Generate Any Scene: Evaluating and Improving Text-to-Vision Generation with Scene Graph Programming

Anonymous CVPR submission

Abstract

Generative models like DALL-E and Sora have gained attention by producing implausible images, such as "astronauts riding a horse in space." Despite the proliferation of text-to-vision models that have inundated the internet with synthetic visuals, from images to 3D assets, current benchmarks predominantly evaluate these models on real-world scenes paired with captions. We introduce GENERATE ANY SCENE*, a framework that systematically enumerates scene graphs representing a vast array of visual scenes, spanning realistic to imaginative compositions.* GENERATE ANY SCENE *leverages 'scene graph programming,' a method for dynamically constructing scene graphs of varying complexity from a structured taxonomy of visual elements. This taxonomy includes numerous objects, attributes, and relations, enabling the synthesis of an almost infinite variety of scene graphs. Using these structured representations,* GENERATE ANY SCENE *translates each scene graph into a caption, enabling scalable evaluation of text-to-vision models through standard metrics. We conduct extensive evaluations across multiple text-to-image, text-to-video, and text-to-3D models, presenting key findings on model performance. We find that DiT-backbone text-to-image models align more closely with input captions than UNet-backbone models. 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. Additionally, we demonstrate the effectiveness of* GENERATE ANY SCENE *by conducting three practical applications leveraging captions generated by* GENERATE ANY SCENE*: (1) a self-improving framework where models iteratively enhance their performance using generated data, (2) a distillation process to transfer specific strengths from proprietary models to open-source counterparts, and (3) improvements in content moderation by identifying and generating challenging synthetic data.*

1. Introduction

Artist Marc Chagall said "*Great art picks up where nature ends.*" The charm of visual content generation lies in the realm of imagination. Since their launch, Dall-E [4, 47] and Sora [6] have promoted their products with implausible generated images of "astronauts riding a horse in space" and "cats playing chess". With the proliferation of text-to-vision generation models, the internet is now flooded with generated visual content—images, videos, and 3D assets—most generated from user-provided captions [4, 6, 47]. While there are numerous benchmarks designed for evaluating these text-to-vision models, they are typically collections of *real*-world visual content paired with captions [8, 31, 60]. To quote Marc Chagall again, "If I create from heart, nearly everything works; if from the head, almost nothing." There is a need for evaluation benchmarks that go beyond real-world scenes and evaluate how well generative models can represent the entire space of imaginary scenes.

Such a comprehensive evaluation requires that we first define *the space of the visual content*. A long list of prior work [23–25, 29, 40] has argued that scene graphs [29] are a cognitively grounded [5] representation of the visual space. 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 [37]. 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. Make it an ideal structure for GENERATE ANY SCENE to systematically define the diverse space of the visual scenes.

We introduce GENERATE ANY SCENE, a system capable of efficiently enumerating the space of scene graphs representing a wide range of visual scenes, from realistic to

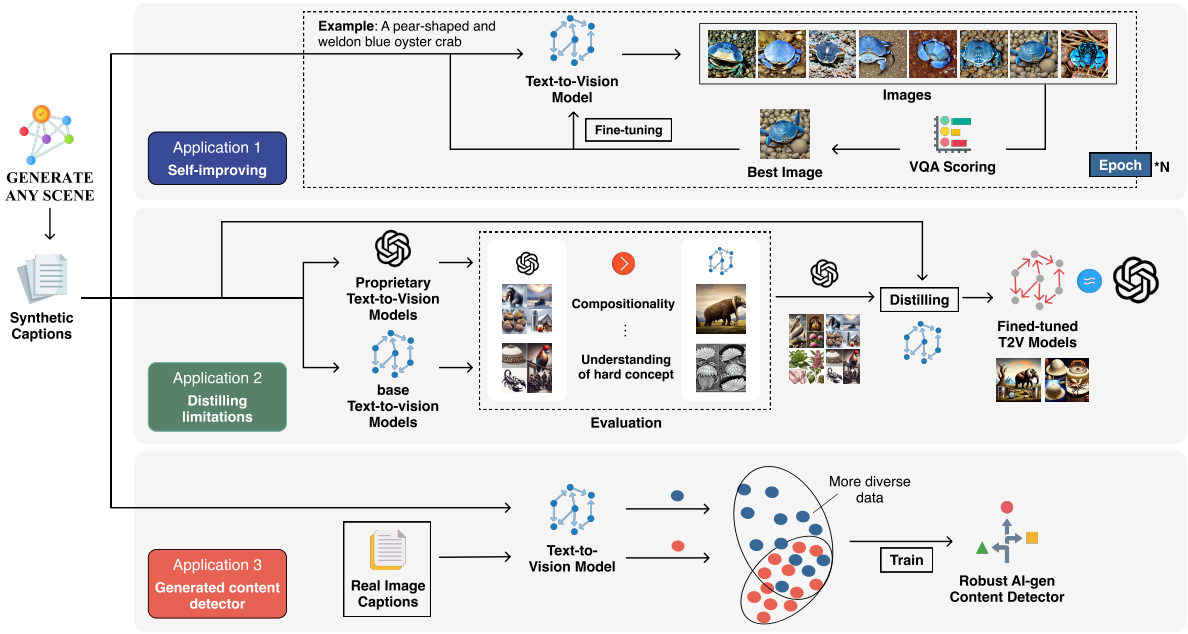


Figure 1. Overview of applications with GENERATE ANY SCENE captions: **Application 1**: (**Self-improving**): Iteratively enhances a model by generating images with GENERATE ANY SCENE captions, selecting the best, and fine-tuning, yielding a performance boost. **Application 2: (Distilling limitations)**: Distills strengths from proprietary models, such as better compositionality and hard concept understanding, into open-source models. **Application 3: (Generated content detector)**: Robustify AI-generated content detection by training on diverse synthetic data generated by GENERATE ANY SCENE's captions.

highly imaginative. GENERATE ANY SCENE is powered by what we call *scene graph programming*, a programmatic approach for composing scene graphs of any complexity using a rich taxonomy of visual elements, and for translating each scene graph into a caption. With a space of synthetically diverse captions, we use GENERATE ANY SCENE to prompt *Text-to-Vision generation* models and evaluate their generations. Like any other representation, scene graphs are also limited: they don't represent tertiary relationships (*e.g.* "three people playing frisbee"). Nonetheless, they account for a large space of possibilities. To systematically define and scalably explore the space of user captions, we adopt the scene graph representation [29] to comprehensively evaluate and improve text-to-vision models.

We construct a rich taxonomy of visual concepts consisting of $28,787$ objects, $1,494$ attributes, $10,492$ relations, $2,193$ image/video/3D scene attributes from various sources. Based on these assets, GENERATE ANY SCENE can programmatically synthesize an almost infinite number of scene graphs of varying complexity [70]. 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. Given an initial scene graph, GENERATE ANY SCENE programmatically translates it into a caption, which, when combined with existing text-to-vision metrics, *e.g.*, *Clip Score* [46] and *VQA Score* [34], can be used to evaluate any text-to-vision model [53]. 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.

With GENERATE ANY SCENE's programmatic generation capability, we release a dataset featuring 10 million diverse and compositional captions, each paired with a corresponding scene graph. This extensive dataset spans a wide range of visual scenarios, from realistic to highly imaginative compositions, providing an invaluable resource for researchers and practitioners in the *Text-to-Vision generation* field. We also conduct extensive evaluations of 12 text-to-image, 9 text-to-video and 5 text-to-3D models across a broad spectrum of visual scenes. We have several crucial findings: (1) DiT-backbone models show superior faithfulness and comprehensiveness to input captions than UNet-backbone models, with human-alignment data training helping to bridge some of these gaps. (2) *Text-to-Video generation* face challenges in balancing dynamics and consistency. (3) All *Text-to-Video* and *Text-to-3D* models we evaluate show negative *ImageReward Score* scores, highlighting a substantial gap in human preference alignment.

