Evaluating Text-to-Visual Generation with Image-to-Text Generation

Supplementary Material

Outline

This document supplements the main paper with benchmark and method details. Below is the outline:

- **Section A** details the skill definitions of GenAI-Bench and compares the skill coverage across popular benchmarks.
- **Section B** describes how GenAI-Bench is collected and shows VQAScore's strong agreement with human judgments.
- **Section C** describes how we compute VQAScore with equations and pseudocode.
- **Section D** includes the implementation details of CLIP-FlanT5 and ablation studies of training data, model size, and question-answer templates.
- **Section E** provides details on the baseline methods, including more failure cases of divide-and-conquer approaches.
- **Section F** provides details on the benchmarks and evaluation metrics, and ablates sampling methods for video and 3D.

A Visio-Linguistic Compositional Reasoning Skills

This section describes how we define and label the compositional reasoning skills for text-to-visual generation, and compare the skill coverage across benchmarks.

Skill definitions. Prior literature on text-to-visual generation [8,25,27,68,90] focuses on generating "basic" objects, attributes, relations, and scenes. However, user prompts often require "advanced" compositional reasoning, including comparison, differentiation, counting, and logic [38,49]. For example, user prompts may require counting not just objects, but also attribute-object pairs and even object-relation-object triplets, like "`one person wearing a white shirt and the other five wearing blue shirts`". To this end, after thoroughly reviewing relevant literature [27,57,77,90], we work with professional designers to design a taxonomy of compositional reasoning skills common in real-world prompts, categorizing them into "basic" and "advanced", where the latter builds upon the former. We provide detailed definitions for "basic" skills in Table 6 and "advanced" skills in Table 7.

Comparing skills across benchmarks. We find the skill categorization in benchmarks like PartiPrompt [90] to be ambiguous or even confusing. For example, PartiPrompt introduces two categories "*complex*" and "*fine-grained detail*".

Table 6: Skill definitions and examples for basic compositions.

Skill Type	Definition	Examples
	Basic Compositions	
`Object`	Basic entities within an image, such as person, animal, food, items, vehicles, or text symbols (e.g., "A", "1+1").	*a* **dog**, *a* **cat** *and a* **chicken** *on a* **table**; *a young* **man** *with a green* **bat** *and a blue* **ball**; *a* **'No Parking'** *sign on a busy street.*
`Attribute`	Visual properties of entities, such as color, material, emotion, size, shape, age, gender, state, and so on.	*a* **silver** *spoon lies to the left of a* **golden** *fork on a* **wooden** *table; a* **green** *pumpkin is smiling* **happily**, *a* **red** *pumpkin is sitting* **sadly**.
`Scene`	Backgrounds or settings of an image, such as weather and location.	*A child making a sandcastle on a* **beach in a cloudy day**; *a grand fountain surrounded by historic buildings in a* **town square**.
`Spatial Relation`	Physical arrangements of multiple entities relative to each other, e.g., on the right, on top, facing, towards, inside, outside, near, far, and so on.	*a bustling city street, a neon 'Open 24 Hours' sign glowing* **above** *a small diner; a teacher standing* **in front of** *a world map in a classroom; tea steams* **in** *a cup,* **next to** *a closed diary with a pen resting* **on** *its cover.*
`Action Relation`	Action interactions between entities, e.g., pushing, kissing, hugging, hitting, helping, and so on.	*a dog* **chasing** *a cat; a group of children* **playing** *on the beach; a boat* **glides** *across the ocean, dolphins* **leaping** *beside it and seagulls* **soaring** *overhead.*
`Part Relation`	Part-whole relationships between entities – one entity is a component of another, such as body part, clothing, and accessories.	*a pilot* **with aviator sunglasses**; *a baker* **with a cherry pin on a polka dot apron**.; *a young lady* **wearing a T-shirt** *puts* **her hand** *on a* **puppy's head**.

The former refers to "*...fine-grained, interacting details or relationships between multiple participants*", while the latter refers to "*...attributes or actions of entities or objects in a scene*". Upon closer examination, the categorization of spatial, action, and part relations into these categories appears arbitrary. To address this, we compare the skill coverage across all alignment and generation benchmarks. For benchmarks (PartiPrompt/T2I-CompBench) with defined skill categories, we map their skills to our definitions. For benchmarks (Winoground/EqBen/Pick-a-pic/DrawBench/EditBench/COCO-T2I/HPDv2-Test/EvalCrafter) without a comprehensive skill set, we manually annotate the samples. Finally, we calculate the skill proportions in each benchmark, identifying skills that constitute more than 2% as genuinely present. Table 8 shows that our GenAI-Bench comprehensively covers all essential skills in real-world prompts like those of [77].

Table 7: Skill definitions and examples for advanced compositions.

Skill Type	Definition	Examples
	Advanced Compositions	
`Counting`	Determining the quantity, size, or volume of entities, e.g., objects, attribute-object pairs, and object-relation-object triplets.	**two** *cats playing with a* **single** *ball;* **five** *enthusiastic athletes and* **one** *tired coach;* **one** *pirate ship sailing through space, crewed by* **five** *robots;* **three** *pink peonies and* **four** *white daisies in a garden.*
`Differentiation`	Differentiating objects within a category by their attributes or relations, such as distinguishing between "old" and "young" people by age, or "the cat on top of the table" versus "the cat under the table" by their spatial relations.	**one cat** *is sleeping on the table and* **the other** *is playing under the table; there are two men in the living room,* **the taller one** *to the left of* **the shorter one***; a notebook lies open in the grass, with sketches on* **the left page** *and blank space* **on the right***; there are two shoes on the grass,* **the one without laces** *looks newer than* **the one with laces**.
`Comparison`	Comparing characteristics like number, attributes, area, or volume between entities.	*there are* **more** *people standing than sitting; between the two cups on the desk, the* **taller** *one holds* **more** *coffee than the* **shorter** *one, which is half-empty; a small child on a skateboard has* **messier** *hair than the person next to him; three little boys are sitting on the grass, and the boy in the middle looks the* **strongest**.
`Negation`	Specifying the absence or contradiction of elements, as indicated by "no", "not", or "without", e.g., entities not present or actions not taken.	*a bookshelf with* **no** *books, only picture frames.; a person with short hair is crying while a person with long hair* **is not***; a smiling girl with short hair and* **no** *glasses*; a cute dog **without** a collar.
`Universality`	Specifying when every member of a group shares a specific attribute or is involved in a common relation, indicated by words like "every", "all", "each", "both".	*in a room,* **all** *the chairs are occupied except one; a bustling kitchen where* **every** *chef is preparing a dish; in a square, several children are playing,* **each** *wearing a red T-shirt; a table laden with apples and bananas, where* **all** *the fruits are green; the little girl in the garden has roses in* **both** *hands.*

B GenAI-Bench

This section describes how we collect GenAI-Bench and showcases VQAScore's superior agreement with human ratings.

