MARVEL-40M+: Multi-Level Visual Elaboration for High-Fidelity Text-to-3D Content Creation

Supplementary Material

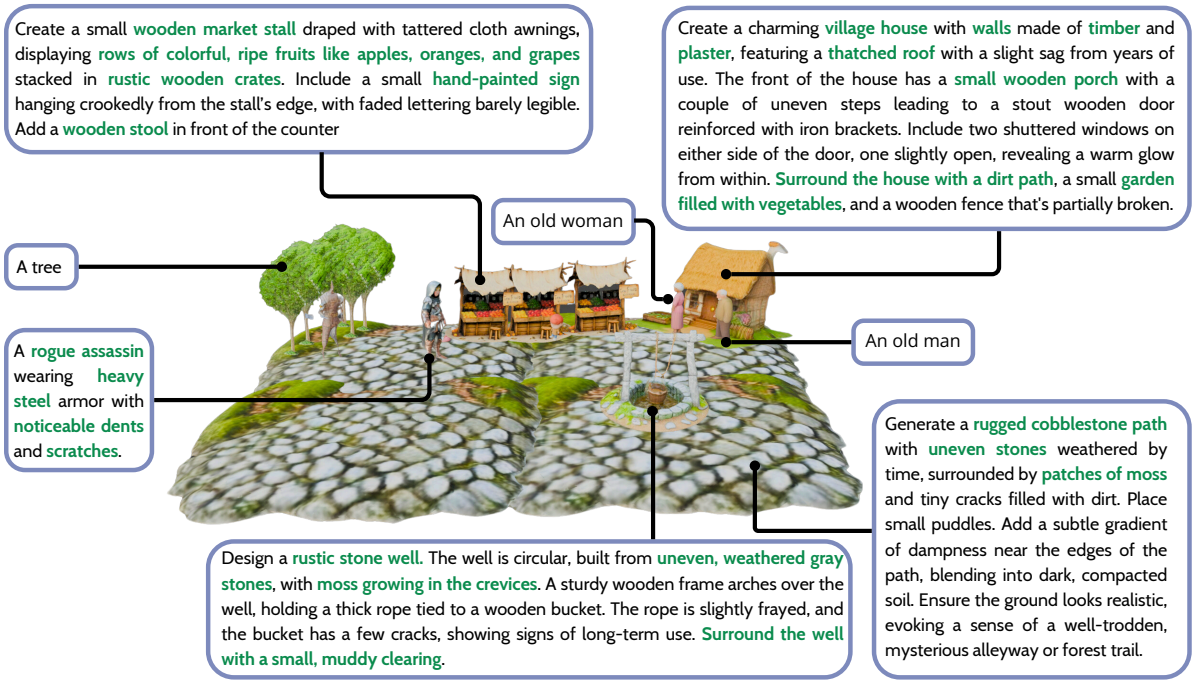


Figure 1. An example use case of MARVEL-FX3D, demonstrating how multiple prompts can be combined to create a detailed and complex 3D scene, with each prompt contributing specific elements such as characters, structures, and environmental details (Zoom in for details).

This supplementary material provides additional details and results to support the main paper. Section 1 outlines the captioning process, including dataset preparation and implementation specifics. Sections 2 and 3 delve deeper into MARVEL annotations and MARVEL-FX3D results, offering more examples, discussions, and insights into their applications and limitations.

1. Additional Details on Captioning Process

1.1. Dataset Preparation

Objaverse: Objaverse[1] [7] contains $798,759$ 3D assets, with metadata (e.g., *name, tags, description*) available for $\sim 93\%$ samples after filtering. From ObjaverseXL [6], we rendered $8,031,637$ assets, of which ~ 3.7M included metadata. After filtering, around 3M samples are retained as valid metadata.

ShapeNet: For the ShapeNet dataset, which contains 52,472 samples, we use the ShapeNet taxonomy as its metadata (e.g., *airplane*, *bowl*, *cap*, *clock*, etc.).

Pix3D: For the Pix3D[2][21] dataset, which contains 374 samples, we use the associated category tag as its metadata (e.g., *bed*, *table*, *desk*, *chair*, etc.).

OmniObject3D: The Omni-Object-3D[3][22] dataset, which contains 5,878 samples, we use the folder names (e.g., *bed*, *table*, *desk*, *chair*, etc.) as our metadata.

Toys4K: For the Toys4K[4][20] dataset, which contains 4,000 samples, we use the folder names (e.g., *car*, *airplane*, *train*, *robot*, etc.) as our metadata.

GSO: The GSO (Google Scanned Objects)[5][8] dataset, which contains 1,030 samples, we use the folder names (e.g., *lamp*, *sofa*, *vase*, *refrigerator*, etc.) as our metadata.

ABO: The ABO (Amazon Berkeley Objects)[6][5] dataset,

[1] https://objaverse.allenai.org/objaverse-1.0

[2] http://pix3d.csail.mit.edu/

[3] https://omniobject3d.github.io/

[4] https://github.com/rehg-lab/lowshot-shapebias/tree/main/toys4k

[5] https://goo.gle/scanned-objects

[6] https://amazon-berkeley-objects.s3.amazonaws.com/index.html

which contains 7,953 samples, provides metadata through listings information. Since these listings are multilingual, we first use the `nllb-200`[7] model to translate the listings to English. The translated English listings are then used as our metadata.

1.2. Implementation Details

For human metadata filtering, we use the Mistral-Nemo-Instruct-2407 model with a temperature of 0.3 and a top-p value of 0.95. For dense description generation, we employ InternVL2-40B, configured with a temperature of 0.70, a top-p value of 0.95, and a repetition penalty of 1.10, with multinomial sampling enabled. For multi-level visual elaboration, we utilize Qwen2.5-72B with 8-bit quantization, a temperature of 0.70, a top-p value of 0.80, and a repetition penalty of 1.05. Finally, the Qwen2.5-14B model, used for the ethical filtering stage, is configured with a temperature of 0 and a top-p value of 0.90.

For human evaluations in our paper, we developed a Gradio app to compare our captions with those from baseline datasets, including Cap3D, 3DTopia, and Kabra, as well as to evaluate FX3D results against text-to-3D baselines. The evaluations were conducted by a panel of 5 human experts.

1.3. Compute and GPU Hours

MARVEL's annotation pipeline utilizes one NVIDIA H100-80GB GPU, one RTX-4090 GPU, and one RTX-A6000 GPU, achieving a throughput of approximately 24,000 samples per day. Annotating the entire Objaverse dataset (800,000 samples) would thus require about 33 days, incurring an estimated total computational cost of approximately $2,700–$3,000, based on publicly available GPU `pricing`[8].

For comparison, sequential human annotation has a considerably lower throughput (1,400 samples/day) and higher cost ($87.18 per 1,000 annotations), resulting in approximately 572 days (about 1.57 years) and a total cost of roughly $69,744 for annotating the complete Objaverse dataset. In contrast, the automated Cap3D pipeline—leveraging BLIP2, CLIP, and GPT-4 models on cloud-hosted NVIDIA A40 GPUs—achieves significantly higher throughput (65,000 samples/day) at a lower cost ($8.35 per 1k annotations), requiring only about 13 days and totaling approximately $6,680 for the entire dataset [16].

Our pipeline annotates the Objaverse dataset at approximately half the total cost of Cap3D, although with a lower throughput (33 days vs. Cap3D's 13 days). Both automated methods substantially outperform sequential human annotation in terms of speed and cost. Importantly, our pipeline delivers annotations of significantly higher quality compared to Cap3D, making it particularly advantageous when balancing annotation quality and cost efficiency. All comparisons assume sequential (non-parallelized) processing; parallelization would further reduce annotation time for all methods.

Method	Throughput (samples/day)	Total Days (800k samples)	Cost per 1k annotations	Total Cost (800k samples)
Human	1,400	572	$87.18	$69,744
Cap3D	65,000	13	$8.35	$6,680
MARVEL	24,000	33	$3.38–$3.75	$2,700–$3,000

Table 1. Comparison of annotation pipelines based on throughput, annotation time, and cost for annotating the Objaverse dataset (800k samples). All estimates assume sequential annotation without parallelization.

2. Additional details on MARVEL annotations

2.1. More Results on Effects of Human Metadata

Figure 2 showcases examples where human-provided metadata from source datasets reduce VLM hallucination and enhances annotations with domain-specific information. To generate captions using InternVL2 [3, 4] and GPT-4 [18], we input the same multi-view images used for MARVEL annotations, instructing them to produce concise descriptions that include names, shapes, textures, colors, and contextual environments.

Examples 1, 2, and 3 demonstrate how the inclusion of simple metadata (*e.g. "La Cava Window"*, *"Mount St. Helens"*) significantly reduces VLM hallucination, resulting in more accurate captions. Example 4 illustrates how metadata can support the generation of highly domain-specific information (*e.g. "alpha-helices and beta sheets"*, *"N-terminus, middle, and C-terminus"*).

2.2. More 3D Captioning Results

We provide more qualitative comparisons of annotations, highlighting differences between the baseline [12, 14, 16] and our proposed MARVEL-40M+ dataset. For consistency, we used only Level 4 annotations, as their length closely matches that of the baselines. To improve clarity, we further categorized examples into distinct domains.

