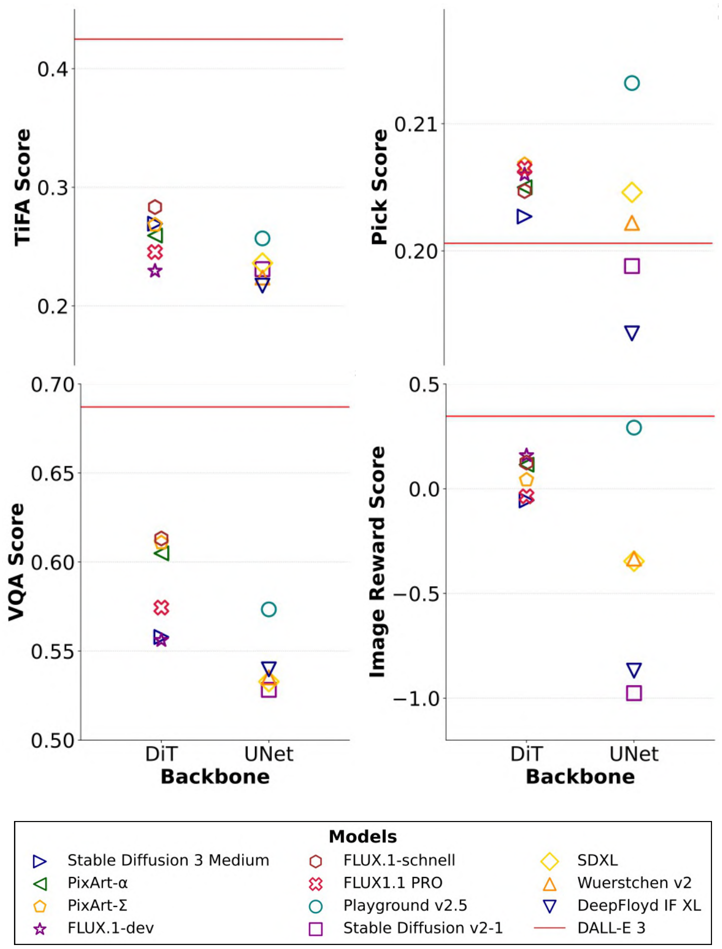


Figure 7: Comparative evaluation of *Text-to-Image generation* models across different backbones (DiT and UNet) using multiple metrics: *TIFA Score*, *Pick Score*, *VQA Score*, and *ImageReward Score*.

A.2 Overall results

We evaluate *Text-to-Image generation*, *Text-to-Video generation*, and *Text-to-3D generation* models on Generate Any Scene.

Table 5: Overall performance of *Text-to-Image generation* models over 10K Generate Any Scene captions. †Evaluated on a 1K caption subset due to inference cost constraints.

Model	clip score	pick score	vqa score	tifa score	image reward score
Playground v2.5 [51]	0.2581	0.2132	0.5734	0.2569	0.2919
Stable Diffusion v2-1 [22]	0.2453	0.1988	0.5282	0.2310	-0.9760
SDXL [50]	0.2614	0.2046	0.5328	0.2361	-0.3463
Wuerstchen v2 [52]	0.2448	0.2022	0.5352	0.2239	-0.3339
DeepFloyd IF XL [54]	0.2396	0.1935	0.5397	0.2171	-0.8687
Stable Diffusion 3 Medium [55]	0.2527	0.2027	0.5579	0.2693	-0.0557
PixArt-α [56]	0.2363	0.2050	0.6049	0.2593	0.1149
PixArt-Σ [57]	0.2390	0.2068	0.6109	0.2683	0.0425
FLUX.1-dev [58]	0.2341	0.2060	0.5561	0.2295	0.1588
FLUX.1-schnell [58]	0.2542	0.2047	0.6132	0.2833	0.1251
FLUX1.1 PRO [58]†	0.2315	0.2065	0.5744	0.2454	-0.0361
Dalle-3 [3]	0.2518	0.2006	0.6871	0.4249	0.3464

***Text-to-Image generation* results. (Figure 7, Table 5)**

1. DiT-backbone models outperform UNet-backbone models on *VQA Score* and *TIFA Score*, indicating greater faithfulness and comprehensiveness to input captions.
2. Despite using a UNet architecture, *Playground v2.5* achieves higher *Pick Score* and *ImageReward Score* scores than other open-source models. We attribute this to *Playground v2.5* 's alignment with human preferences achieved during training.
3. The closed-source model *DaLL-E 3* maintains a significant lead in *VQA Score*, *TIFA Score*, and *ImageReward Score*, demonstrating strong faithfulness and alignment with captions across generated content.

***Text-to-Video generation* results. (Table 6,7)**

Table 6: Overall performance of open-source *Text-to-Video generation* models over 10K GENERATE ANY SCENE captions. Red Cell is the highest score. Yellow Cell is the second highest score.†Close-source models are evaluated on a 1K caption subset due to high inference cost.

Model	clip score	pick score	image reward score	VQA score	TiFA score
VideoCraft2 (65)	0.2398	0.1976	-0.4202	0.5018	0.2466
AnimateLCM (61)	0.2450	0.1987	-0.5754	0.4816	0.2176
AnimateDiff (60)	0.2610	0.1959	-0.7301	0.5255	0.2208
Open-Sora 1.2 (67)	0.2259	0.1928	-0.6277	0.5519	0.2414
FreeInit (64)	0.2579	0.1950	-0.9335	0.5123	0.2047
ModelScope (63)	0.2041	0.1886	-1.9172	0.3840	0.1219
Text2Video-Zero (62)	0.2539	0.1933	-1.2050	0.4753	0.1952
CogVideoX-2B (66)	0.2038	0.1901	-1.2301	0.4585	0.1997
ZeroScope (83)	0.2289	0.1933	-1.1599	0.4892	0.2388
KLING 1.6 (88)†	0.2215	0.1985	-0.3419	0.5307	0.2802
Wanx 2.1 (89)†	0.2308	0.1969	-0.1418	0.5970	0.3328

Table 7: Overall performance of open-source *Text-to-Video generation* models over 10K GENERATE ANY SCENE captions with VBench metrics. Red Cell is the highest score. Blue Cell is the lowest score.

Model	subject consistency	background consistency	motion smoothness	dynamic degree	aesthetic quality	imaging quality
Open-Sora 1.2	0.9964	0.9907	0.9973	0.0044	0.5235	0.6648
Text2Video-Zero	0.8471	0.9030	0.8301	0.9999	0.4889	0.7018
VideoCraft2	0.9768	0.9688	0.9833	0.3556	0.5515	0.6974
AnimateDiff	0.9823	0.9733	0.9859	0.1406	0.5427	0.5830
FreeInit	0.9581	0.9571	0.9752	0.4440	0.5200	0.5456
ModelScope	0.9795	0.9831	0.9803	0.1281	0.3993	0.6494
AnimateLCM	0.9883	0.9802	0.9887	0.0612	0.6323	0.6977
CogVideoX-2B	0.9583	0.9602	0.9823	0.4980	0.4607	0.6098
ZeroScope	0.9814	0.9811	0.9919	0.1670	0.4582	0.6782

1. Open-source text-to-video models face challenges in balancing dynamics and consistency (Table 7). This is especially evident in *Open-Sora 1.2*, which achieves high consistency but minimal dynamics, and *Text2Video-Zero*, which excels in dynamics but suffers from frame inconsistency.
2. All models exhibit negative *ImageReward Score* (Table 6), suggesting a lack of human-preferred visual appeal in the generated content, even in cases where certain models demonstrate strong semantic alignment.

3. As expected, SOTA close-source text-to-video models outperform others overall, particularly in image reward, VQA score, and TIFA score. This indicates their superior alignment with human preferences, as well as stronger faithfulness and compositional capabilities in generation.

4. Among open-source models, *VideoCrafter2* strikes a balance across key metrics, leading in human-preference alignment, faithfulness, consistency, and dynamic.

Text-to-3D generation **results. (Table 8)**

Table 8: Overall performance of *Text-to-3D generation* models over 1K GENERATE ANY SCENE captions. †Evaluated on a 100 caption subset due to high inference cost.

Model	**clip score**	**pick score**	**vqa score**	**tifa score**	**image reward score**
Latent-NeRF (70)	0.2115	0.1910	0.4767	0.2216	-1.5311
DreamFusion-sd (68)	0.1961	0.1906	0.4421	0.1657	-1.5582
Magic3D-sd (71)	0.1947	0.1903	0.4193	0.1537	-1.6327
SJC (69)	0.2191	0.1915	0.5015	0.2563	-1.4370
DreamFusion-IF (68)	0.1828	0.1857	0.3872	0.1416	-1.9353
Magic3D-IF (71)	0.1919	0.1866	0.4039	0.1537	-1.8465
ProlificDreamer (4)	0.2125	**0.1940**	**0.5411**	0.2704	-1.2774
Meshy-4 (90)†	**0.2163**	0.1922	0.5290	**0.2908**	**-1.0496**

1. Among open-source models, *ProlificDreamer* outperforms other models, particularly in *ImageReward Score*, *VQA Score* and *TIFA Score*.

2. All models receive negative *ImageReward Score* scores, highlighting a significant gap between human preference and current *Text-to-3D generation* generation capabilities.

3. Meshy-4 demonstrates overall superior performance compared to all open-source models, especially in terms of *Clip Score*, *TIFA Score* and *ImageReward Score*, reflecting its strengths in semantic generation and human preference alignment.

