TITScore: Towards Long-Tail Effects in Text-to-Visual Evaluation with Generative Foundation Models

Pengliang Ji
Carnegie Mellon University
pengliaj@andrew.cmu.edu

Junchen Liu
UC Berkeley
liujunchen0214@berkeley.edu

Abstract

Evaluation of generative foundation models (GenFMs) for text-to-visual tasks has been enhanced by automatic alignment metrics such as CLIPScore, complementing human feedback. However, existing evaluation methods suffer from a severe long-tail effect, where the balance between token count and semantic validity in the initial step, hinders the accurate evaluation of advanced aspects such as composition. We analyze this drawback and attribute it to a lack of symbolic reasoning attention, while GenFMs demonstrate strong discriminative abilities in handling symbolism. To this end, we propose a pioneering paradigm for evaluating GenFMs' text-to-visual (T2V) generation using neuro-symbolic thinking to mitigate the long-tail effect. By explicitly embedding Mixture-of-experts (MoE) Large Vision Models (LVMs), we introduce symbolic-level understanding while maintaining the strong neuro-level reasoning capability. Through the fusion of semantic and compositional knowledge at the neuro-to-symbolic level, our approach outperforms state-of-the-art T2V evaluation methods, exhibiting stronger compositional reasoning ability on Winoground and better alignment with human judgment. We also demonstrate our impressive effectiveness on diverse tasks, including text-to-3D and text-to-video. To further advance the T2V evaluation of GenFMs, we propose a challenging benchmark that includes richer and more diverse compositional and semantic information compared to Winoground. Overall, our work opens a new direction for neuro-to-symbolic visio-linguistic evaluation of GenFMs and aims to drive further progress in the field.

1. Introduction

The rapid advancement of large language models (LLMs) has propelled generative foundation models (GenFMs) to become one of the most exciting achievements in modern artificial intelligence. Impressive text-to-visual generative foundation models from both industry [3, 5, 7, 10] and the open-source community [2, 8, 26, 52, 68] have showcased the ability to generate highly creative visual content, ranging from simple animations to lifelike scenes.

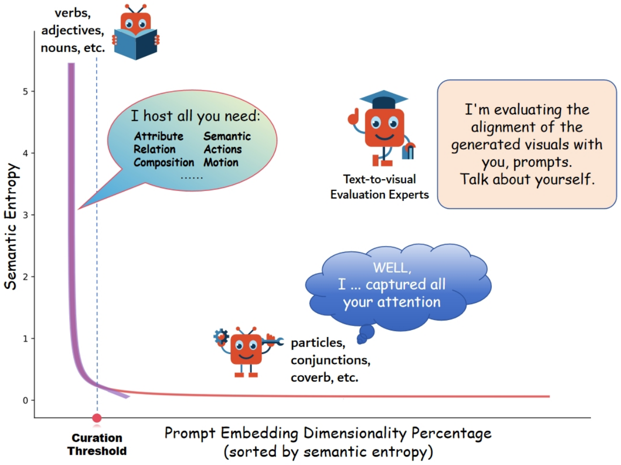


Figure 1. **Long-tail effects between semantic entropy and embedding sizes.** The discrepancy between essential and non-essential elements hampers the extraction of instructive knowledge from prompts. This imbalance shifts focus away from crucial components, diminishing their significance and leading to evaluations that overlook advanced compositional and semantic complexities.

However, despite the emergence of GenFMs, the generative AI community still lacks a robust metric that effectively assesses alignment between generated visual contents and text prompts. Existing methods often rely on subjective human evaluations [11, 33, 43, 56], which are costly and difficult to replicate consistently. Recent studies shift towards automatic metrics such as CLIPScore [19], which assesses the cosine similarity on the latent space. Despite these efforts, achieving precise vision-language evaluation remains a significant challenge, as it requires evaluation methods to possess advanced semantic and compositional reasoning capabilities [23]. Our findings suggest that these capabilities are largely lost during the encoding process, as evidenced by the presence of a long-tail effect.

Long-tails Effects in Text-to-visual Evaluation. Existing comprehensive evaluation methodologies for text-to-image generation [30] and language tasks [6] suffer from severe long-tailed effects. This issue arises from an overemphasis on non-essential elements in prompts, resulting in an imbalance between knowledge representation and embedding dimensionality. Specifically, encoding entire prompts without considering the irrelevance of many grammatical components compresses key knowledge into a small portion of the input while underutilizing the remainder of the model's capacity. Additionally, text-to-image model evaluations typically incorporate unnecessary positional encoding, decreasing accuracy and causing computational inefficiency. As shown in Fig.1, our analysis of 1,000 prompts from diverse benchmarks reveals a substantial imbalance, with a significant portion of the semantic focus misallocated to less essential words, resulting in the loss of critical details such as compositional knowledge essential for accurate evaluation. These findings highlight the need for better evaluation methods that focus on semantic relevance and compositional integrity to improve assessment accuracy.

Neuro-symbolic Paradigm. Neuroscience insights indicate that human cognition is profoundly shaped by both neuro and symbolic factors, notably through the comprehension of semantics and composition [48]. This integration is embodied within the neuro-symbolic paradigm. In the domain of model evaluation, a recent triumph is the utilization of visual question-answering models [21, 57] for generative models evalution, signifying a substantial advancement. This success is attributable to the harmonious interplay between neuro (semantics) and symbolic (composition) aspects. The paradigm assesses compositional understanding through human-generated queries, while semantic comprehension is evaluated using sophisticated visual language foundation models.