Details of GenAI-Bench. GenAI-Bench consists of 1,600 diverse prompts that cover advanced skills not addressed in previous benchmarks [27, 68, 90]. To source prompts relevant to real-world applications, we employ two graphic designers who use Midjourney [57] in their profession. First, we introduce them to our skill definitions and examples. Then, we ask them to craft prompts for each skill, collaborating with ChatGPT to brainstorm prompt variants across diverse visual domains. Importantly, these designers ensure that the prompts are

Table 8: Comparing skill coverage across benchmarks. Compared to existing alignment and generation benchmarks, GenAI-Bench comprehensively covers essential skills (especially advanced ones) in real-world prompts [57] like those in Winoground [77]. Note that SeeTrue is an alignment benchmark proposed in [89] that collects 6,930 human labels for DrawBench [68], EditBench [80], and COCO-T2I [45].

Benchmarks	Basic Compositions					Advanced Compositions				
	Attribute	Scene	Action	Spatial	Part	Counting	Negation	Universal	Comparison	Differentiation
Alignment benchmarks										
Winoground [77]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
EqBen [81]	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗
TIFA160 [25]	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗
SeeTrue [45,68,80,89]	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗
Pick-a-pic [33]	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗
Generation benchmarks										
PartiPrompt (P2) [90]	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗
DrawBench [68,89]	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗
EditBench [80,89]	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗
COCO-T2I [45,89]	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗
T2I-CompBench [27]	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗
HPDv2-Test [86]	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗
EvalCrafter [52]	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗
Our benchmark for both alignment and generation										
GenAI-Bench (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

objective. This contrasts with T2I-CompBench [27], whose prompts are almost entirely auto-generated. For example, in T2I-CompBench's "*texture*" category, an overwhelming 40% of the 1000 programmatically-generated prompts use "metallic" as the attribute, which limits their diversity. Other T2I-CompBench's prompts generated by ChatGPT often contain subjective (non-visual) phrases. For instance, in the prompt "`the delicate, fluttering wings of the butterfly signaled the arrival of spring, a natural symbol of rebirth and renewal`", the "rebirth and renewal" can convey different meanings to different people. Similarly, in "`the soft, velvety texture of the rose petals felt luxurious against the fingertips, a romantic symbol of love and affection`", the "love and affection" is also open to diverse interpretations. Thus, we carefully guide the designers to avoid such prompts. Lastly, each prompt in GenAI-Bench is tagged with its associated visio-linguistic skills. We streamline this process by using GPT4 for automatic tagging, providing it the skill definitions and in-context exemplars. Later, we manually verify and correct all tags for accuracy. This results in over 5,000 human-verified tags.

Collecting human ratings. We evaluate six text-to-image models: Stable Diffusion [66] (SD v2.1, SD-XL, SD-XL Turbo), DeepFloyd-IF [13], Midjourney v6 [57], DALL-E 3 [2]; along with four text-to-video models: ModelScope [79], Floor33 [15], Pika v1 [62], Gen2 [18]. In this preliminary study, we use a subset of 527 prompts from GenAI-Bench. This already exceeds the scale of human annotations in previous work [25,89]. We will extend our benchmark to all 1,600 prompts in a subsequent study. Due to the lack of APIs for Floor33 [15], Pika v1 [62], and Gen2 [18], we manually download videos from their websites. We plan to release our codebase for automatically generating visuals with the rest of the

models. Finally, we collect 1-5 Likert scale human ratings using the recommended annotation protocol of [59]:

> How well does the image (or video) match the description?
> 1. Does not match at all.
> 2. Has significant discrepancies.
> 3. Has several minor discrepancies.
> 4. Has a few minor discrepancies.
> 5. Matches exactly.

Our collected human ratings indicate a high level of inter-rater agreement, with Krippendorff's Alpha reaching 0.72 for image ratings and 0.70 for video ratings, suggesting substantial agreement [25]. Further, we show that VQAScore achieves the state-of-the-art correlation to human ratings in Table 9.

Table 9: Evaluating VQAScore on GenAI-Bench. We report Pairwise accuracy, Pearson, and Kendall, with higher scores indicating better performance for all metrics. VQAScore sets a new SOTA on both the image and video alignment benchmarks of GenAI-Bench (with 527 prompts each), significantly surpassing popular metrics like CLIPScore [21] and PickScore [33].

Method	Pairwise Acc [14]	Old Metrics	
		Pearson	Kendall
Baselines			
CLIPScore [21]	52.2	19.9	14.5
BLIPv2Score [43]	55.1	25.0	20.7
Finetuned on human feedback			
ImageReward [87]	58.7	39.2	28.3
PickScore [33]	57.7	36.3	26.2
HPSv2 [86]	49.8	14.5	10.0
VQAScore w/ open-source models			
InstructBLIP	62.4	43.9	36.0
LLaVA-1.5	62.1	**48.3**	35.6
VQAScore w/ our model			
CLIP-FlanT5 (Ours)	**63.3**	46.9	**38.0**

(a) GenAI-Bench-527 (Image)

Method	Pairwise Acc [14]	Old Metrics	
		Pearson	Kendall
Baselines			
CLIPScore [21]	54.5	26.4	19.1
BLIPv2Score [43]	55.6	27.4	21.5
Finetuned on human feedback			
ImageReward [87]	61.0	44.7	32.7
PickScore [33]	56.8	33.5	24.0
HPSv2 [86]	51.6	18.5	13.2
VQAScore w/ open-source models			
InstructBLIP	62.6	46.9	36.2
LLaVA-1.5	64.3	**54.0**	39.7
VQAScore w/ our model			
CLIP-FlanT5 (Ours)	**64.4**	53.3	**39.9**

(b) GenAI-Bench-527 (Video)

GenAI-Bench performance. We analyze the performance of the ten generative models across all skills in Table 10. Both human ratings and VQAScores prefer DALL-E 3 [2] over the other models in nearly all skills except for negation. In addition, prompts requiring "advanced" compositions are rated significantly lower by both humans and VQAScores. Lastly, current video models do not perform as well as image models, suggesting room for improvement.