A.3 VALIDATION OF PHRASING ROBUSTNESS AND HUMAN ALIGNMENT

To assess robustness to linguistic variation and to verify that automated metrics reflect human preferences, we conduct two focused studies.

A.3.1 PHRASING ROBUSTNESS VIA PARAPHRASING

Setup. We sample 100 scene graphs from the 10K benchmark while preserving the distribution of object counts, relation density, and attribute complexity. For each graph, GPT-4o generates a linguistically varied yet graph-faithful caption using the prompt below.

Paraphrasing Prompt

```
You are given a scene graph in JSON format, where:
- "nodes" contain objects and their attributes,
- "edges" describe relationships between objects or link attributes
to objects.

Your task:
1.  Understand the semantic meaning of each node and edge.
2.  Convert the graph into a natural language caption that describes
the entire scene.
3.  Include all objects, attributes, and relations from the graph,
and strictly follow the graph structure.
4.  Do not introduce new objects or relationships not present in the
graph.
Input:  {scene_graph}
```

We then re-score all models with *VQA Score* under these paraphrased captions. Results are listed in Table 9.

Table 9: Paraphrase robustness: VQA Score and ranks on 100 graphs.

Model	Orig. Score	Para. Score	Diff	Orig. Rank	Para. Rank
DALLE-3	0.6871	0.7542	+0.0671	1	1
FLUX.1-schnell	0.6132	0.6648	+0.0516	2	2
PixArt-Σ	0.6109	0.6159	+0.0050	3	3
PixArt-α	0.6049	0.6043	-0.0006	4	4
Playground v2.5	0.5734	0.5075	-0.0659	5	8
Stable Diffusion 3	0.5579	0.5140	-0.0439	6	7
FLUX.1-dev	0.5561	0.5024	-0.0537	7	9
DeepFloyd IF XL	0.5397	0.5606	+0.0209	8	5
Wuerstchen v2	0.5352	0.5014	-0.0338	9	10
SDXL	0.5328	0.5322	-0.0006	10	6
SD v2-1	0.5282	0.4961	-0.0321	11	11

Findings. The Pearson correlation coefficient between model rankings on programmatic versus paraphrased captions is **0.9232**, indicating a very strong positive correlation.

This validation study demonstrates strong consistency between the two approaches. Importantly, the top-performing models (*DaLL-E 3*, *FLUX.1-schnell*, *PixArt-Σ*, *PixArt-α*) maintain their rankings across both evaluation conditions, while the relative ordering of models remains largely consistent. This high correlation validates that our programmatic approach produces rankings that are generalizable and not artifacts of the templated caption generation. The slight variations observed (e.g., some mid-tier models showing small rank changes) are within expected bounds and do not affect the overall conclusions about model capabilities.

A.3.2 Human Alignment Study

Setup. We evaluate six representative models (*DaLL-E 3*, *FLUX.1-schnell*, *PixArt-Σ*, *Playground v2.5*, *SD3 Medium*, *SDv2.1*) with diverse performance characteristics and recruit 3 human evaluators. Three independent evaluators each assess 40 caption–image groups, with 10 shared overlapping groups across all evaluators to measure inter-annotator agreement. Evaluators ranked the generated images based on both relevance to the caption and overall visual quality. We show the rankings in Table 10.

Findings

Inter-annotator reliability. The 3 evaluators showed strong agreement on the 10 shared samples, with a Spearman correlation coefficient of **0.962**, demonstrating consistent human judgment criteria.

Table 10: Human vs. VQA rankings (lower is better).

Model	VQA Rank	Human Avg. Rank
DaLL-E 3	1	1
FLUX.1-schnell	2	2
PixArt-Σ	3	4
Playground v2.5	4	3
SD3 Medium	5	5
SDv2.1	6	6

Human–metric alignment. The correlation between human rankings and our *VQA Score* rankings is **0.918**, indicating strong alignment between automated and human evaluation:

This study validates that our VQA Score-based rankings closely align with human preferences. The consistency between automated metrics and human judgment strengthens confidence in our benchmark's ability to assess model performance in a manner that reflects human perception.

A.4 More Analysis with Generate Any Scene

With Generate Any Scene, we can generate infinitely diverse and highly controllable captions. Using Generate Any Scene, we conduct several analyses to provide insights into the performance of today's *Text-to-Vision generation* models.

A.4.1 Performance analysis across caption properties

In this section, we delve into how model performance varies with respect to distinct properties of Generate Any Scene captions. While Generate Any Scene is capable of generating an extensive diversity of captions, these outputs inherently differ in key characteristics that influence model evaluation. Specifically, we examine three properties of the caption: Commonsense, Perplexity, and Scene Graph Complexity (captured as the number of elements in the captions). These properties are critical in understanding how different models perform across a spectrum of linguistic and semantic challenges presented by captions with varying levels of coherence, plausibility, and compositional richness.

Perplexity. (Figure 8) Perplexity is a metric used to measure a language model's unpredictability or uncertainty in generating a text sequence. A higher perplexity value indicates that the sentences are less coherent or less likely to be generated by the model.

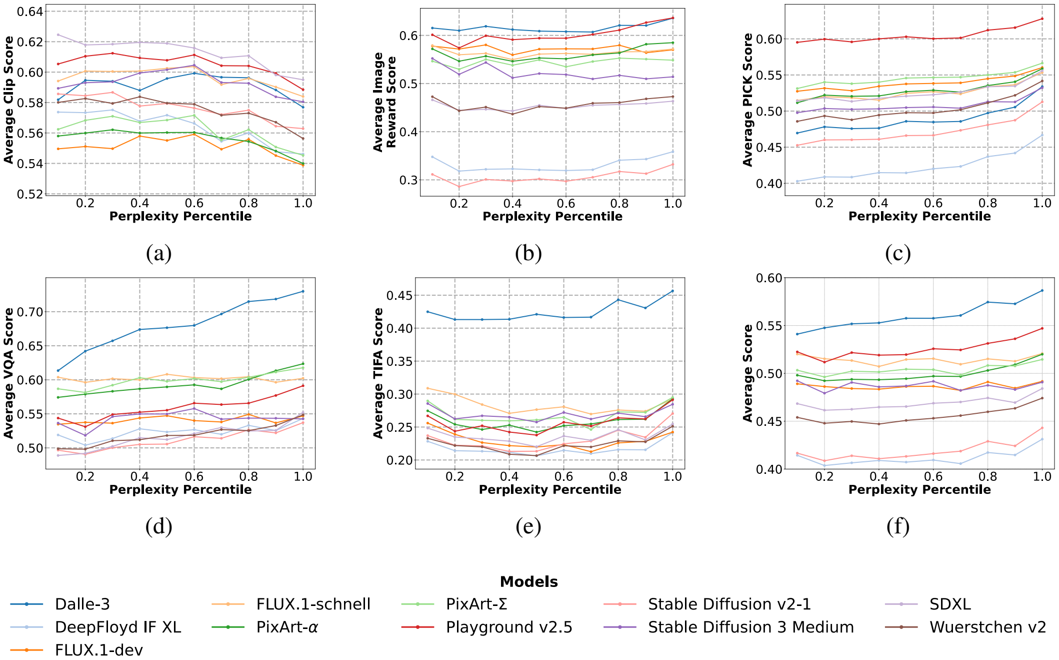
As shown in Figure 8, From left to right, when perplexity increases, indicating that the sentences become less reasonable and less typical of those generated by a language model, we observe no clear or consistent trends across all models and metrics. This suggests that the relationship between perplexity and model performance varies depending on the specific model and evaluation metric.

Commonsense. (Figure 9) Commonsense is an inherent property of text. We utilize the Vera Score ([91]), a metric generated by a fine-tuned LLM to evaluate the text's commonsense level.

As shown in Figure 9, from left to right, as the Vera Score increases—indicating that the captions exhibit greater commonsense reasoning—we observe a general improvement in performance across all metrics and models, except for *Clip Score*. This trend underscores the correlation between commonsense-rich captions and enhanced model performance.

Element Numbers (Complexity of Scene Graph). (Figure 10) Finally, we evaluate model performance across total element numbers in the captions, which represent the complexity of scene graphs (objects + attributes + relations).