Mixture of Experts in Evaluation. Motivated by the aforementioned problem and our insights into the neuro-symbolic paradigm, we introduce a novel Mixture-of-Experts (MoE) framework tailored for generative foundation models to enhance evaluation by addressing the long-tail effects and improving reasoning capabilities. This framework weights essential input components, utilizes expertise large vision models (LVMs) for diverse evaluation facets with neuro-symbolic thinking, and fosters a self-improving cycle by bootstrapping generative models. As the paradigm of our evaluation pipeline in Fig.2 shows, We propose an effective prompt curation module to mitigate the long-tail effects on prompt semantics. Subsequently, we leverage our specialized models for in-depth semantic and compositional analysis, integrating the findings across both symbolic scalar and neuro-embedding spaces to enhance knowledge acquisition. Ultimately, Our approach calculates similarity scores, TlTScore, in latent space and offers the option to include a GPT model for enhanced language understanding. TlTScore surpasses existing VQA-based and divide-and-conquer-based evaluation methods as the state-of-the-art in semantic and compositional reasoning on the challenging benchmark Winoground [50]. Unlike divide-and-conquer-based methods [57, 64] that naively naively split prompts and lose compositionality, our solution essentially solves long-tail issues and significantly enhances reasoning capabilities through Mixture-of-experts. Furthermore, our evaluation pipeline includes models with fixed parameters that are open-source, ensuring stable and consistent evaluations in practice. In contrast, the current state-of-the-art models depend on closed-source systems such as GPT-4Vision [69] or GPT-4 [38], which suffer from fluctuating performance due to API updates.

Comprehensive Evaluation Benchmark. Developing a robust evaluation benchmark for generative models is crucial for providing valuable feedback and driving improvement. The primary challenge is that current benchmarks focus mainly on prompts with simple semantics and composition, with only a few methods [23, 50] considering complex scenarios, while being limited by insufficient evaluation aspects and samples. To address this problem, we introduce a comprehensive, fine-grained, and semantically rich evaluation benchmark covering 16 aspects and featuring 2,400 prompts to enable in-depth insights into generative model evaluation. Unlike all existing benchmarks, our benchmark targets both the encoder and decoder processes of generative models and includes an assessment of metrics to determine their effectiveness in providing accurate and valuable evaluation.

Easy-to-use API to assist evaluation. To comprehensively assess the faithfulness of generated visuals across all prompts, we have developed the first API that enables evaluation with just a single line of code. This advancement aligns with our core mission to push the boundaries of generative model evaluation beyond cherry-picked analysis.

In conclusion, our contributions are summarized below:

- We identify a severe long-tail effect in the evaluation of generative models often overlooked by existing methods, and propose a prompt curation module to solve this issue.
- We introduce TlTScore, an effective state-of-the-art method employing mixture-of-experts models combined with neuro-symbolic reasoning for visio-linguistic evaluation across diverse tasks.
- TlTBench, a comprehensive benchmark for text-to-visual evaluation and metric validation, includes over 2,400 diverse prompts enriched with semantics and compositional knowledge and annotations across 16 aspects.
- An easy-to-use one-line-code API that can be effortlessly embedded into existing pipelines for efficient evaluation of generative vision-language foundation models.

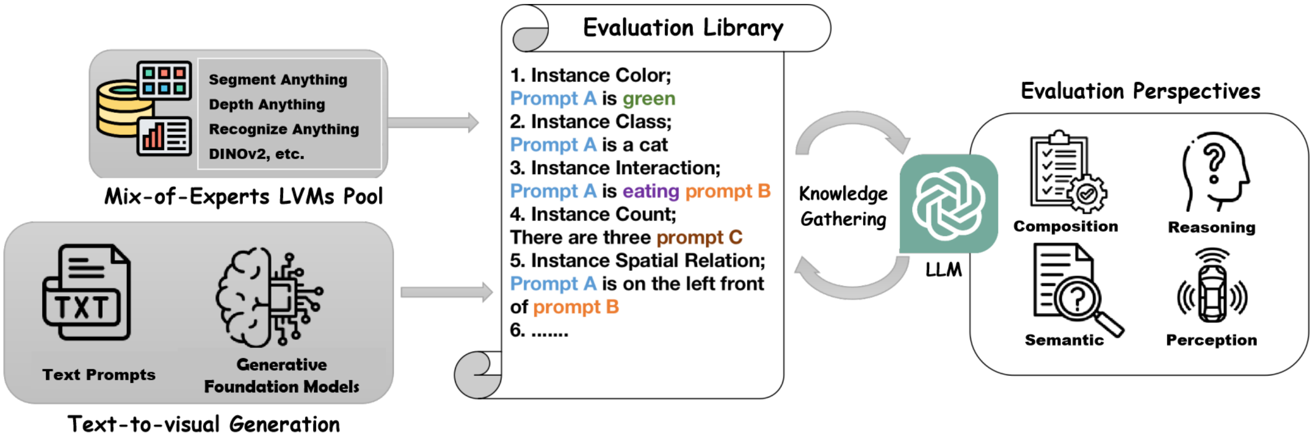


Figure 2. **A Comprehensive Framework for Text-to-Visual Evaluation in GenFMs**. We propose an effective paradigm for comprehensively evaluating text-to-visual generation. Text prompts and generated visuals are fed into a diverse pool of expert large vision models (LVMs) specializing in various evaluation aspects like segmentation, perception, recognition, etc. The outputs are then analyzed against a semantically-rich evaluation library. Finally, the knowledge is integrated into an evaluation system with diverse aspects including composition, reasoning, semantics, and so on, creating a comprehensive framework for evaluating GenFMs' text-to-visual capabilities.

2. Related Works

2.1. Text-to-visual Generative Models Evaluation

Initially, the evaluation of generative foundation models heavily relied on human ratings through user cases, and stood as the primary means of quantitative assessment [11, 33, 43, 56]. The high cost of human evaluations led to the adoption of automated metrics like the Fréchet Inception Distance [20], Inception Score [47], and CLIPScore [19], which gauge the feature similarity between text prompts and generated visuals. Although these metrics are good at assessing visual quality, they struggle to measure the intricate text-visual alignment in visio-linguistic content, leading to mismatches with human preferences [23]. To address these limitations, researchers have turned to multimodal large language models (MLLMs), such as LLaVA [34], Llama-2 [51], and BLIP [32], for text-to-visual evaluation. These approaches involve strategies like Visual Question Answering [21], fine-tuning based on human feedback [28, 59], and applying the Chain-of-Thought technique to models like GPT-4V [29, 69]. Most methods encode text prompts directly into latent space without prioritizing importance, creating a significant gap between the number of tokens and semantic entropy. Furthermore, using proprietary models like GPT-4Vision [1] hinders scalability and consistent evaluation due to practical constraints. The critical absence of an accessible and effective evaluation method impedes the progress of generative foundation models.