Table 10: Performance breakdown on GenAI-Bench. We present the averaged human ratings and VQAScores (based on CLIP-FlanT5) for "basic" and "advanced" prompts. Human ratings use a 1-5 Likert scale, and VQAScore ranges from 0 to 1, with higher scores indicating better performance for both. Generally, both human ratings and VQAScores favor DALL-E 3 over other models, with DALL-E 3 preferred across almost all skills except for negation. In addition, we find that video models receive significantly lower scores than image models. Overall, VQAScore closely matches human ratings.

Method	Attribute	Scene	Relation			Overall
			Spatial	Action	Part	
Image models						
SD v2.1	3.1	3.2	2.9	3.2	3.1	3.1
SD-XL	3.7	3.7	3.4	3.7	3.6	3.6
SD-XL Turbo	3.6	3.7	3.3	3.5	3.5	3.5
DeepFloyd-IF	3.6	3.7	3.4	3.7	3.6	3.6
Midjourney v6	3.9	3.9	3.7	4.0	4.0	3.9
DALL-E 3	4.3	4.5	4.3	4.3	4.3	4.3
Video models						
ModelScope	3.0	3.1	2.8	3.1	3.2	2.9
Floor33	3.1	3.2	2.9	3.3	3.2	3.1
Pika v1	3.3	3.5	3.1	3.3	3.3	3.2
Gen2	3.4	3.6	3.3	3.6	3.5	3.5

(a) Human ratings on "basic" prompts

Method	Attribute	Scene	Relation			Overall
			Spatial	Action	Part	
Image models						
SD v2.1	0.80	0.79	0.76	0.77	0.80	0.78
SD-XL	0.84	0.84	0.82	0.83	0.89	0.83
SD-XL Turbo	0.83	0.83	0.80	0.81	0.84	0.82
DeepFloyd-IF	0.83	0.85	0.80	0.82	0.89	0.83
Midjourney v6	0.88	0.87	0.87	0.87	0.91	0.87
DALL-E 3	0.91	0.90	0.92	0.89	0.91	0.90
Video models						
ModelScope	0.67	0.68	0.65	0.64	0.71	0.65
Floor33	0.69	0.70	0.65	0.66	0.69	0.67
Pika v1	0.77	0.79	0.74	0.71	0.76	0.74
Gen2	0.77	0.79	0.73	0.76	0.84	0.76

(b) VQAScores on "basic" prompts

Method	Count	Differ	Compare	Logical		Overall
				Negate	Universal	
Image models						
SD v2.1	2.4	2.5	2.3	2.9	3.0	2.7
SD-XL	2.5	2.6	2.5	2.7	3.5	2.8
SD-XL Turbo	2.5	2.8	2.4	3.0	3.4	2.8
DeepFloyd-IF	2.8	2.9	2.6	2.9	3.6	3.0
Midjourney v6	3.2	3.3	3.2	2.9	3.9	3.2
DALL-E 3	3.3	3.4	3.4	2.8	4.0	3.3
Video models						
ModelScope	2.1	2.3	2.0	2.7	3.0	2.5
Floor33	2.6	2.8	2.4	3.0	3.4	2.8
Pika v1	2.5	2.7	2.4	3.0	3.6	2.9
Gen2	2.5	2.8	2.4	3.1	3.5	2.9

(c) Human ratings on "advanced" prompts

Method	Count	Differ	Compare	Logical		Overall
				Negate	Universal	
Image models						
SD v2.1	0.68	0.70	0.68	0.54	0.64	0.62
SD-XL	0.71	0.73	0.69	0.50	0.66	0.63
SD-XL Turbo	0.72	0.74	0.70	0.52	0.65	0.65
DeepFloyd-IF	0.74	0.74	0.71	0.53	0.68	0.66
Midjourney v6	0.78	0.78	0.79	0.50	0.76	0.69
DALL-E 3	0.82	0.78	0.82	0.48	0.80	0.70
Video models						
ModelScope	0.56	0.61	0.56	0.51	0.55	0.55
Floor33	0.66	0.69	0.61	0.53	0.56	0.58
Pika v1	0.65	0.67	0.63	0.56	0.68	0.62
Gen2	0.71	0.69	0.65	0.53	0.61	0.61

(d) VQAScores on "advanced" prompts

C Implementing VQAScore

In this section, we describe how we compute VQAScore.

Computing VQAScore as an auto-regressive product. Recall that VQAScore calculates the alignment score of an image $\mathbf{i}$ and text $\mathbf{t}$ directly from a VQA model. We first use a simple QA template to convert the text $\mathbf{t}$ to a question and an answer (denoted as $\mathbf{q}(\mathbf{t})$ and $\mathbf{a}(\mathbf{t})$), for example:

$$\begin{aligned}\mathbf{t} &= \text{The moon is over the cow}\\ \mathbf{q}(\mathbf{t}) &= \text{Does this figure show "The moon is over the cow"?}\\ &\quad\ \text{Please answer yes or no.}\\ \mathbf{a}(\mathbf{t}) &= \text{Yes}\end{aligned}$$

We later demonstrate that such a straightforward question-answer pair is sufficient for good performance. In language modeling [1], a piece of text is pre-processed (or tokenized) into a token sequence, e.g., $\mathbf{a}(\mathbf{t}) = \{a_1, \cdots, a_m\}$. Although "`Yes`" usually counts as a single token, we include the EOS (end-of-sentence) token at the end of the text sequence for a simpler implementation. We find that the EOS token only marginally affects the VQAScore results. Next, the generative likelihood of the answer (conditioned on both the question and image) can be naturally factorized as an auto-regressive product [1]:

$$\text{VQAScore}(\mathbf{i}, \mathbf{t}) := P(\mathbf{a}(\mathbf{t})|\mathbf{i}, \mathbf{q}(\mathbf{t})) = \prod_{k=1}^{m} P(a_k|a_{<k}, \mathbf{i}, \mathbf{q}(\mathbf{t})) \tag{3}$$

The answer decoders of VQA models [11,48] return back m softmax distributions corresponding to the m terms in the above expression. Computing VQAScore is more efficient than generating answer token-by-token. Since the entire sequence of tokens $\{a_k\}$ is already available as input for VQAScore, the above m terms can be efficiently computed in *parallel*. In contrast, answer generation as done by [7,25] requires *sequential* token-by-token prediction, as token a_k must be generated before it can serve as input to generate the softmax distribution for the subsequent token a_{k+1}.