- **Figure** 4 showcases 3D models of automotive designs (e.g., *cars, planes*) and CAD models.
- **Figure** 5 features iconic characters from *anime, movies, and video games*.
- **Figure** 6 illustrates biological elements such as *animals, plants, and molecules*.
- **Figure** 7 includes diverse items ranging from *everyday objects, essentials, food to luxury items*.
- **Figures** 8 and 9 depict historical artifacts (e.g., *statues, memorials*) and various scenes (*e.g. digital elevation maps, realistic and animated scenes*) respectively.

[7] https://huggingface.co/facebook/nllb-200-distilled-600M

[8] https://tinyurl.com/gpu-usage-pricing (Original)

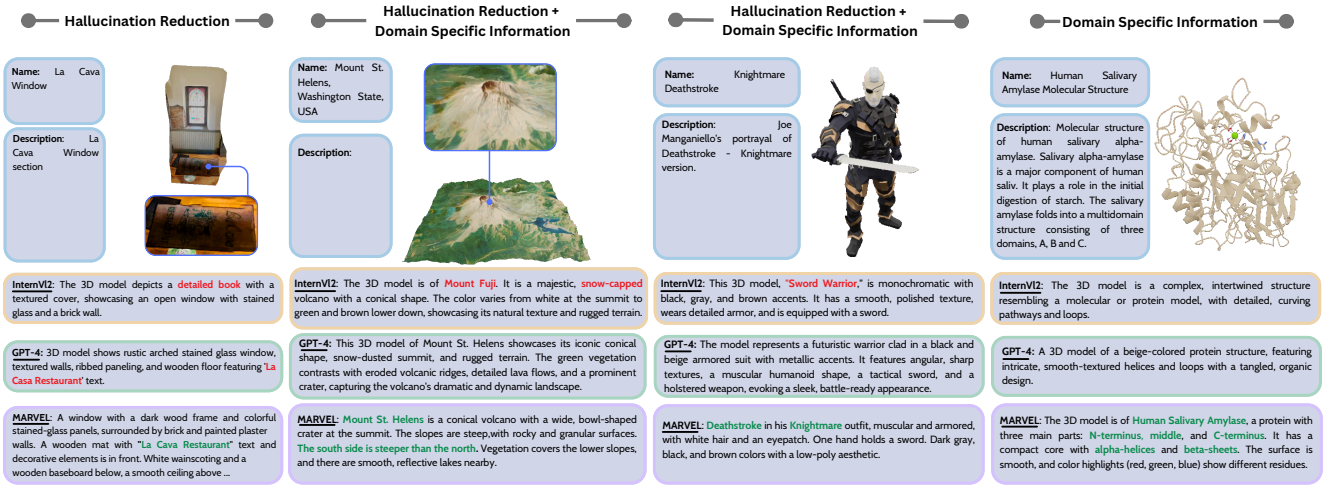


Figure 2. Effect of including human metadata, highlighting improvements in descriptive accuracy and contextual relevance compared to outputs generated without metadata, even when using state-of-the-art models like GPT-4 [18] and InternVL2 [3]. Metadata inclusion helps reduce hallucinations and enhances domain-specific understanding.

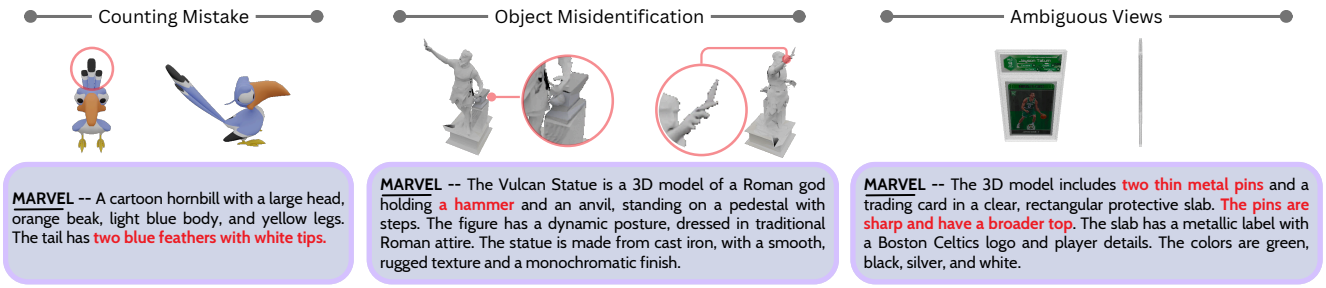


Figure 3. Failure cases of the MARVEL annotation pipeline. From left to right, the examples illustrate errors such as counting mistakes, object misidentification, and challenges with ambiguous views.

As illustrated in the figures, MARVEL annotations offer more precise and domain-specific descriptions, leveraging accurate nomenclature and contextual terminology, surpassing the quality of the baseline datasets.

2.3. More Multi-Level Examples

We present additional qualitative results showcasing our multi-level annotations across all seven datasets [2, 5, 7, 8, 20–22], with two examples per dataset - **Objaverse** (Figure 10), **OmniObject3D** (Figure 11), **ShapeNet** (Figure 12), **Toys4k** (Figure 13), **ABO** (Figure 14) and **GSO** (Figure 15).

2.4. More on Simple and Textureless Models

Our annotation pipeline is robust and adaptable, effectively handling both simple and texture-less models by dynamically adjusting its descriptive verbosity. As illustrated in Figure 17, the pipeline generates concise yet accurate annotations for texture-less models—such as the smooth, monochromatic humanoid figure and the symmetrical lemon head with sunglasses—by emphasizing their geometric precision and structural symmetry. Additionally, for simpler models like the low-poly tree with geometric leaves and the realistically textured orange, the pipeline adeptly captures essential details, highlighting subtle irregularities and primary shape characteristics. This flexibility ensures consistent annotation quality across diverse modeling scenarios.

2.5. Need for Multi-Level-Structure

Hierarchical structures are essential in AI. Examples include multi-resolution models in computer vision, such as feature pyramids, and hierarchical embeddings in NLP, such as document summarization. In 3D modeling, ShapeNet uses a hierarchical taxonomy [2] to improve adaptability across tasks. MARVEL similarly adopts a hierarchical design, ensuring task-specific granularity. By using a predefined hierarchy, MARVEL eliminates the need for repeated prompting. This reduces latency, computational costs, and inconsistencies associated with dynamically ad-

justing verbosity. Such dynamic adjustments would require multiple inference steps and additional processing, making them impractical for large-scale pipelines—even for future LLMs/VLMs. Additionally, dynamic generation introduces risks like semantic drift, information loss, and verbosity imbalance, decreasing annotation reliability. Our ablation study (Section 4.3B and Table 5) confirms that MARVEL's structured verbosity effectively maintains essential details. It optimizes verbosity levels according to task requirements, as validated by cosine similarity and compression ratio.

2.6. Failure Cases

Figure 3 presents examples of the failure cases discussed in Section 5 of the main paper, illustrating the challenges associated with using pretrained VLMs to generate dense descriptions of 3D models.

2.7. More on MTLD Scores

The **MTLD (Measure of Textual Lexical Diversity)** algorithm quantifies vocabulary diversity by segmenting a text whenever the *Type-Token Ratio (TTR)*—the ratio of unique words to total words—drops below a fixed threshold (commonly 0.72). It processes the text both forwards and in reverse to reduce positional bias, and calculates the final MTLD score as the total number of words divided by the number of segments (called *factors*). A low MTLD score indicates a repetitive vocabulary and low lexical diversity, while a high score reflects a rich and varied vocabulary. For instance, the repetitive string "`hello hello hello hello hello hello`" results in a low MTLD score of approximately 2.02, due to the lack of word variation. In contrast, the diverse sentence "`the quick brown fox jumps over the lazy dog`" yields a high MTLD score of around 22.68, as it contains many unique words. The pseudo-code for the algorithm is given in Algorithm 1 as seen in [17].

Algorithm 1 MTLD Score [17]

function MTLD(text, min = 10)
  $forward \leftarrow$ MTLDPROCESS$(text, min)$
  $reverse \leftarrow$ MTLDPROCESS$(Reverse_text, min)$
  return $(forward + reverse)/2$
end function
function MTLDPROCESS(text, min)
  $factor \leftarrow 0$
  $factor\_lengths \leftarrow 0$
  $start \leftarrow 0$
  for $x \leftarrow 0$ **to** $length(text) - 1$ **do**
    $segment \leftarrow text[start : x + 1]$
    if $x + 1 = length(text)$ **then**
      $partial \leftarrow \frac{1-\text{TTR}(segment)}{1-0.72}$
      $factor \leftarrow factor + partial$
      $factor\_lengths \leftarrow factor\_lengths + length(segment)$
    else if $\text{TTR}(segment) < 0.72$ **and** $length(segment) \geq min$ **then**
      $factor \leftarrow factor + 1$
      $factor\_lengths \leftarrow factor\_lengths + length(segment)$
      $start \leftarrow x + 1$
    end if
  end for
  return $\frac{factor\_lengths}{factor}$
end function

3. Additional results of MARVEL-FX3D

3.1. More Implementation Details

As discussed in the main paper, MARVEL-FX3D is a two-stage pipeline. In the first stage, Stable Diffusion 3.5 [1, 9] is fine-tuned. During each epoch, one annotation is sampled from five levels and paired with a randomly selected multi-view image for MSE loss calculation. During inference, CFG [11] is set to 7.5, and 30 steps are used to balance speed and output diversity.