Further, we demonstrate the effectiveness of GENERATE ANY SCENE by conducting three practical applications leveraging captions generated by GENERATE ANY SCENE (Figure 1):

Application 1: Self-improving. We show that 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 *Stable Diffusion v1-5* and achieve an average of 5% 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 [8] over different benchmarks.

Application 2: Distilling limitations. 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* excels particularly in generating composite images with multiple parts. We distill this capability into *Stable Diffusion v1-5*, effectively bridging the gap between *DaLL-E 3* and *Stable Diffusion v1-5*.

Application 3: Generated content detector. Content moderation is a vital application, especially as *Text-to-Vision generation* models improve. We identify which kinds of data content moderation models are bad at detecting, generate more of such content, and retrain the detectors. We train a ViT-T with our generated data and boost its detection capabilities across benchmarks.

2. Generate Any Scene

We present our implementation of GENERATE ANY SCENE system. (Figure 2) It programmatically 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 their relationships (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 [29] and GQA [22] for visual question answering (VQA). Their utility has expanded to various *Text-to-Vision generation* tasks. For example, the DSG score [12] leverages MLMs to evaluate how well captions align with generated scenes by analyzing scene graphs.

Metadata Type	Number	Source
Objects	28,787	WordNet [39]
Attributes	1,494	Wikipedia [61], etc.
Relations	10,492	Robin [41]
Scene Attributes	2,193	Places365 [36], etc.

Table 1. Summary of the quantities and source of visual elements.

Taxonomy of visual elements. To construct a scene graph, we use three main metadata types: **objects**, **attributes**, and **relations**. We also have **scene attributes** that capture the board aspect of the caption, such as art style, to create a complete visual caption. The numbers and the source of our metadata are illustrated in Table 1. Additionally, we build a taxonomy that categorizes metadata into distinct levels and types, enabling fine-grained analysis. This structure allows for detailed assessments, such as evaluating model performance on "flower" as a general concept and on specific subcategories like "daisy." More details in Appendix C.

2.1. Scene graph programming

Step 1: Scene graph structure enumeration and query. Our system first generates and stores a variety of scene graph structures based on a specified level of **complexity**, defined by the total number of objects, relationships, and attributes in each graph. The process begins by determining the number of object nodes, and then by systematically enumerating different combinations of relationships among these objects and their associated attributes. Once all graph structures meeting the complexity constraint are enumerated, they are stored in a database for later use. This enumeration process is executed only once for each level of complexity, allowing us to efficiently query the database for suitable templates when needed.

Step 2: Populate the scene graph structure with metadata. Given a 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: users can define the topics and types of metadata to be included (e.g., selecting only common metadata or specifying particular relationships between particular objects, among other options). By determining the scope of metadata sampling, we can precisely control the final content of the captions and easily extend the diversity and richness in the scene graphs by incorporating new datasets.

Step 3: Sampling scene attributes. In addition to scene graphs that capture the visual content of the image, 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 an algorithm that converts scene graphs and a list of scene attributes into captions. The algorithm processes the scene graph in topological order, transforming each object, its attributes, and relational edges into descriptive text. To maintain coherence, it tracks each concept's occurrence, distinguishing objects with identical names using terms like "the first" or "the second." Objects that have been previously referenced without new relations are skipped to avoid misreferencing. This approach enhances caption clarity by pre-

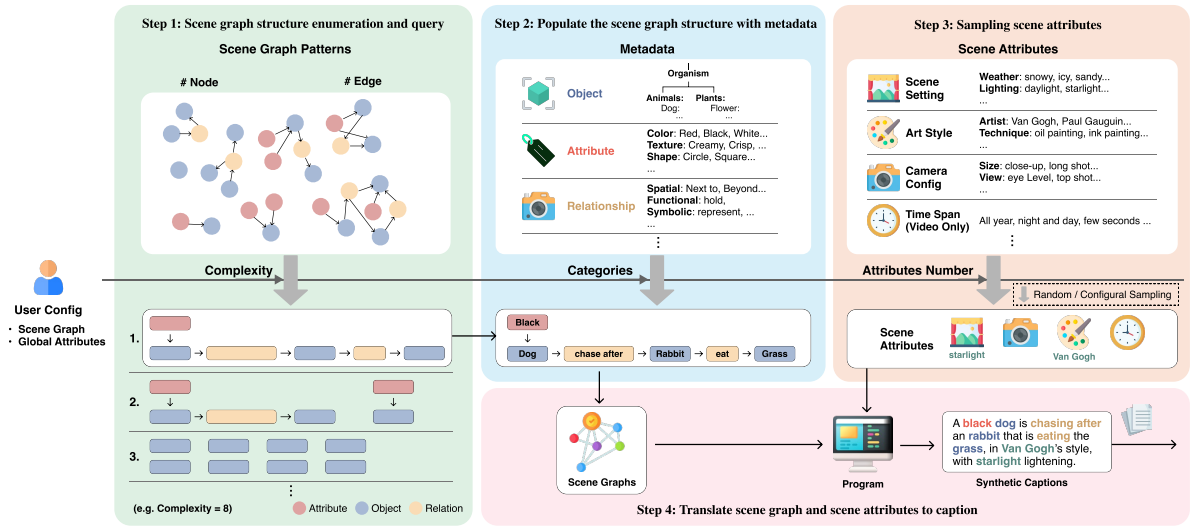


Figure 2. The generation pipeline of GENERATE ANY SCENE: **Step 1**: The system enumerates scene graph structures that contain objects, attributes, and relations based on complexity, and queries the corresponding scene graph structure that satisfies the needs. **Step 2**: It populates these structures with metadata, assigning specific content to each node. Scene graphs are completed in this step. **Step 3**: In addition to the scene graph, scene attributes—such as art style and camera settings—are sampled to provide contextual depth beyond the scene graph. **Step 4**: The GENERATE ANY SCENE system combines the scene graph and scene attributes, such as art style and camera settings, and then translates them into a coherent caption by organizing the elements into structured text.

venting repetition and maintaining a logical reference.

3. Evaluating *Text-to-Vision generation* models

3.1. Experiment Settings

Details of experiment settings are in Appendix D.

Models. We conduct experiments on 12 *Text-to-image* models [1, 4, 10, 11, 15, 30, 32, 43, 44, 48], 9 *Text-to-Video* models [9, 17, 27, 52, 55, 57, 63, 69, 71], and 5 *Text-to-3D* models [33, 38, 45, 56, 59]. *Text-to-image* models are evaluated at a resolution of 1024 × 1024 pixels. We standardize the frame length to 16 across all *Text-to-Video* models for fair comparisons. For *Text-to-3D*, we generate videos by rendering from 120 viewpoints.

Metrics. Across all *Text-to-Vision generation* tasks, we use *Clip Score* [7] (semantic similarity), *VQA Score* [34] (faithfulness), *TIFA Score* [12, 20] (faithfulness), *Pick Score* [28] (human preference), and *ImageReward Score* [66] (human preference) as general metrics, and for *Text-to-Video generation*, VBench [21] for fine-grained video analysis like consistency and dynamics.

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. 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.

3.2. Overall results

We evaluate *Text-to-Image generation*, *Text-to-Video generation*, and *Text-to-3D generation* models on GENERATE ANY SCENE. Here, we only list key findings; more details and raw results can be found in Appendix D.

***Text-to-Image generation* results. (Figure 3)**

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 prompts across generated content.