Pseudocode of VQAScore. To better explain how VQAScore works, we attach the pseudocode in algorithm 1. We will release a pip-installable API to compute VQAScore using one-line of Python code.

D Training CLIP-FlanT5

In this section, we detail the training procedure of CLIP-FlanT5, and ablate design choices including training data, model size, and prompting strategies.

Training CLIP-FlanT5. For a fair comparison, we adhere to the training recipe of the state-of-the-art LLaVA-1.5 [47]. We adopt the same (frozen) CLIP visual encoder (ViT-L-336) [64] and the 2-layer MLP projector for image

Algorithm 1: PyTorch-style pseudocode for VQAScore.

```
# tokenize(): text tokenizer that converts texts to a list of token indices
# vqa_model(): VQA model returns logits for predicted answer

def vqa_score(image, text):
    # Format the text into the below QA pair
    question = f"Does this figure show '{text}'? Please answer yes or no."
    answer = "Yes"

    # Tokenize the QA pair into tokens
    question_tokens = tokenize(question)
    answer_tokens = tokenize(answer)

    # Extract logits for predicted answer of shape [len(answer_tokens), vocab_size]
    # answer_tokens is a required input for auto-regressive decoding
    logits = vqa_model(image, question_tokens, answer_tokens)

    # labels must skip the first BOS (Begin-Of-Sentence) token
    labels = answer_tokens[1:]
    # logits must skip the last EOS (End-Of-Sentence) token
    logits = logits[:-1]

    # Compute the log likelihood of the answer
    log_likelihood = -torch.nn.CrossEntropyLoss()(logits, labels)
    # (Optional) Cancel the log to obtain P("Yes" | image, question)
    score = log_likelihood.exp()
    return score
```

tokenization. We also follow LLaVA-1.5's two-stage finetuning procedure and datasets. In stage-1 training, we finetune the MLP projector on 558K captioning data (LAION-CC-SBU with BLIP captions [43]). To accommodate FlanT5's encoder-decoder architecture, we adopt the split-text training method proposed in BLIPv2 [43]. This involves splitting a caption into two parts at a random position, with the first part sent to the encoder and the second part to the decoder. In stage-2 training, we finetune both the MLP projector and the language model (FlanT5) on 665K mixture of public VQA datasets (e.g., VQAv2 [19] and GQA [28]). To efficiently train the encoder-decoder architecture, we convert all multi-turn VQA samples into single-turn, resulting in 3.4M image-question-answer pairs. We also retrain LLaVA-1.5 on the same single-turn VQA samples and observe the same VQAScore results. We borrow hyperparameters of LLaVA-1.5 (see Table 11), such as the learning rate schedule, optimizer, number of epochs, and weight decay. We use 8 A100 (80Gbs) GPUs to train all our models. Our largest CLIP-FlanT5-XXL (11B) takes 5 hours for the stage-1 and 80 hours for the stage-2. For stage-2 training, we adhere to the system (prefix) prompt of LLaVA-1.5 during training [3]:

> A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.
> **USER:** <image> \n <question> **ASSISTANT:** <answer>

[3] By default, we also use the system prompt during inference. Interestingly, removing the system prompt ("A chat between a curious user ... answers to the user's questions") during inference does not affect CLIP-FlanT5 but will hurt LLaVA-1.5's performance.

Table 11: Training hyperparameters for CLIP-FlanT5.

Hyperparameter	Stage-1	Stage-2
dataset size	558K	665K
batch size	256	96
lr	1e-2	2e-5
lr schedule	cosine decay	
lr warmup ratio	0.03	
weight decay	0	
epoch	1	
optimizer	AdamW	
DeepSpeed stage	2	3

Ablating language models and training data. We evaluate four language models: the encoder-decoder FlanT5 (11B and 3B) and the decoder-only Llama-2 (13B and 7B). We also ablate finetuning strategies: using both captioning and VQA data (stage-2) against only captioning data (stage-1). We report overall performance across 7 image-text alignment benchmarks in Table 12. We highlight three key observations:

1. **Finetuning on VQA data** is crucial (whereas captioning data only helps a little).
2. **Scaling up language models** consistently boosts performance.
3. **Encoder-decoder FlanT5** significantly outperforms decoder-only Llama-2.

Figure 5 shows more VQAScore results of different models. We hope our ablations can help future work develop stronger models for VQAScore. We will make all model checkpoints and data available for reproducibility.

Table 12: Ablation on language model and training data. We show overall performance on seven benchmarks: group score on Winoground/EqBen, AUROC on DrawBench/EditBench/COCO-T2I, pairwise accuracy on TIFA160, and binary accuracy on Pick-a-pic, with higher scores indicating better performance for all metrics. We highlight that scaling up the size of LLMs and finetuning on VQA data consistently improve the performance. In addition, the encoder-decoder FlanT5 is stronger than the decoder-only Llama-2, likely because FlanT5 benefits from bidirectional image-question encoding [76] and extensive training on challenging QA datasets [10].