2.2. Benchmark for Comprehensive Evaluation

The Evaluation Benchmark Framework comprises two primary categories: alignment benchmarks, represented by Winoground [50], EqBen [54], TIFA160 [21], and Pick-a-pic [28], assess models' ability to maintain consistency and faithfulness between text and generated visuals during the encoding process. In contrast, generation benchmarks, including PartiPrompt [65], DrawBench [46], EditBench [53], and EvalCrafter [36], evaluate a model's generative capabilities during the decoding process. However, to the best of our knowledge, no current benchmark addresses both alignment and generation, which relate to semantic and compositional knowledge, respectively. On the other hand, a comprehensive evaluation benchmark requires high-quality prompts from multiple angles and offers precise human assessments. Unfortunately, such benchmarks are also still very scarce. Winoground [50] stands out as a relatively semantically rich benchmark for evaluating models' advanced compositional abilities. However, its scope is narrow and it is limited in scale, containing only 400 prompts. Moreover, many evaluations rely on uncurated online datasets [28], arbitrary user ratings [28], and unverified GPT-generated prompts [29, 58], leading to erroneous and unreliable results in the evaluation of generative models.

3. Datasets

3.1. Challenges of Multimodal GenFMs Evaluation

Multimodal GenFMs Evaluation is highly correlated with benchmark quality and evaluation scope [6]. Currently, the main challenges in evaluating the text-to-visual generation of generative models can be categorized into two aspects:

Abundance of uninformative texts in Evaluation Benchmarks. Current benchmarks assess text-to-visual alignment using formats such as multiple-choice or question-answering [23, 37, 57, 66, 67]. However, a significant portion of these benchmarks include text samples that lack informativeness, allowing the correct answer to be identified without consulting the visually rich semantic content [24, 25, 39, 61–63]. This results in both inaccurate and misleading evaluations. Regrettably, more than half of the existing benchmarks are affected by this problem.

Limited Evaluation Scope in Existing Techniques. Most current evaluation methods for generated visual content focus on basic semantics [6, 37, 55, 66], lacking robustness in critical aspects like challenging composition due to limited diverse samples coverage. When assessing out-of-distribution content with varied styles or qualities, like focus and shadow, existing methods miss crucial nuances, underscoring the necessity for more comprehensive evaluation techniques. To address the aforementioned limitations of text-to-visual generative models' evaluation, we propose a new targeted benchmark TlTBench.

3.2. TlTBench: A Comprehensive Benchmark for Semantically-rich Text-to-Visual Evaluation

TlTBench is designed to mitigate long-tail effects in text semantics, characterized by its rich semantic and comprehensive coverage of evaluation aspects. It features prompt-visual pairs across 16 essential evaluation aspects in Tab.1, supported by human feedback annotations. This benchmark facilitates comprehensive generative model evaluation and metrics validation, emphasizing often overlooked composition, relation, and semantic fidelity.

TlTBench-A includes a set of 2,400 diverse, high-quality prompts, specifically focusing on 16 nuanced evaluation dimensions that encompass visual and compositional reasoning abilities. An illustrative prompt, such as ”Two real bears playing with a brown teddy bear in front of the tree, rather than one behind it,” assesses the model's composition, distinction, counting, and recognition ability.

TlTBench-B targets the evaluation of metrics by addressing the shortcomings, especially in compositional reasoning. It features over 1,000 question-choice pairs with visuals and human ratings to measure the reasoning accuracy of evaluation metrics and their alignment with human judgments. For instance, a question like ”How many bears are playing with the brown teddy bear in front of the tree?” with choices ranging from zero to three, tests the metrics' capability to assess the quality of generated content.

Creation of High-quality prompts. Our prompt development process is grounded in identifying key evaluation dimensions, including composition (spatial, relations), semantics(consistency, concept), etc., as shown in Tab.1. Each prompt, manually crafted, covers two to five of these aspects to ensure relevance and comprehensiveness.

Human Judgments. We generated visual content by employing five prominent text-to-image models, including Stable Diffusion [45], Midjourney [4], DALLE 3 [7], and others. The human ratings, scaled from 1 to 5, were compiled following an established annotation methodology [42], yielding human preference scores.

Eval. Aspects	Description
Alignment	Match prompt styles.
Category	Correct genre classification.
Color	Accurate color representation.
Concept	Understands abstract ideas.
Consistency	Maintains thematic coherence.
Counting	Precise object count.
Customization	Adaptation to specific preferences.
Differentiation	Distinguishes similar concepts.
Logic	Follows logical structures.
Quality	Exhibits high aesthetic value.
Relationship	Depicts element interactions.
Semantic	Interprets meaning accurately.
Size	Represents true sizes.
Spatial	Correct spatial arrangement.
Symbol	Recognizes symbolic meanings.
Texture	Captures surface qualities.

Table 1. **Evaluation Dimensions in TlTBench.** TlTBench represents a comprehensive benchmark, encompassing critical evaluation aspects for a fine-grained generative multimodal evaluation.

4. Evaluating Generative Models with Generative Models

Our method integrates three primary components, including prompt curation(4.1), a mixture of experts(4.2), and knowledge gathering(4.3), within a neuro-symbolic framework to facilitate diverse tasks for text-to-visual evaluation.

4.1. Prompt Curation

To mitigate severe long-tail effect in T2V evaluation, we propose a hierarchical methodology for prompt curation. This approach encompasses both high-level classification of evaluation aspects and low-level decomposition, enhancing the overall effectiveness of the evaluation process.

High-level Evaluation Aspect Delineation. We introduce a delineation module, M, for the high-level curation of prompts. Given an input prompt, P, the classifier M yields a set of related evaluation aspects, $\{A_i\}$, where $i \in [1, 16]$, corresponding to the aspects listed in Table 1. For a given prompt, such as ”a campervan parked under the stars in the desert,” the module identifies its evaluation aspects via multi-class classification as $\{A_{\text{category}}, A_{\text{counting}}, A_{\text{relationship}}, A_{\text{spatial}}, A_{\text{semantic}}\}$.