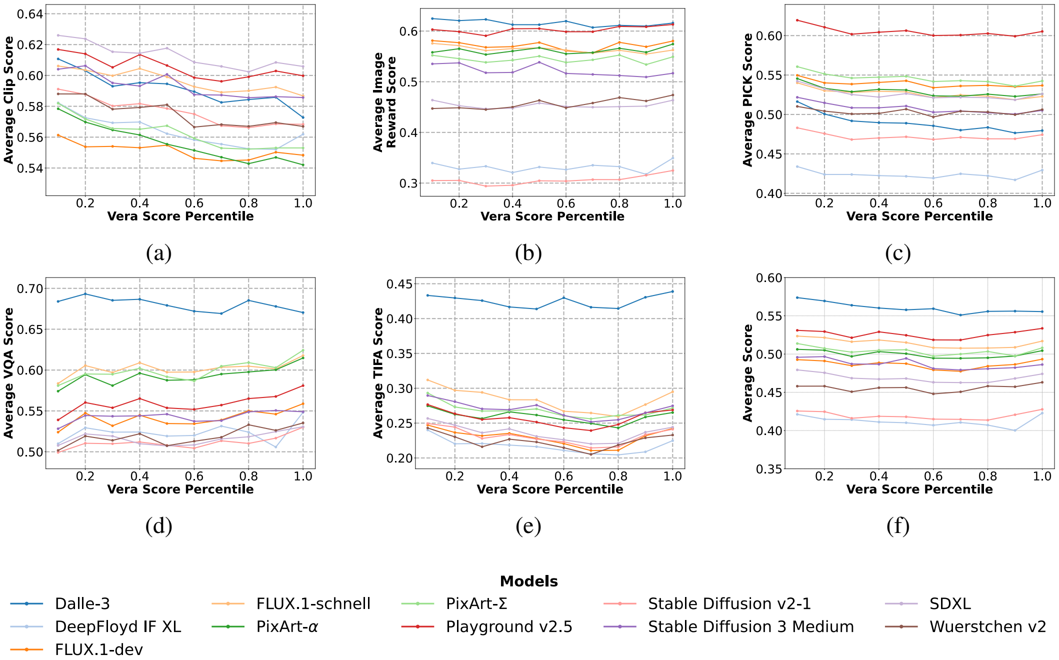
From left to right, the complexity of scene graphs becomes higher, reflecting more compositional and intricate captions. Across most metrics and models, we observe a noticeable performance decline as the scene graphs become more complex. However, an interesting exception is observed in the



(a) (b) (c)

(d) (e) (f)

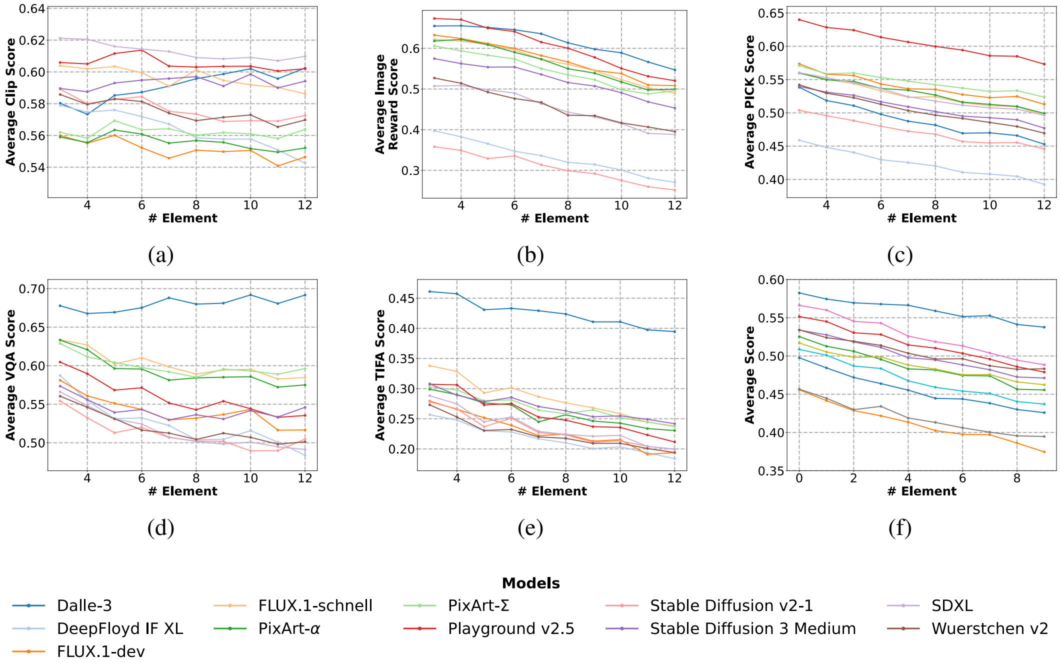
Figure 8: Average performance of models across different percentiles of perplexity of captions, evaluated on various metrics. From left to right, the perplexity decreases, indicating captions that are progressively more reasonable and easier for the LLM to generate.



(a) (b) (c)

(d) (e) (f)

Figure 9: Average performance of models across different percentiles of Vera Score for captions, evaluated on various metrics. From left to right, the Vera Score decreases, indicating captions that exhibit less commonsense reasoning and are more likely to describe implausible scenes.



(a) (b) (c)

(d) (e) (f)

Figure 10: Average performance of models across different numbers of elements (objects + attributes + relations) in the scene graph (complexity of the scene graph) of the captions, evaluated on various metrics. From left to right, as the number of elements (complexity) increases, the scene graphs become more complicated and compositional.

performance of *DaLL-E 3*. Unlike other models, *DaLL-E 3* performs exceptionally well on *VQA Score* and *TIFA Score*, particularly on *VQA Score*, where it even shows a slight improvement as caption complexity increases. This suggests that *DaLL-E 3* may have a unique capacity to handle complex and compositional captions effectively.

A.4.2 Analysis on different metrics

Compared with most LLM and VLM benchmarks that use multiple-choice questions and accuracy as metrics. There is no universal metric in evaluating *Text-to-Vision generation* models. Researchers commonly used model-based metrics like *Clip Score*, *VQA Score*, etc. Each of these metrics is created and fine-tuned for different purposes with bias. Therefore, we also analysis on different metrics.

***Clip Score* isn't a universal metric.** *Clip Score* is one of the most widely used metrics in *Text-to-Vision generation* for evaluating the alignment between visual content and text. However, our analysis reveals that *Clip Score* is not a perfect metric and displays some unusual trends. For instance, as shown in Figures 8, 9, and 10, we compute the perplexity across 10K captions used in our study, where higher perplexity indicates more unpredictable or disorganized text. Interestingly, unlike other metrics, *Clip Score* decreases as perplexity lowers, suggesting that *Clip Score* tends to favor more disorganized text. This behavior is counterintuitive and highlights the potential limitations of using *Clip Score* as a robust alignment metric.

Limitations of human preference-based metrics. We use two metrics fine-tuned using human preference data: *Pick Score* and *ImageReward Score*. However, we found that these metrics exhibit a strong bias toward the data on which they were fine-tuned. For instance, as shown in Table 5, *Pick Score* assigns similar scores across all models, failing to provide significant differentiation or meaningful insights into model performance. In contrast, *ImageReward Score* demonstrates clearer preferences, favoring models such as *DaLL-E 3* and *Playground v2.5*, which incorporated human-alignment techniques during their training. However, this metric shows a significant drawback:

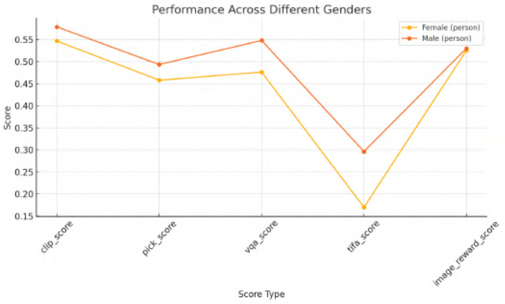


Figure 11: Average performance scores of all models across different genders evaluated using various metrics.

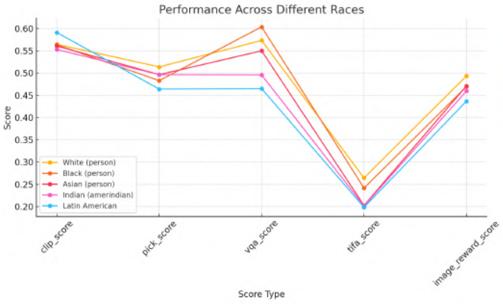


Figure 12: Average performance scores of all models across different races evaluated using various metrics.

it assigns disproportionately large negative scores to models like *SDv2.1*, indicating a potential over-sensitivity to alignment mismatches. Such behavior highlights the limitations of these metrics in providing fair and unbiased evaluations across diverse model architectures.

***VQA Score* and *TIFA Score* are relative reliable metrics.** Among the evaluated metrics, *VQA Score* and *TIFA Score* stand out by assessing model performance on VQA tasks, rather than relying solely on subjective human preferences. This approach enhances the interpretability of the evaluation process. Additionally, we observed that the results from *VQA Score* and *TIFA Score* show a stronger correlation with other established benchmarks. Based on these advantages, we recommend prioritizing these two metrics for evaluation. However, it is important to note that their effectiveness is constrained by the limitations of the VQA models utilized in the evaluation.

A.4.3 Fairness analysis

We evaluate fairness by examining the model's performance across different genders and races. Specifically, we calculate the average performance for each node and its associated child nodes within the taxonomy tree constructed for objects. For example, the node "females" includes child nodes such as "waitresses," and their combined performance is considered in the analysis.

Gender. In gender, we observe a notable performance gap between females and males, as could be seen from Figure 11, Models are better at generating male concepts.

Race. There are also performance gaps in different races. From Figure 12, we found that "white (person)" and "black (person)" perform better than "asian (person)", "Indian (amerindian)", and "Latin American".

A.4.4 Correlation of Generate Any Scene with other *Text-to-Vision generation* benchmarks

The Generate Any Scene benchmark uniquely relies entirely on synthetic captions to evaluate models. To assess the transferability of these synthetic captions, we analyzed the consistency in model rankings across different benchmarks ([79]; [37]; [92]). Specifically, we identified the overlap of models evaluated by two benchmarks and computed the Spearman correlation coefficient between their rankings.

As shown in the figure 13, Generate Any Scene demonstrates a strong correlation with other benchmarks, such as Conceptmix ([79]) and GenAI Bench ([37]), indicating the robustness and reliability of Generate Any Scene's synthetic caption-based evaluations. This suggests that the synthetic captions generated by Generate Any Scene can effectively reflect model performance trends, aligning closely with those observed in benchmarks using real-world captions or alternative evaluation methods.