***Text-to-Video generation* results. (Table 2,3)**

1. Text-to-video models face challenges in balancing dynamics and consistency (Table 3). 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 2), suggesting a lack of human-preferred visual ap-

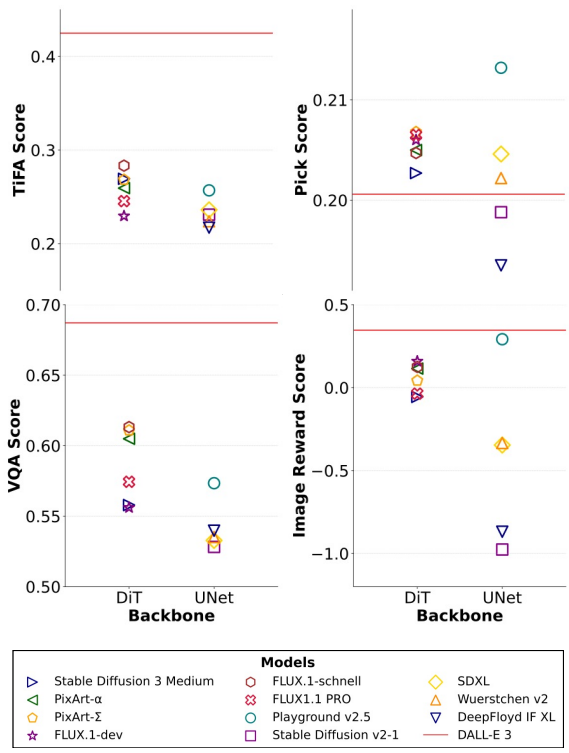


Figure 3. Comparative evaluation of *Text-to-Image generation* models across different backbones (DiT and UNet) using multiple metrics: *TIFA Score*, *Pick Score*, *VQA Score*, and *ImageReward Score*.

Model	clip score	pick score	image reward score	VQA score	TiFA score
VideoCraft2 [9]	0.2398	[RGB]255, 255, 2000.1976	[RGB]255, 200, 200-0.4202	0.5018	[RGB]255, 200, 2000.2466
AnimateLCM [55]	0.2450	[RGB]255, 200, 2000.1987	[RGB]255, 255, 200-0.5754	0.4816	0.2176
AnimateDiff [17]	[RGB]255, 200, 2000.2610	0.1959	-0.7301	[RGB]255, 255, 2000.5255	0.2208
Open-Sora 1.2 [71]	0.2259	0.1928	-0.6277	[RGB]255, 200, 2000.5519	[RGB]255, 255, 2000.2414
FreeInit [63]	[RGB]255, 200, 2000.2579	0.1950	-0.9335	0.5123	0.2047
ModelScope [57]	0.2041	0.1886	-1.9172	0.3840	0.1219
Text2Video-Zero [27]	0.2539	0.1933	-1.2050	0.4753	0.1952
CogVideoX-2B [69]	0.2038	0.1901	-1.2301	0.4585	0.1997
ZeroScope [52]	0.2289	0.1933	-1.1599	0.4892	0.2388

Table 2. Overall performance of *Text-to-Video generation* models over 10K GENERATE ANY SCENE captions. Red Cell is the highest score. Yellow Cell is the second highest score.

Model	subject consistency	background consistency	motion smoothness	dynamic degree
Open-Sora 1.2	[RGB]255, 200, 2000.9964	[RGB]255, 200, 2000.9907	[RGB]255, 200, 2000.9973	[RGB]200, 220, 2550.0044
Text2Video-Zero	[RGB]200, 220, 2550.8471	[RGB]200, 220, 2550.9030	[RGB]200, 220, 2550.8301	[RGB]255, 200, 2000.9999
VideoCraft2	0.9768	0.9688	0.9833	0.3556
AnimateDiff	0.9823	0.9733	0.9859	0.1406
FreeInit	0.9581	0.9571	0.9752	0.4440
ModelScope	0.9795	0.9831	0.9803	0.1281
AnimateLCM	0.9883	0.9802	0.9887	0.0612
CogVideoX-2B	0.9583	0.9602	0.9823	0.4980
ZeroScope	0.9814	0.9811	0.9919	0.1670

Table 3. Overall performance of *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.

peal in the generated content, even in cases where certain models demonstrate strong semantic alignment.

3. *VideoCrafter2* strikes a balance across key metrics, leading in human-preference alignment, faithfulness, consistency, and dynamic.

***Text-to-3D generation* results. (Table 4)**

Model	clip score	pick score	vqa score	tifa score	image reward score
Latent-NeRF [38]	0.2115	0.1910	0.4767	0.2216	-1.5311
DreamFusion-sd [45]	0.1961	0.1906	0.4421	0.1657	-1.5582
Magic3D-sd [33]	0.1947	0.1903	0.4193	0.1537	-1.6327
SJC [56]	0.2191	0.1915	0.5015	0.2563	-1.4370
DreamFusion-IF [45]	0.1828	0.1857	0.3872	0.1416	-1.9353
Magic3D-IF [33]	0.1919	0.1866	0.4039	0.1537	-1.8465
ProlificDreamer [59]	**0.2125**	**0.1940**	**0.5411**	**0.2704**	**-1.2774**

Table 4. Overall performance of *Text-to-3D generation* models over 10K GENERATE ANY SCENE captions.

1. *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.

4. Application 1: Self-Improving Models

In this section, we explore how GENERATE ANY SCENE facilitates a self-improvement framework for model generation capabilities. By programmatically 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 [53], we designed an iterative self-improving framework using GENERATE ANY SCENE with *Stable Diffusion v1-5* as the baseline model. With *VQA Score*, which shows strong correlation with human evaluations on compositional images [34], we guide the model's improvement throughout the process.

Specifically, GENERATE ANY SCENE generates 3 × 10K captions across three epochs. For each caption, *Stable Diffusion v1-5* 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 [19] for parameter-efficient fine-tuning. Additional details are available in Appendix E.

To evaluate the effectiveness of self-improvement using synthetic data generated by GENERATE ANY SCENE, we conduct comparative experiments with the CC3M dataset, which comprises high-quality and diverse real-world image-caption pairs [51]. We randomly sample 3 × 10K captions from CC3M, applying the same top-score selection strategy for iterative fine-tuning of *Stable Diffusion*

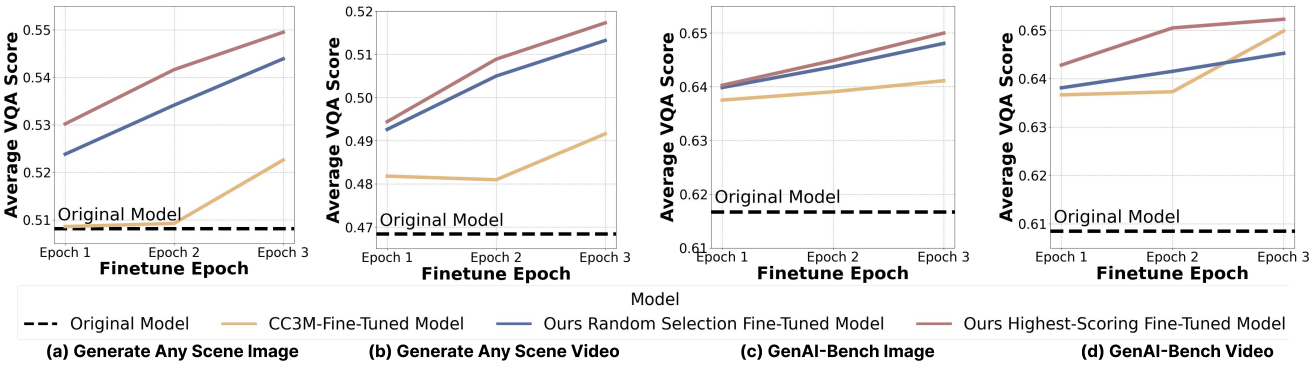


Figure 4. **Results for Application 1: Self-Improving Models**. Average VQA score of *Stable Diffusion v1-5* fine-tuned on different data across 1K GENERATE ANY SCENE image/video evaluation set and GenAI-Bench image/video benchmark [31].

v1-5. Additionally, we include a baseline using random-sample fine-tuning strategy to validate the advantage of our highest-scoring selection-based strategy.

Results. We evaluate our self-improving pipeline on *Text-to-Vision generation* benchmarks, including GenAI Bench [31]. For the *Text-to-Video generation* task, we use *Text2Video-Zero* as the baseline model, substituting its backbone with the original *Stable Diffusion v1-5* and our fine-tuned *Stable Diffusion v1-5* models.

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 4), 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. Additional results are in Appendix F.

Takeaway for application 1

Iterative self-improving *Text-to-Vision generation* models with compositional and diverse synthetic captions can surpass fine-tuning with real-world image-caption data.