3.2. Baseline Adaptation

We use the official implementations and pretrained models for Shap-E [13] and Luciddreamer [15], training the latter for 3k steps. Dreamfusion [19] and HIFA [23] are trained using the open-source threestudio [10] implementation, with 10,000 and 24,000 steps, respectively, under default settings.

3.3. More Text-to-3D Results

Figures 18 and 19 showcase visual results of TT3D generation on unseen prompts. Using GPT-4 [18], we generated 10 random prompts focused on shape and scene descriptions. As demonstrated, MARVEL-FX3D produces higher-fidelity 3D models from text prompts compared to the baseline methods.

4. Discussion on Application of MARVEL

The MARVEL-40M+ dataset, with its scale and diversity, serves as a powerful resource for text-to-3D tasks such as reconstruction, multi-view consistency, and compositional scene generation. A notable real-world use case, illustrated in Figure 1, demonstrates how MARVEL-FX3D which is trained on MARVEL dataset enables rapid prototyping of diverse 3D objects from complex, fine-grained or simple text prompts. This capability facilitates the creation of intricate scenes, making it particularly valuable for applications in gaming, AR, and VR.

Figure 4. Qualitative comparison of 3D annotations across baselines [12, 14, 16] and the proposed MARVEL-40M+ for *automotive (cars, planes, etc) and CAD models*. MARVEL-40M+ provides more accurate and domain-specific annotations, compared to the baselines. Incorrect captions are highlighted in red, while important captions are highlighted in green.

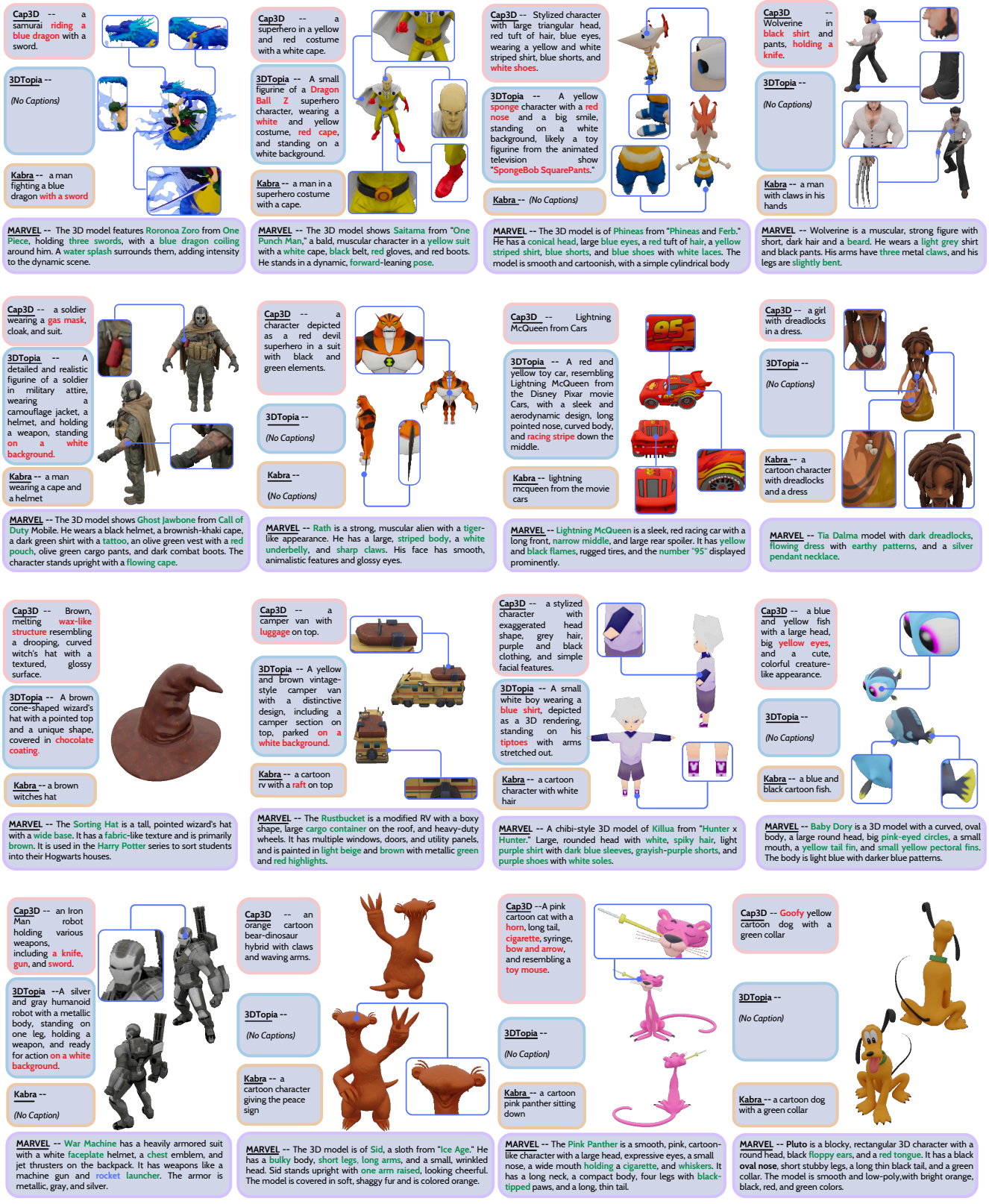


Figure 5. Qualitative comparison of 3D annotations across baselines [12, 14, 16] and the proposed MARVEL-40M+ for *popular anime, movie, and cartoon characters*. MARVEL-40M+ provides more accurate and domain-specific annotations, compared to the baselines. Incorrect captions are highlighted in red, while important captions are highlighted in green.

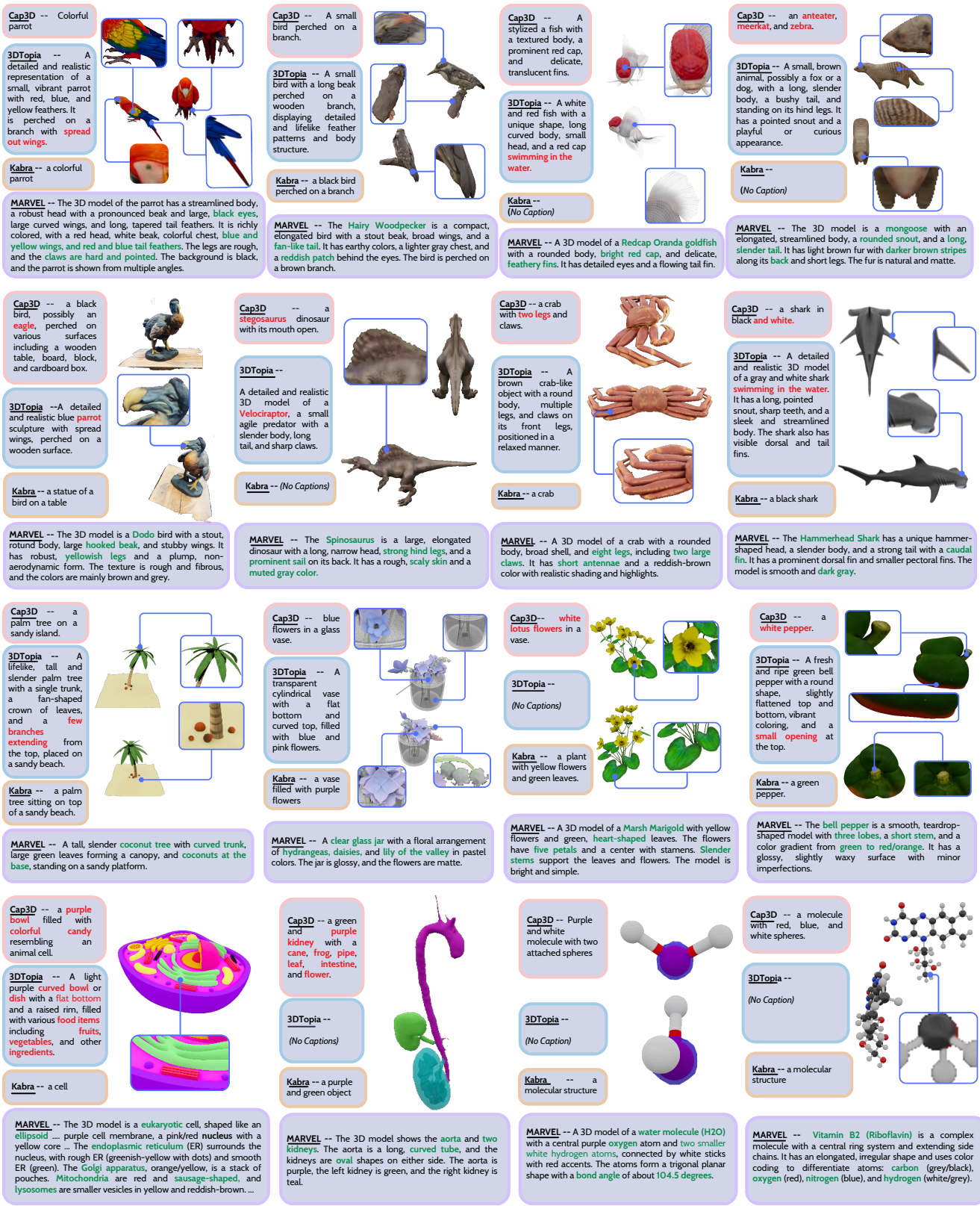


Figure 6. Qualitative comparison of 3D annotations across baselines [12, 14, 16] and the proposed MARVEL-40M+ for *biological objects, including animals, plants, and molecular models*. MARVEL-40M+ provides more accurate and domain-specific annotations, compared to the baselines. Incorrect captions are highlighted in **red**, while important captions are highlighted in **green**.