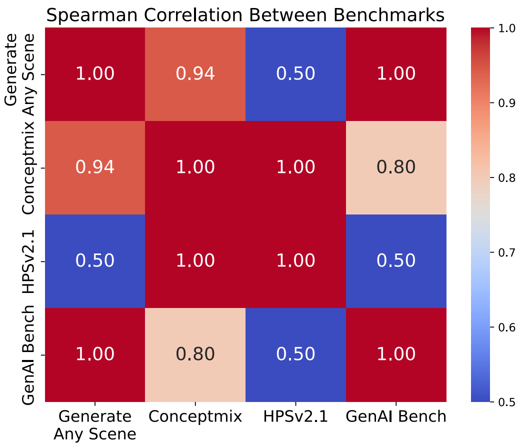
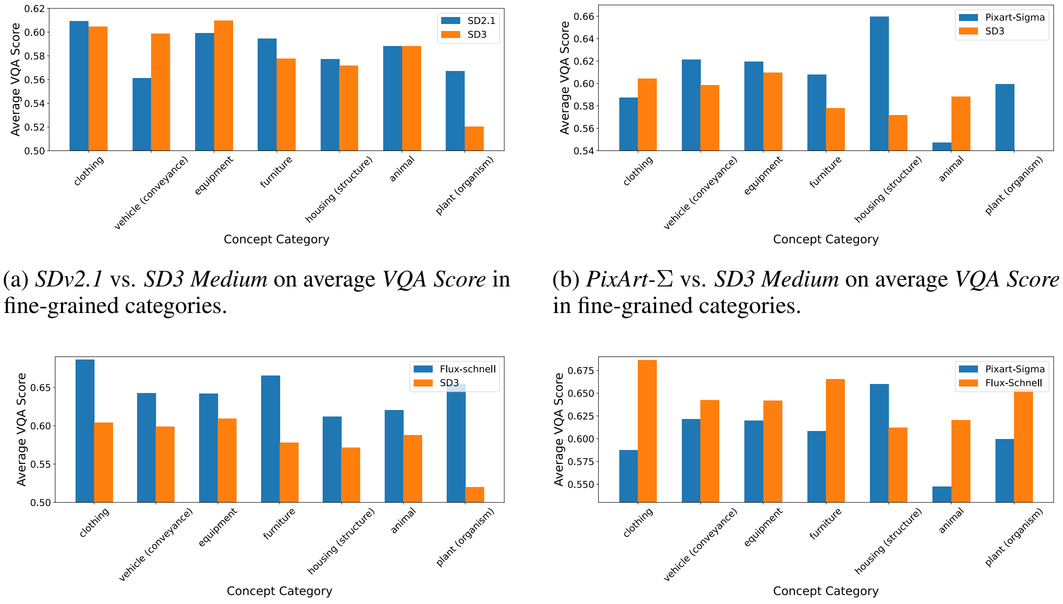


Figure 13: Correlation of GENERATE ANY SCENE with other popular *Text-to-Vision generation* benchmarks.



(a) *SDv2.1* vs. *SD3 Medium* on average *VQA Score* in fine-grained categories.

(b) *PixArt-Σ* vs. *SD3 Medium* on average *VQA Score* in fine-grained categories.

(c) *FLUX.1-schnell* vs. *SD3 Medium* on average *VQA Score* in fine-grained categories.

(d) *PixArt-Σ* vs. *FLUX.1-schnell* on average *VQA Score* in fine-grained categories.

Figure 14: Pairwise comparison on average *VQA Score* in fine-grained categories.

A.4.5 Case study: Pairwise fine-grained model comparison

Evaluating models using a single numerical average score can be limiting, as different training data often lead models to excel in generating different types of concepts. By leveraging the taxonomy we developed for Generate Any Scene, we can systematically organize these concepts and evaluate each model's performance on specific concepts over the taxonomy. This approach enables a more detailed comparison of how well models perform on individual concepts rather than relying solely on an overall average score. Our analysis revealed that, while the models may achieve similar average performance, their strengths and weaknesses vary significantly across different concepts. Here we present a pairwise comparison of models across different metrics.

B Details of Taxonomy of Visual Concepts

To construct a scene graph, we utilize three primary types of metadata: objects, attributes, and relations, which represent the structure of a visual scene. Additionally, scene attributes—which include factors like image style, perspective, and video time span—capture broader aspects of the visual content. Together, the scene graph and scene attributes form a comprehensive representation of the scene.

Our metadata is further organized using a well-defined taxonomy, enhancing the ability to generate controllable captions. This hierarchical taxonomy not only facilitates the creation of diverse scene graphs, but also enables fine-grained and systematic model evaluation.

Objects. To enhance the comprehensiveness and taxonomy of object data, we leverage noun synsets and the structure of WordNet (32). In WordNet, a *physical object* is defined as *"a tangible and visible entity; an entity that can cast a shadow."* Following this definition, we designate the *physical object* as the root node, constructing a hierarchical tree with all *28,787* hyponyms under this category as the set of objects in our model.

Following WordNet's hypernym-hyponym relationships, we establish a tree structure, linking each object to its primary parent node based on its first-listed hypernym. For objects with multiple hypernyms, we retain only the primary parent to simplify the hierarchy. Furthermore, to reduce ambiguity, if multiple senses of a term share the same parent, we exclude that term itself and reassign its children to the original parent node. This approach yields a well-defined and disambiguated taxonomy.

Attributes. The attributes of a scene graph represent properties or characteristics associated with each object. We classify these attributes into *nine* primary categories. For *color*, we aggregate *677* unique entries sourced from Wikipedia (33). The *material* category comprises *76* types, referenced from several public datasets (93; 94; 95). The *texture* category includes *42* kinds from the Describable Textures Dataset (96), while the *architectural style* encompasses *25* distinct styles (97). Additionally, we collect *85 states*, *41 shapes*, and *24 sizes*. For *human descriptors*, we compile 59 terms across subcategories, including body type and height. Finally, we collect *465* common *adjectives* covering general characteristics of objects to enhance the descriptive richness of our scene graphs.

Relationships. We leverage the Robin dataset (34) as the foundation for relationship metadata, encompassing six key categories: spatial, functional, interactional, social, emotional, and symbolic. With 10,492 relationships, the dataset provides a comprehensive and systematic repository that supports modeling diverse and complex object interactions. Its extensive coverage captures both tangible and abstract connections, forming a robust framework for accurate scene graph representation.

Scene Attributes. In *Text-to-Vision generation* tasks, people mainly focus on creating realistic images and art from a text description (98; 2; 3). For artistic styles, we define scene attributes using *76* renowned *artists*, *41 genres*, and *126 painting styles* from WikiArt (99), along with *29* common *painting techniques*. For realistic imagery, we construct camera settings attributes across 6 categories: camera models, focal lengths, perspectives, apertures, depths of field, and shot scales. The camera models are sourced from the 1000 Cameras Dataset (100), while the remaining categories are constructed based on photography knowledge and common captions in *Text-to-Vision generation* tasks (1; 101). To control scene settings, we categorize location, weather and lighting attributes, using 430 diverse locations from Places365 (35), alongside *76 weathers* and *57 lighting conditions*. For video generation, we introduce attributes that describe dynamic elements. These include 12 types of camera rig, 30 distinct camera movements, 15 video editing styles, and 27 temporal spans. The comprehensive scene attributes that we construct allow for the detailed and programmatic *Text-to-Vision generation* generation.

C Details of self-improving models with synthetic captions (Section 3)

C.1 Experiment details

C.1.1 Captions Preparation

To evaluate the effectiveness of our iterative self-improving *Text-to-Vision generation* model, we generated three distinct sets of 10K captions using Generate Any Scene, covering a sample complexity range from 3 to 12. These captions were programmatically created to reflect a spectrum of structured scene graph compositions, designed to challenge and enrich the model's learning capabilities.

For comparative analysis, we leveraged the Conceptual Captions (CC3M) ([102]) dataset, a large-scale benchmark containing approximately 3.3 million image-caption pairs sourced from web alt-text descriptions. CC3M is renowned for its diverse visual content and natural language expressions, encompassing a wide range of styles, contexts, and semantic nuances.

To ensure fair comparison, we randomly sampled three subsets of 10K captions from the CC3M dataset, matching the Generate Any Scene-generated caption sets in size. This approach standardizes data volume while enabling direct performance evaluation. The diversity and semantic richness of the CC3M captions serve as a robust benchmark to assess whether Generate Any Scene-generated captions can match or exceed the descriptive quality of real-world data across varied visual contexts.

C.1.2 Dataset Construction and Selection Strategies

For the captions generated by Generate Any Scene, we employed a top-scoring selection strategy to construct the fine-tuning training dataset, using a random selection strategy as a baseline for comparison. Specifically, for each caption, the model generated eight images. Under the top-scoring strategy, we evaluated the generated images using the VQA score and selected the highest-scoring image as the best representation of the caption. This process yielded 10K top-ranked images per iteration, from which the top 25% (approximately 2.5k images) with the highest VQA scores were selected to form the fine-tuning dataset.

In the random selection strategy, one image was randomly chosen from the eight generated per caption, and 25% of these 10K randomly selected images were sampled to create the fine-tuning dataset, maintaining parity in data size.

For the CC3M dataset, each caption was uniquely paired with a real image. From the 10K real image-caption pairs sampled from CC3M, the top 25% with the highest VQA scores were selected as the fine-tuning training dataset. This ensured consistency in data size and selection criteria across all methods, facilitating a rigorous and equitable comparison of fine-tuning strategies.

