GPT-4V(ision) is a Human-Aligned Evaluator for Text-to-3D Generation
– Supplementary Materials –

A. Overview

This supplementary material includes additional details about our experiment and methods, which can not be fully covered in the main paper due to limited space. We first provide more details about the method, including our meta-prompts and comparison prompts, in Section B. Experiment details, such as baselines and data, are included in Section C. We also provide detailed ablation studies about the effectiveness of different ways to prompt GPT-4V and ensemble its output (Section D). More experimental results are provided in Section E. Finally, we demonstrate some failure cases of our methods in Section F.

B. Method Details

We will include detailed descriptions about how we implement the two components of our method: prompt generator (in Sec. B.1) and 3D assets evaluator (in Sec B.2). Section B.3 provides additional details about how we use the elo rating.

B.1. Prompt Generator

Our prompt generation pipeline includes a conversation with GPT-4V. We aim to design a conversation pipeline that can provide GPT-4V with the necessary context of the prompt generation task while remaining customizable so that evaluators can adapt this conversation to their need.

Our first prompt describes the task of generating text prompt for text-to-3D generative models. For example, what's the typical length of each prompt and how do we want the distribution of the final collection of prompts look like. We also include a description of how an evaluator might want to control this generator in this prompt. Please see Figure S8-S9 for this opening prompt.

Now that GPT-4V has some basic understanding of the task, we will further provide it with necessary components needed to composed the prompt. First, we provide with GPT-4V a list of ten main categories from which it's able to select subjects of the prompt from (see Figure S1, first box). These categories are chosen to include the most common classes of objects users are interested in generating in text-to-3D tasks. The goal is to make our generator aware of a comprehensive list of potential options, reducing the risk of our prompt generator being unintentionally biased toward certain categories when performing in-context learning with exemplary prompts. These ten categories focuses on what one might wish to generate (*i.e.* the *subjects* mentioned in Section 4.1 of the main paper).

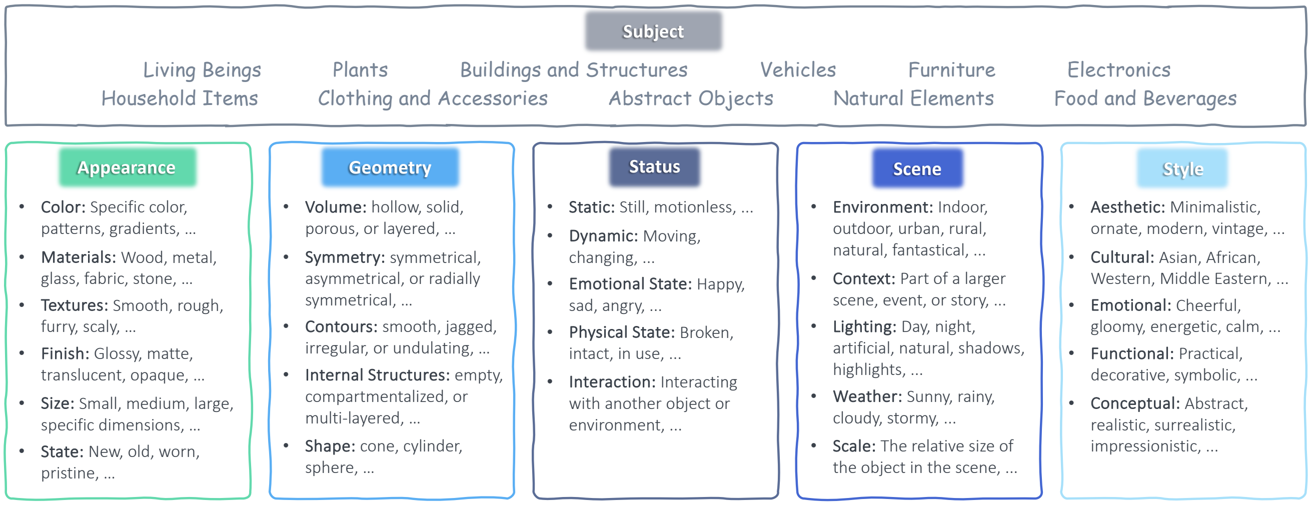
In addition to choosing what to generate, the text-to-3D generative model user also might want to specify a certain state the subject is in (*e.g.* "a sleeping cat" as opposed to "a cat"). Such description is referred to as the **properties** in Section 4.1 of the main paper. To achieve this, we additionally provide GPT-4V a list of descriptions of different properties a user of text-to-3d model might be interested in including (see Figure S1, second row, where each column include list of properties under one of the five aspects we listed). Note that this specific instruction will base on the subjects and a given level of creativity and complexity the evaluator specifies, which will be attached to the beginning of prompt. Please see Figure S8-S9 for the complete detailed prompt.

Finally, we will provide our prompt generator with a list of exemplary prompt where it can model the output from. Users can curate this list of prompt according to their need. In our case, we include the prompts by Dreamfusion [19].

With these exemplary prompts and our provided instruction designed to be customizable according to different scenarios, GPT-4V is now able to provide a list of prompts according to the evaluator's input. Examples of the generated prompts from different levels of creativity and complexity are shown in Figure S2. We can see that users can create prompts with different difficulties and focus using our prompt generator. While we only focus on only two different axes (*i.e.* creativity and complexity) in this work, but our prompt generating pipeline can be easily adapted to other needs as the evaluators can change various part of our conversations accordingly.

B.2. 3D Assets Evaluator

In the instruction for the GPT-4V evaluator, we first explain the task and the definition of each criterion. Then, we asked to take a close look at the input pairs of multi-view images and normal maps before providing its analysis



Appearance

- **Color:** Specific color, patterns, gradients, ...
- **Materials:** Wood, metal, glass, fabric, stone, ...
- **Textures:** Smooth, rough, furry, scaly, ...
- **Finish:** Glossy, matte, translucent, opaque, ...
- **Size:** Small, medium, large, specific dimensions, ...
- **State:** New, old, worn, pristine, ...

Geometry

- **Volume:** hollow, solid, porous, or layered, ...
- **Symmetry:** symmetrical, asymmetrical, or radially symmetrical, ...
- **Contours:** smooth, jagged, irregular, or undulating, ...
- **Internal Structures:** empty, compartmentalized, or multi-layered, ...
- **Shape:** cone, cylinder, sphere, ...

Status

- **Static:** Still, motionless, ...
- **Dynamic:** Moving, changing, ...
- **Emotional State:** Happy, sad, angry, ...
- **Physical State:** Broken, intact, in use, ...
- **Interaction:** Interacting with another object or environment, ...