Figure 7. Qualitative comparison of 3D annotations across baselines [12, 14, 16] and the proposed MARVEL-40M+ for *diverse items including daily objects, essentials*. MARVEL-40M+ provides more accurate and domain-specific annotations, compared to the baselines. Incorrect captions are highlighted in **red**, while important captions are highlighted in **green**.

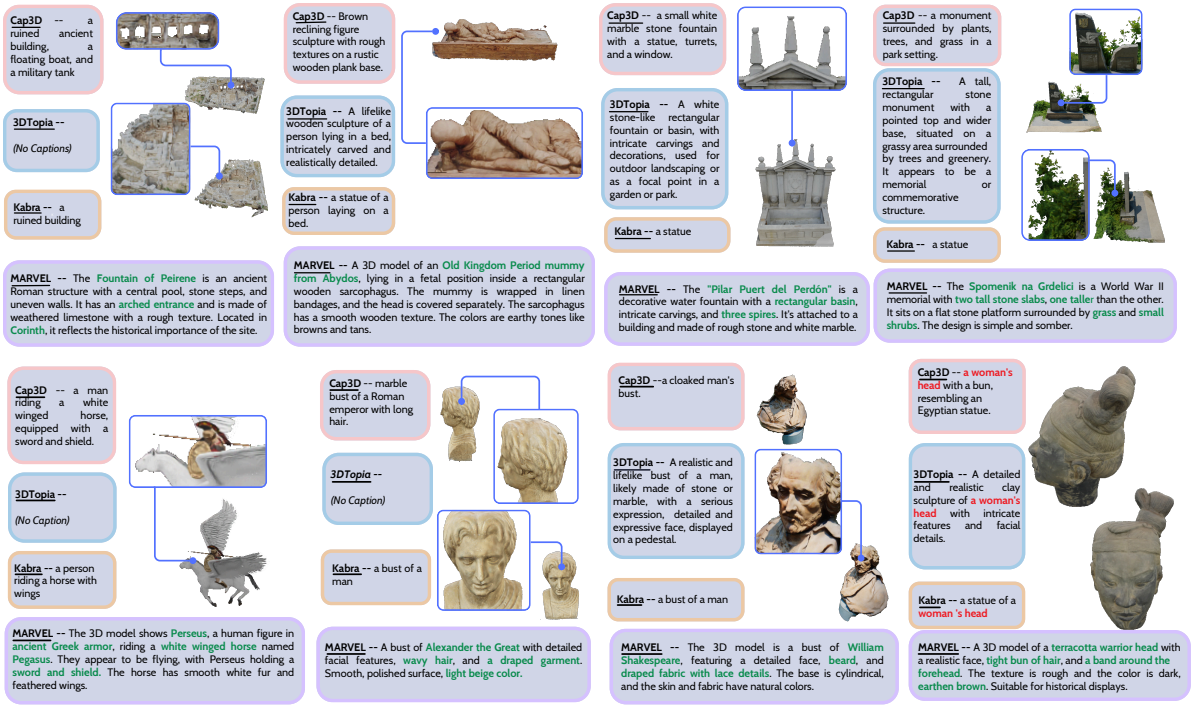


Figure 8. Qualitative comparison of 3D annotations across baselines [12, 14, 16] and the proposed MARVEL-40M+ for *historical elements including statues, places, memorials, etc*. Incorrect captions are highlighted in **red**, while important captions are highlighted in **green**.

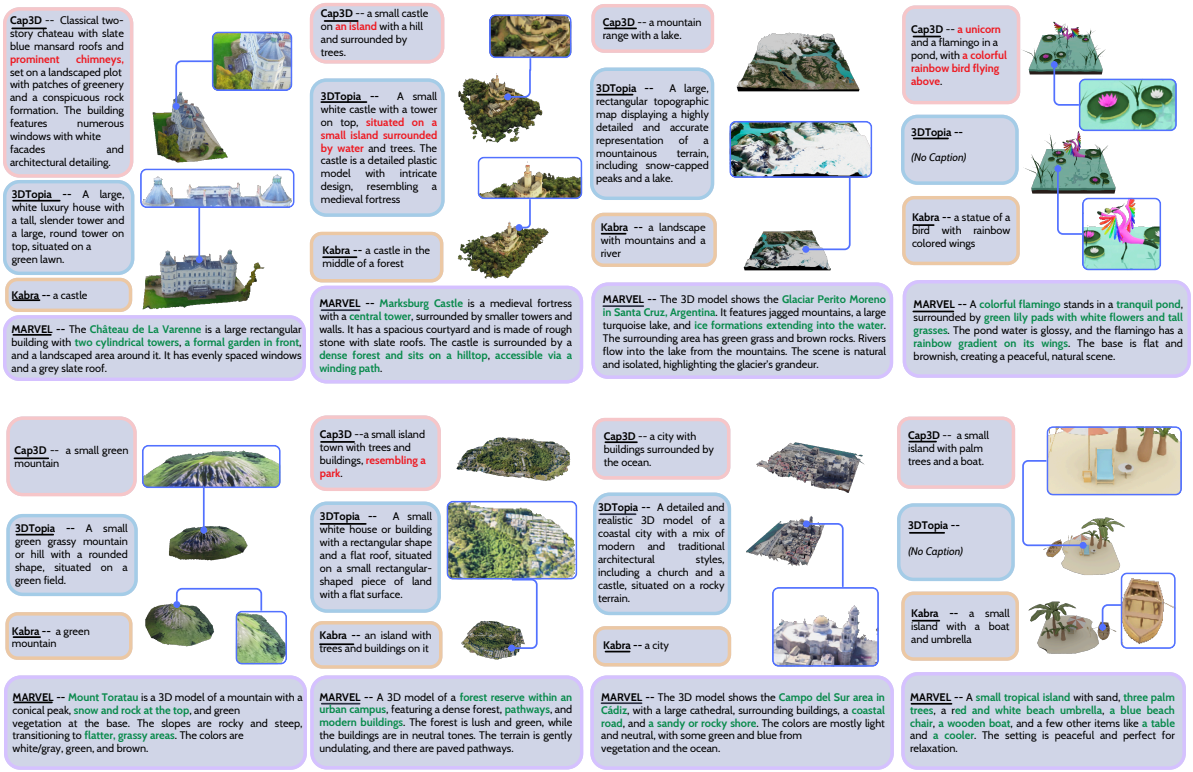


Figure 9. Qualitative comparison of 3D annotations across baselines [12, 14, 16] and the proposed MARVEL-40M+ for *diverse scenes including digital elevation maps, places, realistic or animated scenes*. Incorrect captions are in **red**, while important captions are in **green**.

Blue hair, red eyes, anime style, blue-to-white hoodie, black shorts, knee-high socks, multicolored sneakers, neutral expression, smooth skin, matte hoodie, shiny shoes.

NEARCHAN is a 3D anime character with blue hair, red eyes, a blue-to-white hoodie, black shorts, and multicolored sneakers.

NEARCHAN is a 3D model of an anime-style character with blue, short hair and large red eyes. She wears a blue-to-white hoodie with a hood, black shorts, and knee-high socks with blue cuffs. Her hands are open, and she has multicolored sneakers.

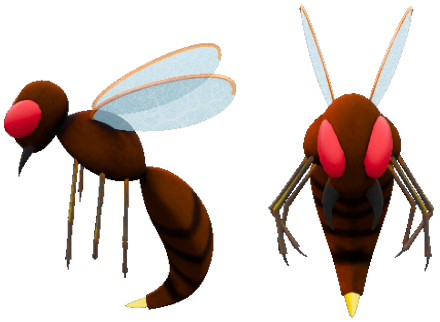
NEARCHAN is a 3D model of an anime-style character with a large head and exaggerated proportions. She has blue, short bob cut hair and large, red eyes. Her torso is covered in a hoodie that transitions from blue at the top to white at the bottom, with a hood resting on her head. Her arms extend downward, with hands slightly away from the body and fingers spread. She wears black shorts and knee-high socks with blue cuffs. Her feet are in multicolored sneakers. The skin is smooth, the hoodie is matte, and the shoes are shiny.

NEARCHAN is a 3D model of an anime-style character with a humanoid form and exaggerated proportions. Her head is larger than her body, featuring blue, short bob cut hair with straight bangs. Her anime-styled head includes large, red eyes and a neutral expression. The neck is slim, connecting to a torso clad in a hoodie jacket that transitions from blue at the top to white at the bottom. The hood rests on her head, adding to the casual look. Her upper torso is light blue, while the lower torso is white with no distinct patterns. The arms extend downward from the shoulders, with hands positioned slightly away from the body, fingers spread out in an open gesture. She wears black shorts over knee-high socks, with blue cuffs and black fabric above. Her feet are adorned with multicolored sneakers featuring blue, pink, and light green. The skin is smooth and slightly glossy, the hoodie has a matte finish, and the shoes have a shiny appearance.

Elongated body, striped abdomen, semi-transparent wings, red eyes, thin legs, yellow stinger, metallic sheen.