C.1.3 Fine-tuning details

We fine-tuned the *SDv1.5* using the LoRA technique. The training was conducted with a resolution of 512 × 512 for input images and a batch size of 8. Gradients were accumulated over two steps. The optimization process utilized the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, an ϵ value of 1×10^{-8}, and a weight decay of 10^{-2}. The learning rate was set to 1×10^{-4} and followed a cosine scheduler for smooth decay during training. To ensure stability, a gradient clipping threshold of 1.0 was applied. The fine-tuning process was executed for one epoch, with a maximum of 2500 training steps. For the LoRA-specific configurations, we set the rank of the low-rank adaptation layers and the scaling factor α to be 128.

After completing fine-tuning for each epoch, we set the LoRA weight to 0.75 and integrate it into *SDv1.5* to guide image generation and selection for the next subset. For the CC3M dataset, images from the subsequent subset are directly selected.

In the following epoch, the fine-tuned LoRA parameters from the previous epoch are loaded and used to resume training on the current subset, ensuring continuity and leveraging the incremental improvements from prior iterations.

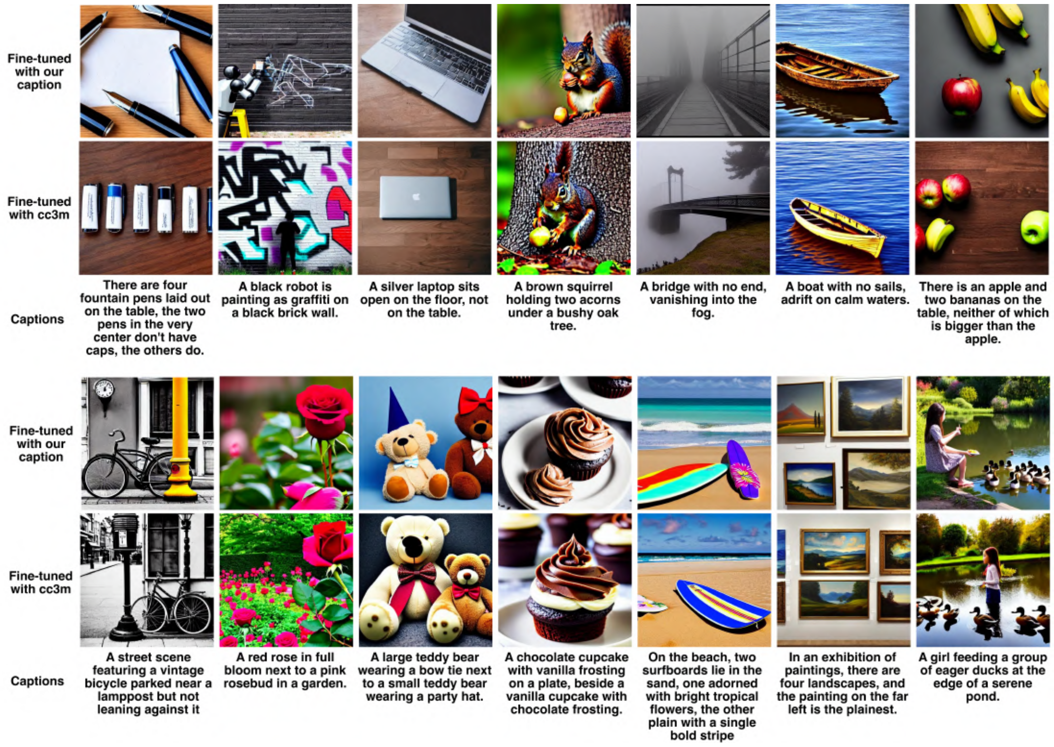

Fine-tuned with our caption
Fine-tuned with cc3m
Captions
There are four fountain pens laid out on the table, the two pens in the very center don't have caps, the others do.
A black robot is painting as graffiti on a black brick wall.
A silver laptop sits open on the floor, not on the table.
A brown squirrel holding two acorns under a bushy oak tree.
A bridge with no end, vanishing into the fog.
A boat with no sails, adrift on calm waters.
There is an apple and two bananas on the table, neither of which is bigger than the apple.
Fine-tuned with our caption
Fine-tuned with cc3m
Captions
A street scene featuring a vintage bicycle parked near a lamppost but not leaning against it
A red rose in full bloom next to a pink rosebud in a garden.
A large teddy bear wearing a bow tie next to a small teddy bear wearing a party hat.
A chocolate cupcake with vanilla frosting on a plate, beside a vanilla cupcake with chocolate frosting.
On the beach, two surfboards lie in the sand, one adorned with bright tropical flowers, the other plain with a single bold stripe
In an exhibition of paintings, there are four landscapes, and the painting on the far left is the plainest.
A girl feeding a group of eager ducks at the edge of a serene pond.

Figure 15: **Visualization of Different Caption Fine-Tuning.**

In Figure 15, we present results using our captions and the CC3M captions. The model fine-tuned with captions generated by GENERATE ANY SCENE demonstrates superior performance in terms of text semantic relevance and the generation of complex compositional scenes.

C.2 EVALUATION ON TIFA BENCH

Aside from our own test set and GenAI benchmark, we also evaluated our fine-tuned *Text-to-Image generation* models on the Tifa Bench (Figure 16), where we observed the same trend: models fine-tuned with our captions consistently outperformed the original *SDv1.5* and CC3M fine-tuned models.

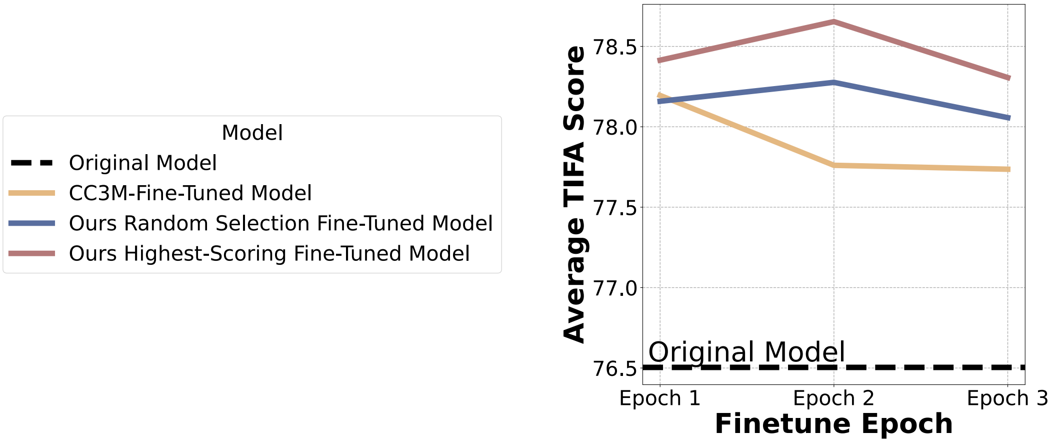


Figure 16: **Results for Application 1: Self-Improving Models**. Average TIFA score of *SDv1.5* fine-tuned with different data over TIFA Bench.

C.3 Additional real-data baselines

Setup. We conduct more experiments comparing Generate Any Scene synthetic captions to other real-world caption sources. We sampled 10K captions from MS-COCO-2017 and LAION-COCO for one-epoch LoRA fine-tuning under same experimental settings. The results on Generate Any Scene test set are summarized in Table 11.

Table 11: Self-improvement on Generate Any Scene Test (VQA). One-epoch finetuning, equal budget.

Method	VQA ↑
Baseline (*SDv1.5*)	0.508
MS-COCO-2017	0.508
LAION-COCO	0.510
CC3M	0.508
GAS (Random)	0.524
GAS (Top-Score)	**0.530**

Findings. Fine-tuning with MS-COCO-2017 and LAION-COCO captions yields results similar to CC3M, with none surpassing the significant improvements achieved by our Generate Any Scene captions. We think that although MS-COCO-2017 and LAION captions are generally high-quality and well-aligned with images, they offer limited compositional diversity. These additional results confirm that the observed gains are not specific to CC3M but generalize across other widely used real-caption datasets. This further supports our claim that the compositional diversity of Generate Any Scene synthetic captions drives the improvement.

C.4 Full fine-tuning vs. LoRA fine-tuning

Setup. We replicate the self-improvement pipeline with *full fine-tuning* and compare three strategies: Generate Any Scene captions with high-score selection, Generate Any Scene captions with random selection, and CC3M captions as the real-data baseline. The results are shown in Tables 12 and 13.

Table 12: Results on Generate Any Scene test set under full fine-tuning. (*VQA Score*)

Method	Iter-1	Iter-2	Iter-3
Baseline	0.508	—	—
CC3M (Full FT)	0.496	0.518	0.519
GAS (Rand, Full FT)	0.510	0.519	0.520
GAS (Top, Full FT)	**0.510**	**0.534**	**0.540**

Table 13: Results on GenAI-Bench under full fine-tuning. (*VQA Score*)

Method	Iter-1	Iter-2	Iter-3
Baseline	0.617	—	—
CC3M (Full FT)	0.589	0.619	0.622
GAS (Rand, Full FT)	0.599	0.621	0.617
GAS (Top, Full FT)	**0.620**	**0.626**	**0.634**

Findings. Using our Generate Any Scene captions with high score selection not only improves performance consistently across iterations but also surpasses CC3M at every stage. The full fine-tuning results confirm that our captions and strategy's effectiveness is not dependent on the specific training approach (LoRA vs. full fine-tuning). The consistent improvement patterns across both evaluation benchmarks validate the robustness of our iterative self-improvement framework.