Low-level Conditional Semantic Decomposition. Upon identifying the evaluation aspects, we perform a low-

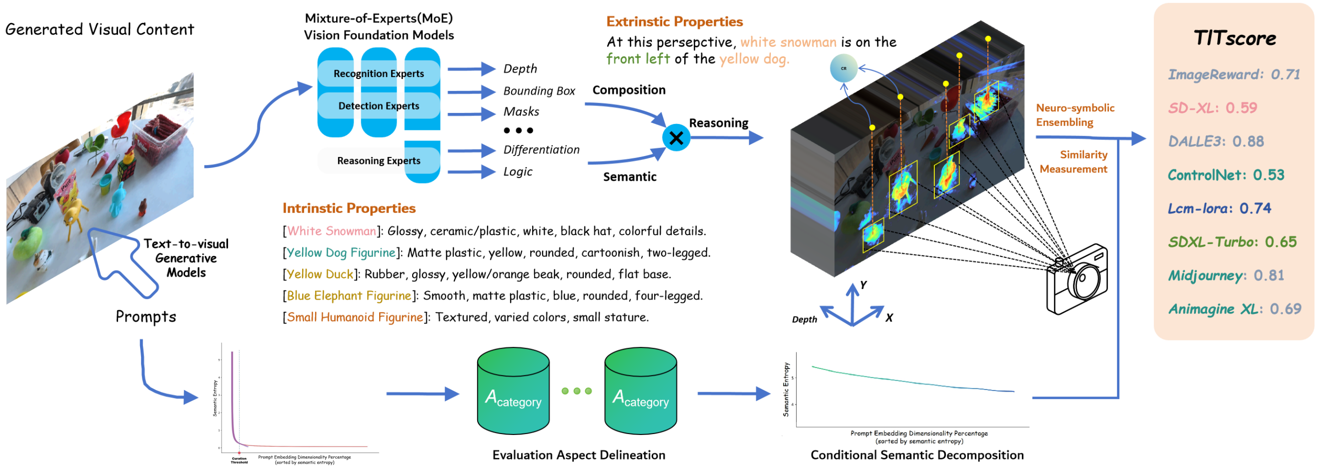

Generated Visual Content
Text-to-visual Generative Models
Prompts
Mixture-of-Experts(MoE) Vision Foundation Models
Recognition Experts
Detection Experts
Reasoning Experts
Depth
Bounding Box
Masks
• • •
Differentiation
Logic
Composition
Semantic
Reasoning
Extrinstic Properties
At this persepctive, white snowman is on the front left of the yellow dog.
Intrinstic Properties
[White Snowman]: Glossy, ceramic/plastic, white, black hat, colorful details.
[Yellow Dog Figurine]: Matte plastic, yellow, rounded, cartoonish, two-legged.
[Yellow Duck]: Rubber, glossy, yellow/orange beak, rounded, flat base.
[Blue Elephant Figurine]: Smooth, matte plastic, blue, rounded, four-legged.
[Small Humanoid Figurine]: Textured, varied colors, small stature.
CR
Y
X
Depth
Neuro-symbolic Ensembling
Similarity Measurement
TITscore
ImageReward: 0.71
SD-XL: 0.59
DALLE3: 0.88
ControlNet: 0.53
Lcm-lora: 0.74
SDXL-Turbo: 0.65
Midjourney: 0.81
Animagine XL: 0.69
Semantic Entropy
Curation Threshold
Prompt Embedding Dimensionality Percentage (sorted by semantic entropy)
$A_{category}$
$A_{category}$
Evaluation Aspect Delineation
Semantic Entropy
Prompt Embedding Dimensionality Percentage (sorted by semantic entropy)
Conditional Semantic Decomposition

Figure 3. **Overview of the Mixture-of-Experts (MoE) Framework for Enhanced Text-to-Visual Evaluation**. We illustrate the MoE framework tailored to refine text-to-visual evaluation for generative models, addressing long-tail effects by emphasizing semantic relevance and compositional integrity. The framework integrates a prompt curation module, minimizing focus on non-essential elements, and employs specialized large vision models (LVMs) for nuanced evaluation. The TITScore leverages both symbolic and neural reasoning for enhanced language understanding and compositional analysis, achieving effective evaluation across diverse text-to-visual tasks.

level decomposition to generate tokens for each identified aspect, resulting in a dictionary representation, $\mathcal{D}$, for the prompt P. Each aspect A_i within $\mathcal{D}$ is associated with a set of tokens that best represent its semantic meaning. Subsequently, we integrate the original prompt embedding, E_p, with the embeddings of the decomposed tokens, $E_{\mathcal{D}}$, to maximize the conditional probability $p(E_{\mathcal{D}}|E_P)$, enhancing the final prompt curation output. The process employs a pre-trained, Robustly Optimized BERT model [35] for aspect identification and semantic breakdown, utilizing the TITBench-A dataset for refinement.

Unlike existing methodologies [21, 57] that split prompts directly into entity pairs (neglecting underlying semantics and comprehensive evaluation aspects) and subsequently generate questions for a single general VQA model (introducing noise back), our approach dissects these semantics in the embedding space. It designates specific experts for evaluating each aspect, yielding enhanced accuracy and coverage with a richer semantic understanding.

4.2. Mixture-of-Expert Structure

Our primary objective is to deliver both precise and efficient evaluations for diverse tasks. To this end, we meticulously develop expert models tailored to rich evaluation aspects, allowing for a thorough evaluation. We facilitate this by dividing our sixteen criteria into two distinct paradigms: the explicit symbolic level and the implicit neuro-level.