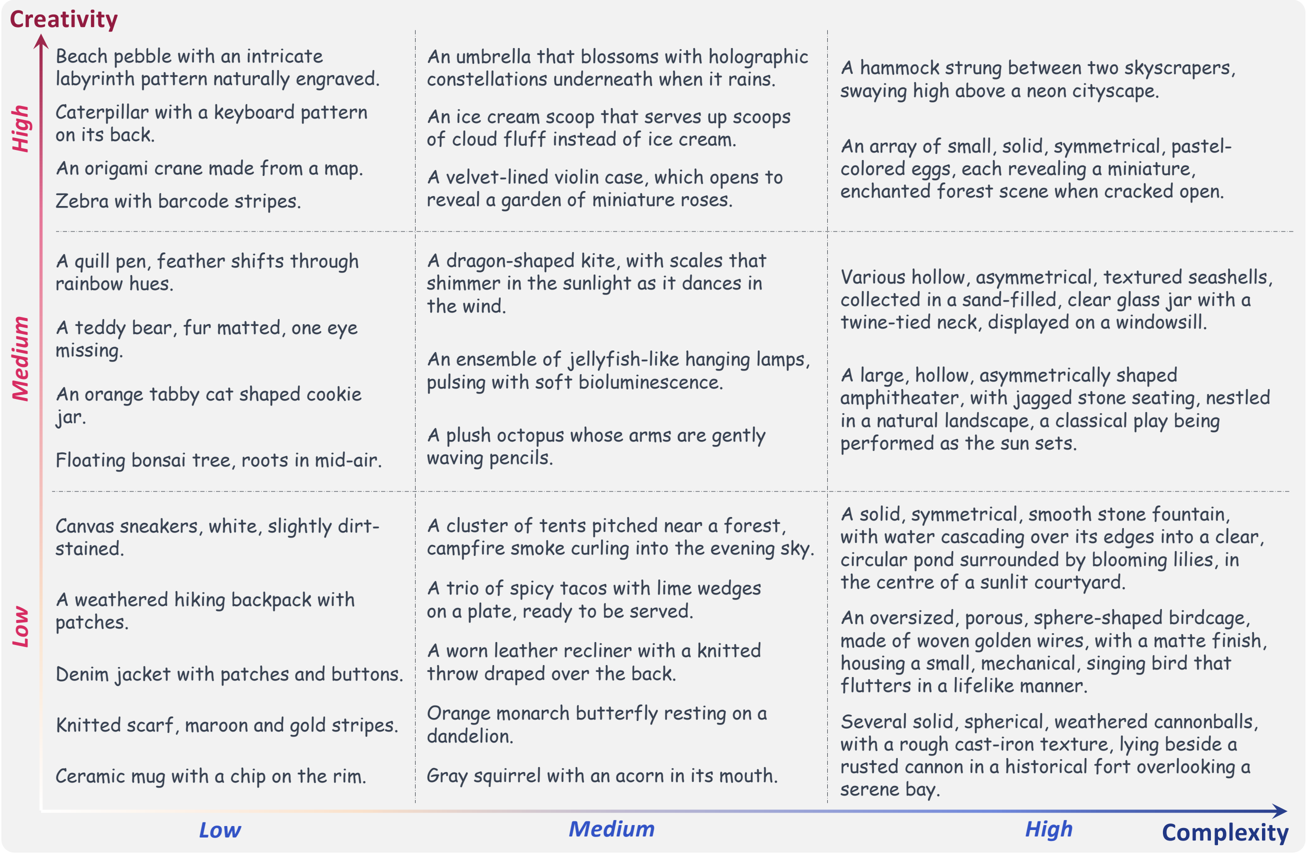
Scene

- **Environment:** Indoor, outdoor, urban, rural, natural, fantastical, ...
- **Context:** Part of a larger scene, event, or story, ...
- **Lighting:** Day, night, artificial, natural, shadows, highlights, ...
- **Weather:** Sunny, rainy, cloudy, stormy, ...
- **Scale:** The relative size of the object in the scene, ...

Style

- **Aesthetic:** Minimalistic, ornate, modern, vintage, ...
- **Cultural:** Asian, African, Western, Middle Eastern, ...
- **Emotional:** Cheerful, gloomy, energetic, calm, ...
- **Functional:** Practical, decorative, symbolic, ...
- **Conceptual:** Abstract, realistic, surrealistic, impressionistic, ...

Figure S1. **Subjects and properties provided for the prompt generator.**


Creativity
Complexity

	Low	Medium	High
High	Beach pebble with an intricate labyrinth pattern naturally engraved. Caterpillar with a keyboard pattern on its back. An origami crane made from a map. Zebra with barcode stripes.	An umbrella that blossoms with holographic constellations underneath when it rains. An ice cream scoop that serves up scoops of cloud fluff instead of ice cream. A velvet-lined violin case, which opens to reveal a garden of miniature roses.	A hammock strung between two skyscrapers, swaying high above a neon cityscape. An array of small, solid, symmetrical, pastel-colored eggs, each revealing a miniature, enchanted forest scene when cracked open.
Medium	A quill pen, feather shifts through rainbow hues. A teddy bear, fur matted, one eye missing. An orange tabby cat shaped cookie jar. Floating bonsai tree, roots in mid-air.	A dragon-shaped kite, with scales that shimmer in the sunlight as it dances in the wind. An ensemble of jellyfish-like hanging lamps, pulsing with soft bioluminescence. A plush octopus whose arms are gently waving pencils.	Various hollow, asymmetrical, textured seashells, collected in a sand-filled, clear glass jar with a twine-tied neck, displayed on a windowsill. A large, hollow, asymmetrically shaped amphitheater, with jagged stone seating, nestled in a natural landscape, a classical play being performed as the sun sets.
Low	Canvas sneakers, white, slightly dirt-stained. A weathered hiking backpack with patches. Denim jacket with patches and buttons. Knitted scarf, maroon and gold stripes. Ceramic mug with a chip on the rim.	A cluster of tents pitched near a forest, campfire smoke curling into the evening sky. A trio of spicy tacos with lime wedges on a plate, ready to be served. A worn leather recliner with a knitted throw draped over the back. Orange monarch butterfly resting on a dandelion. Gray squirrel with an acorn in its mouth.	A solid, symmetrical, smooth stone fountain, with water cascading over its edges into a clear, circular pond surrounded by blooming lilies, in the centre of a sunlit courtyard. An oversized, porous, sphere-shaped birdcage, made of woven golden wires, with a matte finish, housing a small, mechanical, singing bird that flutters in a lifelike manner. Several solid, spherical, weathered cannonballs, with a rough cast-iron texture, lying beside a rusted cannon in a historical fort overlooking a serene bay.

Figure S2. **Examples of the generated prompts with different levels of creativity and complexity.**

and final comparisons. Finally, we specify a specific output format for GPT-4V to follow. In our output format, we also require GPT-4V to provide a short reasoning of why it commit to certain choices, inspired by prompting technique such as chain-of-thought [24]. Specifying an output format can also help us better collect results at a larger scale. An example of such an instruction is shown in Figure S3.

Note that the description of the criteria can be adjusted according to evaluation task. In additional, one can also include in-context learning by breaking down this single text-prompt into multiple pieces, where the following pieces provide exemplary answer of the comparison task in hand. While we believe this can potentially provide a better performance, we only use a single text-prompt and a single image for each pairwise comparison due to limited API access.

In additional to the text instruction, we also need to render images from different view-points to input into GPT-4V. For each text-to-3D model, we provide one, four, or nine renders with camera evenly distributed surroundingly. Note that different text-to-3D generative models might use different coordinate systems, and GPT-4V might prefer certain views of the object than others. We leave it to the evaluator or the text-to-3D model developer to adjust the coordinate system to allow GPT-4V to make the best judgment.

Finally, we create several slightly perturbed version of the input instruction and images to obtain multiple outcomes from GPT-4V. We will ensemble these outcomes to provide a more accurate estimate, as suggested by Yang et al. [25]. Specifically, we perturb the visual information with three versions: including pure RGB renders, pure surface normal maps, and containing both. For text instructions, we experiment with three versions: 1) GPT-4V is required to evaluate a full list of six criteria, 2) include only the criteria of interest, and 3) explicitly ask GPT-4V to evaluate geometry first. We also perturb the number of render views in three versions: 1) only a single view-point; 2) four view-points in a 2x2 layout, and 3) 9 view-points in a 3x3 layout. Finally, we also augment the visual information in three ways: 1) horizontally flipping the left-right order of the objects; 2) vertically flipping the up-down order of the RGB images and normal maps and 3) adding visual watermark into the image [25]. A detailed ablation will be provided in the later section (Sec R1).

B.3. ELO Rating

Comparing two 3D assets can yield different outputs depending on the evaluator or the criteria. This is because human preference is highly subjective and stochastic. We need to model such stochasticity when designing a rating that correctly reflects how much a text-to-3D method is preferred by human evaluators. Such stochasticity is also encountered when two chess players are playing against each other. Since a stronger player can sill lose to a weaker player when the stronger player he or she is having a bad day. In other words, the outcome of comparisons can be noisy. Fortunately, the ELO rating system is designed to create a rating that can reflect the true player's performance while the comparison results are noisy. We will provide some basic intuition of how our ELO system worked for completeness and please refer to prior literatures for more details [4, 15].