A 3D wasp with a long, striped body, large transparent wings, thin dark legs, and a yellow stinger.

The 3D model is a wasp with a long, striped abdomen, small head, and thorax. It has large, see-through wings with vein patterns, thin, dark legs, and a yellow stinger at the end of the abdomen. The head features red eyes, and the body has a smooth texture with a metallic sheen on the legs.



The 3D model represents a stylized wasp with an elongated body divided into a head, thorax, and abdomen. The head and thorax are small and dark brown, with the head featuring red eyes. The abdomen is long and striped with dark brown and tan, tapering at the end. Large, semi-transparent wings with visible veins extend from the thorax, and the legs are thin and dark with a metallic sheen. A pointed yellow stinger is located at the tip of the abdomen.

The 3D model of a wasp features an elongated body divided into three main sections: head, thorax, and abdomen. The head is small with smooth, dark brown surfaces and red eyes. The thorax, similarly smooth and dark brown, supports two pairs of semi-transparent wings with intricate veining patterns. The abdomen is long and slender, tapering at the end with a slight upward curve, and is marked by alternating dark brown and tan stripes, creating a segmented appearance. The wings are large relative to the body, reflecting light differently and adding depth. The legs are thin, jointed, and have a dark metallic sheen, while the stinger at the end of the abdomen is pointed and yellow, contrasting sharply against the darker segments.

Figure 10. Multi-level annotation examples of MARVEL for the Objaverse [7] dataset. Words corresponding to **Object and Components** are highlighted in violet, **Shape and Geometry** in green, **Texture and Materials** in orange, **Colors** in blue, and **Contextual Environment** in purple. From top to bottom, we go from level-5 (Concise Tags) captions to level-1 (Comprehensive Description) captions.

Tennis paddle. circular or oval blade, smooth and glossy surface, dark brown handle, matte finish, gradient brown hues, "TABLE TENNIS MATCH" text, cat's face illustration, slight edge curvature.

A table tennis paddle with a smooth, brown blade featuring a cat's face and the text "TABLE TENNIS MATCH" and a dark brown, matte handle.

The 3D model is a table tennis paddle with a circular or oval blade and a handle. The blade is smooth and glossy, with a gradient of brown hues, darker at the edges and lighter in the center. The text "TABLE TENNIS MATCH" is in white with a blue shadow, and there's a cat's face illustration in orange and white. The handle is dark brown and matte.

The 3D model is a table tennis paddle with a blade and a handle. The blade is circular or oval, slightly curved at the edges for better control. The blade has a smooth, glossy surface, contrasting with the matte handle. The blade features a gradient of brown hues, with darker edges and lighter centers. The text "TABLE TENNIS MATCH" is prominently displayed in white with a blue shadow, and there's a central illustration of a cat's face in orange and white. The handle is dark brown.

The 3D model represents a table tennis paddle, consisting of a blade and a handle. The blade is circular or oval, with a symmetrical design and a slight curvature at the edges, enhancing maneuverability. The proportions ensure a wide striking surface. The blade is smooth and glossy, likely made of high-quality wood laminate, while the handle has a matte finish, indicating a grip material over a wooden core. The blade features a gradient of brown hues, darker at the edges and lighter towards the center. The text "TABLE TENNIS MATCH" is displayed in white uppercase letters with a blue shadow, adding a 3D effect. An illustration of a cat's face, using orange and white, is centrally located. The handle is uniformly dark brown, matching the blade's darker edges.

Round shape, wide opening, tapered base, glossy finish, white base color, central strawberry design, scattered colorful dots, thin orange rim.

A round ceramic bowl with a wide opening, glossy white surface, central strawberry design, and colorful dots. The rim has a thin orange border.

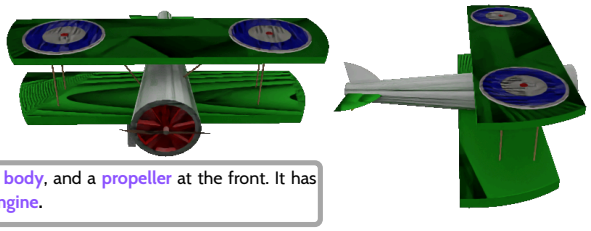
The ceramic bowl is round with a wide opening that tapers down to a smaller base. It has a smooth, glossy finish and is primarily white. A central strawberry design in red and green is featured at the bottom, with scattered colorful dots. The rim is outlined with a thin orange border.

The ceramic bowl has a smooth, symmetrical shape with a wide circular opening that narrows towards the base. Both the interior and exterior follow a gentle conical curve, creating a balanced form. The bowl is made of ceramic or porcelain and has a glossy, reflective finish. The base color is white, with a central strawberry design in red and green, and scattered decorative dots in orange, green, and yellow. The rim is accented with a thin orange border.

The ceramic bowl features a smooth, symmetrical shape with a wide circular opening that tapers down to a narrower base. The interior is conically shaped, transitioning smoothly from the rim to the base. The exterior mirrors this curvature, creating a balanced and harmonious form. The bowl is made of ceramic or porcelain, with a glossy finish that reflects light, giving it a smooth, polished appearance. The base color is white, serving as a neutral backdrop for the decorative elements. A central strawberry design, rendered in red and green, is prominently featured at the bottom, surrounded by scattered dots and splashes in orange, green, and yellow, adding a playful touch. The rim is highlighted with a thin orange border, framing the bowl elegantly.

Figure 11. Multi-level annotation examples of MARVEL for the Omni-Object [22] dataset. Words corresponding to **Object and Components** are highlighted in violet, **Shape and Geometry** in green, **Texture and Materials** in orange, **Colors** in blue, and **Contextual Environment** in purple. From top to bottom, we go from level-5 (Concise Tags) captions to level-1 (Comprehensive Description) captions.

Biplane, stacked wings, circular insignias, elongated fuselage, propeller, landing gear, horizontal tailplane, vertical stabilizers, green body, blue and red markings, silver engine.



A classic biplane with two sets of wings, a long cylindrical body, and a propeller at the front. It has blue and red wing insignias, two small wheels, and a silver engine.

The 3D model is a biplane with two sets of wings stacked on top of each other. The fuselage is long and cylindrical, narrowing at the back, with a propeller at the front. The wings have blue circular insignias with red dots. The landing gear consists of two small wheels under the fuselage. The control surfaces include a horizontal tailplane and vertical stabilizers at the rear. The plane is mainly green, with blue and red wing markings and a silver engine and propeller.

The 3D model is a biplane with a traditional layout. It features two sets of horizontally stacked wings, each adorned with circular blue insignias and red dots. The fuselage is elongated and cylindrical, tapering towards the rear, with a cylindrical engine and propeller at the front. The landing gear includes two small wheels at the bottom center of the fuselage. Control surfaces consist of a horizontal tailplane and vertical stabilizers at the rear. Struts and braces provide structural support between the upper and lower wings. The aircraft is primarily green, with blue and red wing insignias and a silver engine and propeller.

The 3D model represents a classic biplane with a detailed and symmetrical design. The main wings are horizontally stacked in a biplane configuration, each featuring circular blue insignias with red dots at their centers. The fuselage is elongated and cylindrical, tapering slightly towards the rear, with a cylindrical engine mounted on top and a prominent propeller at the front. The landing gear consists of two small wheels positioned near the bottom center of the fuselage. The control surfaces include a horizontal tailplane and vertical stabilizers at the rear. Struts and braces connect the upper and lower wings for structural support. The wings are rectangular with rounded tips, and the fuselage and engine housing exhibit a semi-matte finish, simulating lightweight materials like aluminum or wood. The engine and propeller blades have a metallic, slightly glossy appearance. The primary color of the aircraft is green, with blue and red accents on the wing insignias and silver for the engine and propeller.

Hammock, fabric body, wooden spreader bars, black pillow, vibrant colors, woven pattern, ropes, hooks, outdoor use, rectangular form, bilateral symmetry.

A colorful hammock with wooden spreader bars, a black pillow, and ropes for hanging. Suitable for outdoor relaxation.

The 3D model is a hammock with a colorful fabric body, two wooden spreader bars, and a black pillow. The fabric is an elongated rectangle that curves slightly and features vibrant patterns. The spreader bars keep the hammock open and rigid. The pillow, attached at one end, is soft and comfortable. Ropes and hooks are provided for hanging the hammock. The spreader bars are wood-colored, and the pillow is brownish-yellow. It is designed for outdoor use.

The 3D model is a hammock with a main fabric body, two wooden spreader bars, and a black pillow. The fabric is an elongated rectangle that curves slightly, with a woven pattern and vibrant colors including red, yellow, pink, and green. The spreader bars maintain the rectangular form and provide rigidity. The pillow, attached at one end, adds comfort. Ropes and hooks are included for secure suspension. The spreader bars are a natural wood color, and the pillow is brownish-yellow. The design is suitable for outdoor settings.

The 3D model represents a hammock with a main fabric body, two wooden spreader bars, and a black pillow. The fabric is an elongated rectangle that curves slightly, with a woven pattern suggesting a synthetic blend. The spreader bars, attached at the ends, maintain the rectangular form and provide rigidity. The hammock exhibits bilateral symmetry along its central axis, with open ends revealing the inner fabric and attachment points. The pillow, attached at one end, has a plush texture. Ropes and hooks are included for suspension, likely made of nylon or polyester. The fabric features vibrant red, yellow, pink, and green patterns, with the spreader bars in a natural wood color and the pillow in a brownish-yellow hue. The design is versatile for outdoor use, such as patios, gardens, or camping.