Explicit Compositional Reasoning. First, we adopt explicit symbolic approaches for evaluations involving compositional reasoning, representing outcomes as tokens. As our evaluation pipeline Fig.3 shown, the selection of robust visual reasoning models is deliberate, targeting specific tasks such as segmentation (Segment Anything Model [27]), detection (DINOv2 [41]), recognition (LART [44]), depth estimation (Depth Anything Model [60]), as well as their integrative application across diverse tasks. This strategy enables us to comprehensively reach ten compositional dimensions: Alignment, Category, Color, Counting, Differentiation, Relationship, Size, Spatial, Symbol, and Texture.

Implicit Semantic Understanding. In semantic evaluation's evolving landscape, we transition from symbolic visual models to advanced multimodal models, anchored in implicit neural processes effectively addressing nine evaluation dimensions: Alignment, Concept, Consistency, Customization, Logic, Quality, Semantic, Relationship, and Texture. Notably, certain evaluation aspects may intersect across both paradigms, facilitating a richer synthesis of knowledge, and evaluations are quantified as scores, paralleled by encoded representations in the latent space. To be more specific, we fine-tune an adapter after the ViT-Large [14] vision encoder, which consists of two MLP layers [17], for each evaluation aspect. Using human evaluations, the adapter is trained with text embeddings of curated prompts obtained from conditional semantic decomposition (Sec.4.1). Additionally, we directly adopt the official pre-trained weights from MetaCLIP[22], leveraging its strong performance on multimodal tasks, as the starting point for our fine-tuning process.

4.3. Knowledge Gathering

Neuro-symbolic Ensembling. We integrate implicit knowledge, represented by tokens from visual foundation models, with explicit knowledge, manifested as scalar scores and latent embeddings from multimodal models. Prior to integration, implicit knowledge is embedded into structured templates, such as ”there are $\{counting\}$ $\{attribute\}$ $\{category\}$ in this scenario.”, which aligns the knowledge with the respective evaluation aspects. These symbolic embeddings are then encoded using the same text encoder as 4.2. Subsequently, for each evaluation aspect, the symbolic branch embeddings are fused with the neuro-branch embeddings in the latent space, thereby enhancing the compositional and semantic depth of the representation.

Calculation of TITScore. TITScore, our evaluation metric, is computed by comparing the evaluation embeddings, ϕ, with the curated prompt embeddings, ψ, as elaborated in Sec.4.1. The TITScore is calculated by:

$$\mathcal{T} = \text{CosineSimilarity}(\phi, \psi)$$

$$\mathfrak{U} = s(\mathbf{i}, \mathbf{t}; \theta)$$

$$\text{TITScore} = \begin{cases} \sigma(\mathcal{T} + \mathfrak{U}), & \text{if } \mathfrak{U} \text{ is defined,} \\ \mathcal{T}, & \text{otherwise.} \end{cases}$$

where θ represents the aspect of evaluation, $\mathbf{i}$ denotes the image, $\mathbf{t}$ signifies the text in the evaluated pairs, σ is a merge function, and $s(\mathbf{i}, \mathbf{t}; \theta)$ is a scalar score pertinent to each aspect from the segment branch. TITScores range from 0 to 1, with higher scores indicating superior performance.

4.4. Towards a Better Evaluation

Enhancing Evaluation with GPT-4. Our framework includes a more advanced version that incorporates GPT-4 [40] for language tasks, particularly in Prompts Curation. Specifically, we utilize GPT-4 for the zero-shot decomposition of prompts during the initial phase of our methodology (Sec.4.1). In detail, GPT-4 is tasked with dissecting semantics pertinent to each evaluative dimension, thereby providing inputs for Implicit Semantic Analysis. This approach leverages GPT-4’s strong language understanding to improve the precision of our evaluation metric. It is pertinent to mention that our GPT integration is confined to textual analysis, given GPT’s occasional oversight of visual semantics in multimodal evaluations (as illustrated in Sec.5). This decision ensures that our evaluation process remains straightforward and user-friendly.

The Scalability of Evaluation Pipeline. Our evaluation techniques exhibit strong scalability due to the design of the neuro-symbolic mechanism, which enables the seamless integration of additional generative foundation models (GenFMs) into our pipeline. This scalability offers the potential for developing a unified evaluation pipeline in this domain. The incorporation of these models facilitates robust evaluation for diverse tasks, achieving a more impressive interpretation of symbolism and enhancing the analytical capabilities of the pipeline.

5. Experiments

This section details the experimental methodology and presents the results, demonstrating that TITScore outperforms state-of-the-art evaluation metrics represented by CLIPScore across a range of evaluation aspects.

Table 2. **TITScore achieves SOTA performance on challenging image-text matching benchmarks that require advanced compositional reasoning.** We thoroughly compare our proposed TITScore with popular recent approaches on the Winoground dataset. We adhere to the original evaluation protocols and report text, image, and group scores.

Methods	**Publications**	**Winoground**		
		Text	Image	Group
Random Chance	–	25.0	25.0	16.7
Human Evaluation	–	89.5	88.5	85.5
CLIP-Score [19]	EMNLP’21	26.3	11.0	7.5
BLIPv2-Score [32]	ICML’23	41.3	20.3	16.8
PickScore [28]	NeurIPS’23	22.5	11.0	6.0
ImageReward [59]	NeurIPS’23	41.3	14.8	12.5
VisProg [16]	CVPR’23	3.5	3.5	3.5
ViperGPT [49]	ICCV’23	7.5	7.3	7.3
VPEval [13]	NeurIPS’23	12.5	9.8	5.8
VQ2 [64]	NeurIPS’23	13.3	26.8	9.8
TIFA [21]	ICCV’23	17.5	11.5	10.3
Davidsonian [12]	ICLR’24	20.8	16.5	15.3
VIEScore [29]	Arxiv’2312	39.5	39.3	34.3
GPT4V-Eval [69]	Arxiv’2311	43.8	48.5	35.3
TITScore	Ours	**52.5**	**55.3**	**44.9**
TITScore-GPT	Ours	**54.5**	**56.0**	**46.8**

5.1. Metric Validation on Visual Reasoning

Dataset and experimental settings. We evaluate the visio-linguistic ability of our evaluation method on image-text matching tasks, which is essential for high-quality text-to-visual generation evaluation. To access their compositional reasoning ability, we select Winoground[50], a challenging benchmark that focuses on compositional information, including attribute, spatial, counting, and differentiation evaluation aspects covering 400 image and caption pairs. We calculate the text, image, and group scores, following the original dataset’s setting.