The ELO rating system tries to assign a number, σ_i for the ith player, which reflects the strength of this player. Let's examine the case when two players, i and j, play together and assume without loss of generality that $\sigma_i > \sigma_j$. The probability of ith player winning the game should be larger when the difference between σ_i and σ_j is larger. One way to define such probability of winning is the following:

$$\Pr(\text{“}i \text{ beats player } j\text{”}) = \frac{1}{1 + 10^{\frac{\sigma_j - \sigma_i}{c}}}, \tag{1}$$

where c controls the scale of the difference (*i.e.* if the difference is c then the probability of winning will be $1/11$). From this formula, if $\sigma_i = \sigma_j$, then the probability that player i wins over player j is exactly $1/2$. It's also easy to see that

$$\Pr(\text{“}i \text{ beats } j\text{”}) + \Pr(\text{“}j \text{ beats } i\text{”}) \tag{2}$$

$$= \frac{1}{1 + 10^{\frac{\sigma_j - \sigma_i}{c}}} + \frac{1}{1 + 10^{\frac{\sigma_i - \sigma_j}{c}}} = 1. \tag{3}$$

During experiment, we observe that i beats j for A_{ij} number of times and j beats i for A_{ji} number of times. The idea is to find the assignment of σ_i and σ_j such that it best reflects the empirical observation. Specifically, we achieve this by maximum likelihood estimation:

$$\begin{aligned}
&\arg\max_{\sigma_{i,j}} \ \Pr(A_{ij}, A_{ji} | \sigma_i, \sigma_j) \\
=&\arg\max_{\sigma_{i,j}} \ \log\left(\Pr(\text{“}i \text{ beats } j\text{”})^{A_{ij}} \Pr(\text{“}j \text{ beats } i\text{”})^{A_{ji}}\right) \\
=&\arg\max_{\sigma_{i,j}} \ A_{ij} \log\left(\Pr(\text{“i beats j”})\right) + A_{ji} \log\left(\Pr(\text{“j beats i”})\right) \\
=&\arg\min_{\sigma_{i,j}} \ A_{ij} \log\left(1 + 10^{\frac{\sigma_j - \sigma_i}{c}}\right) + A_{ji} \log\left(1 + 10^{\frac{\sigma_i - \sigma_j}{c}}\right).
\end{aligned}$$

Equation (1) in the main paper can be directly derived from the equation above by summing over all pairs of $i \neq j$. In practice, we will initialize $\sigma_i = 1000$ for all i and use Adam optimizer to optimize σ_i for this loss for 10000 iterations with a learning rate of 0.1. Since Elo score is invariant to adding or subtracting a constant, we further calibrate our scores by setting Dreamfusion [19] to have an Elo of 1000 for all criteria.

Note that this method consider the setting where the outcomes include only i wins or j wins. In our case, there are non trivial number of 3D assets pairs from which the human annotator cannot determine which is better. To handle these cases, we follow Nichol et al. [15] to add a win to both methods. This can effectively dilate the number of winning times. Our A_{ij} counts the number of times text-to-3D generative model i wins over model j over any captions. This can be adapted in this theoretical framework by considering the distribution of input text-prompts:

$$\Pr(\text{“}i \text{ beats player } j\text{”}) = \int \Pr(\text{“}i \text{ beats player } j\text{”} | t) P(t) dt,$$

Our task here is the compare two 3D objects, both generated from the same text description. We want to decide which one is better according to the provided criteria.

# Instruction

1. Text prompt and Asset Alignment. Focus on how well they correspond to the given text description. An ideal model should accurately reflect all objects and surroundings mentioned in the text prompt, capturing the corresponding attributes as described. Please first describe each of the two models, and then evaluate how well it covers all the attributes in the original text prompt.

2. 3D Plausibility. Look at both the RGB and normal images and imagine a 3D model from the multi-view images. Determine which model appears more natural, solid, and plausible. Pay attention to any irregularities, such as abnormal body proportions, duplicated parts, or the presence of noisy or meaningless 3D structures. An ideal model should possess accurate proportions, shapes, and structures that closely resemble the real-world object or scene.

3. Geometry-Texture Alignment. This examines how well the texture adheres to the geometry. The texture and shape should align with each other locally. For instance, a flower should resemble a flower in both the RGB and normal map, rather than solely in the RGB. The RGB image and its corresponding normal image should exhibit matching structures.

4. Low-Level Texture Details. Focus on local parts of the RGB images. Assess which model effectively captures fine details without appearing blurry and which one aligns with the desired aesthetic of the 3D model. Note that overly abstract and stylized textures are not desired unless specifically mentioned in the text prompt.

5. Low-Level Geometry Details. Focus on the local parts of the normal maps. The geometry should accurately represent the intended shape. Note that meaningless noise is not considered as high-frequency details. Determine which one has a more well-organized and efficient structure, which one exhibits intricate details, and which one is more visually pleasing and smooth.

6. Considering all the degrees above, which one is better overall?

Take a really close look at each of the multi-view images for these two 3D objects before providing your answer.

When evaluating these aspects, focus on one of them at a time.

Try to make independent decisions between these criteria.

# Output format

To provide an answer, please provide a short analysis for each of the abovementioned evaluation criteria. The analysis should be very concise and accurate.

For each of the criteria, you need to make a decision using these three options:

1. Left (object 1) is better;

2. Right (object 2) is better;

3. Cannot decide.

IMPORTANT: PLEASE USE THE THIRD OPTION SPARSELY.

Then, in the last row, summarize your final decision by "<option for criterion 1> <option for criterion 2> <option for criterion 3> <option for criterion 4> <option for criterion 5> <option for criterion 6>".

# Example

"
Analysis:

1. Text prompt and Asset Alignment: The left one ...; The right one ...; The left/right one is better or cannot decide.

2. 3D Plausibility. The left one ...; The right one ...; The left/right one is better or cannot decide.

3. Geometry-Texture Alignment. The left one ...; The right one ...; The left/right one is better or cannot decide.

4. Low-Level Texture Details. The left one ...; The right one ...; The left/right one is better or cannot decide.

5. Low-Level Geometry Details. The left one ...; The right one ...; The left/right one is better or cannot decide.

6. Overall, ... The left/right one is better or cannot decide.

Final answer:

x x x x x x (e.g., 1 2 2 3 2 1 / 3 3 3 2 1 3 / 3 2 2 1 1 1)
"

Figure S3. **An example of prompts used to guide the GPT-4V evaluator.**

where t denotes an input text prompt. Following most Elo systems, we choose $c = 400$.

C. Experimental Details

In this section we will provide some additional experiment details. We provide a detailed list of text-to-3d generative models for benchmarking in Section C.1. Additional detail about our user studies is provided in Section C.2.

C.1. Text-to-image Models

We involve 13 generative models in the benchmark, including ten optimization-based methods and three recently proposed feed-forward methods. Please refer to the supplementary for the complete list of methods. For optimized-base methods, we include DreamFusion [19], SJC [22], Latent-Nerf [14], Magic3D [9], Fantasia3D [2], Prolific Dreamer [23], DreamGaussian [21], MVDream [20], SyncDreamer [12] and Wonder3D [13]. For feed-forward methods, we include PointE [16], Shap-E [6], and Instant3D [8]. We leverage each method's official implementations when available. Alternatively, we turn to Threestudio's implementation [5]. For methods designed mainly for image-to-3D, we utilize Stable Diffusion XL [18] to generate images conditioned on text as input to these models. Experiments are conducted with default hyper-parameters.

C.2. User study details

In this paper, we mainly rely on labels provided by expert annotators. Our expert annotators are all graduate students with computer graphic background (*e.g.* they all have experience looking at surface normal maps). We recruit twenty such expert annotators from this background. For 13 methods, we create pairwise comparison between each pair of methods (so 78 method pairs). For each method pairs of methods, we sample 3 captions from the caption generators. For each of these 234 comparisons, we assign three different experts to rank all criteria. The experts are asked to pick a shape that performs better according to certain criteria or indicate that these two shapes are indistinguishable. Each user will fill out a query form that includes the same description we provided with GPT-4V. Different from what we provided with GPT-4V, expert annotators are able to see video of 360 rotated render of the object in RGB and in surface normals. The annotator can control the video (*e.g.* replaying to certain point) to exame the 3D shape in more details. The video is rendered in 1024x2048 resolution. In our setting, expert annotators are provided with more information comparing to what we provided to GPT-4V. Our expert annotators achieve reasonable agreement with greater than 0.53 Cohen kappa [3].

One disadvantage of expert annotation is that it's difficult to get by so we are only able to obtain a small scale of such annotation. On the contrary, one can obtain larger-scale annotation from general users. We've also conducted some preliminary exploration conducting user preference studies with general users. Unfortunately, we found that data collected from general users are very noisy. Specifically, we recruited about 53 users and performed the same user studies as done with the expert. In addition to the instruction provided with GPT-4V, we also provide examples comparison and requires the user to go through a short training session. The average user agreement (measured by Cohen's kappa) among these users can barely reach 0.3. This can potentially be caused by the fact that general users do not have experience reasoning about 3D asset information, so computer graphics terms (*e.g.* texture, surface normal, or geometry), become harder to understand. As a result, we leverage expert user data as reference labels in this paper. How to conduct large-scale user studies with more available annotators remains a very interesting future work direction.

