BEYOND RECOGNITION: BENCHMARKING AND ADVANCING FINE-GRAINED OBJECT UNDERSTANDING

Anonymous authors
Paper under double-blind review

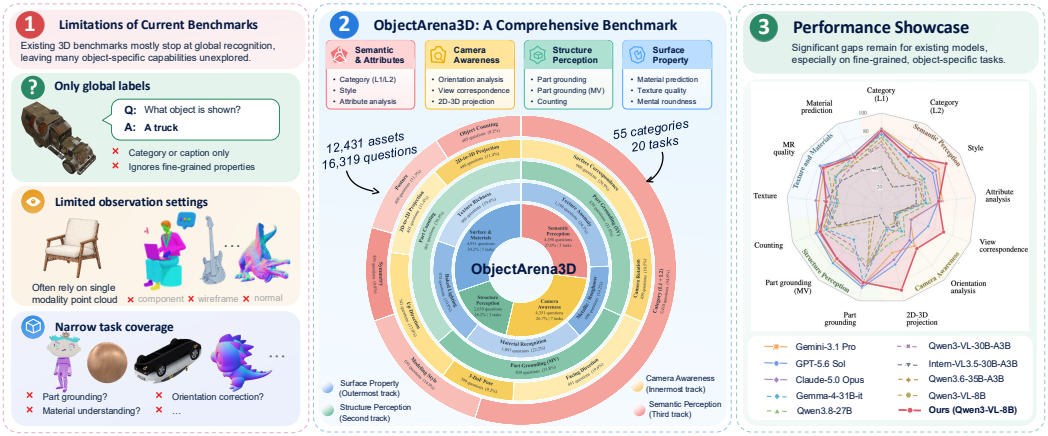


Figure 1: **Overview of fine-grained 3D asset understanding.** Existing probes often stop at a global label. ObjectArena3D evaluates 3D assets across four capability families and 20 tasks, while Objective3D fuses RGB, normal, materials, and dense 3D evidence for task-conditioned answers.

ABSTRACT

Object-centric understanding is fundamental to how intelligent systems perceive, reason about, and interact with the 3D world. Yet 3D vision-language models (VLMs) have largely been evaluated on category recognition, captioning, and broad question answering, which reveals limited insight about whether they understand the fine-grained properties of individual objects. General-purpose VLMs offer a promising alternative by reasoning over multi-view renderings, but it remains unclear whether they can answer fine-grained, object-specific questions reliably, and whether that reliability persists when an asset's appearance or orientation changes. We introduce ***ObjectArena3D***, a benchmark of 16,319 questions over 12,431 assets spanning 20 tasks across semantic attributes, camera awareness, part structure, and surface properties. Beyond standard evaluation, it tests robustness by randomizing object orientations and replacing textures with random solid colors. Our evaluation reveals weak viewpoint and orientation reasoning, with category recognition further degraded under randomized orientations. To narrow these gaps, we propose ***Objective3D***, an framework that fine-tunes a pretrained VLM using view augmentation and intra-view attention across aligned rendering modalities. Compared to baseline Qwen3-VL-8B, Objective3D raises the average benchmark score from 42.95 to 70.26, surpassing the strongest proprietary baseline GPT-5.6-Sol by 4.34 points.

1 INTRODUCTION

A 3D asset is more than an instance of a category: it has a particular appearance, orientation, part composition, and surface properties. For systems that retrieve, inspect, or interact with assets, recognizing the object is only the starting point. A model may recognize a chair yet fail to identify its style, determine which way it faces, localize its armrest, count its legs, or distinguish its material from its color. These properties matter increasingly as 3D assets become the basic units of

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: **Scale, diversity, and capability coverage of object-level 3D resources.** ObjectArena3D combines seven modeling styles (scanned, hard surface, ai-generated, organic, hybrid and procedural generation) and twenty fine-grained subtasks across all four dimensions.