LLM-Type	Model-Size	Training-Data	Winoground	EqBen	DrawBench	EditBench	COCO-T2I	TIFA160	Pick-a-Pic
Llama-2	7B	Caption Only	3.8	7.9	42.5	45.0	46.2	46.6	53.0
		Caption+VQA	21.8	20.7	81.7	65.6	80.5	64.9	81.0
	13B	Caption Only	0.8	1.4	56.5	47.0	51.5	49.7	44.0
		Caption+VQA	29.8	35.0	82.2	70.6	79.4	66.4	76.0
FlanT5	3B	Caption Only	7.3	9.3	71.9	58.3	59.9	52.8	67.0
		Caption+VQA	34.8	39.3	82.8	74.5	80.7	68.8	**84.0**
	11B	Caption Only	11.0	15.0	68.1	55.1	66.5	56.4	72.0
		Caption+VQA	**46.0**	**47.9**	**85.3**	**77.0**	**85.0**	**71.2**	**84.0**

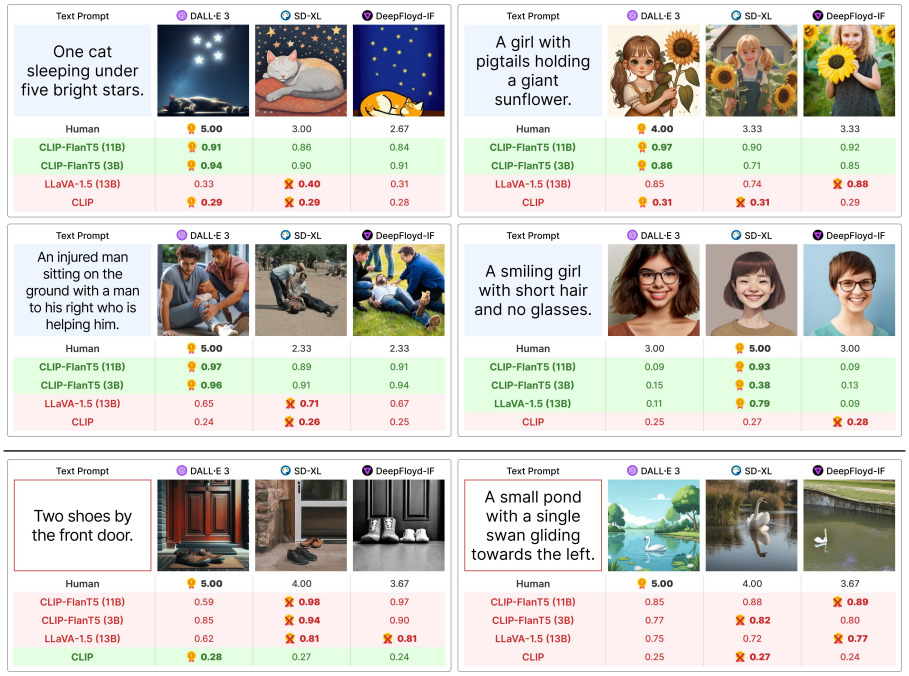

Fig. 5: More qualitative results. We compare VQAScores using CLIP-FlanT5 (11B and 3B) against those using LLaVA-1.5 (13B), highlighting correct predictions in green and incorrect ones in red. We present four successful predictions from CLIP-FlanT5 at the top and two failures at the bottom.

VQAScore is effective with simple question-answers. Table 13 shows that VQAScore consistently performs well across various question templates. Notably, on the challenging Winoground and EqBen benchmarks, simple yet clear questions tend to yield the best results for all VQA models. Interestingly, Table 14 shows that computing the negative answer likelihood (e.g., –P("No")) often yields comparable results. Furthermore, concise answers like P("Yes") perform better than longer responses such as P("Yes it does"). We believe that VQAScore's simplicity makes it a strong alternative to the widely adopted divide-and-conquer approaches [7, 9, 25, 27, 83], which depend on carefully crafted in-context prompts.

E Details of Baseline Methods

In this section, we detail the implementation of the baseline methods and explore the reasons behind their failures.

Metrics based on vision-language models (CLIPScore/BLIPv2Score). To calculate CLIPScore, we use the same CLIP-L-336 model [21] of CLIP-FlanT5 and LLaVA-1.5. For BLIPv2Score, we use the ITM (image-text-matching) head [43] from the largest BLIPv2-ViT-G variant. For an in-depth analysis of

Table 13: Ablating question templates for VQAScore. We ablate 16 question templates across the three VQA models on the challenging Winoground and EqBen benchmarks. We report the group score, where higher scores indicate better performance. We highlight that most questions yield comparable performance, with clearer questions (e.g., those ending with ".. Please answer yes or no.") outperforming more ambiguous ones like "{}?". We also note that CLIP-FlanT5 and InstructBLIP tend to be more stable across different question templates, while LLaVA-1.5 varies more.

Question Template	CLIP-FlanT5		LLaVA-1.5		InstructBLIP	
	Winoground	EqBen	Winoground	EqBen	Winoground	EqBen
Our default question						
Does this figure show "{}"? Please answer yes or no.	46.0	47.9	29.8	35.0	28.5	**38.6**
Paraphrased yes-or-no questions						
Is this figure showing "{}"? Please answer yes or no.	**46.5**	48.6	26.8	35.0	28.2	35.0
Does this photo show "{}"? Please answer yes or no.	44.0	49.3	30.5	31.4	28.7	33.6
Does this picture show "{}"? Please answer yes or no.	44.5	48.6	30.2	38.6	29.5	32.9
Does this image show "{}"? Please answer yes or no.	43.2	47.9	29.2	30.7	28.2	32.9
Does it show "{}"? Please answer yes or no.	43.8	49.3	24.5	28.6	28.2	35.7
Does "{}"? Please answer yes or no.	43.8	49.3	31.8	37.1	28.7	32.1
Is "{}" an accurate description of this figure? Please answer yes or no.	43.5	47.9	27.5	30.0	27.3	**38.6**
Can "{}" be seen in this figure? Please answer yes or no.	40.8	49.3	25.8	27.9	26.8	32.9
"{}"? Please answer yes or no.	44.8	**52.1**	32.5	30.0	**30.2**	35.7
Other questions						
"{}"?	41.0	47.9	24.0	19.3	25.8	27.1
Does this figure show "{}"?	44.8	49.3	25.8	27.1	27.5	37.1
Does this figure show "{}"? Answer the question using a single word or phrase.	44.8	47.1	**35.0**	39.3	26.8	37.1
What is the answer to the following question? "Does this figure show "{}"?"	42.0	45.0	20.8	32.1	27.8	35.7
Based on the image, respond to this question with a short answer: "Does this figure show "{}"?"	42.5	45.7	33.2	**42.9**	27.8	35.0
The question "Does this figure show "{}"?" can be answered using the image. A short answer is	42.8	46.4	18.2	31.4	27.3	36.4

how these discriminatively pre-trained VLMs behave as bags-of-words models, we refer readers to previous studies [30,46,77,91].