D. Ablation Studies

In this section, we will examines key technical design choices of our methods. Specifically, we will focus our ablation on different ways to perturb the input data to create ensemble results (Section 5.2 of main paper). We first carry out an ablation study on different ways to perturb the input (Section D.1). Then we show an ablation on how to ensemble these perturbations together (Section D.2).

Due to limited API access, we are not able to obtain enough GPT-4V queries to compute method-level human alignment scores as in Table 1 or the pair-level alignment score in Table 2 in the main papers. To address this limitation, we use an alternative metric in the ablation. For each variant of our methods, we randomly sample 78 different comparisons of 3D assets. For each comparison, we compute the probability that our GPT-4V 3D asset evaluator would select one shape versus the other. We denote such empirically estimated probability as p_i for our method's variant to choose one shape at the ith comparison. Let q_i be the same probability estimated from expert annotation data. We return the following L1-distance as an estimation of how our method is *mis*-alignment with human judgment:

$$\text{L1-dist}(p, q) = \frac{2}{N} \sum_{i=1}^{N} |p_i - q_i|, \tag{4}$$

where N is the number of comparisons here. Note that the lower this metric is, the better alignment with human judgement.

D.1. Ablation for GPT-4V Input Perturbation

We conduct ablations on different ways to perturn the inputs for GPT-4V. We will investigate four different categories of perturbations, including *visual information*, *text*

Table R1. **Ablation studies on different visual and textual input to GPT-4V.** We mark rank one, rank two, and rank three in each criterion with increasingly lighter shades of blue, and the same baseline (RGB + Normal, 2x2, Joint) is marked in gray.

Methods		Alignment (↓)	Plausibility (↓)	Color-Geo (↓)	Texture (↓)	Geometry (↓)
Visual Information	Pure RGB	0.523	**0.564**	-	**0.354**	-
	Pure Normal	0.674	0.654	-	-	0.579
	RGB + Normal	0.518	0.672	0.628	0.444	0.510
Text Instruction	Joint	0.518	0.672	0.628	0.444	0.510
	Separate	**0.451**	0.597	0.610	0.433	0.528
	Geo-first	0.682	0.646	0.662	0.487	0.505
View number	1	0.592	0.644	0.603	0.423	**0.438**
	2x2	0.518	0.672	0.628	0.444	0.510
	3x3	0.546	0.582	0.654	0.503	0.559
Augmentation	Horizontal Flip	0.615	0.702	0.676	0.522	0.651
	Vertical Flip	0.764	0.695	0.754	0.738	0.708
	Watermark	0.605	**0.559**	**0.597**	0.577	0.492

instruction, *view number*, and *augmentation*. The main results are summarized in Table R1.

Visual Information The visual information can include RGB images, normal maps, or both of them. Purely using RGB as input can benefit the perception of texture details, probably because GPT-4V can spend more computational power on the presentation of the textures. However, not all the criteria can be evaluated merely from the RGB information. As a result, skip evaluation of alignment on those criteria, namely “Color-Geo” and “Geometry”. Only presenting the normal maps to GPT-4V does not bring much improvement for its alignment to human choices even for “Geometry Details”. We can see that RGB renders seem to play the most important role for GPT-4V to make human-aligned decisions. Surface normal render is required to perform many evaluations of criteria about geometries.

Text Instruction We experiment different way to input user criteria into the text instruction. Jointly inputting all criteria into a same text prompt can significantly reduce the number of API calls required. An example of this kind of text instruction can be seen in Figure S3. We also try to evaluate only one criterion at a time. One can see a clear improvement for most of the degrees thanks to the more focused analysis, especially for “Text-Asset Alignment”. This presents a trade-off between compute and accuracy.

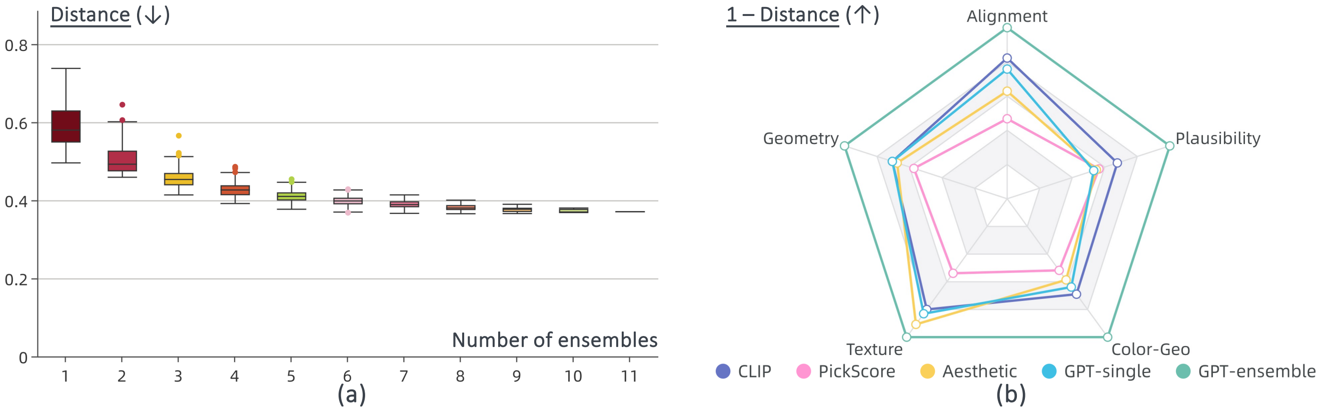
View number The number of views denotes how many multi-view images are shown at the same time. Given the assumption that the visual context size is fixed for GPT-4V [17], this ablation explores the trade-off between the perception of global coherency and local details. Presenting only one view at a time can largely improve GPT-4V’s ability in evaluating low-level criteria like “Texture-Geometry Alignment”, “Texture Details”, and “Geometry Details”. However, the scarcity of views leads to challenges in evaluating the global criteria like “Text-Asset Alignment” and “3D Plausibility”. Increasing view numbers to four or nine will largely alleviate this problem.

Augmentation In the study, we have also experimented with various visual augmentation techniques [25], which refers to changing the visual input slightly without changing the key information contained in the image. We experiment with three augmentation methods: horizontal flipping of object positions (*i.e.* “Horizontal Flip”), the rearrangement of RGB images in conjunction with their respective normal maps (*i.e.* “Verticle Flip”), and the inclusion of watermark annotations to indicate the “left” and “right” objects (*i.e.* “Watermark”). Adding watermarks slightly improves the alignment. This can be a result of watermarks reducing the ambiguity happened when we refer to certain image positions from the text instruction.

Other findings. Another interesting finding is that the results get worse when the normal maps are emphasized in the inputs. For example, in the setting of “Geo-first”, we first provide the surface normal and ask questions involving geometries before providing the full RGB renders and asking the rest of questions. The setting of “Pure Normal” also emphasizes surface normal map by not including the RGB renders. These settings both lead to slightly worse result than the baseline method.

D.2. Ablation for Output Ensemble.

In this section, we want to explore what is the best way to combine different perturbations together. The ex-



(a)

(b)

Figure S4. **Ablation studies of the robust ensemble.** **(a)** highlights a consistent improvement in performance with an increase in ensemble size, together with the decrease in the differences among various ensemble strategies. **(b)** shows how the robust ensemble significantly improves human alignment across all dimensions.

Methods	Alg.	Pla.	T-G	Tex.	Geo.	Avg.
BLIP	0.56	0.44	0.39	0.44	0.46	0.46
Combined	0.81	0.48	**0.58**	0.81	0.74	0.67
Ours-llava	-0.13	0.03	-0.18	-0.08	0.21	-0.03
Ours-mini	0.28	0.33	0.13	0.28	0.05	0.22
Ours-GPT	**0.82**	**0.64**	0.56	**0.82**	**0.80**	**0.71**

Table R2. **Human-alignment of results by more baselines and other open-source LMMs.** We show the Kendell's τ-correlation between rankings provided by a method and those provided by human experts. "Combined" denotes combining all the existing metrics in Table 1 into a single metric by selecting the best one for each criterion. "Ours-llava" and "Ours-mini" denote that we replace the GPT-4V agent in our pipeline with LLaVA-1.5 and MiniGPT, respectively.

ploration space of such combinations is huge as there are n^k numbers of ways to combine n ways to perturb the input together to create k ensembles. Interestingly, we observed that the variance of the L1-distance reduces as we increase the number of ensembles increases. This is depicted from Figure S4-(a) shows the distribution of the L1-distance when ensembling different number of results together. We can see that the alignment performance is not sensitive to particular choice of input perturbations when the number of ensembles is large enough. To achieve the most effective results, we will incorporate a combination of various view numbers, augmentations, visual information, and individual queries for specific criteria, selecting 4-5 strategies overall for the final combination. Figure S4-(b) that ensembling these strategies together results in a metric outperforms all the previous metrics when measured in L1-distance.