Potential reason: The compositional, synthetic captions generated by GENERATE ANY SCENE exhibit greater diversity than real-world data.

5. Application 2: Distilling limitations

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

5.1. Fine-Grained Analysis of DaLL-E 3's Exceptional Performance

As shown in Figure 3, *DaLL-E 3* achieves *TIFA Score* **1.5** to **2** times higher than those of other models. When we compare *TIFA Score* across varying numbers of elements (objects, relations, and attributes per caption) in Figure 6b, *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 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.

5.2. Distilling compositionality from DaLL-E 3

Observations. We find that *DaLL-E 3* tends to produce straightforward combinations of multiple objects (Figure 5). In contrast, open-source models like *Stable Diffusion v1-5* often omit some objects from the captions, even though they are capable of generating each object individually.

This difference suggests that *DaLL-E 3* may be trained on datasets emphasizing multi-object presence without rigorous attention to image layout or object interaction. Such training likely underpins *DaLL-E 3*'s stronger performance on metrics like *TIFA Score* and *VQA Score*, prioritizing object inclusion over detailed compositional arrangement.

Finetuning. To encourage *Stable Diffusion v1-5* to learn compositional abilities similar to those of *DaLL-E 3*., we select a set of 778 images generated by *DaLL-E 3*, each containing multiple objects, and utilize this dataset to fine-tune *Stable Diffusion v1-5*. For the baseline, we randomly sam-

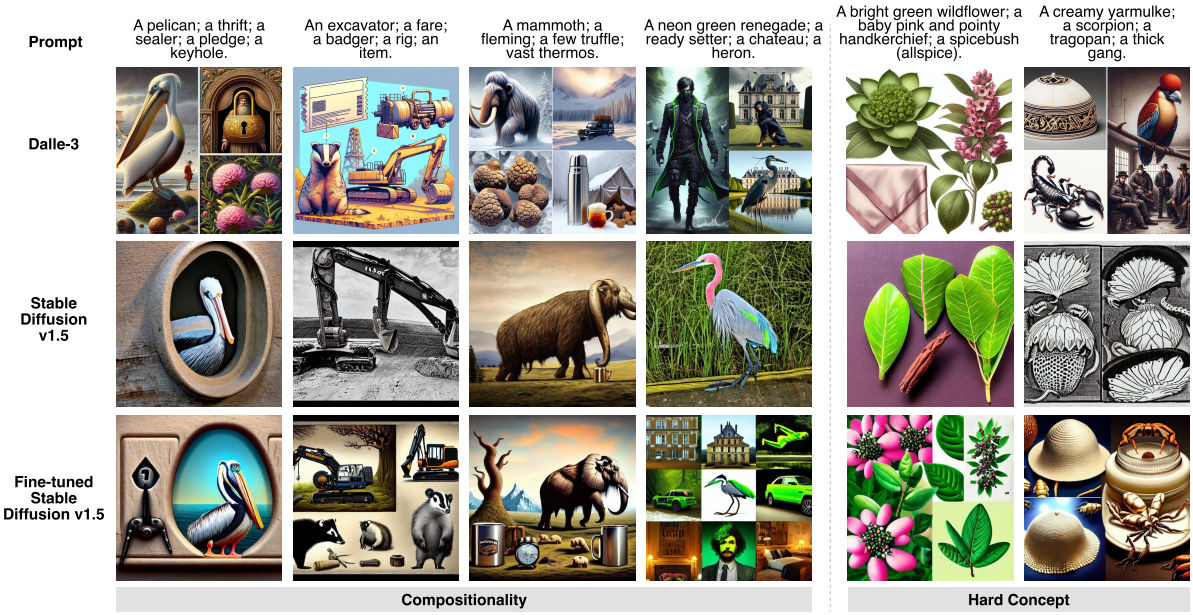


Figure 5. **Examples for Application 2: Distilling limitations**. Examples of images generated by *DaLL-E 3*, the original *Stable Diffusion v1-5*, and the fine-tuned versions. The left four captions demonstrate fine-tuning with multi-object captions generated by GENERATE ANY SCENE for better compositionality, while the right two columns focus on understanding hard concepts.

pled an equivalent set of *DaLL-E 3*-generated images paired with generated captions from GENERATE ANY SCENE.

Results. To evaluate compositional improvements, we generate 1K multi-object captions. Figure 6b shows a 10% *TIFA Score* increase after fine-tuning, random fine-tuning by an average of 3%. These results indicate enhanced compositional abilities in handling complex generation tasks.

We analyze images generated by *Stable Diffusion v1-5* before and after fine-tuning on high-complexity image-caption pairs (Figure 5). It is surprising to see that, with only 1K LoRA fine-tuning steps, *Stable Diffusion v1-5* effectively learn *DaLL-E 3* 's capability to arrange and compose multiple objects within a single image,. This fine-tuning strategy notably enhances alignment between generated images and their given captions.

On a broader set of 10K GENERATE ANY SCENE-generated captions, the fine-tuned model consistently outperformed the randomly fine-tuned model (Figure 6a), confirming the generalizability and superiority of targeted fine-tuning for improving model performance.

5.3. Learning hard concepts from DaLL-E 3

Observation. Figure 5 shows that 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.

Finetuning. Using the taxonomy of GENERATE ANY SCENE, we compute model performance on each conceptby averaging scores across captions containing that concept.Accumulating results through the taxonomy, we identify the 100 concepts where *Stable Diffusion v1-5* shows the largest performance gap relative to *DaLL-E 3*. For fine-tuning, we generate 778 captions incorporating these concepts with others, using *DaLL-E 3* to produce corresponding images. As a baseline, we randomly select 778 GENERATE ANY SCENE-generated captions for fine-tuning and compare these with the original *Stable Diffusion v1-5* model.

Results. The results in Figure 6c show that our targeted fine-tuning led to improved model performance, reflected in higher average scores across captions and increased scores for each challenging concept.

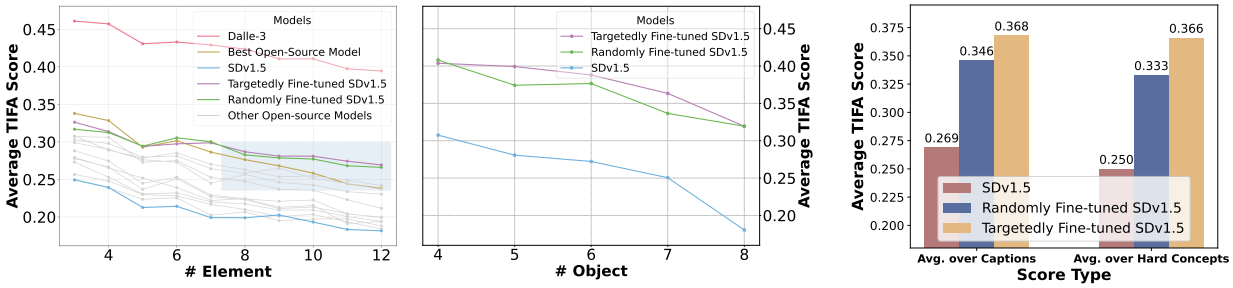
Takeaway for application 2

Targeted fine-tuning can distill proprietary model strengths, effectively bridging gaps in compositionality and concept handling for open-source models.

Potential Reason: GENERATE ANY SCENE facilitates fine-grained analysis to identify specific performance gaps, enabling targeted data selection to distill limitations.