Metrics finetuned on human feedback (PickScore/ImageReward/HPSv2). We use the official code and model checkpoints to calculate these metrics. Specifically, PickScore [33] and HPSv2 [86] finetune the CLIP-H model, and ImageReward [87] finetunes the BLIPv2, using costly human feedback from either random web users or expert annotators. Our experiments on the Winoground and EqBen benchmarks (Table 1) show that these metrics perform no better than random chance, likely because the discriminative pre-trained VLMs bottleneck their performance due to bags-of-words behaviors. In addition, their finetuning datasets may lack compositional texts. Finally, we observe that human annotations can be noisy or subjective, especially when these annotators are not well trained (e.g., random web users used by Pick-a-pic [33]). Appendix F discusses these issues.

Visual programming methods (VisProg/ViperGPT/VPEval). We follow the official implementation of these methods. For VisProg [20] and ViperGPT [75], we apply the same VQAScore prompt ("Does this figure show "{text}"? Please answer yes or no."). However, these methods struggle with compositional texts, e.g., Winoground [77]. For instance, given the text "`someone talks on the phone happily while another person sits angrily`", VisProg simply requests a yes-or-no answer from a VQA model, without decomposing. ViperGPT generates the below program that overlooks the action relation:

Table 14: Ablating answer formats for VQAScore. Our analysis of the Winoground and EqBen benchmarks shows that extracting the negative answer likelihood yields comparable results, e.g., P("Yes") performs similarly to the negation of P("No"). Furthermore, concise answers are more effective than longer responses like "Yes it does".

Question Template	Answer	CLIP-FlanT5		LLaVA-1.5		InstructBLIP	
		Winoground	EqBen	Winoground	EqBen	Winoground	EqBen
Does this figure show "{}"? Please answer yes or no.	P(Yes)	46.0	**47.9**	**29.8**	35.0	**28.5**	**38.6**
	–P(No)	**46.3**	**47.9**	27.5	**37.1**	28.0	32.9
Does this figure show "{}"? Please answer correct or wrong.	P(Correct)	18.0	30.7	21.8	32.9	24.8	30.7
	–P(Wrong)	36.0	31.4	18.3	20.0	**28.5**	35.0
Does this figure show "{}"? Please answer true or false.	P(True)	29.8	39.3	31.0	34.3	25.8	32.9
	–P(False)	42.5	37.9	27.0	30.0	**28.5**	33.6
Does this figure show "{}"?	P(Yes it does)	17.0	25.7	15.5	22.9	17.8	25.7
	–P(No it does not)	30.3	23.6	16.8	30.7	23.0	22.9

```
# Text is "someone talks on the phone happily while another person sits angrily"
# Below is the incorrect program generated by ViperGPT that ignores action relation
def execute_command(image) -> int:
        image_patch = ImagePatch(image)
        person_patches = image_patch.find("person")
        if len(person_patches) < 2:
                return 0
        person_patches.sort(key=lambda x: x.horizontal_center)
        person1_patch = person_patches[0]
        person2_patch = person_patches[1]
        person1_happy = person1_patch.verify_property("person", "happy")
        person2_angry = person2_patch.verify_property("person", "angry")
        if person1_happy and person2_angry:
                return 1
        else:
                return 0
```

For VPEval [9], we follow its "open-ended evaluation program" designed for compositional texts. Nonetheless, we observe that it occasionally generates erroneous or nonsensical programs, like asking a VQA model "what is the person doing while talking on the phone?" and expecting an answer of "happily".

Divide-and-conquer using VQA (TIFA/VQ2/Davidsonian). We first note that divide-and-conquer methods are the most popular in recent text-to-visual evaluation [2, 27, 74, 83]. Therefore, we comprehensively analyze all open-source methods, ensuring fair comparison by using the same VQA models as for VQAScore. Specifically, Table 2 already shows that our simple VQAScore surpasses the more complex TIFA [25], VQ2 [89], and Davidsonian [7] across all VQA models (e.g., InstructBLIP-FlanT5-11B, LLaVA-1.5-13B, CLIP-FlanT5-11B). TIFA uses a finetuned Llama-2 to generate multiple-choice QA pairs, returning the answer accuracy of a VQA model as the alignment score. Davidsonian uses a more sophisticated pipeline by prompting ChatGPT to generate yes-or-no QA pairs while avoiding inconsistent questions. For example, given the text "the moon is over the cow", if a VQA model already answers "No" to "Is there a cow?", it then skips the follow-up question "Is the moon over the cow?". VQ2 [89] uses a finetuned FlanT5 to generate free-form QA pairs and computes the average

score of P(answer | image, question). However, these methods often generate nonsensical QA pairs, as shown in Table 16. Lastly, Table 15 confirms that using (a) a single question template *without decomposition* and (b) the *likelihood* of "Yes" is much more effective than decomposition using Davidsonian [7] or checking if the model can directly generate "Yes".

Table 15: Ablation on question decomposition and answer generation versus likelihood. For a fair comparison, we apply all methods to the same CLIP-FlanT5 model. Our end-to-end VQAScore (using the default question template) outperforms question decomposition using Davidsonian [7] or direct answer generation (i.e., checking if the generated answer is "Yes").

VQA Model	Question Template(s)	Scoring	Winoground			EqBen		
			Text	Image	Group	Text	Image	Group
CLIP-FlanT5-11B	Davidsonian [7]	Generation	16.3	11.5	9.8	17.1	11.4	11.4
		VQAScore	41.0	38.3	28.3	45.7	47.9	35.0
	Does this figure show "{}"? Please answer yes or no.	Generation	15.3	15.3	15.3	21.4	21.4	21.4
		VQAScore	**60.0**	**57.5**	**46.0**	**59.3**	**63.6**	**47.9**

GPT4-Vision-based methods (GPT4-Eval/VIEScore). We follow the official prompts from GPT4-Eval [94] and VIEScore [34] to ask GPT4-Vision [58] to directly generate an alignment score (in text format) for an image-text pair (e.g., 0 to 100). For detailed prompts, we direct readers to the respective papers or codebases. Note that we cannot use GPT4-Vision for VQAScore because its API currently does not expose likelihoods of generated answers. Nonetheless, we posit that using VQAScore on stronger VQA models like GPT4-Vision can outperform text-based alignment score generation as done by [34,94].