E. Additional Results

In this section, we will provide additional results which do not fit in the original paper.

E.1. Experiments with More Baselines.

We involve some more baselines for further comparisons, as shown in Table R2. First, we involve a combination of baseline methods in Table 1, where we manually select the "best" baseline method for each criterion. However, the combined baseline is impractical since when evaluating a new criterion, we need to collect user annotations to compute alignment scores on all existing metrics and report only the best one. This makes the combined metrics either not versatile or not automatic. Moreover, our metric still outperforms this impractical combined metric by a large margin in the 3D plausibility criterion.

We then involve the feature similarity between the text and image features encoded by BLIP [7], which is slightly weaker than the CLIP results in Table 1. We also add two open-sourced LMMs, LLaVA-1.5 [10, 11] and MiniGPT [1] to our pipeline in place of GPT-4V for comparisons. However, they frequently struggle with comprehending multi-view renderings and following textual instructions accurately. Additionally, they exhibit a pronounced bias towards favoring the left option. These together lead to a poor performance.

E.2. Holistic Evaluation

The versatility of our method lands the provision to paint a holistic picture of each text-to-3D model's performance. Here we present preliminary results, which may differ slightly from the final results released on GitHub. Specifically, we compute each model's average Elo scores across each criterion and present the Radar charts of the models achieving the top averaged Elo scores in Figure S5. Table R3 shows the top four methods according to GPT-4V in all the criteria. According to our metric, MVDream [20] won first place on all five criteria. MVDream achieves about

Table R3. **Top-4 methods for different criteria according to our metrics.**

	Alignment	Plausibility	Color-Geo	Texture	Geometry
1st	MVDream	MVDream	MVDream	MVDream	MVDream
2nd	Prolific Dreamer	Latent-NeRF	Latent-NeRF	Prolific Dreamer	Prolific Dreamer
3rd	Latent-NeRF	Instant3D	Prolific Dreamer	Latent-NeRF	Latent-NeRF
4th	Instant3D	Dreamfusion	Instant3D	Instant3D	Instant3D

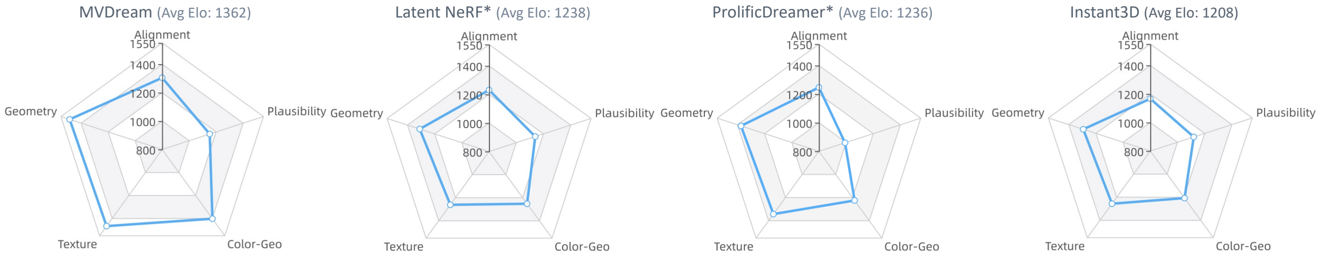


Figure S5. **Holistic evaluation.** Since our evaluation metric is human-aligned in multiple criteria, we can evaluate text-to-3D models more holistically. In this figure, we listed the Radar charts of the top four text-to-3D models according to their averaged Elo scores across all five criteria we evaluated. The Radar charts report the Elo rating for each of the five criteria. These radar charts can provide relative strengths and weaknesses among these models, providing guidance to improve these models. * indicates results from Threestudio implementation.

100 more ELO scores than its runner-ups. The second, the third, and the fourth places are taken by Latent-NeRF [14], ProlificDreamer [23], and Instant3D [8]. These models achieve similar averaged Elo scores, with differences of less than 30 Elos. These three models achieve about 100 Elos more than the next tiers of models, which score about 1100 Elos.

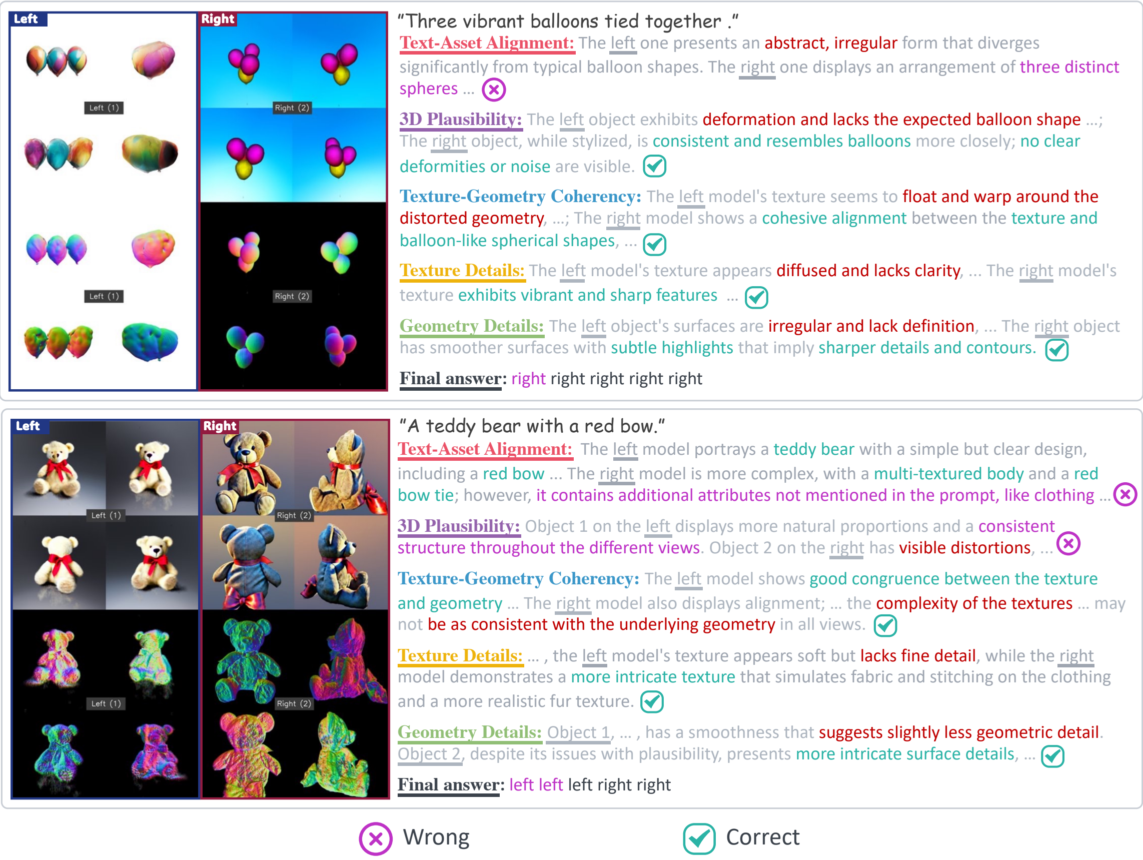
While Latent-NeRF, ProlificDreamer, and Instant3D achieve similar overall scores, our metrics allow further analysis into the relative strengths and weakness of each models. For example, ProlificDreamers show strong performance in three criteria: alignment, geometry details, and texture details. However, its performance in 3D Plausibility is lagging behind when comparing with the other top-performing models. Among these three models, Instant3D [8] is a feed-forward method that takes a much shorter time compared to the top two methods. While our evaluation metrics' reliability can still be limited, we hope that such a holistic picture can provide essential guidance for developing future text-to-3D algorithms.

E.3. GPT-4V Comparison Examples

In Figure S7, we show some more examples of GPT-4V's analysis and answers to this task (using "left / right / equal" instead of "1 / 2 / 3" for better clarity. These examples demonstrate how the evaluator is able to make comparisons of the paired visual input based on close observations and detailed analysis.