TITScore outperforms existing evaluation metrics. We compare our model with baseline models originating from five different paradigms, including the widely adopted CLIPScore [19] and BLIPScore [32], as well as the most recent state-of-the-art model assisted by human feedback [59] and GPT-4V [69]. As shown in Table 2, our TlTScore showcases impressive multimodal reasoning and compositional knowledge-gathering abilities with a high image score boost, surpassing these established baselines and achieving state-of-the-art results on Winoground. With the assistance of GPT, our TlTscore-GPT further improves multimodal understanding, as evidenced by a higher text score.

TlTScore excels at evaluating compositional scenarios. We analyze TlTScore on a fine-grained version of Winoground, a dataset divided into subsets with rich compositional knowledge based on four crucial evaluation aspects: attribute, composition, relationship, and semantics. We choose the most general evaluation method, CLIPScore, as the baseline. As shown in Table 3, TlTScore consistently outperforms CLIPScore across all aspects, achieving results that are, on average, six times better. This demonstrates TlTScore's superior ability to handle compositional text prompts effectively and understand complex linguistic prompt structures.

Table 3. **Fine-grained Analysis on Winoground.** We report group scores per skill category. Each sample can naturally incorporate multiple skills. Our metrics show superior reasoning ability compared to the baseline CLIPScore [19].

Method	Subsets				Overall
	Attribute	Composition	Relation	Semantics	
CLIPScore [19]	12.5	3.8	8.4	15.1	8.7
TlTScore (Ours)	**64.0**	**68.9**	**56.9**	**59.6**	**63.2**
TlTScore-GPT (Ours)	**66.3**	**69.6**	**57.0**	**60.8**	**64.1**

Effiency of the mixture-of-experts paradigm for text-to-visual evaluation. We compare the evaluation time of metrics for T2V generation. As shown in Table 4, our evaluation pipeline, which includes prompt curation, MoE evaluation, and knowledge gathering, takes only 0.447s on a single text-image pair using our well-packaged APIs with just one line of code and well-achieved parallelization. We achieve this efficiency by mitigating the long-tail effect through the removal of useless tokens in the input to foundation models, resulting in a smaller embedding space for inference. Parallelization ensures reliance only on the slowest GemFMs within the mixture of experts. It is also worth noting that our evaluation speed is more than 100 times faster than GPT4V-Eval, which achieves similar performance.

5.2. Evaluation with TlTBench

Here we demonstrate how our benchmarks solve two main challenges in generative model evaluation mentioned in Sec.3 with TlTBench-A and TlTBench-B.

TlTBench contains more informative content. Many existing evaluation benchmarks contain texts that are not particularly meaningful [67]. We set aside the visual content and provide only the text (question and choices) to GPT-4, reporting its accuracy (percent of meaningless samples) in Table 5. These popular datasets contain much text that can be easily solved, rendering the evaluation meaningless, while the carefully designed TlTBench-A serves as a more informative benchmark for generative models.

Table 4. **TlTScore maintains superior efficiency with well-Packaged APIs.** This comparison highlights the evaluation time required for a single (image, text) pair, utilizing our metric with inference performed on a single NVIDIA A100 GPU.

Metric	Models	# Parameters	Eval Time (s/pair)
CLIPScore [19]	CLIP-ViT-L-14	416 M	0.218
PickScore [28]	CLIP-ViT-H-14	986 M	0.233
BLIPv2Score [32]	BLIP-2	2.7 B	0.259
ImageReward [59]	BLIP-2	2.7 B	0.336
GPT4V-Eval [69]	MoE (GPT-4V)	1.76 T	20.403
TlTScore (Ours)	MoE (Visual GenFMs)	2313 M	0.447

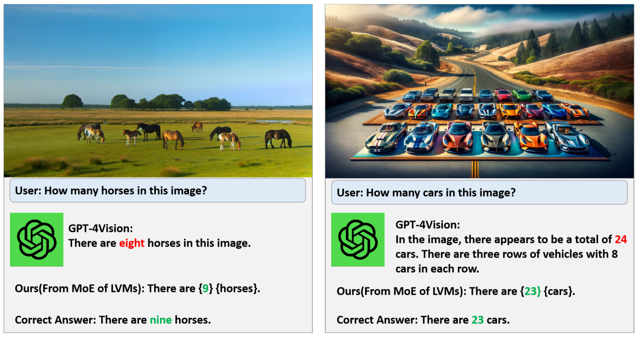
Table 5. **Information entropy analysis among existing benchmarks.** Multimodal benchmarks such as ARO [67] contain uninformative samples that GPT4 can easily solve without visual information, while the balanced importance of textual and visual elements in TlTBench-A yields results comparable to random chance.

Benchmark	GPT-4Vision	Random Chance
ARO [67]	72%	20% (1 of 5)
SEED-Bench [31]	43%	25% (1 of 4)
Mme [15]	56%	50% (1 of 2)
TlTBench-A (Ours)	**27%**	**25%** (1 of 4)

TlTBench introduces complex scenarios for compositional evaluation. Table 6 presents an analysis of TlTScore's performance across prevailing generative models using prompts from TlTBench-A and validates the metric's compositional ability through TlTBench-B. The results demonstrate the challenge posed by our benchmark to existing generative foundation models and showcase TlTScore's superior compositional discriminative power as a more reliable evaluation tool compared to the state-of-the-art method, GPT4V-Eval. Moreover, Fig.4 illustrates an example where TlTScore excels in handling complex visual compositional scenarios, while GPT4V-Eval exhibits inaccuracies in its evaluations.

5.3. Alignment with Human Preference

An effective evaluation of generative foundation models necessitates assessing their alignment with human preferences. Utilizing the newly introduced TlTBench-B dataset, we employed Pearson and Kendall correlation [9] analy-



User: How many horses in this image?

GPT-4Vision:
There are eight horses in this image.