Figure 12. Multi-level annotation examples of MARVEL for the ShapeNet [2] dataset. Words corresponding to **Object and Components** are highlighted in violet, **Shape and Geometry** in green, **Texture and Materials** in orange, **Colors** in blue, and **Contextual Environment** in purple. From top to bottom, we go from level-5 (Concise Tags) captions to level-1 (Comprehensive Description) captions.

Triceratops, large head, eye horns, nasal horn, frill with spikes, robust body, strong legs, long tapering tail, rough scaly skin, green body, lighter horns.

The 3D model is a Triceratops, a large dinosaur with a big head, two eye horns, a nasal horn, and a frill with spikes. It has a robust body, strong legs, and a long, tapering tail. The skin is rough and scaly, with a green body and lighter horns.

The 3D model is a Triceratops, known for its large head with two eye horns and a big nasal horn, and a frill with spikes. The body is robust and supported by strong legs. The tail is long and tapers. The skin is rough and scaly, with a green body and lighter horns. This model captures the essential features of a Triceratops, making it suitable for both educational and creative projects.

The 3D model is a Triceratops, featuring a large head with two eye horns and a prominent nasal horn, set against a frill with spikes. The body is robust and tapers into a shorter tail. Strong, sturdy legs support the heavy frame. The skin is rough and scaly, with raised bumps along the back. The horns and frill are smooth and slightly glossy, contrasting with the green body and lighter horns.

The 3D model represents a Triceratops, a large ornithopod dinosaur. The head features two prominent horns above the eyes and a larger nasal horn, with a frill at the back adorned with spikes. The body is bulky and robust, transitioning into powerful hindquarters and a relatively short but thick neck. The tail is long and tapers backward, providing balance. The skin texture is rough and scaly, with raised bumps along the back, suggesting osteoderms. The horns and frill have a smooth, slightly glossy surface, possibly covered with keratinous material. The main body color is predominantly green with darker patches, while the horns and frill are a lighter, almost white shade.

Pug, traffic cone hat, round body, folded ears, small round eyes, flat snout, short sturdy legs, curled tail, bright yellow, brown eyes, matte texture, yellow and white stripes.

A pug dog with a round body and a traffic cone hat. The pug has folded ears, small round eyes, and a flat snout. The hat is yellow with white stripes. The body is bright yellow with brown eyes and nose. The texture is matte.

The 3D model is a pug dog wearing a traffic cone hat. The pug has a round, short body with a broad face and folded ears. Its eyes are small and round, and the snout is flat and large. The legs are short and sturdy, and the tail is small and curled. The **traffic cone hat** is yellow with white stripes and fits around the pug's head. The pug's body is mostly bright yellow, with brown areas for the eyes and nose. The texture is matte, and the model is simple.

The 3D model depicts a pug dog wearing a traffic cone hat. The pug has a rounded, short-statured body with a broad face and compact build. Its ears are folded downward, and the eyes are small and round, positioned slightly above mid-face level. The snout is flat and large, typical of pugs. The legs are short and sturdy, and the tail is small and curled up over the back. The traffic cone hat is conical with alternating yellow and white stripes, fitting around the pug's head. The pug's body is primarily bright yellow, with brownish areas for the eyes and nose. The texture is matte, and the model is in a low poly style.

The 3D model represents a pug dog adorned with a traffic cone hat. The pug's body is rounded and short-statured, with a broad face and compact build. The ears are folded downward, close to the sides of the face. The eyes are small and round, positioned slightly above mid-face level, with a glossy appearance, possibly indicating a glass or polished plastic material. The snout is flat and large, characteristic of pugs. The legs are short and sturdy, supporting the round body, while the tail is small and curled up over the back. The traffic cone hat is conical, featuring alternating yellow and white stripes, with a circular base that fits snugly around the pug's head. The pug's body is a uniform bright yellow, with brownish areas for the eyes, nose, and some parts of the face. The texture of the pug and the cone hat is matte, with no shiny highlights or reflections. The model is designed in a low poly style, with subtle geometric facets.

Figure 13. Multi-level annotation examples of MARVEL for the Toys4K dataset. Words corresponding to **Object and Components** are highlighted in violet, **Shape and Geometry** in green, **Texture and Materials** in orange, **Colors** in blue, and **Contextual Environment** in purple. From top to bottom, we go from level-5 (Concise Tags) captions to level-1 (Comprehensive Description) captions.

Blue ceramic cup, embossed flower patterns, tapered shape, smooth interior, deep blue color, lighter interior, gradient effect, symmetrical design, stable base.

A blue ceramic cup with embossed flower patterns on the outside and a smooth interior. It tapers from a wide top to a narrow base and is primarily deep blue with a lighter interior.

The 3D model is a blue ceramic cup with flower patterns. It has a standard cup shape, tapering slightly from a wide top to a narrow base. The exterior is textured with embossed flowers, while the interior and base are smooth. The cup is primarily deep blue, with a lighter shade inside. The flowers have a subtle gradient effect.

The 3D model is a blue ceramic cup with a flower pattern. It has a standard cup shape, slightly tapering from a wide opening at the top to a narrower base at the bottom. The exterior features embossed flower patterns, evenly distributed around the cylindrical body, creating a textured look. The interior and base are smooth, making it easy to clean and stable on surfaces. The primary color is deep blue, with a lighter shade inside. The flower patterns have a subtle gradient effect, with darker centers and lighter petals.

The 3D model represents a blue ceramic cup with a flower pattern. The cup has a standard shape, slightly tapering from a wider opening at the top to a narrower base at the bottom, ensuring stability. The geometry is symmetrical along its vertical axis, with equal proportions on all sides. The exterior surface features multiple embossed flower patterns, evenly distributed around the cylindrical body. Each pattern consists of concentric petals radiating outward from a central point, resembling a sunflower. The texture of the exterior is raised, providing a tactile quality. The interior and base surfaces are smooth, facilitating easy cleaning and enhancing stability. The primary color is deep blue, with a lighter shade of blue or off-white on the interior. The flower patterns have a subtle gradient effect, with darker centers and lighter petals, creating a harmonious visual contrast.

Ergonomic seat, high backrest, headrest, horizontal armrests, five-spoke base, caster wheels, smooth white upholstery, gray metal accents, minimalistic design.

Modern office chair with a curved seat, high backrest, and horizontal armrests. Five-spoke base with wheels. Smooth white upholstery, gray metal accents. Clean, minimalist design.

The office chair has a comfortable, slightly curved seat and a high, curved backrest with a headrest. Armrests are horizontal with slight upward curves. The base has five spokes with wheels. The chair is covered in smooth, white leather-like material with gray metal accents. It has a clean, modern look.

The office chair has an ergonomic design with a slightly curved, rectangular seat that tapers at the front. The high backrest curves gently and includes a headrest. Armrests are positioned mid-width and extend horizontally with slight upward curves. The base features five spokes with caster wheels. The chair is upholstered in a smooth, white leather-like material, with gray accents on the metal parts. The design is clean and minimalistic, ideal for modern offices.

The modern office chair features an ergonomic design with a slightly curved, rectangular seat that tapers at the front for leg comfort. The high backrest provides substantial lumbar support and gently curves from top to bottom, integrating a headrest. Armrests, positioned near the midpoint of the backrest width, extend horizontally with slight upward curves for optimal forearm rest and shoulder alignment. The base consists of five spokes converging into a central hub, each ending in a caster wheel for mobility. The chair is upholstered in a smooth, leather-like material, predominantly white, with subtle gray accents on the metal components, including the base and adjustment mechanisms. The design is minimalistic, with uniform colors and no patterns, making it suitable for modern office settings.

Figure 14. Multi-level annotation examples of MARVEL for the ABO (Amazon Berkeley Objects) [5] dataset. Words corresponding to **Object and Components** are highlighted in violet, **Shape and Geometry** in green, **Texture and Materials** in orange, **Colors** in blue, and **Contextual Environment** in purple. From top to bottom, we go from level-5 (Concise Tags) captions to level-1 (Comprehensive Description) captions.

Cylindrical shape, beige fabric, horse print, light beige zipper, corner reinforcements, blue tag, flat base, stands upright, vibrant colors, symmetrical design.

A cylindrical pencil case with a beige fabric body and colorful horse print. Features a light beige zipper, corner reinforcements, and a blue tag. Stands upright on a flat base.

The Horse Print Pencil Case is a cylindrical pencil holder with a beige fabric body and a colorful horse print. It has a light beige zipper, corner reinforcements, and a blue tag near the zipper. The flat base allows it to stand upright. The design is clean and functional.

The Horse Print Pencil Case is a cylindrical, symmetrical object with a beige fabric body featuring a vibrant horse print. It includes a light beige zipper mechanism for opening and closing, and corner reinforcements to prevent damage. A blue rectangular tag is attached near the zipper. The flat base allows the case to stand upright.