F. Failure cases

We present some typical failure cases in Figure S6. In the first instance, GPT-4V fails to detect the mismatch in the number of balloons between the two objects, thereby highlighting its limitations in accurately counting specific quantities. In the second instance, both objects exhibit significant issues with their underlying geometries, The object on the left presents severe 3D plausibility problems, as evidenced by the presence of multiple faces and leg. Both objects are plagued by low-level geometric noise. GPT-4V demonstrates sensitivity to low-level noise, consequently overlooking the overarching issue with the object on the left. Such problems can potentially be rectified if one can provide a larger number of views to GPT-4V.



"Three vibrant balloons tied together ."

Text-Asset Alignment: The left one presents an abstract, irregular form that diverges significantly from typical balloon shapes. The right one displays an arrangement of three distinct spheres ... ⊗

3D Plausibility: The left object exhibits deformation and lacks the expected balloon shape ...; The right object, while stylized, is consistent and resembles balloons more closely; no clear deformities or noise are visible. ✓

Texture-Geometry Coherency: The left model's texture seems to float and warp around the distorted geometry, ...; The right model shows a cohesive alignment between the texture and balloon-like spherical shapes, ... ✓

Texture Details: The left model's texture appears diffused and lacks clarity, ... The right model's texture exhibits vibrant and sharp features ... ✓

Geometry Details: The left object's surfaces are irregular and lack definition, ... The right object has smoother surfaces with subtle highlights that imply sharper details and contours. ✓

Final answer: right right right right right

"A teddy bear with a red bow."

Text-Asset Alignment: The left model portrays a teddy bear with a simple but clear design, including a red bow ... The right model is more complex, with a multi-textured body and a red bow tie; however, it contains additional attributes not mentioned in the prompt, like clothing ... ⊗

3D Plausibility: Object 1 on the left displays more natural proportions and a consistent structure throughout the different views. Object 2 on the right has visible distortions, ... ⊗

Texture-Geometry Coherency: The left model shows good congruence between the texture and geometry ... The right model also displays alignment; ... the complexity of the textures ... may not be as consistent with the underlying geometry in all views. ✓

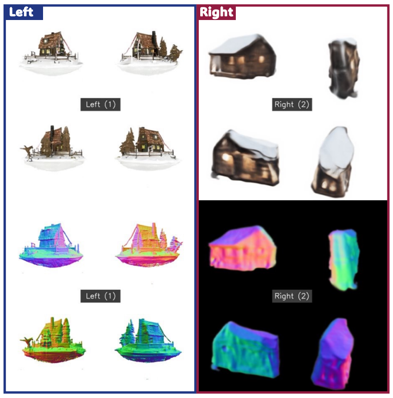
Texture Details: ... , the left model's texture appears soft but lacks fine detail, while the right model demonstrates a more intricate texture that simulates fabric and stitching on the clothing and a more realistic fur texture. ✓

Geometry Details: Object 1, ... , has a smoothness that suggests slightly less geometric detail. Object 2, despite its issues with plausibility, presents more intricate surface details, ... ✓

Final answer: left left left right right

⊗ Wrong ✓ Correct

Figure S6. **Failure cases.** The analysis of GPT-4V can sometimes be *partially* wrong. We show several typical examples for the "Text-Asset Alignment" and "3D Plausibility" criteria.



"A small, rustic cabin sits alone in a peaceful, snow-covered forest."

Text-Asset Alignment: The left one … a cabin … aligning with the "snow-covered forest" description. The right one … a cabin with an ambiguous environment and less visible snow.

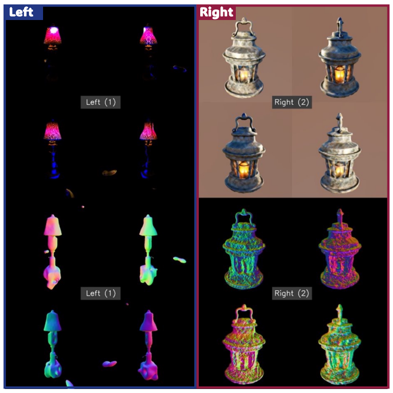
3D Plausibility: The left one .. a clear cabin with discernible roof and walls, appearing more cohesive overall; The right one has parts that are less defined, with some areas looking melted or incomplete.

Texture-Geometry Coherency: The texture of the left cabin aligns well with the underlying geometry, … ; The right cabin … not as precise as the left cabin.

Texture Details: The left cabin has a sharper, more detailed texture with visible wood grains and a snow effect that looks realistic. The right cabin has a softer texture, lacking the same level of crisp detail.

Geometry Details: The left one … more recognizable architectural elements such as windows, a door, and a chimney; The right one has less distinct features, overall appearing smoother and lacking the same level of detail.

Final answer: left left left left left



"A lamp casting a warm glow."

Text-Asset Alignment: The left object is a table lamp with a colorful lampshade and an irregularly shaped stand… The right … that accurately adheres to "casting a warm glow".

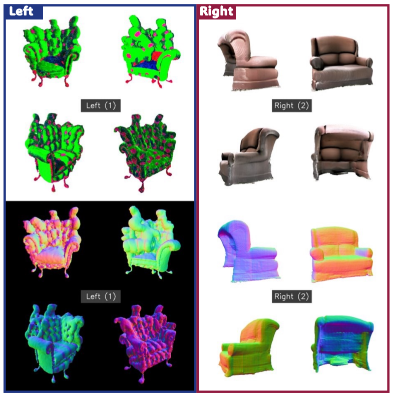
3D Plausibility: Object 1 appears to have a more plausible structure with fewer artifacts than Object 2. Object 2 demonstrates several protrusions and disjointed elements that detract from its plausibility as a lamp.

Texture-Geometry Coherency: The left object shows some inconsistencies, especially at the base. The right one has a coherent alignment between its texture and geometry …

Texture Details: The textures on the left object … are less sharp and do not convey a clear material quality. The texture of the right object appears more detailed and realistic, with materials that reflect light akin to metal and glass, ….

Geometry Details: Object 2 exhibits a higher level of complexity and detail in the geometry of its components, suggesting sharper detail, especially around the lantern's details and the ornamental aspects.

Final answer: right left right right right



"A crying sofa."

Text-Asset Alignment: The left one … do not succinctly convey the concept of a "crying sofa". The right one … but also does not clearly represent the idea of "crying". Neither model has attributes that explicitly suggest crying.

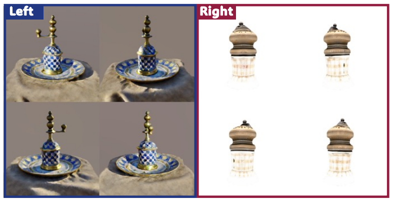
3D Plausibility: The left … a high degree of warping and distortion, with elements that are malformed and far from resembling a stable structure. The right object … maintains a more consistent shape and structure …

Texture-Geometry Coherency: The left one appears … with no discernible patterns that relate to a typical sofa texture. The right one has a texture that aligns with its geometry …

Texture Details: The left one … lacking any realistic sofa material qualities. The right one presents a texture with a degree of realism, depicting shadows and highlights consistent with the geometry of cushions and armrests.

Geometry Details: The left object's geometry is extremely warped, being difficult to discern any clear details. The right object has a more defined structure, …

Final answer: equal right right right right

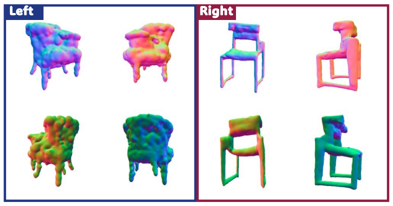


"A compact, cylindrical, vintage pepper mill, with a polished, ornate brass body, slightly worn from use, placed beside a porcelain plate on a checkered tablecloth."

3D Plausibility: Both models seem solid and plausible. The left one has a decorative complexity that is coherent, while the right one, although simpler, also represents a consistent form.