Ours(From MoE of LVMs): There are {9} {horses}.

Correct Answer: There are nine horses.

User: How many cars in this image?

GPT-4Vision:
In the image, there appears to be a total of 24 cars. There are three rows of vehicles with 8 cars in each row.

Ours(From MoE of LVMs): There are {23} {cars}.

Correct Answer: There are 23 cars.

Figure 4. **TITScore serves as a better evaluation metric for complex compositional scenarios compared to GPT-4Vision.** The symbolism-to-neural thinking paradigm significantly enhances visual reasoning capabilities, showcasing its potential for evaluating generative foundation models precisely.

Table 6. **Generative Models Evaluation and Metrics Validation on TITBench.** TITScores generally favor DALLE3 [7] over other models due to its superior generated composition, although it still exhibits limitations when given challenging prompts. On the other hand, TITScore serves as a more reliable metric compared to others, demonstrating the highest accuracy during validation and showcasing strong compositional reasoning ability.

Methods	**TITScore**
SD [45]	0.46
SD-XL [45]	0.52
SD-XL Turbo [45]	0.54
Midjourney [4]	0.63
DALLE3 [7]	0.69

(a) Models Evaluation on TITBench-A.

Metrics	**Accuracy**
CLIPScore [19]	0.38
BLIPv2Score [32]	0.43
PickScore [28]	0.63
GPT4-Eval [69]	0.71
TITScore (Ours)	**0.83**

(b) Metrics Validation on TITBench-B.

ses to quantify the agreement between human ratings and scores generated by TITScore. As demonstrated in Table 7, TITScore emerges as a highly potent metric for assessing generative models, exhibiting stronger congruence with human judgments compared to previously established metrics.

5.4. Evaluation on Diverse Generative Tasks

To demonstrate the versatility of our text-to-visual evaluation methodology, we conduct experiments not only on text-to-image generation but also on text-to-3D and text-to-video tasks. For text-to-3D evaluation, we capture 2D views of 3D assets from various camera angles on challenging benchmark T3Bench [18], while for text-to-video assessment, we select video frames at different time points on FETV [37]. As shown in Table 8, TITScore outperforms T2VScore and GPT4-Eval, which utilizes the specialized GPT4-Vision model. Furthermore, our score surpasses well-established text-to-video benchmarks, such as CLIPScore and PickScore, demonstrating the effectiveness of our approach across diverse generative tasks.

Table 7. **Evaluating TITScore on TITBench-B.** This table presents Pearson and Kendall correlation scores, where higher values indicate better performance. TITScore establishes a new SOTA by achieving greater alignment with human judgments, significantly surpassing existing metrics such as CLIPScore.

Method	**Pearson**	**Kendall**
CLIPScore [19]	16.1	10.8
BLIPv2Score [32]	20.5	17.2
PickScore [28]	14.5	10.0
ImageReward [59]	31.2	29.2
GPT4-Eval [69]	38.2	33.5
TITScore (Ours)	**45.7**	**41.0**

Table 8. **Evaluation of Text-to-Visual Generation Metrics on Diverse Tasks**. The results underscore TITScore's outstanding performance in text-to-3D and text-to-video tasks on the T3Bench and FETV datasets, surpassing both T2VScore and GPT4-Eval. Moreover, TITScore outperforms well-established benchmarks such as CLIPScore and PickScore, showcasing its versatility and effectiveness across various generative modalities.

Method	**Pearson**	**Kendall**
CLIPScore [19]	46.4	32.0
BLIPv2Score [32]	21.2	13.3
PickScore [28]	39.4	29.2
ImageReward [59]	45.4	33.9
GPT4-Eval [69]	52.1	42.2
TITScore (Ours)	**56.3**	**45.8**
TITScore-GPT (Ours)	**56.7**	**46.0**

(a) Text-to-3D Evaluation on T3Bench.

Method	**Pearson**	**Kendall**
CLIPScore [19]	33.9	24.3
BLIPv2Score [32]	26.3	17.5
PickScore [28]	32.1	24.7
ImageReward [59]	37.2	27.9
GPT4-Eval [69]	43.4	33.7
T2VScore [57]	46.4	37.3
TITScore (Ours)	**47.5**	**36.9**
TITScore-GPT (Ours)	**49.8**	**37.7**

(b) Text-to-video Evaluation on FETV.

6. Discussion and Conclusion

In this paper, we propose TITScore, a novel paradigm for evaluating generative foundation models in text-to-visual tasks. Our approach solved the severe long-tail effects in existing evaluation methodologies and innovatively combined neuro-symbolic thinking with mixture-of-experts LVMs. We demonstrate the effectiveness of our method by achieving state-of-the-art performance across various challenging scenarios and diverse tasks. Furthermore, we present TITBench, a comprehensive benchmark designed to be semantically rich and compositionally diverse. Our user-friendly API simplifies the evaluation process and enables possible refinement through our reward. In the future, we aim to further optimize our MoE architecture to find a unified solution for the evaluation of generative models.

References

[1] Gpt-4v(ision) system card. 2023. 3

[2] Zeroscope, 2023. 1

[3] Gen-2: Gen-2: The next step forward for generative ai, 2023. 1

[4] midjourney, 2023. 4, 8

[5] Pika, 2023. 1

[6] Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 19984–19996. IEEE, 2023. 2, 3, 4

[7] James Betker, Gabriel Goh, Li Jing, † TimBrooks, Jianfeng Wang, Linjie Li, † LongOuyang, † JuntangZhuang, † JoyceLee, † YufeiGuo, † WesamManassra, † PrafullaDhariwal, † CaseyChu, † YunxinJiao, and Aditya Ramesh. Improving image generation with better captions. 1, 4, 8

[8] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 22563–22575. IEEE, 2023. 1

[9] Douglas G. Bonett and Thomas A. Wright. Sample size requrements for estimating pearson, kendall and spearman correlations. *Psychometrika*, 65(1):23–28, 2000. 7

[10] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 1

[11] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 22189–22199. IEEE, 2023. 1, 3

[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. *CoRR*, abs/2310.18235, 2023. 6