6. Application 3: Generated content detector

Advances in *Text-to-Vision generation* underscore the need for effective content moderation [42]. Major challenges include the lack of high-quality and diverse datasets and the difficulty of generalizing detection across models *Text-to-Vision generation* [26, 58]. GENERATE ANY SCENE ad-

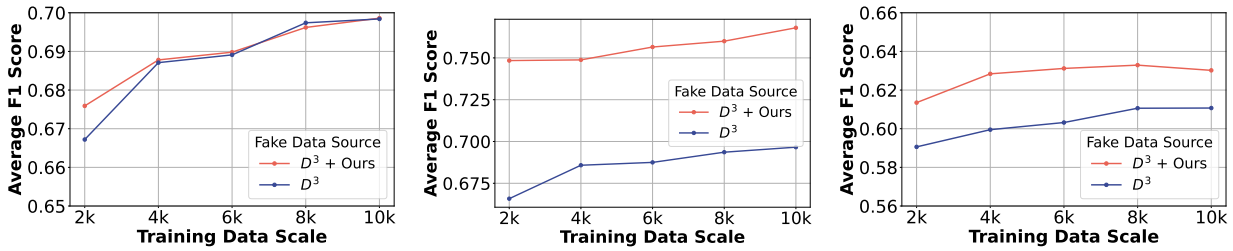


(a) **Distilling compositionality from DaLL-E 3**: Model results on TIFA vs. total element numbers in captions in 10K general GENERATE ANY SCENE captions.

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

Figure 6. **Results for Application 2: Distilling limitations**. The left two figures show the results for **Distilling compositionality from DALL-E 3**, while the rightmost figure shows the results for **Learning hard concepts from DALL-E 3**.



(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 7. **Results for Application 3: 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.

dresses these issues by enabling scalable, programmatic 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.

To demonstrate GENERATE ANY SCENE's effectiveness in training generated content detectors, we used the D^3 dataset [2] 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 [62] model with a single-layer linear classifier, varying dataset sizes with N real and N synthetic images. For synthetic data, we compared N samples solely from D^3 with a mixed set of $N/2$ from GENERATE ANY SCENE and $N/2$ from D^3, keeping the same training size.

We evaluate the detector's generalization on the GenImage [72] validation set and images generated using GENERATE ANY SCENE captions. Figure 7 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 G.

Takeaway for application 3

Compositional synthetic captions robustify generated content detectors.

Potential reason: GENERATE ANY SCENE can generate more diverse captions to complement real-world image-caption training data by enriching compositional variety and imaginative scope.

7. 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 imaginary scenes and even implausible scenarios. To demonstrate the effectiveness of GENERATE ANY SCENE, we explore three applications: (1) self-improvement by iteratively optimizing models, (2) distillation of proprietary model strengths into open-source models, and (3) robust content moderation with diverse synthetic data. GENERATE ANY SCENE highlights the importance of synthetic data in evaluating and improving *Text-to-Vision generation*, and addresses the need to systematically define and scalably produce the space of visual scenes.

References

[1] 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. 4, 16, 17

[2] 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. 8

[3] 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. 16

[4] 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. 1, 4, 16, 17

[5] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2): 115, 1987. 1

[6] 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. 1, 16

[7] Tuhin Chakrabarty, Kanishk Singh, Arkadiy Saakyan, and Smaranda Muresan. Learning to follow object-centric image editing instructions faithfully. *ArXiv*, abs/2310.19145, 2023. 4

[8] 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. 1, 3, 18

[9] 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. 4, 5, 16, 17

[10] 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. 4, 16, 17

[11] 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. 4, 16, 17

[12] 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. 3, 4

[13] 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. 16

[14] Colby Crawford. 1000 cameras dataset. https://www.kaggle.com/datasets/crawford/1000-cameras-dataset, 2018. Accessed: 2024-11-09. 16

[15] 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. 4, 16, 17

[16] 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. 16

[17] 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. 4, 5, 16, 17

[18] 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. 20

[19] 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. 5

[20] 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. 4

[21] 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. 4, 17

[22] 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. 3

[23] 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. 1

[24] 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.

[25] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE con-*

ference on computer vision and pattern recognition, pages 1219–1228, 2018. 1

[26] 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. 7

[27] 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. 4, 5, 16, 17

[28] 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. 4

[29] 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. 1, 2, 3

[30] Black Forest Labs. Flux.1: Advanced text-to-image models, 2024. Accessed: 2024-11-10. 4, 16, 17

[31] 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. 1, 6, 14

[32] 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. 4, 16, 17

[33] 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. 4, 5, 16, 17

[34] 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. 2, 4, 5, 17

[35] 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. 12

[36] Alejandro López-Cifuentes, Marcos Escudero-Vinolo, Jesús Bescós, and Álvaro García-Martín. Semantic-aware scene recognition. *Pattern Recognition*, 102:107256, 2020. 3, 16

[37] 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. 1

[38] 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. 4, 5, 16, 17

[39] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 3, 14

[40] 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. 1

[41] Jae Sung Park, Zixian Ma, Linjie Li, Chenhao Zheng, Cheng-Yu Hsieh, Ximing Lu, Khyathi Chandu, Norimasa Kobori Quan Kong, Ali Farhadi, and Ranjay Krishna Yejin Choi. Robin: Dense scene graph generations at scale with improved visual reasoning. 2024. 3, 16

[42] 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. 7

[43] 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. 4, 16, 17

[44] 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. 4, 16, 17

[45] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 4, 5, 16, 17

[46] 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. 2

[47] 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. 1, 16

[48] 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, 2022. 4, 16, 17

[49] 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. 16

[50] 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. 16

[51] 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. 5

[52] Spencer Sterling. zeroscope_v2_576w, 2023. Accessed: 2024-11-10. 4, 5, 16, 17

[53] 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. 2, 5

[54] 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. 16

[55] 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. 4, 5, 16, 17

[56] 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. 4, 5, 16, 17

[57] 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. 4, 5, 16, 17

[58] 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. 7

[59] 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. 4, 5, 16, 17

[60] 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. 1, 16

[61] Wikipedia Contributors. Lists of colors. https://en.wikipedia.org/wiki/Lists_of_colors, 2024. Accessed: 2024-11-09. 3, 16

[62] 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. 8, 20

[63] 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. 4, 5, 16, 17

[64] 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. 14

[65] 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. 14

[66] 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. 4

[67] 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. 16

[68] 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. 16

[69] 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. 4, 5, 16, 17

[70] 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. 2

[71] 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, 2024. 4, 5, 16, 17

[72] 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. 8

Generate Any Scene: Evaluating and Improving Text-to-Vision Generation with Scene Graph Programming

Supplementary Material

A. More Analysis with GENERATE ANY SCENE

With GENERATE ANY SCENE, we can generate infinitely diverse and highly controllable prompts. Using GENERATE ANY SCENE, we conduct several analyses to provide insights into the performance of today's *Text-to-Vision generation* models.

A.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 [35], 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 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.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 prompts 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:

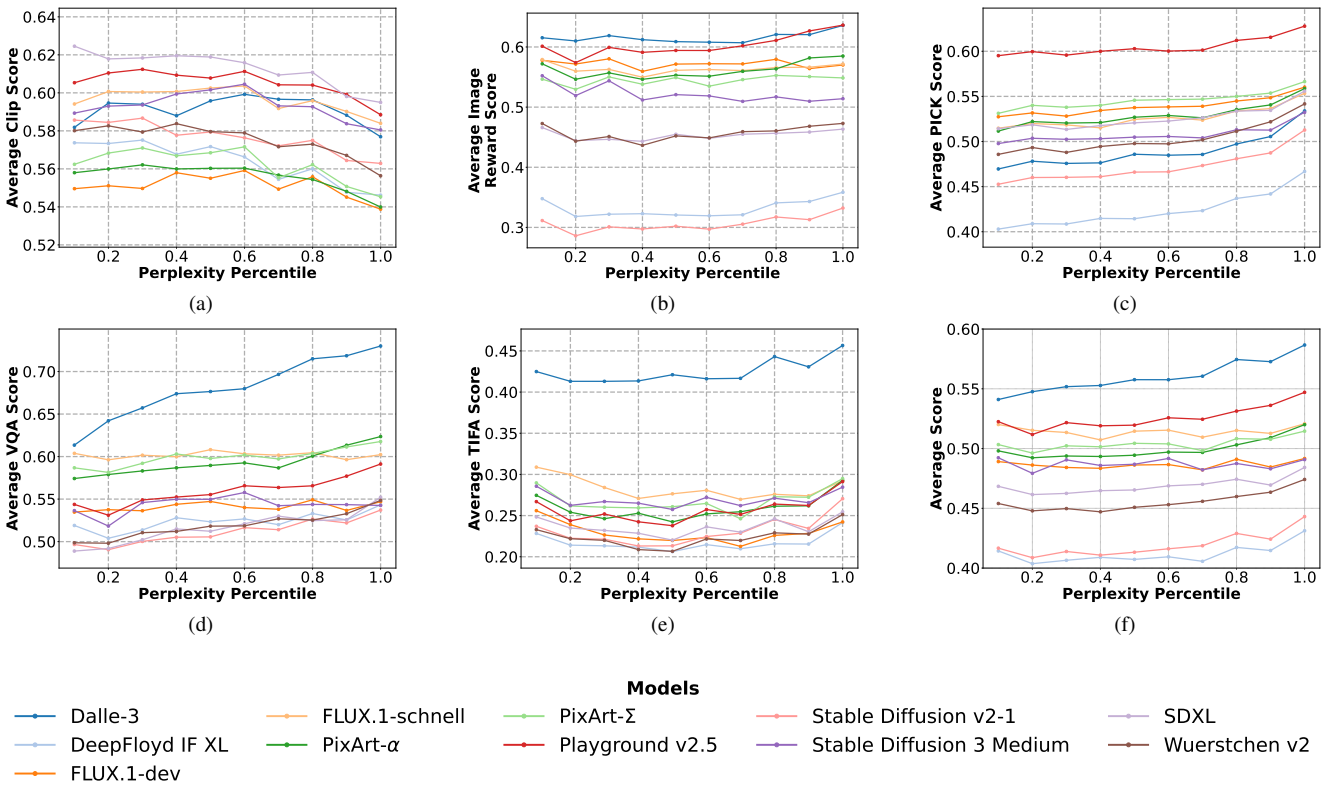


Figure 8. Average performance of models across different percentiles of perplexity of captions, evaluated on various metrics. From left to right, the perplexity decreases, indicating captions that are progressively more reasonable and easier for the LLM to generate.

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: it assigns disproportionately large negative scores to models like *Stable Diffusion v2-1*, 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.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".

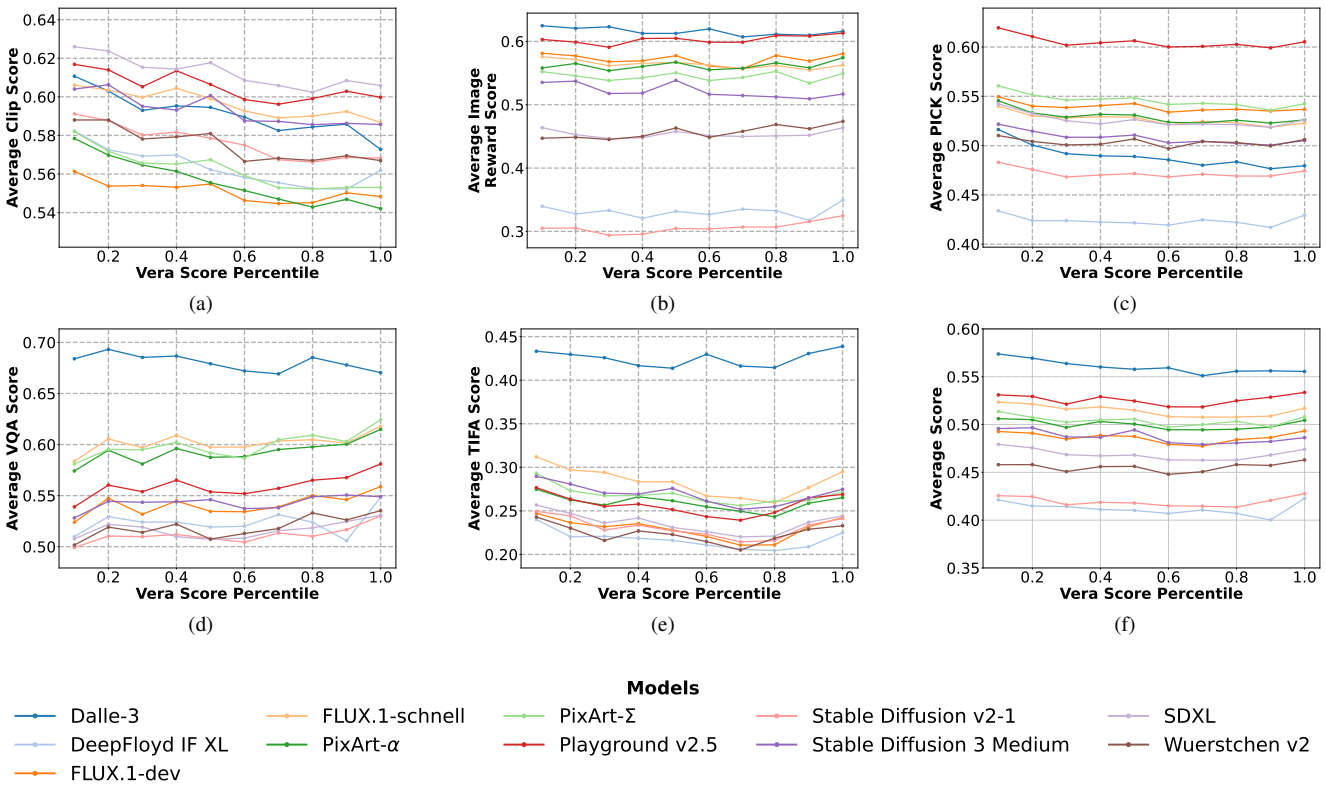


Figure 9. Average performance of models across different percentiles of Vera Score for captions, evaluated on various metrics. From left to right, the Vera Score decreases, indicating captions that exhibit less commonsense reasoning and are more likely to describe implausible scenes.

B. 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 [31, 64, 65]. 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 [65] and GenAI Bench [31], 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.

C. 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 prompts. 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 [39]. 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