Texture Details: The left model shows a high level of detail … effectively conveys the porcelain and brass textures suggested in the prompt. The right model has a more subdued texture appearance, …

Final answer: equal right



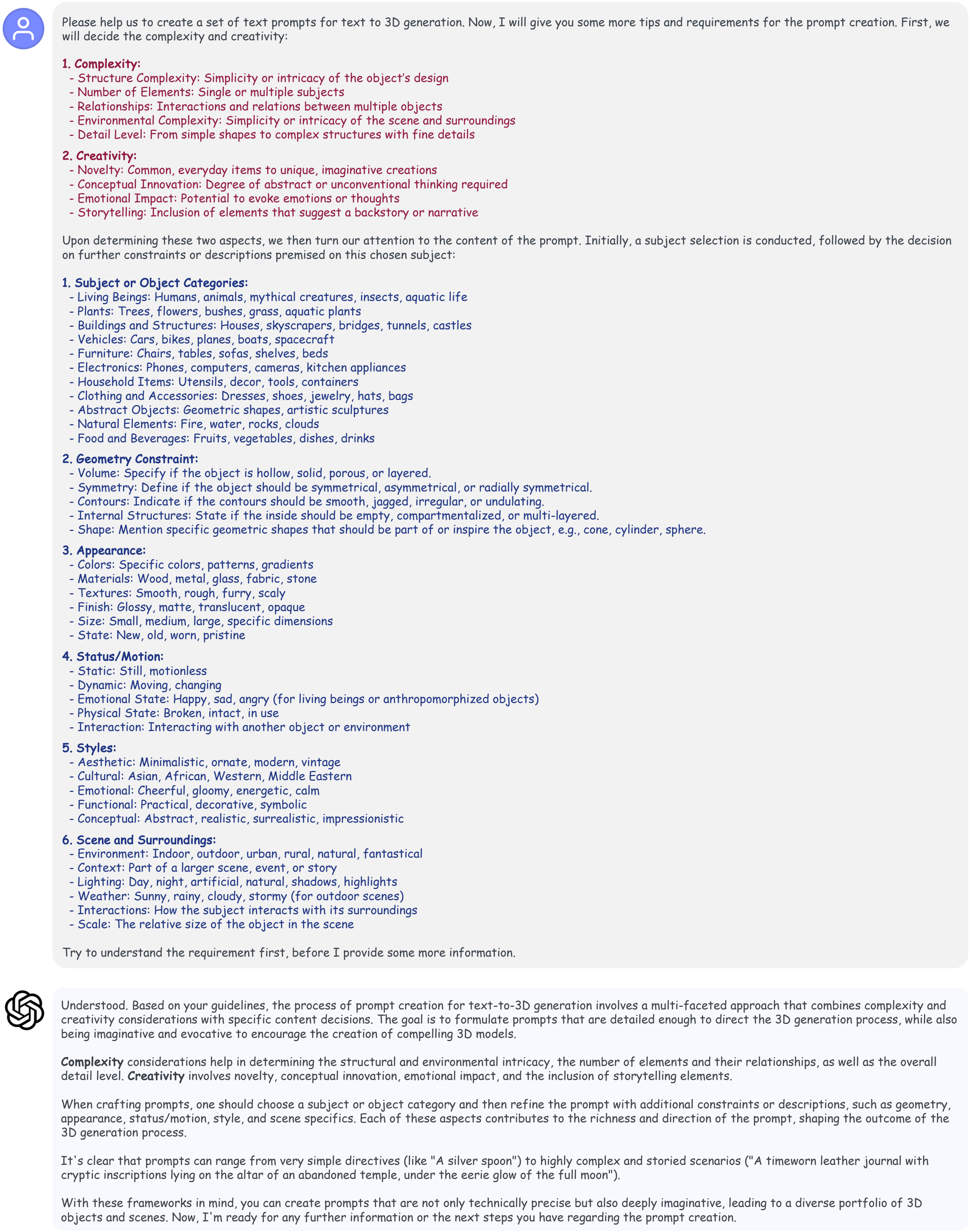
"A chair made from polished oak ."

3D Plausibility: The left model shows implausible distortions and uneven surfaces, while the right model has a consistent structure that is plausible as a chair.

Geometry Details: The normal maps of the left … chaotic arrangement of colors with little to no discernible consistent pattern or structure. The right .. more orderly and structured appearance, with more defined shapes and patterns.

Final answer: right right

Figure S7. **Additional examples of the analysis by GPT-4V.**



Please help us to create a set of text prompts for text to 3D generation. Now, I will give you some more tips and requirements for the prompt creation. First, we will decide the complexity and creativity:

1. Complexity:
- Structure Complexity: Simplicity or intricacy of the object's design
- Number of Elements: Single or multiple subjects
- Relationships: Interactions and relations between multiple objects
- Environmental Complexity: Simplicity or intricacy of the scene and surroundings
- Detail Level: From simple shapes to complex structures with fine details

2. Creativity:
- Novelty: Common, everyday items to unique, imaginative creations
- Conceptual Innovation: Degree of abstract or unconventional thinking required
- Emotional Impact: Potential to evoke emotions or thoughts
- Storytelling: Inclusion of elements that suggest a backstory or narrative

Upon determining these two aspects, we then turn our attention to the content of the prompt. Initially, a subject selection is conducted, followed by the decision on further constraints or descriptions premised on this chosen subject:

1. Subject or Object Categories:
- Living Beings: Humans, animals, mythical creatures, insects, aquatic life
- Plants: Trees, flowers, bushes, grass, aquatic plants
- Buildings and Structures: Houses, skyscrapers, bridges, tunnels, castles
- Vehicles: Cars, bikes, planes, boats, spacecraft
- Furniture: Chairs, tables, sofas, shelves, beds
- Electronics: Phones, computers, cameras, kitchen appliances
- Household Items: Utensils, decor, tools, containers
- Clothing and Accessories: Dresses, shoes, jewelry, hats, bags
- Abstract Objects: Geometric shapes, artistic sculptures
- Natural Elements: Fire, water, rocks, clouds
- Food and Beverages: Fruits, vegetables, dishes, drinks

2. Geometry Constraint:
- Volume: Specify if the object is hollow, solid, porous, or layered.
- Symmetry: Define if the object should be symmetrical, asymmetrical, or radially symmetrical.
- Contours: Indicate if the contours should be smooth, jagged, irregular, or undulating.
- Internal Structures: State if the inside should be empty, compartmentalized, or multi-layered.
- Shape: Mention specific geometric shapes that should be part of or inspire the object, e.g., cone, cylinder, sphere.

3. Appearance:
- Colors: Specific colors, patterns, gradients
- Materials: Wood, metal, glass, fabric, stone
- Textures: Smooth, rough, furry, scaly
- Finish: Glossy, matte, translucent, opaque
- Size: Small, medium, large, specific dimensions
- State: New, old, worn, pristine

4. Status/Motion:
- Static: Still, motionless
- Dynamic: Moving, changing
- Emotional State: Happy, sad, angry (for living beings or anthropomorphized objects)
- Physical State: Broken, intact, in use
- Interaction: Interacting with another object or environment

5. Styles:
- Aesthetic: Minimalistic, ornate, modern, vintage
- Cultural: Asian, African, Western, Middle Eastern
- Emotional: Cheerful, gloomy, energetic, calm
- Functional: Practical, decorative, symbolic
- Conceptual: Abstract, realistic, surrealistic, impressionistic

6. Scene and Surroundings:
- Environment: Indoor, outdoor, urban, rural, natural, fantastical
- Context: Part of a larger scene, event, or story
- Lighting: Day, night, artificial, natural, shadows, highlights
- Weather: Sunny, rainy, cloudy, stormy (for outdoor scenes)
- Interactions: How the subject interacts with its surroundings
- Scale: The relative size of the object in the scene

Try to understand the requirement first, before I provide some more information.

Understood. Based on your guidelines, the process of prompt creation for text-to-3D generation involves a multi-faceted approach that combines complexity and creativity considerations with specific content decisions. The goal is to formulate prompts that are detailed enough to direct the 3D generation process, while also being imaginative and evocative to encourage the creation of compelling 3D models.