T2VScore-A(lignment). T2VScore-A [83] is a divide-and-conquer method specifically designed for video-text alignment. When reporting T2VScore-A [83] (based on GPT4-Vision), we calculate the pairwise accuracy [14] using scores released by the authors. However, the authors do not provide the corresponding T2VScore-A outputs for other VQA models (e.g., InstructBLIP).

F Details of Alignment Benchmarks

In this section, we provide details on evaluation metrics and benchmarks in the main paper.

(Meta-)evaluation metrics for human agreement (Pairwise accuracy/Pearson/Kendall). To meta-evaluate metrics (e.g., VQAScore) on benchmarks that provide 1-5 Likert scale ratings (e.g., TIFA160 [25]), we primarily report the pairwise accuracy (with tie calibration) as advocated by Deutsch et al. [14]. Pairwise accuracy effectively addresses ties common in human ratings, unlike the classic Kendall metric which ignores ties. We direct readers to [14] for detailed equations and provide a brief overview below. For a dataset containing

Table 16: Failure cases of divide-and-conquer methods (TIFA, VQ2, and Davidsonian). We show generated question-and-answer pairs of TIFA, VQ2, and Davidsonian on three Winoground texts. These methods often generate irrelevant or erroneous QA pairs (highlighted in red), especially with more compositional texts.

Method	Generated questions	Candidate answers (correct answer choice in bold)
	Text: "`the moon is over the cow`"	
TIFA	Is the moon over the cow?	**yes**, no
	Is the moon over or under the cow?	**over**, under, next to, behind
VQ2	What part of the sun is above the cow?	**the moon**
	What is the name of the moon over the cow?	**the moon**
Davidsonian	Is there a moon?	**yes**, no
	Is there a cow?	**yes**, no
	Is the moon over the cow?	**yes**, no
	Text: "`someone talks on the phone happily while another person sits angrily`"	
TIFA	Who is talking on the phone?	**someone**, no one, everyone, someone else
	Who is sitting angrily?	**person**, animal, robot, alien
VQ2	Who has a good time on the phone?	**someone**
	What part of the life does someone talk to?	**the phone**
Davidsonian	Is the someone happy?	**yes**, no
	Is there another person?	**yes**, no
	Is there a phone?	**yes**, no
Text: "`all paper airplanes fly on a curved path except for one which takes a straight one`"		
TIFA	Are the paper airplanes flying on a curved path?	**yes**, no
	Are the paper airplanes flying on a curved path or a straight path?	**curved path**, straight path, wavy path, zigzag path
VQ2	What type of airplanes fly on a straight path?	**all paper airplanes**
	All paper airplanes fly on what?	**a straight path**
Davidsonian	Do paper airplanes fly on a curved path?	**yes**, no
	Is there one paper airplane?	**yes**, no
	Do paper airplanes fly?	**yes**, no

M image-text pairs, there are two score vectors of size M each: one for human ratings and one for metric scores. [14] evaluates pairwise rankings to determine if human and metric scores agree, i.e., if one image-text pair scores higher, lower, or ties with another image-text pair across both human and metric scores. Additionally, [14] performs tie calibration to optimize for the best tie threshold in metric scores. We emphasize that Pairwise accuracy (with tie calibration) is more reliable and interpretable. Unlike the Pearson coefficient, [14] does *not* assume linear correspondence between human ratings and metric scores. Furthermore, when compared to the Kendall coefficient (which also measures correct pairwise ranking decisions), [14] provides an accuracy value ranging from 0 to 1, making it easier to interpret. For completeness, Table 17 and Table 18 report all three metrics on TIFA160 [25] and Flickr8K [21].

TIFA160 [25]. TIFA160 collects 160 text prompts from four sources: MSCOCO captions [45], DrawBench [68], PartiPrompts [90], and PaintSkill [8]. Each text prompt is paired with five text-to-image models, generating a total of 800 image-text pairs. Furthermore, Davidsonian [7] labels these image-text pairs using 1-5

Likert scale for human evaluation. Table 17 shows that our VQAScore consistently surpasses prior methods across all three meta-evaluation metrics.

Table 17: Evaluating agreement with human judgment on text-to-image benchmark TIFA160 [7,25]. We report Pairwise accuracy, Pearson, and Kendall(-b), with higher scores indicating stronger agreement between human and metric scores. VQAScore based on our CLIP-FlanT5 consistently surpasses all other methods.

Method	Pairwise Acc [14]	Old metrics	
		Pearson	Kendall
Baselines			
CLIPScore [21]	55.8	29.6	19.9
BLIPv2Score [43]	57.5	35.6	23.3
HumanFeedback-based			
ImageReward [87]	67.3	61.5	43.8
PickScore [33]	59.4	39.8	27.4
HPSv2 [86]	55.2	30.1	19.1
GPT4-Vision-based			
GPT4V-Eval [94]	64.0	58.9	46.8
VIEScore [34]	63.9	61.2	47.4
InstructBLIP-based			
TIFA [25]	60.0	56.5	44.0
VQ2 [89]	50.8	12.1	9.4
Davidsonian [7]	61.8	63.4	48.5
VQAScore (Ours)	70.1	58.5	49.7
LLaVA-1.5-based			
TIFA [25]	60.4	49.3	38.1
VQ2 [89]	48.7	4.7	5.1
Davidsonian [7]	54.3	55.6	45.4
VQAScore (Ours)	66.4	58.9	41.9
CLIP-FlanT5-based (Ours)			
TIFA [25]	60.4	46.3	36.0
VQ2 [89]	49.0	3.9	5.6
Davidsonian [7]	61.4	49.0	37.0
VQAScore (Ours)	**71.2**	**66.2**	**51.9**

Flickr8K [21]. We report on the image-to-text evaluation benchmark Flickr8K-CF to show that VQAScore can evaluate image captions in a *reference-free* manner like CLIPScore [21] (without using reference captions of each image). Specifically, Flickr8K-CF contains 145K binary quality judgments collected via CrowdFlower for 48K (image, caption) pairs. Each pair receives at least 3 binary judgments, with human ratings calculated as the mean proportion of "yes" annotations for each pair. Table 18 demonstrates that our VQAScore outperforms all prior art, including reference-based metrics such as BLEU-4, CIDEr, and RefCLIPScore [21].