D Details of distilling targeted capabilities (Section 4)

D.1 Collecting hard concepts

We evaluate both models on 10K Generate Any Scene captions and select 81 challenging object concepts where *SDv1.5* and *DaLL-E 3* exhibit the largest gap. To determine the score for each concept, we calculated the average *TIFA Score* of the captions containing that specific concept. For each targeted-generated caption, we generate four images and use the one with the highest *VQA Score*. The full list of hard concepts is shown below:

1. cloverleaf
2. aerie (habitation)
3. admixture
4. webbing (web)
5. platter
6. voussoir
7. hearthstone
8. puttee
9. biretta
10. yarmulke
11. surplice
12. overcoat
13. needlepoint
14. headshot
15. photomicrograph
16. lavaliere
17. crepe
18. tureen
19. bale
20. jetliner
21. square-rigger
22. supertanker
23. pocketcomb
24. filament (wire)
25. inverter
26. denture
27. lidar
28. volumeter
29. colonoscope
30. synchrocyclotron
31. miller (shaper)
32. alternator
33. dicer
34. trundle
35. paddle (blade)
36. harmonica

37. piccolo
38. handrest
39. rundle
40. blowtorch
41. volleyball
42. tile (man)
43. shuttlecock
44. jigsaw
45. roaster (pan)
46. maze
47. belt (ammunition)
48. gaddi
49. drawer (container)
50. tenter
51. pinnacle (steeple)
52. pegboard
53. afterdeck
54. scaffold
55. catheter
56. broomcorn
57. spearmint
58. okra (herb)
59. goatsfoot
60. peperomia
61. ammobium
62. gazania
63. echinocactus
64. birthwort
65. love-in-a-mist (passionflower)
66. ragwort
67. spicebush (allspice)
68. leadplant
69. barberry
70. hamelia
71. jimsonweed
72. undershrub
73. dogwood
74. butternut (walnut)
75. bayberry (tree)
76. lodestar
77. tapa (bark)
78. epicalyx
79. blackberry (berry)
80. stub
81. shag (tangle)

D.2 Experiment details

We conducted targeted fine-tuning experiments on *SDv1.5* to evaluate Generate Any Scene's effectiveness in distilling model compositionality and learning hard concepts. For each task, we selected a dataset of 778 Generate Any Scene captions paired with images generated by *DaLL-E 3*. For compositionality, we selected multi-object captions from the existing dataset of 10K Generate Any Scene captions and paired them with the corresponding images generated by *DaLL-E 3*. To address hard concept learning, we first used *SDv1.5* to generate images based on the 10K Generate Any Scene captions and identified the hard concepts with the lowest VQA scores. These concepts were then used to create a subset of objects, which we recombined into our scene-graph based captions with complexity levels ranging from 3 to 9. Finally, we used *DaLL-E 3* to generate corresponding images for these newly composed captions.

The fine-tuning configurations were consistent with those used in the self-improving setup (Appendix C.1.3). To accommodate the reduced dataset size, the maximum training steps were set to 1000.

As a baseline, we randomly selected 778 images from 10K Generate Any Scene-generated images, using captions produced by Generate Any Scene. This ensured a controlled comparison between the targeted and random fine-tuning strategies.

D.3 Benchmark against web-crawled caption–image pairs

Setup. We conduct additional experiments to benchmark against alternative data sourcing strategies, specifically comparing our *DaLL-E 3* distillation approach with web-scraped real images. Using the Bing Image Search API, we retrieve images matching our multi-object and hard-concept captions and constructed two datasets of equivalent scale for comparison. We then apply the same fine-tuning setup described in Application 2. The results are shown in Table 14:

Table 14: Comparison of VQA scores from targeted fine-tuning on different data sources. (*SDv1.5*)

Test Set	Original	*DaLL-E 3* Distill	Web-crawled
Hard Concept	0.303	**0.361**	0.258
Multi-object	0.271	**0.325**	0.264

Findings. The results show that web-scraped images not only failed to improve performance but actually degraded model capabilities.

Upon examination of the retrieved images, we identify several critical issues. The web-crawled images contain significant noise, including watermarks, overlaid text, and irrelevant visual element. Our hard concept and multi-object captions feature high compositional complexity and novel object combinations that rarely exist in real-world photographs. The retrieved images show poor relevance to our systematically designed compositional scenarios, as real-world images cannot adequately represent the diverse and controlled compositional variations we programmatically generate. Thus, training on such misaligned data appears to introduce incorrect visual-textual associations, leading to performance degradation rather than improvement.

Table 15: *VQA Score* of targeted distillation on *FLUX.1-dev*.

Test Set	Original	Fine-tuned
Hard Concept	0.303	**0.361**
Multi-object	0.271	**0.325**

D.4 DISTILLATION ON FLUX.1-DEV

Setup. We further apply our distillation framework to *FLUX.1-dev*, a current SOTA open-source model, using *DaLL-E 3* -generated images of hard concepts and multi-object captions to distill these capabilities into *FLUX.1-dev*. The results are shown in the Table 15:

Findings. The results demonstrate that our approach's effectiveness extends to state-of-the-art models (*FLUX.1-dev*). The distillation approach yields substantial improvements on challenging compositional tasks.

E DETAILS OF REINFORCEMENT LEARNING WITH A SYNTHETIC REWARD FUNCTION (SECTION 5)

E.1 TRAINING DATA PREPARATION

We adopt SimpleAR-0.5B-SFT ([26]) as our base model. Given that SImpleAR-0.5B-SFT is pretrained on high-quality real image datasets such as LAION ([11]) and CC3M ([12]), we aim to mitigate potential distributional shift between the original training data and the reinforcement learning phase. To this end, we perform metadata pre-selection for GENERATE ANY SCENE by analyzing the frequency of each object category appearing in the LAION dataset. Leveraging the controllable compositional capabilities of GENERATE ANY SCENE, we filter object categories by selecting the top 10% most frequent entries and constrain scene complexity to 3–6 objects per scene. Based on these conditions, we synthesize a set of 10K captions, ensuring semantic alignment with the base model's pretraining distribution while maintaining structural and content diversity.

E.2 EXPERIMENT DETAILS

The detailed training configuration is provided in Table 16. We utilize 8 × NVIDIA H100 GPUs (80GB HBM3), with one GPU allocated for online generation using vLLM. The total training time is approximately 14 hours.

Table 16: Scene-graph based GRPO Fine-tuning Configuration for SimpleAR

Component	**Details**
Model Name	SimpleAR-0.5B-SFT
Model Size	$\sim$0.5B parameters
Training Policy	GRPO
Inference Engine	vLLM (GPU utilization = 0.7)
Completion Length	4096 tokens
Training Epochs	1
Batch Size per Device	4
Learning Rate	1×10^{-5}
Scheduler	Cosine Annealing (min lr rate = 0.1)
Warm-up Ratio	0.1
Gradient Accumulation	1

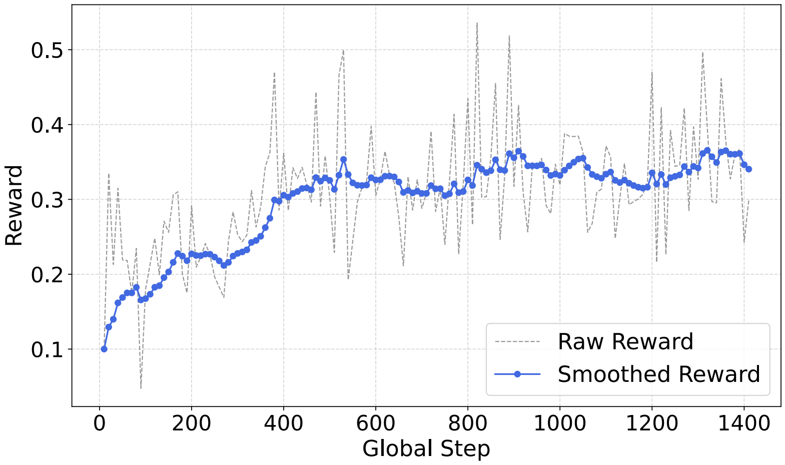


Figure 17: Reward progression during scene-graph based GRPO training.

Figure 17 illustrates the reward progression during training. A noticeable improvement in reward is observed following the application of a learning rate of 1e-5 combined with a warm-up strategy.

Overall, the reward increases by approximately 0.2, indicating effective learning under the adjusted training configuration.

In Table 4, we observe that the reproduced results of baseline models on DPG-Bench and GenEval Bench are slightly lower than those reported in the original paper. Considering the inherent stochasticity in generative model outputs, we cite the original results for comparison. For GenAI-Bench, all reported results are based on our own experimental evaluations.

E.3 Reward variants and ablations

Setup. To verify the observed gains arise specifically from the scene-graph–generated QA reward, rather than simply from using any QA-based reward, we conduct experiments incorporating manually annotated QA datasets, VQAv2, as additional reward signals under the same RLHF framework. We sample 10K images from VQAv2, with corresponding QA pairs, matched them to COCO2017 captions, and apply same training frameworks to SimpleAR-0.5B-SFT with RL training. The results on GenAI Bench are shown in the table:

Table 17: GenAI Bench performance (VQA) under RLHF with different reward sources. All models start from *SimpleAR-0.5B-SFT*.