The Horse Print Pencil Case is a cylindrical, symmetrical object designed to hold writing instruments. It features a beige fabric body with a vibrant horse print pattern, showcasing horses in various poses and colors such as black, brown, white, and blue. The case has a light beige zipper mechanism, likely made of metal with plastic components for ease of use. Corner reinforcements at both ends are made of sturdy material to prevent fraying and tearing, matching the main body's color for a seamless look. A blue rectangular tag, possibly fabric or plastic, is sewn onto one end near the zipper, providing additional branding or information. The flat base allows the case to stand upright, and the overall proportions are consistent throughout its length.

Blue, Nintendo 3DS XL, handheld gaming console, rounded rectangle, matte finish, two screens, black bezels, touchscreen, directional pad, action buttons, start/select buttons, strap holes, hinges, detachable upper cover, speaker slots, branding.

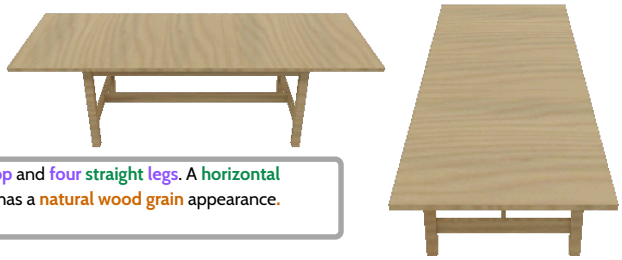
The 3D model is a blue **Nintendo 3DS XL** handheld gaming console with a rounded rectangular shape. It has two screens, a larger touchscreen on the bottom and a smaller screen above, both with black bezels. Controls include a directional pad, action buttons, and start/select buttons. The upper case is a detachable cover, and the backside has "Nintendo 3DS XL" branding.

The 3D model is a blue Nintendo 3DS XL handheld gaming console. It has a slightly rounded rectangular shape with a matte finish. The console features two screens: a larger touchscreen on the bottom and a smaller screen above, both with black bezels. Controls include a directional pad, action buttons ('A', 'B', 'X', 'Y'), and start/select buttons. There are strap holes on the sides and hinges connecting the upper and lower parts. The upper case is a detachable cover, and the backside has "Nintendo 3DS XL" branding. The primary material is plastic, and the color scheme is bright blue with black accents.

The 3D model is a blue Nintendo 3DS XL handheld gaming console. The main body is a slightly rounded rectangle with a matte finish. It features two screens: a larger touchscreen on the bottom and a smaller screen above, both with black bezels. Controls include a directional pad, action buttons ('A', 'B', 'X', 'Y'), and start/select buttons, with a power button near the top-left corner. Strap holes are on the upper edges, and hinges connect the upper and lower parts. The upper case is a detachable cover, and speaker slots are on the sides of the upper case. The backside has "Nintendo 3DS XL" branding and regulatory text. The primary material is plastic, and the color scheme is bright blue with black accents.

The 3D model represents a blue Nintendo 3DS XL handheld gaming console. The main body is a slightly rounded rectangular shape with a matte finish, housing all internal components. Two screens are present: a larger touchscreen on the bottom and a smaller screen above, both surrounded by black bezels. Controls include a directional pad on the left, action buttons ('A', 'B', 'X', 'Y') on the right, and start/select buttons at the center, with a power button near the top-left corner. Strap holes are located on the upper edges of both sides. The closure mechanism features visible hinges where the upper and lower parts meet. The upper case is a detachable cover that folds over the main body, and the lower case houses the screens and controls. Speaker slots are visible on either side of the upper case, just below the hinge area. The backside displays "Nintendo 3DS XL" branding, along with regulatory text and logos. The primary material is plastic with a matte texture, offering a soft tactile feel, while control areas may have a slightly glossier finish. The predominant color is bright blue, with black accents for contrast.

Figure 15. Multi-level annotation examples of MARVEL for the GSO (Google Scanned Objects) [8] dataset. Words corresponding to **Object and Components** are highlighted in violet, **Shape and Geometry** in green, **Texture and Materials** in orange, **Colors** in blue, and **Contextual Environment** in purple. From top to bottom, we go from level-5 (Concise Tags) captions to level-1 (Comprehensive Description) captions.



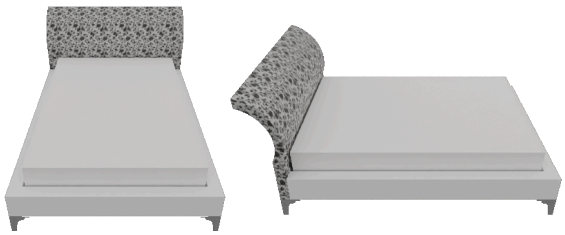
Rectangular tabletop, smooth wood grain, four straight legs, horizontal support beam, light brown color, natural wood appearance.

A rectangular wooden table with a smooth, light brown tabletop and four straight legs. A horizontal support beam runs underneath, connecting the legs. The table has a natural wood grain appearance.

The 3D model is a rectangular wooden table. It has a flat, smooth tabletop with wood grain patterns. Four legs are positioned at each corner, and a horizontal support beam runs underneath, connecting the legs. The table is light brown with subtle wood grain, giving it a natural look.

The 3D model is a rectangular wooden table. The tabletop is flat and smooth with visible wood grain, indicating natural materials. Four legs, positioned at each corner, have a simple, straight shape and a slightly rounded texture. A horizontal support beam connects the legs underneath, running parallel to the longer edges of the table. The table has a light brown color with subtle variations in hue, creating a realistic wood grain effect.

The 3D model represents a rectangular wooden table. The tabletop is a flat, smooth surface with visible wood grain patterns, suggesting natural materials like plywood or solid wood veneer. The table measures significantly longer than it is wide, maintaining symmetry in leg placement and underframe design. Four legs, positioned at each corner, have a simple, straight shape with a slightly rounded texture, showing consistent wood grain patterns. A horizontal support beam connects two pairs of legs underneath the tabletop, running parallel to the longer edges, enhancing stability. All components share a light brown color with subtle variations in hue, creating a realistic wood grain effect.



Modern rectangular bed, white metal frame, slightly curved headboard, grayish floral pattern, plain white mattress, symmetric design, extended headboard, smooth matte finish.

A modern rectangular bed with a slightly curved headboard. The bed has a white metal frame and a grayish floral-patterned headboard. The mattress is plain white. The headboard extends slightly beyond the mattress width.

The bed with headboard is a modern, rectangular design. It has a sturdy white metal frame and a slightly curved headboard covered in grayish fabric with a floral pattern. The mattress is plain white and sits on top of the frame. The design is symmetric and well-proportioned, with the headboard extending slightly beyond the mattress width.

The bed with headboard is a modern, rectangular design. The base frame is made from a sturdy metal with a smooth, matte white finish. The headboard, which is slightly curved, is covered in a grayish fabric with a detailed floral pattern. The mattress is plain white and sits on top of the frame. The design is symmetric, with sharp, clean lines and a balanced proportion between the headboard and the bed's length. The headboard extends slightly beyond the width of the mattress. Ensure the mattress aligns perfectly with the headboard and maintain the smooth, matte finish of the frame.

The bed with headboard is a modern, rectangular design featuring a slightly curved headboard at one end. The base frame is constructed from a sturdy, sleek metal with a smooth, matte white finish. The headboard is covered in a grayish fabric with an intricate floral pattern, featuring small dark flowers and leaves. The fabric has a soft, slightly raised texture, adding depth and detail. The mattress, placed on top of the base frame, is covered in plain white fabric, suggesting a smooth, padded surface for comfort. The design is symmetric along both axes, with sharp, clean lines defining each side. The headboard extends beyond the width of the mattress, maintaining a balanced proportion with the bed's length. Ensure the mattress dimensions match those of the frame, aligning perfectly with the headboard. Pay close attention to the textures and proportions to achieve a faithful recreation.

Figure 16. Multi-level annotation examples of MARVEL for the Pix3D [21] dataset. Words corresponding to **Object and Components** are highlighted in violet, **Shape and Geometry** in green, **Texture and Materials** in orange, **Colors** in blue, and **Contextual Environment** in purple. From top to bottom, we go from level-5 (Concise Tags) captions to level-1 (Comprehensive Description) captions.