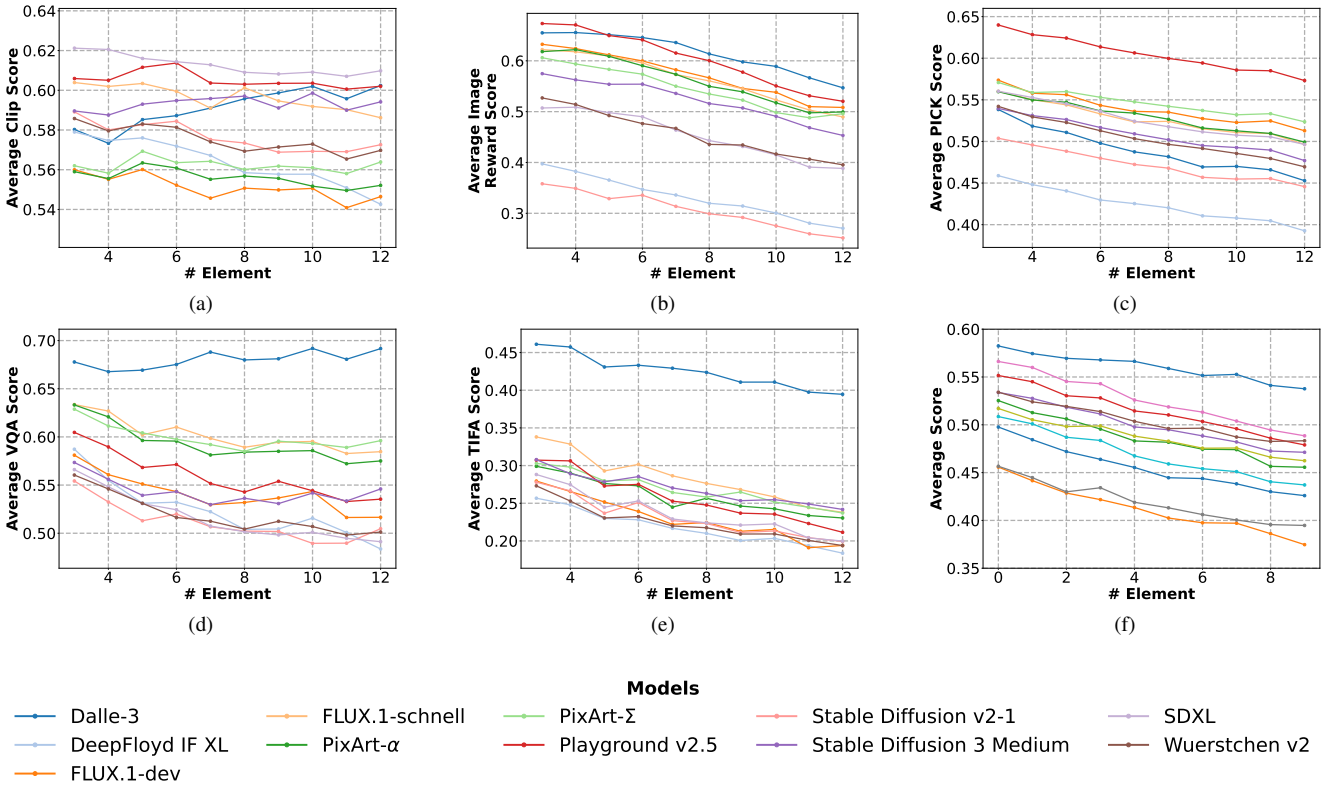


Figure 10. Average performance of models across different numbers of elements (objects + attributes + relations) in the scene graph (complexity of the scene graph) of the captions, evaluated on various metrics. From left to right, as the number of elements (complexity) increases, the scene graphs become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

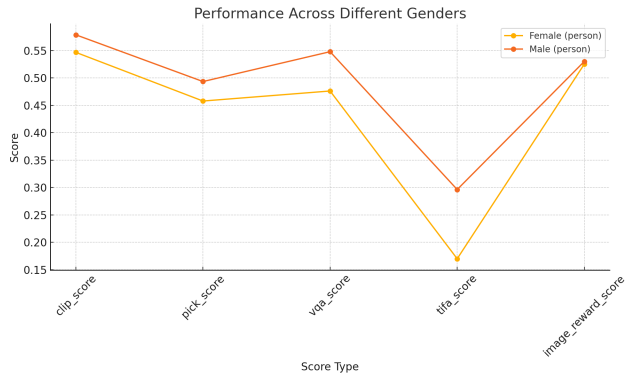


Figure 11. Average performance scores of all models across different genders were evaluated using various metrics.

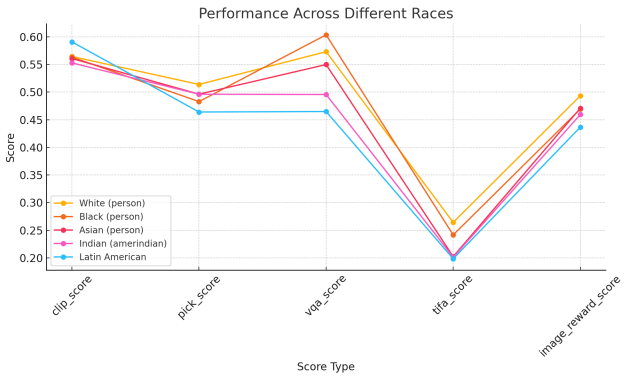


Figure 12. Average performance scores of all models across different races evaluated using various metrics.

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.

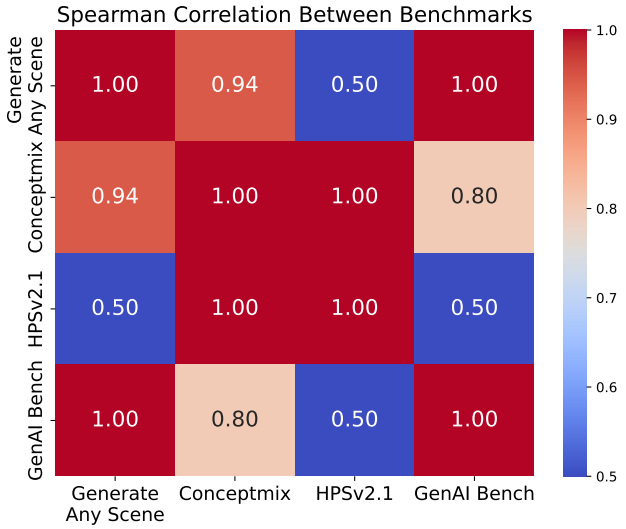


Figure 13. Correlation of GENERATE ANY SCENE with other popular *Text-to-Vision generation* benchmarks.

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 [61]. The *material* category comprises *76* types, referenced from several public datasets [3, 54, 68]. The *texture* category includes *42* kinds from the Describable Textures Dataset [13], while the *architectural style* encompasses *25* distinct styles [67]. 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 [41] 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 [4, 47, 49]. For artistic styles, we define scene attributes using *76* renowned *artists*, *41 genres*, and *126 painting styles* from WikiArt [50], 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 [14], while the remaining categories are constructed based on photography knowledge and common prompts in *Text-to-Vision generation* tasks [6, 60]. To control scene settings, we categorize location, weather and lighting attributes, using 430 diverse locations from Places365 [36], 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.

D. Details of Overall Performance (Section 3)

D.1. Detailed experiment settings

- For *Text-to-Image generation*, we select a range of open-source models, including those utilizing UNet backbones, such as *DeepFloyd IF* [1], *Stable Diffusion v2-1* [48], *SDXL* [44], *Playground v2.5* [32], and *Wuerstchen v2* [43], as well as models with DiT backbones, including *Stable Diffusion 3 Medium* [15], *PixArt-α* [10], *PixArt-Σ* [11], *FLUX.1-schnell* [30], *FLUX.1-dev* [30], and FLUX 1. Closed-source models, such as *DaLL-E 3* [4] and *FLUX1.1 PRO* [30], 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* [57], *ZeroScope* [52], *Text2Video-Zero* [27], *CogVideoX-2B* [69], *VideoCrafter2* [9], *AnimateLCM* [55], *AnimateDiff* [17], *FreeInit* [63], and *Open-Sora 1.2* [71]. 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* [56], *DreamFusion* [45], *Magic3D* [33], *Latent-NeRF* [38], and *Prolific-Dreamer* [59]. We employ the implementation and configurations provided by ThreeStudio [16] 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 *Stable Diffusion v2-1* as their 2D backbones.

Metrics. Across all *Text-to-Image generation*, *Text-to-Video generation*, and *Text-to-3D generation*, we employ five widely used *Text-to-Vision generation* metrics to comprehensively assess model performance:

- *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 [21] 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 [34].

D.2. Detailed overall results

We evaluate *Text-to-Image generation*, *Text-to-Video generation*, and *Text-to-3D generation* models on GENERATE ANY SCENE. The detailed results of each model and each metric are shown in Tabs. 5 to 8

Model	clip score	pick score	vqa score	tifa score	image reward score
Playground v2.5 [32]	0.2581	0.2132	0.5734	0.2569	0.2919
Stable Diffusion v2-1 [48]	0.2453	0.1988	0.5282	0.2310	-0.9760
SDXL [44]	0.2614	0.2046	0.5328	0.2361	-0.3463
Wuerstchen v2 [43]	0.2448	0.2022	0.5352	0.2239	-0.3339
DeepFloyd IF XL [1]	0.2396	0.1935	0.5397	0.2171	-0.8687
Stable Diffusion 3 Medium [15]	0.2527	0.2027	0.5579	0.2693	-0.0557
PixArt-α [10]	0.2363	0.2050	0.6049	0.2593	0.1149
PixArt-Σ [11]	0.2390	0.2068	0.6109	0.2683	0.0425
FLUX.1-dev [30]	0.2341	0.2060	0.5561	0.2295	0.1588
FLUX.1-schnell [30]	0.2542	0.2047	0.6132	0.2833	0.1251
FLUX1.1 PRO [30]	0.2315	0.2065	0.5744	0.2454	-0.0361
Dalle-3 [4]	0.2518	0.2006	0.6871	0.4249	0.3464

Table 5. Overall performance of *Text-to-Image generation* models over 10K GENERATE ANY SCENE prompts.