[13] Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for text-to-image generation and evaluation. In *NeurIPS*, 2023. 6

[14] 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. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 5

[15] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394, 2023. 7

[16] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 14953–14962. IEEE, 2023. 6

[17] Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994. 5

[18] Yuze He, Yushi Bai, Matthieu Lin, Wang Zhao, Yubin Hu, Jenny Sheng, Ran Yi, Juanzi Li, and Yong-Jin Liu. T^{3}bench: Benchmarking current progress in text-to-3d generation, 2023. 8

[19] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7514–7528. Association for Computational Linguistics, 2021. 1, 3, 6, 7, 8

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6626–6637, 2017. 3

[21] 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. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 20349–20360. IEEE, 2023. 2, 3, 5, 6

[22] Xiaoqing Ellen Tan Po-Yao Huang Russell Howes Vasu Sharma Shang-Wen Li Gargi Ghosh Luke Zettlemoyer Hu Xu, Saining Xie and Christoph Feichtenhofer. Demystifying clip data. 2023. 5

[23] 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 preprint arXiv: 2307.06350*, 2023. 1, 2, 3, 4

[24] Pengliang Ji, Angtian Wang, Yi Zhang, Adam Kortylewski, and Alan Yuille. Volumetric neural human for robust pose optimization via analysis-by-synthesis. In *SVRHM 2022 Workshop@ NeurIPS*, 2022. 4

[25] Jiongchao Jin, Huanqiang Xu, Pengliang Ji, and Biao Leng. Imc-net: Learning implicit field with corner attention network for 3d shape reconstruction. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 1591–1595, 2022. 4

[26] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Pnp inversion: Boosting diffusion-based editing with 3 lines of code. *International Conference on Learning Representations (ICLR)*, 2024. 1

[27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and

Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 5

[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. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 3, 6, 7, 8

[29] Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *CoRR*, abs/2312.14867, 2023. 3, 6

[30] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Fei-Fei Li, Jiajun Wu, Stefano Ermon, and Percy Liang. Holistic evaluation of text-to-image models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 2

[31] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 7

[32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 19730–19742. PMLR, 2023. 3, 6, 7, 8

[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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 300–309. IEEE, 2023. 1, 3

[34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 3

[35] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. 5

[36] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond H. Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. *CoRR*, abs/2310.11440, 2023. 3

[37] Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. *arXiv preprint arXiv: 2311.01813*, 2023. 4, 8

[38] Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 2

[39] Chenwei Lyu, Huai Yu, Zhipeng Zhao, Pengliang Ji, Xiangli Yang, and Wen Yang. Self-supervised dense depth estimation with panoramic image and sparse lidar. In *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 6819–6822, 2023. 4

[40] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. 6

[41] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *CoRR*, abs/2304.07193, 2023. 5

[42] Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. Toward verifiable and reproducible human evaluation for text-to-image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 14277–14286. IEEE, 2023. 4

[43] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 1, 3

[44] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, Christoph Feichtenhofer, and Jitendra Malik. On the benefits of 3d pose and tracking for human action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 640–649, 2023. 5

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022. 4, 8

[46] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 3

[47] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques

for training gans. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2226–2234, 2016. 3

[48] George Siemens, Fernando Marmolejo-Ramos, Florence Gabriel, Kelsey Medeiros, Rebecca Marrone, Srecko Joksimovic, and Maarten de Laat. Human and artificial cognition. *Computers and Education: Artificial Intelligence*, 3:100107, 2022. 2

[49] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11854–11864. IEEE, 2023. 6

[50] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5228–5238. IEEE, 2022. 2, 3, 6

[51] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023. 3

[52] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *CoRR*, abs/2308.06571, 2023. 1

[53] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J. Fleet, Radu Soricut, Jason Baldridge, Mohammad Norouzi, Peter Anderson, and William Chan. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 18359–18369. IEEE, 2023. 3

[54] Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11964–11974. IEEE, 2023. 3

[55] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 4

[56] 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. *arXiv preprint arXiv:2305.16213*, 2023. 1, 3

[57] Jay Zhangjie Wu, Guian Fang, Haoning Wu, Xintao Wang, Yixiao Ge, Xiaodong Cun, David Junhao Zhang, Jia-Wei Liu, Yuchao Gu, Rui Zhao, Weisi Lin, Wynne Hsu, Ying Shan, and Mike Zheng Shou. Towards A better metric for text-to-video generation. *CoRR*, abs/2401.07781, 2024. 2, 4, 5, 8

[58] 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. *CoRR*, abs/2306.09341, 2023. 3

[59] 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. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 3, 6, 7, 8

[60] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 5

[61] Jiawei Yao, Chuming Li, Keqiang Sun, Yingjie Cai, Hao Li, Wanli Ouyang, and Hongsheng Li. Ndc-scene: Boost monocular 3d semantic scene completion in normalized devicecoordinates space. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9421–9431. IEEE Computer Society, 2023. 4

[62] Jiawei Yao, Tong Wu, and Xiaofeng Zhang. Improving depth gradientcontinuity in transformers: A comparative study on monocular depth estimation with cnn. *arXiv preprint arXiv:2308.08333*, 2023.

[63] Jiawei Yao, Xiaochao Pan, Tong Wu, and Xiaofeng Zhang. Building lane-level maps from aerial images. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and SignalProcessing (ICASSP)*, pages 3890–3894. IEEE, 2024. 4

[64] Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. What you see is what you read? improving text-image alignment evaluation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 2, 6

[65] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and

Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Trans. Mach. Learn. Res.*, 2022, 2022. 3

[66] Yuanhan Zhang Bo Li-Songyang Zhang Wangbo Zhao Yike Yuan Jiaqi Wang Conghui He Ziwei Liu Kai Chen Dahua Lin Yuan Liu, Haodong Duan. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*, 2023. 4

[67] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023. 4, 7

[68] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *CoRR*, abs/2309.15818, 2023. 1

[69] Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. Gpt-4v(ision) as a generalist evaluator for vision-language tasks. *CoRR*, abs/2311.01361, 2023. 2, 3, 6, 7, 8