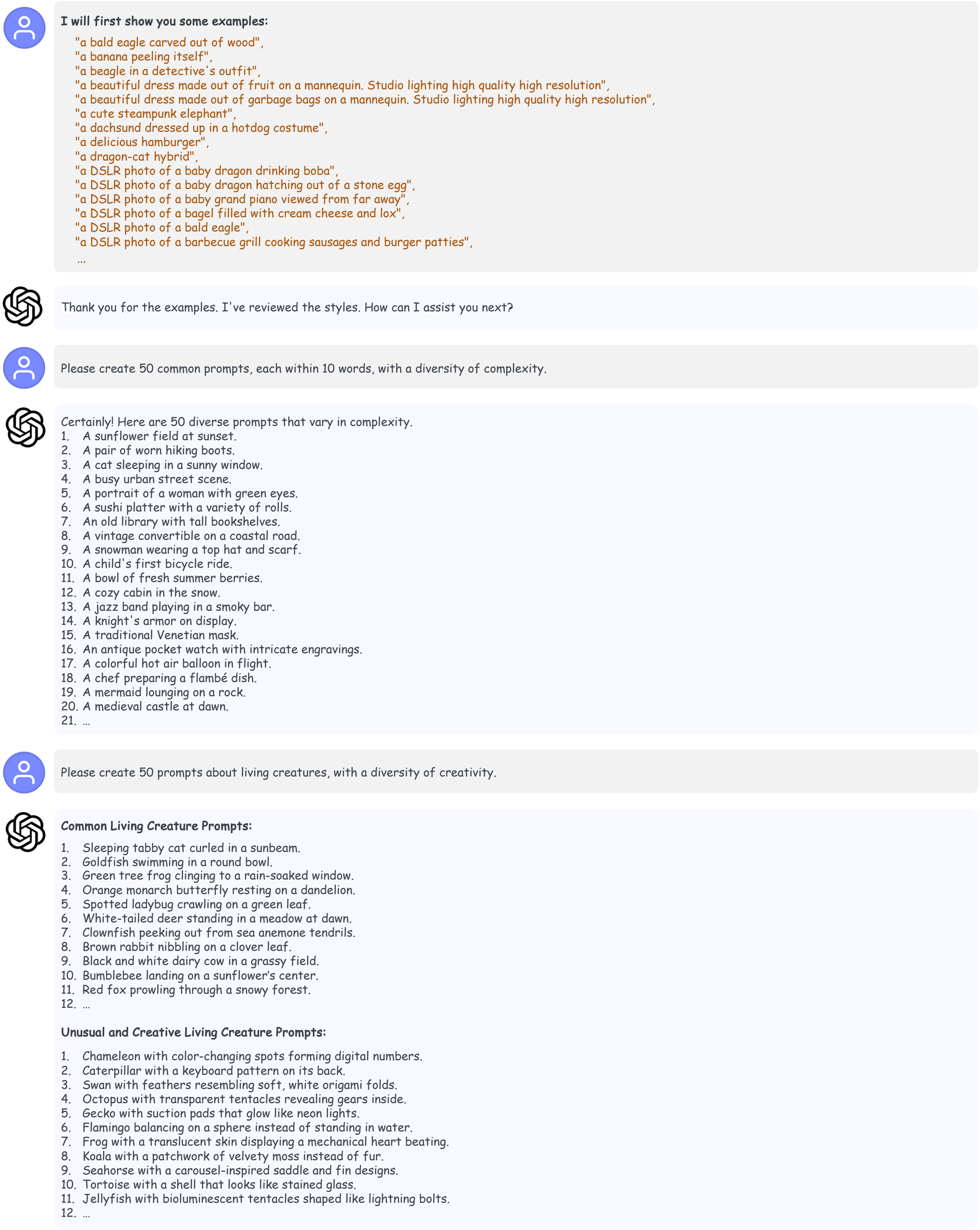
Complexity considerations help in determining the structural and environmental intricacy, the number of elements and their relationships, as well as the overall detail level. **Creativity** involves novelty, conceptual innovation, emotional impact, and the inclusion of storytelling elements.

When crafting prompts, one should choose a subject or object category and then refine the prompt with additional constraints or descriptions, such as geometry, appearance, status/motion, style, and scene specifics. Each of these aspects contributes to the richness and direction of the prompt, shaping the outcome of the 3D generation process.

It's clear that prompts can range from very simple directives (like "A silver spoon") to highly complex and storied scenarios ("A timeworn leather journal with cryptic inscriptions lying on the altar of an abandoned temple, under the eerie glow of the full moon").

With these frameworks in mind, you can create prompts that are not only technically precise but also deeply imaginative, leading to a diverse portfolio of 3D objects and scenes. Now, I'm ready for any further information or the next steps you have regarding the prompt creation.

Figure S8. **Examples of guiding GPT for automatic prompt generation. (Part 1/2)**



I will first show you some examples:

"a bald eagle carved out of wood",
"a banana peeling itself",
"a beagle in a detective's outfit",
"a beautiful dress made out of fruit on a mannequin. Studio lighting high quality high resolution",
"a beautiful dress made out of garbage bags on a mannequin. Studio lighting high quality high resolution",
"a cute steampunk elephant",
"a dachsund dressed up in a hotdog costume",
"a delicious hamburger",
"a dragon-cat hybrid",
"a DSLR photo of a baby dragon drinking boba",
"a DSLR photo of a baby dragon hatching out of a stone egg",
"a DSLR photo of a baby grand piano viewed from far away",
"a DSLR photo of a bagel filled with cream cheese and lox",
"a DSLR photo of a bald eagle",
"a DSLR photo of a barbecue grill cooking sausages and burger patties",
...

Thank you for the examples. I've reviewed the styles. How can I assist you next?

Please create 50 common prompts, each within 10 words, with a diversity of complexity.

Certainly! Here are 50 diverse prompts that vary in complexity.

1. A sunflower field at sunset.
2. A pair of worn hiking boots.
3. A cat sleeping in a sunny window.
4. A busy urban street scene.
5. A portrait of a woman with green eyes.
6. A sushi platter with a variety of rolls.
7. An old library with tall bookshelves.
8. A vintage convertible on a coastal road.
9. A snowman wearing a top hat and scarf.
10. A child's first bicycle ride.
11. A bowl of fresh summer berries.
12. A cozy cabin in the snow.
13. A jazz band playing in a smoky bar.
14. A knight's armor on display.
15. A traditional Venetian mask.
16. An antique pocket watch with intricate engravings.
17. A colorful hot air balloon in flight.
18. A chef preparing a flambé dish.
19. A mermaid lounging on a rock.
20. A medieval castle at dawn.
21. ...

Please create 50 prompts about living creatures, with a diversity of creativity.

Common Living Creature Prompts:

1. Sleeping tabby cat curled in a sunbeam.
2. Goldfish swimming in a round bowl.
3. Green tree frog clinging to a rain-soaked window.
4. Orange monarch butterfly resting on a dandelion.
5. Spotted ladybug crawling on a green leaf.
6. White-tailed deer standing in a meadow at dawn.
7. Clownfish peeking out from sea anemone tendrils.
8. Brown rabbit nibbling on a clover leaf.
9. Black and white dairy cow in a grassy field.
10. Bumblebee landing on a sunflower's center.
11. Red fox prowling through a snowy forest.
12. ...

Unusual and Creative Living Creature Prompts:

1. Chameleon with color-changing spots forming digital numbers.
2. Caterpillar with a keyboard pattern on its back.
3. Swan with feathers resembling soft, white origami folds.
4. Octopus with transparent tentacles revealing gears inside.
5. Gecko with suction pads that glow like neon lights.
6. Flamingo balancing on a sphere instead of standing in water.
7. Frog with a translucent skin displaying a mechanical heart beating.
8. Koala with a patchwork of velvety moss instead of fur.
9. Seahorse with a carousel-inspired saddle and fin designs.
10. Tortoise with a shell that looks like stained glass.
11. Jellyfish with bioluminescent tentacles shaped like lightning bolts.
12. ...

Figure S9. **Examples of guiding GPT for automatic prompt generation. (Part 2/2)**

References

[1] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 7

[2] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023. 5

[3] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960. 5

[4] Arpad E Elo. The proposed uscf rating system, its development, theory, and applications. *Chess Life*, 22(8):242–247, 1967. 3

[5] Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. threestudio: A unified framework for 3d content generation, 2023. 5

[6] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 5

[7] 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. *ArXiv*, abs/2301.12597, 2023. 7

[8] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *https://arxiv.org/abs/2311.06214*, 2023. 5, 8

[9] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 300–309, 2023. 5

[10] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 7

[11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 7

[12] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Learning to generate multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 5

[13] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. Wonder3d: Single image to 3d using cross-domain diffusion, 2023. 5

[14] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. *arXiv preprint arXiv:2211.07600*, 2022. 5, 8

[15] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[16] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 5

[17] OpenAI. Gpt-4v(ision) system card. *OpenAI*, 2023. 6

[18] Dustin Podell, Zion English, Kyle Lacey, A. Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *ArXiv*, abs/2307.01952, 2023. 5

[19] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 1, 3, 5

[20] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv:2308.16512*, 2023. 5, 7

[21] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *ArXiv*, abs/2309.16653, 2023. 5

[22] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. *arXiv preprint arXiv:2212.00774*, 2022. 5

[23] 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. 5, 8

[24] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903, 2022. 2

[25] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9, 2023. 3, 6