EvalCrafter [52,83]. We use the text-to-video evaluation benchmark EvalCrafter with 1-5 Likert scales collected by T2VScore [83] for assessing video-text alignment. This benchmark contains 700 prompts paired with five text-to-video models such as Pika [62], Gen2 [18], and Floor33 [15]. By default, we average the

Table 18: Evaluating agreement on image-to-text benchmark Flickr8K [21]. We report Pairwise accuracy, Pearson, and Kendall, with higher scores indicating better performance for all metrics. In this benchmark, each image-caption pair is rated by at least three annotators. VQAScore achieves superior performance compared to existing methods like RefCLIPScore and CIDEr in a reference-free manner (without using the reference captions of the images as provided by the dataset).

Method	Model	Pairwise Acc [14]	Old metrics	
			Pearson	Kendall
Reference-based metrics				
BLEU-4	-	78.1	19.8	16.9
METEOR	-	78.4	36.8	22.3
ROUGE	-	78.0	32.6	19.9
CIDEr	-	79.3	46.1	24.6
SPICE	-	78.2	35.7	24.4
RefCLIPScore [21]	ViT-B/32	78.2	47.9	36.4
Reference-free metrics using CLIPScore				
CLIPScore [21]	ViT-B/32	77.8	44.4	34.4
	ViT-L/14-336px	78.2	46.5	34.7
Reference-free metrics using VQAScore				
VQAScore (Ours)	InstructBLIP	81.5	58.2	36.0
	LLaVA-1.5	82.4	61.9	36.4
	CLIP-FlanT5 (Ours)	**83.1**	**65.4**	**36.7**

VQAScore of all 36 frames from the 3-second videos. Table 19 also shows that sampling as few as four frames can achieve near-optimal performance.

Table 19: Ablating the number of sampled frames for the text-to-video benchmark EvalCrafter [83]. We report the pairwise accuracy [14] of VQAScore for one, four, and all (36) uniformly sampled frames. VQAScore achieves the best performance with 36 frames and near-optimal performance with as few as four frames.

Model	Sampled Frames		
	One	Four	All
InstructBLIP	65.4	65.8	65.7
LLaVA-1.5	63.2	63.7	63.6
CLIP-FlanT5	**65.8**	**66.5**	**66.5**

StanfordT23D [85]. We use the text-to-3D evaluation benchmark StanfordT23D and collect our own 1-5 Likert scales for assessing 3D-text alignment. We follow the same annotation procedure as GenAI-Bench (Section B) and gather 3 human ratings per 3D-text pair, spanning six text-to-3D models (Latent-Nerf [56]/Magic-3D [44]/MVDream [70]/DreamFusion [63]/Instant3D [42]/Shap-

E [29]) across 60 prompts. For human annotators, we provide a 3x3 grid view of each 3D asset, with 9 views sampled uniformly across camera angles. By default, we average the VQAScore of all 120 provided views. However, Table 20 shows that using the same 3x3 grid view (that requires only a single pass) can achieve near-optimal performance.

Table 20: Ablating the number of sampled views and input formats for text-to-3D benchmark StanfordT23D [85]. We report the pairwise accuracy [14] with higher scores indicating better performance. Interestingly, using a single grid layout (2x2 or 3x3) image often performs almost as well as averaging VQAScores across 4 or 9 views.

Model	Sampled Views				
	Uniform (4)	Grid (2x2)	Uniform (9)	Grid (3x3)	All (120)
InstructBLIP	67.4	67.4	68.0	68.1	68.1
LLaVA-1.5	64.5	64.8	64.9	64.9	64.9
CLIP-FlanT5	**68.1**	**67.8**	**68.5**	**68.4**	**68.6**

Pic-a-pick [33]. We find that the text-to-image evaluation benchmark, Pic-a-pick, contains an excessive amount of NSFW (sexual/violent) content and incorrect labels, likely due to an inadequate automatic filtering procedure. Specifically, after manually reviewing the test set of 500 samples, we find that 10% contain inappropriate content (e.g., "*zentai*" and "*Emma Frost as an alluring college professor wearing a low neckline top*") and approximately 50% had incorrect labels. This may also account for the inferior performance of PickScore. As a result, we manually filter the test set to obtain a clean subset of 100 prompts paired with 200 images for evaluating binary accuracy. We also remove all tied labels due to their subjective nature. We will release this subset of Pick-a-pic for reproducibility.

SeeTrue [89] (DrawBench/EditBench/COCO-T2I). We utilize the binary match-or-not labels collected by SeeTrue [89] for the three benchmarks. These benchmarks consist of individual image-text pairs, where some pairs are correctly paired and others are not. We follow their original evaluation protocols to report the AUROC (Area Under the Receiver Operating Characteristic curve), taking into account all possible classification thresholds.

Winoground [77] and EqBen [81]. In our study, we use the entire Winoground dataset consisting of 400 pairs of image-text pairs. For EqBen, because the official test set includes low-quality images (e.g., very dark or blurry pictures), we analyze the higher-quality EqBen-Mini subset of 280 pairs of image-text pairs, as recommended by their official codebase. These two benchmarks evaluate image-text alignment via matching tasks: each sample becomes 2 image-to-text matching tasks with one image and two candidate captions, and 2 text-to-image matching tasks with one caption and two candidate images. The text (and image) score is awarded 1 point only if *both* matching tasks are correct. The final group score is awarded 1 point only if *all* 4 matching tasks are correct.

Importantly, we discover that these benchmarks (especially Winoground) test advanced compositional reasoning skills crucial for understanding real-world prompts, such as counting, comparison, differentiation, and logical reasoning. These advanced compositions operate on basic visual entities, which themselves can be compositions of objects, attributes, and relations.