Model	clip score	pick score	image reward score	VQA score	TiFA score
VideoCraft2 [9]	0.2398	0.1976	-0.4202	0.5018	0.2466
AnimateDiff [17]	0.2610	0.1959	-0.7301	0.5255	0.2208
Open-Sora 1.2 [71]	0.2259	0.1928	-0.6277	0.5519	0.2414
FreeInit [63]	0.2579	0.1950	-0.9335	0.5123	0.2047
ModelScope [57]	0.2041	0.1886	-1.9172	0.3840	0.1219
Text2Video-Zero [27]	0.2539	0.1933	-1.2050	0.4753	0.1952
AnimateLCM [55]	0.2450	0.1987	-0.5754	0.4816	0.2176
CogVideoX-2B [69]	0.2038	0.1901	-1.2301	0.4585	0.1997
ZeroScope [52]	0.2289	0.1933	-1.1599	0.4892	0.2388

Table 6. Overall performance of *Text-to-Video generation* models over 10k GENERATE ANY SCENE prompts.

Model	subject consistency	background consistency	motion smoothness	dynamic degree	aesthetic quality	imaging quality
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
Open-Sora 1.2	0.9964	0.9907	0.9973	0.0044	0.5235	0.6648
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
Text2Video-Zero	0.8471	0.9030	0.8301	0.9999	0.4889	0.7018
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

Table 7. Overall performance of *Text-to-Image generation* models over 10k GENERATE ANY SCENE prompts with VBench metrics.

Model	clip score	pick score	vqa score	tifa score	image reward score
ProlificDreamer [59]	0.2125	0.1940	0.5411	0.2704	-1.2774
Latent-NeRF [38]	0.2115	0.1910	0.4767	0.2216	-1.5311
DreamFusion-sd [45]	0.1961	0.1906	0.4421	0.1657	-1.5582
Magic3D-sd [33]	0.1947	0.1903	0.4193	0.1537	-1.6327
SJC [56]	0.2191	0.1915	0.5015	0.2563	-1.4370
DreamFusion-IF [45]	0.1828	0.1857	0.3872	0.1416	-1.9353
Magic3D-IF [33]	0.1919	0.1866	0.4039	0.1537	-1.8465

Table 8. Overall performance of *Text-to-3D generation* models over 10k GENERATE ANY SCENE prompts.

D.3. 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.

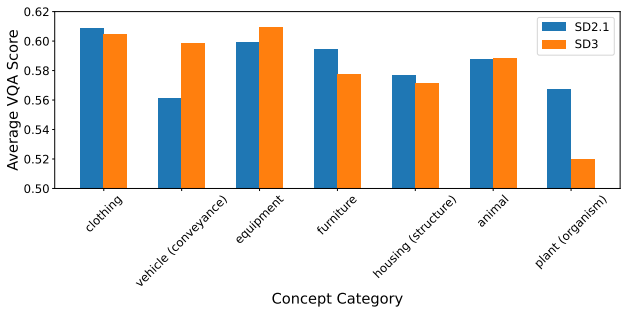


Figure 14. *Stable Diffusion v2-1* vs. *Stable Diffusion 3 Medium* on average *VQA Score* in fine-grained categories.

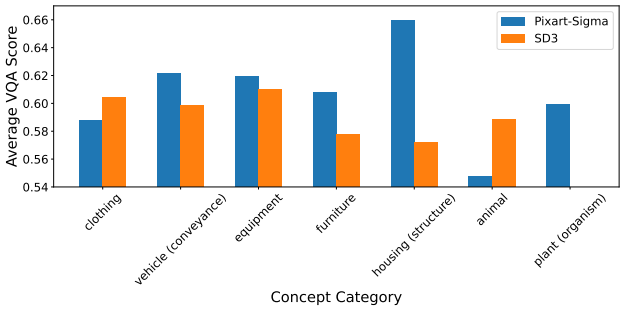


Figure 15. *PixArt-Σ* vs. *Stable Diffusion 3 Medium* on average *VQA Score* in fine-grained categories.

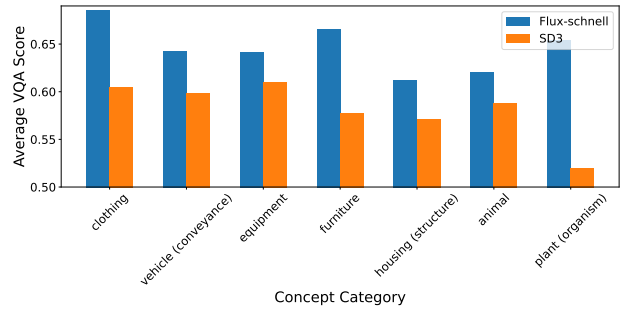


Figure 16. *FLUX.1-schnell* vs. *Stable Diffusion 3 Medium* on average *VQA Score* in fine-grained categories.

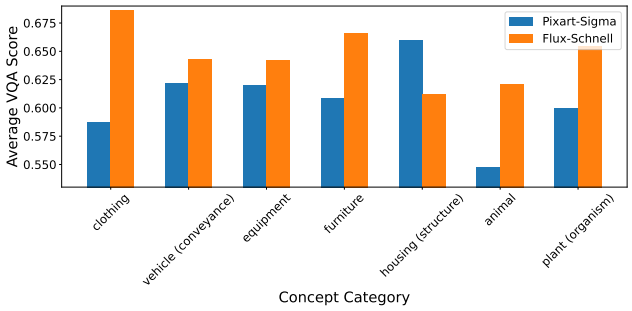


Figure 17. *PixArt-Σ* vs. *FLUX.1-schnell* on average *VQA Score* in fine-grained categories.

E. Details of Application 1: Self-Improving Models (Section 4)

E.1. Experiment details

E.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) [8] 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.

E.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 prompt, 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 prompt. 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 prompt, 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 prompt 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.

E.1.3 Fine-tuning details

We fine-tuned the *Stable Diffusion v1-5* 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 *Stable Diffusion v1-5* 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.

E.2. More results of fine-tuning models

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 18), where we observed the same trend: models fine-tuned with our prompts consistently outperformed the original *Stable Diffusion v1-5* and CC3M fine-tuned models.

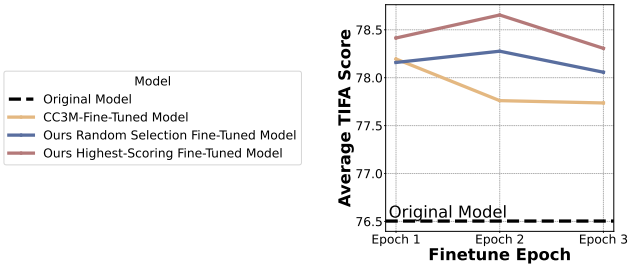


Figure 18. **Results for Application 1: Self-Improving Models**. Average TIFA score of *Stable Diffusion v1-5* fine-tuned with different data over TIFA Bench.

F. Details of Application 2: Distilling limitations (Section 5)

F.1. Collecting hard concepts

We selected 81 challenging object concepts where *Stable Diffusion v1-5* and *DaLL-E 3* exhibit the largest gap in *VQA Score*. To determine the score for each concept, we calculated the average VQA score of the captions containing that specific concept. 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)

F.2. Experiment details

We conducted targeted fine-tuning experiments on *Stable Diffusion v1-5* 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 *Stable Diffusion v1-5* 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 E.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.

G. Details of Application 3: Generated content detector (Section 3)

G.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 [62] and ResNet-18 [18] 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: *Stable Diffusion v2-1*, *PixArt-α*, *Stable Diffusion 3 Medium*, and *Playground v2.5*. This created distinct test sets for evaluating model performance across different synthetic data sources.

G.2. Results

Table 10 and Table 9 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.

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

Table 9. 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)	SDv1.5		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

Table 10. 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