Method	Basic ↑	Advanced ↑	All ↑
SimpleAR-0.5B-SFT	0.74	0.60	0.66
SimpleAR-0.5B-RL (CLIP)	**0.75**	0.60	0.67
SimpleAR-0.5B-RL (VQAv2)	0.73	0.59	0.66
SimpleAR-0.5B-RL (Ours)	**0.75**	**0.61**	**0.68**

Findings. The results show that using VQAv2 captions and QA pairs as rewards yields even lower performance than CLIP-based RL training. Furthermore, we observe minimal reward improvement from VQA signals throughout training. We attribute this to the fact that, although VQAv2 QA pairs are rich, the underlying image captions fail to cover enough visual elements, leading to a mismatch between QA pairs and captions that undermines RLHF reward alignment.

This highlights the inherent difficulty and cost of constructing high-quality image-caption and QA annotations, whereas our method leverages scene-graph structures to systematically generate synthetic caption-QA pairs at minimal cost with unique advantages.

F Details of improving generated-content detection (Section 6)

F.1 Experiment details

In this section, our goal is to validate that the more diverse captions generated by Generate Any Scene can complement existing datasets, which are predominantly composed of real-world images paired with captions. By doing so, we aim to train AI-generated content detectors to achieve greater robustness.

Dataset preparation We conducted comparative experiments between captions generated by Generate Any Scene and entries from the D^3 dataset. From the D^3 dataset, we randomly sampled 10K entries, each including a caption, a link to a real image, and an image generated by SD v1.4. Due to some broken links, we successfully downloaded 8.5K real images and retained 10K SD v1.4-generated images. We also used SD v1.4 to generate images based on 10K Generate Any Scene captions.

We varied the training data sizes based on the sampled dataset. Specifically, we sampled N real images from the 10K D^3 real images. For synthetic data, we compared N samples exclusively from D^3 with a mixed set of N/2 samples from 10K Generate Any Scene images and N/2 sampled from D^3, ensuring a total of N synthetic samples. Combined, this resulted in 2N training images. We tested 2N across various sizes, ranging from 2K to 10K.

Detector architecture and training We employed ViT-T (47) and ResNet-18 (103) as backbones for the detection models. Their pretrained parameters on ImageNet-21K were frozen, and the final classification head was replaced with a linear layer using a sigmoid activation function to predict the probability of an image being AI-generated. During training, We used Binary Cross-Entropy (BCE) as the loss function, and the AdamW optimizer was applied with a learning rate of $2e^{-3}$. Training was conducted with a batch size of 256 for up to 50 epochs, with early stopping triggered after six epochs of no improvement in validation performance.

Testing To evaluate the performance of models trained with varying dataset sizes and synthetic data combinations, we tested them on both GenImage and Generate Any Scene datasets to assess their in-domain and out-of-domain performance under different settings.

For GenImage, we used validation data from four models: SD v1.4, SD v1.5, MidJourney, and VQDM. Each validation set contained 8K real images and 8k generated images. For Generate Any Scene, we sampled 10K real images from CC3M and paired them with 10K generated images from each of the following models: *SDv2.1*, *PixArt-α*, *SD3 Medium*, and *Playground v2.5*. This created distinct test sets for evaluating model performance across different synthetic data sources.

Table 18: F1-Score Comparison of ResNet-18 and ViT-T Detectors Trained with D^3 and D^3+ Generate Any Scene Across In-Domain Settings

Detector	Data Scale (2N)	SDv1.4 (In-domain, same model)		SDv2.1		Pixart-α		SDv3-medium		Playground v2.5		Average (In-domain, cross model)	
		D^3 + Ours	D^3	D^3 + Ours	D^3	D^3 + Ours	D^3	D^3 + Ours	D^3	D^3 + Ours	D^3	D^3 + Ours	D^3
Resnet-18	2K	0.6561	**0.6663**	0.7682	0.6750	0.7379	0.606	0.7509	0.6724	0.7380	0.5939	**0.7488**	0.6368
	4K	0.6751	**0.6812**	0.7624	0.6853	0.7328	0.6494	0.7576	0.7028	0.7208	0.6163	**0.7434**	0.6635
	6K	0.6780	**0.6995**	0.7886	0.6870	0.7493	0.6586	0.7768	0.7285	0.7349	0.6335	**0.7624**	0.6769
	8K	0.6828	**0.6964**	0.7710	0.6741	0.7454	0.6418	0.7785	0.7186	0.7215	0.6033	**0.7541**	0.6595
	10K	0.6830	**0.6957**	0.7807	0.6897	0.7483	0.6682	0.7781	0.7326	0.7300	0.6229	**0.7593**	0.6784
ViT-T	2K	**0.6759**	0.6672	0.7550	0.6827	0.7585	0.6758	0.7473	0.6941	0.7327	0.6106	**0.7484**	0.6658
	4K	**0.6878**	0.6871	0.7576	0.7000	0.7605	0.7071	0.7549	0.7217	0.7221	0.6144	**0.7488**	0.6858
	6K	**0.6898**	0.6891	0.7663	0.6962	0.7666	0.7164	0.7629	0.7238	0.7303	0.6134	**0.7565**	0.6875
	8K	0.6962	**0.6974**	0.7655	0.6894	0.7712	0.7253	0.7653	0.7253	0.7381	0.6344	**0.7600**	0.6936
	10K	**0.6986**	0.6984	0.7828	0.6960	0.7777	0.7275	0.7786	0.7334	0.7330	0.6293	**0.7680**	0.6966

F.2 Results

Table 19 and Table 18 evaluate the performance of ResNet-18 and ViT-T detection backbones trained on datasets of varying sizes and compositions across in-domain (same model and cross-model) and out-of-domain settings. While models trained with D^3 and Generate Any Scene occasionally underperform compared to those trained solely on D^3 in the in-domain same-model setting, they exhibit significant advantages in both in-domain cross-model and out-of-domain evaluations. These results demonstrate that incorporating our data (GENERATE ANY SCENE) into the training process enhances the detector's robustness. By supplementing existing datasets with GENERATE ANY SCENE under the same training configurations and dataset sizes, detectors achieve stronger cross-model and cross-dataset capabilities, highlighting improved generalizability to diverse generative models and datasets.

Table 19: F1-Score Comparison of ResNet-18 and ViT-T Detectors Trained with D^3 and D^3+ GENERATE ANY SCENE Across Out-of-Domain Settings

Detector	Data Scale (2N)	SDv1.5		VQDM		Midjourney		Average (Out-of-domain)	
		D^3 + Ours	D^3	D^3 + Ours	D^3	D^3 + Ours	D^3	D^3 + Ours	D^3
Resnet-18	2K	0.6515	0.6591	0.5629	0.5285	0.5803	0.5647	**0.5982**	0.5841
	4K	0.6709	0.6817	0.5693	0.5428	0.6016	0.5941	**0.6139**	0.6062
	6K	0.6750	0.6963	0.5724	0.5327	0.6084	0.6072	**0.6186**	0.6121
	8K	0.6792	0.6965	0.5716	0.5282	0.6097	0.5873	**0.6202**	0.6040
	10K	0.6814	0.6955	0.5812	0.5454	0.6109	0.6040	**0.6245**	0.6150
ViT-T	2K	0.6755	0.6685	0.5443	0.4966	0.6207	0.6066	**0.6135**	0.5906
	4K	0.6845	0.6865	0.5591	0.4971	0.6416	0.6149	**0.6284**	0.5995
	6K	0.6900	0.6890	0.5580	0.4948	0.6455	0.6259	**0.6313**	0.6032
	8K	0.6940	0.6969	0.5553	0.4962	0.6495	0.6387	**0.6329**	0.6106
	10K	0.6961	0.6988	0.5499	0.4975	0.6447	0.6358	**0.6302**	0.6107

G ADVANTAGES OF GENERATE ANY SCENE OVER LLM-DRIVEN SCENE GRAPH AND CAPTION GENERATION

GENERATE ANY SCENE is conceptually superior to a well-prompted LLM for large-scale scene graph and corresponding captions generation. While modern LLMs are powerful, they do not provide the guarantees required for systematic, controllable, and reproducible enumeration of compositional structures. In contrast, GENERATE ANY SCENE explicitly enumerates graph topologies under user-specified constraints (e.g., complexity, topics, connectivity) and then deterministically instantiates them, yielding uniform coverage, strict structural validity, and high efficiency.

Controllability and Diversity. GENERATE ANY SCENE explicitly enumerates scene graph structures and populates with user-specified configuration and taxonomy (e.g., complexity, topics, connectivity, etc.), ensuring systematic coverage of rare or unconventional compositions without requiring users to manually write prompts for desired structures. In contrast, an LLM tends to default to common patterns in its training distribution. For example, given only the metadata {book, table, on}, an LLM will prefer the statistically dominant configuration "the book is on the table", and struggle to produce the less common but equally valid “the table is on the book” without extensive prompt engineering. Moreover, such extensive or high-quality prompting for scene graph generation essentially requires the user to manually enumerate graph structures and design multiple templates in natural language, whereas GAS accomplishes this systematically with a single program.

Reduced Bias and Hallucination. Relying on LLMs to generate large-scale captions inherently inherits their internal biases and increases the likelihood of hallucinating unseen or semantically inconsistent object configurations. GENERATE ANY SCENE avoids this by enumerating scene graphs and then deterministically mapping them to captions, producing text that is faithful by construction to the underlying graph structure.