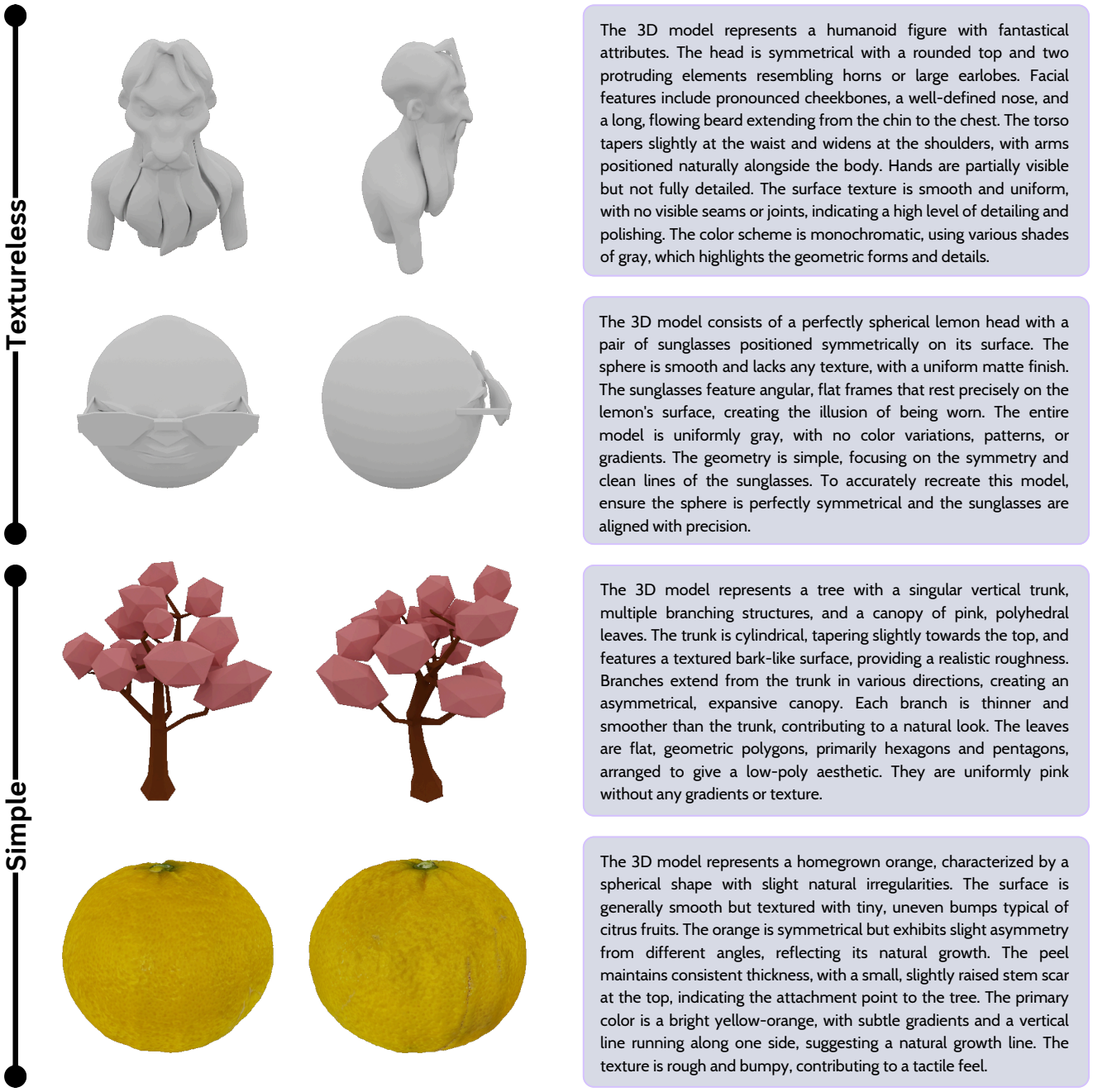


Figure 17. Examples illustrating MARVEL's robustness to simple and textureless models. Our annotation pipeline dynamically adjusts verbosity, ensuring concise yet accurate descriptions even when texture details are minimal or absent. (**Top**) Textureless models: a smooth humanoid figure with fantastical attributes, and a symmetrical, matte-finished spherical lemon head wearing sunglasses. (**Bottom**) Simple yet detailed models: a low-poly tree with geometric pink leaves, and a realistically textured homegrown orange showcasing subtle natural irregularities.

Figure 18. Qualitative Results for high fidelity TT3D generation on unseen prompts. From left to right, 3D models generated using Shap-E [13], DreamFusion [19], LucidDreamer [15], HiFA [23] and MARVEL-FX3D (ours).

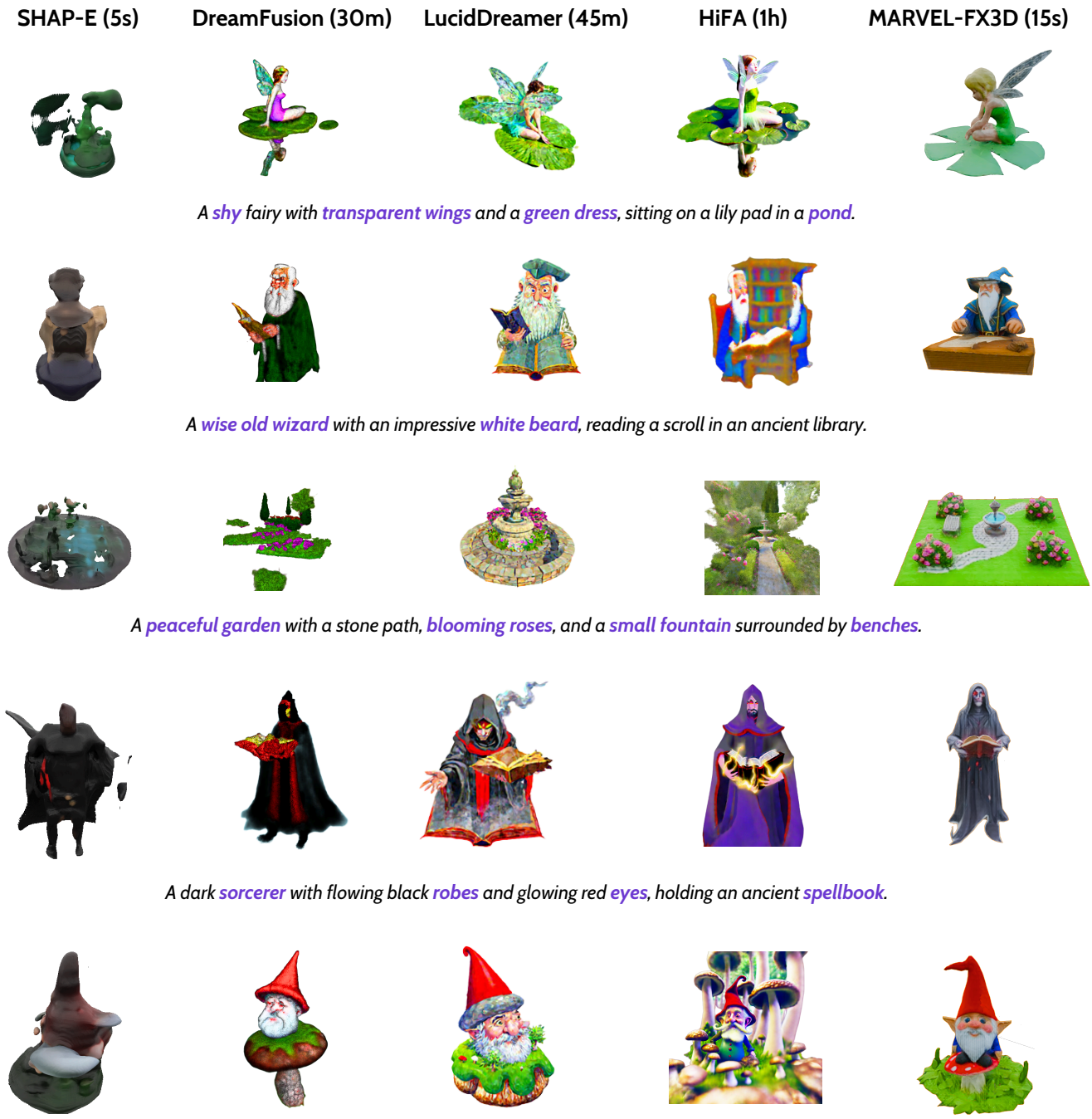


Figure 19. Qualitative Results for high fidelity TT3D generation on unseen prompts. From left to right, 3D models generated using Shap-E [13], DreamFusion [19], LucidDreamer [15], HiFA [23] and MARVEL-FX3D (ours).

References

[1] Stable diffusion 3.5 large - huggingface. https://huggingface.co/stabilityai/stable-diffusion-3.5-large. 4

[2] Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, L. Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *ArXiv*, abs/1512.03012, 2015. 3, 12

[3] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023. 2, 3

[4] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 2

[5] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. Abo: Dataset and benchmarks for real-world 3d object understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 3, 14

[6] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-XL: A universe of 10m+ 3d objects. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 1

[7] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 1, 3, 10

[8] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B. McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items, 2022. 1, 3, 15

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. 4

[10] 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. https://github.com/threestudio-project/threestudio, 2023. 4

[11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 4

[12] Fangzhou Hong, Jiaxiang Tang, Ziang Cao, Min Shi, Tong Wu, Zhaoxi Chen, Shuai Yang, Tengfei Wang, Liang Pan, Dahua Lin, and Ziwei Liu. 3dtopia: Large text-to-3d generation model with hybrid diffusion priors, 2024. 2, 5, 6, 7, 8, 9

[13] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions, 2023. 4, 18, 19

[14] Rishabh Kabra, Loic Matthey, Alexander Lerchner, and Niloy J. Mitra. Leveraging vlm-based pipelines to annotate 3d objects. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 2024. 2, 5, 6, 7, 8, 9

[15] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6526, 2024. 4, 18, 19

[16] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. In *Advances in Neural Information Processing Systems*, pages 75307–75337. Curran Associates, Inc., 2023. 2, 5, 6, 7, 8, 9

[17] Philip M. McCarthy and Scott Jarvis. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42:381–392, 2010. 4

[18] Josh OpenAI, Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 3, 4

[19] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 4, 18, 19

[20] Stefan Stojanov, Anh Thai, and James M. Rehg. Using shape to categorize: Low-shot learning with an explicit shape bias. 2021. 1, 3

[21] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 16

[22] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Liang Pan Jiawei Ren, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 3, 11

[23] Junzhe Zhu and Peiye Zhuang. Hifa: High-fidelity text-to-3d generation with advanced diffusion guidance, 2023. 4, 18, 19