Lower Cost and Higher Reproducibility. In GENERATE ANY SCENE, once a scene graph is enumerated, it is cached and reused across multiple populations, and it can also serve as a seed graph for controllable topological expansion without re-enumerating the entire structure. Combined with our fully programmatic operations, this makes large-scale generation substantially more cost-efficient. In contrast, relying on an LLM would require repeated API calls or prompt redesign for structural variant and new content, making the process both costly and labor-intensive.

To empirically validate these points, we compare GENERATE ANY SCENE against Gemini 2.5-flash on generating 10K scene graphs from our common metadata (3,649 items: 2,591 objects, 551 attributes, 507 relations). Because Gemini becomes increasingly error-prone when prompted with the full metadata list, we adopt a batching strategy: in each batch we randomly sample 5% of the metadata (182 items) and prompt the model to generate 20 scene graphs containing 3–12 elements.

Table 20 shows the distribution quality and diversity of generated elements. GENERATE ANY SCENE achieves near-uniform usage across objects, attributes, and relations, with Gini coefficients between 0.14 and 0.17 and normalized entropy above 99.3%. Gemini, in contrast, exhibits strong concentration (Gini 0.53–0.66) and substantially lower entropy (79.5–92.5%), indicating a tendency to overuse a narrow subset of frequent categories. The top-10% coverage further highlights this imbalance: under Gemini, 37.29% of object occurrences and 50.38% of relation occurrences are concentrated in only 10% of the vocabulary, whereas GENERATE ANY SCENE remains close to the uniform ideal.

Beyond distributional properties, we assess structural validity and data quality using strict schema-level checks (Table 21). GENERATE ANY SCENE produces 100% structurally valid graphs with zero hallucinated elements. In contrast, only 49.1% of Gemini’s outputs satisfy the schema. Common failure modes include treating relations as nodes (34.6% of graphs), and omitting required `value` (31.2%) or `type` (30.3%) fields. Gemini also hallucinates 1,773 “unknown” objects (4.59% of all objects) and 3,638 “unknown” relations.

Finally, GENERATE ANY SCENE is more efficient than LLM-based generation (Table 22). Because GENERATE ANY SCENE uses programmatic enumeration, it generates 10K scene graphs in under one minute, with negligible cost. In contrast, Gemini requires 1.5 hours and incurs over $50 of API cost for the same workload. Overall, GENERATE ANY SCENE provides a 90× speedup and near-zero marginal expense.

Table 20: Distribution quality and diversity of generated scene graphs.

Metric	GAS (Ours)	Gemini 2.5-flash
Gini Coefficient (↓)		
Objects	0.14	0.53
Attributes	0.14	0.57
Relations	0.17	0.66
Normalized Entropy (↑)		
Objects	99.6%	92.5%
Attributes	99.3%	91.7%
Relations	99.3%	79.5%
Top 10% Coverage (↓)		
Objects	14.68%	37.29%
Relations	15.41%	50.38%

Table 21: Structural validity and data quality of generated scene graphs.

Metric	GAS (Ours)	Gemini 2.5-flash
Structurally valid graphs	100%	49.1%
Graphs with relations as nodes (error)	0%	34.6%
Graphs missing `value` field	0%	31.2%
Graphs missing `type` field	0%	30.3%
Hallucinated “unknown” objects	0	1,773 (4.59%)
Hallucinated “unknown” relations	0	3,638

These results confirm that programmatic enumeration in GENERATE ANY SCENE outperforms LLM-based generation, providing the systematic guarantees of uniformity, validity, and efficiency.

H DISCUSSION

H.1 COMMONSENSE AND PLAUSIBILITY FILTERING

GENERATE ANY SCENE enables systematic, controllable, and diverse compositional scene construction through programmatic scene graph enumeration. This allows the synthesized captions to cover not only realistic scenes commonly observed in the real world, but also uncommon, imaginative, and unrealistic scenes. Many widely-used generative models, including DALL-E, Midjourney, and Sora/Sora2, derive much of their practical value from producing surreal, imaginative, or physically unlikely compositions (e.g., "an astronaut riding a horse on the moon," or "a raccoon astronaut with a glowing space donut"). Such prompts are not outliers; they reflect common user intents in art, game design, advertising, and entertainment. Users frequently employ abstract or fantastical combinations precisely to explore the model’s creativity, and the community often discusses and evaluates models based on performance on these highly “unrealistic” prompts. From a research perspective, a broad and systematically controlled compositional space is essential for improving and benchmarking modern generative models. Limiting sampling to only strictly "realistic" combinations would substantially reduce both the training and the evaluation value.

Table 22: Efficiency and cost of generating 10K scene graphs.

Metric	GAS (Ours)	Gemini 2.5-flash
Generation time (10K graphs)	< 1 minute	1.5 hours
Monetary cost	Negligible	> $50

Our approach is specifically designed to meet this need for diverse captions and systematic visual representations. At the same time, GENERATE ANY SCENE differentiates uncommon or unrealistic scenes from nonsensical scenes. The taxonomy enforces strong type-level constraints, e.g., architectural attributes apply only to buildings, human-specific attributes only to the "person" subtree, and attentional relations only between animate entities, ensuring that generated scenes remain meaningful and structurally valid, even when creatively unrealistic. Beyond these inherent structural constraints, GENERATE ANY SCENE additionally provides an optional two-stage commonsense and plausibility filtering mechanism to support use cases that require higher visual realism. (1) Pre-population filtering. We maintain for every object/attribute/relation its LAION-5B ([11]) frequency and embedding representation, and select candidates by jointly enforcing minimum frequency thresholds and semantic coherence: for each newly added relation or attribute attached to a given object, we compute the top-k semantically compatible candidates based on embedding similarity to that object. Likewise, when expanding a relation triple, we compute candidate object similarity to the anchor object within the triplet, including all attributes and relations already attached to the anchor, and then sample from the top-k most semantically compatible objects (where k is user-configurable). Users may further specify complexity limits to avoid highly complex scenes. (2) Post-population filtering. After population, once the scene graph is translated into a caption, we compute its Vera score ([91]) and caption perplexity, and discard captions falling below plausibility or above perplexity thresholds. These mechanisms ensure that GENERATE ANY SCENE preserves meaningfulness while still enabling broad creative coverage.

H.2 SOCIAL BIAS

Assessing social bias is important for understanding whether synthetic data introduces unintended shifts in model behavior. To examine this, we evaluate models on gender-related prompts from the DALL-Eval ([104]) benchmark, comparing SDv1.5, SDv1.5 fine-tuned on CC3M captions, and SDv1.5 fine-tuned on GENERATE ANY SCENE captions. The gender MAD results are shown in Table 23. The experiment shows that fine-tuning with GENERATE ANY SCENE does not amplify gender bias relative to the base model. We attribute this to our design choices. First, GENERATE ANY SCENE does not generate data by propagating textual descriptions or cultural associations from these sources; instead, our metadata is used purely as a structural vocabulary of objects, attributes, and relations. GENERATE ANY SCENE doesn't sample linguistic definitions or corpus-derived stereotypes from WordNet. Second, the systematic, programmatic nature of our scene-graph enumeration further reduces the pathways through which social bias present in real-world distributions could propagate. Also, any more debiased metadata can be plugged into GENERATE ANY SCENE engine seamlessly.

Table 23: Gender MAD Scores on DALLEval

Model	MAD ↓
SDv1.5	0.3602
FT w/ CC3M	0.3476
FT w/ GAS	0.3555

I LIMITATION

Programmatically generated prompts can be unrealistic and biased. Programmatically generated prompts can be unrealistic and biased. Although our system is capable of producing a wide range of rare compositional scenes and corresponding prompts, some of these outputs may violate rules or conventions, going beyond what is even considered imaginable or plausible. We also implement a pipeline to filter the commonsense of the generated prompts using the *Vera score* (a large language model-based commonsense metric) and *Perplexity*, but we make this pipeline **optional**.

Linguistic diversity of programmatic prompts is limited. While GENERATE ANY SCENE excels at generating diverse and compositional scene graphs and prompts, its ability to produce varied language expressions is somewhat constrained. The programmatic approach to generating content ensures diversity in terms of the elements of the scene, but it is limited when it comes to linguistic diversity and the richness of expression. To address this, we introduce a pipeline that leverages large language models (LLMs) to paraphrase prompts, enhancing linguistic variety. However, this addition introduces new challenges. LLMs are prone to biases and hallucinations, which can affect the quality and reliability of the output. Furthermore, the use of LLMs risks distorting the integrity of the original scene graph structure, compromising the coherence and accuracy of the generated content. So we make this LLM paraphrase pipeline **optional** for our paper.

Toward curriculum-aware GRPO training. Our proposed GENERATE ANY SCENE framework plays a central role in GRPO training by providing structured scene graphs that serve as the foundation for a semantically grounded and controllable reward function. This design enables effective optimization by aligning generation objectives with fine-grained visual semantics. Beyond this, we also observe that GENERATE ANY SCENE also offers broader potential: the scene graphs it produces vary in complexity, such as in the number of objects, attributes, relationships and graph degree. These variations naturally correspond to different levels of generation difficulty and reward variance. This property suggests an opportunity for curriculum-based training, where the model could be progressively exposed to increasingly complex scene graphs. Such a strategy may improve training stability and efficiency, especially in the early stages of learning. We identify this as a promising direction for future work, further leveraging the controllability of GENERATE ANY SCENE to guide structured policy learning.