	Scale & Diversity				*Fine-grained Understanding*			
Resource	#QA	#Assets	#Styles	#Tasks	**Semantic**	**Camera**	**Structure**	**Surface**
Datasets								
3D-FUTURE (Fu et al., 2021)	–	9,992	1	5	✓	✓	✗	✓
3DCoMPaT200 (Ahmed et al., 2024)	–	19,051	1	5	✓	✗	✓	✓
OmniObject3D (Wu et al., 2023)	–	6,000	1	4	✓	✗	✗	✗
PartNeXt (Wang et al., 2025b)	–	23,519	1	4	✓	✗	✓	✗
Benchmarks								
Objaverse++ (Lin et al., 2025)	1,971	1,971	NR	6	✓	✗	✗	✓
PointLLM (Xu et al., 2024)	400	200	NR	2	✓	✗	✗	✗
GPT4Point (Qi et al., 2024b)	1000	1000	NR	3				
PointCoT (Zhang et al., 2026)	8,832	2,944	NR	3	✓	✓	✓	✗
ObjectArena3D	**16,319**	**12,431**	**7**	**20**	✓	✓	✓	✓

#QA counts annotated evaluation items; for ObjectArena3D, #QA uses the revised suite and #Assets the original export. Objaverse++ counts labeled test objects; ✓: explicit labels or dedicated evaluation; ✗: neither. Generic captions and input modalities do not count. NR: style counts unreported and not inferable from Objaverse sourcing alone.

embodied interaction (Qi et al., 2024a), robotic manipulation (Xiang et al., 2020), and game and simulation construction (Wang et al., 2026). Yet they are precisely what current models resolve least reliably: Gemini-3.1 Pro (Google DeepMind, 2026) reaches 83.9% on top-level category classification, but only 32.3% on view correspondence, 52.4% on orientation analysis, and 63.7% on material prediction (Table 2).

This gap in fine-grained 3D asset understanding has gone largely unmeasured. As shown in Figure 1, object-level evaluations of 3D vision-language models (VLMs) emphasize category recognition, captioning, and open-ended question answering (Xu et al., 2024; Qi et al., 2024b). A single score over such tasks conflates capabilities that fail for different reasons, and per-question accuracy can conceal language shortcuts and inconsistent behavior on the same object (Huang et al., 2025; Ma et al., 2026). Other resources cover important individual properties, including textured appearance (Fu et al., 2021; Wu et al., 2023; Lin et al., 2025), part–material composition (Li et al., 2022; Ahmed et al., 2024), part structure (Wang et al., 2025b), and orientation and geometric reasoning (Wang et al., 2025c; Jin et al., 2025; Zhang et al., 2026), but each uses its own assets and protocol (Table 1). As a result, semantic attributes, camera awareness, part structure, and surface properties are rarely evaluated on the same asset, and a model's category-level competence cannot be set against its property-level failures. What is missing is a *property-resolved* evaluation that measures these capabilities on shared assets under a single protocol.

We therefore study *fine-grained 3D asset understanding*: the ability to reason about the properties of an individual asset beyond its category. Specifically, we organize this ability into four capability families (Figure 2). *Semantic and attribute understanding* covers category, modeling style, and instance-level attributes such as symmetry, posture, and object count; *camera awareness* covers object orientation, cross-view correspondence, and projection between image and object coordinates; *part-structure understanding* covers part counting and language-conditioned part grounding; and *surface understanding* covers texture quality, material identity, and the plausibility of metallic and roughness maps. This capability profile raises two questions: *where do current VLMs fail across these families, and can object-centric benchmark and targeted fine-tuning close the gaps they leave?*

To answer the first question, we introduce ***ObjectArena3D***, a benchmark of 16,319 questions over 12,431 assets that spans 20 tasks across these four families. General-purpose proprietary and open-source VLMs and specialized 3D object models answer identical questions about the same assets. Beyond this standard setting, ObjectArena3D re-renders assets with randomized orientations and with textures replaced by random solid colors. Accuracy drops under these perturbations expose reliance on the canonical poses in which assets are conventionally rendered and on appearance cues that co-vary with category, rather than on the asset's own geometry.

Our evaluation reveals three gaps in current models. First, camera awareness is the weakest family: no baseline exceeds 53% on view correspondence or orientation analysis, although the strongest models exceed 80% on top-level category classification. Second, apparent structural competence



Figure 2: **Task scope of ObjectArena3D.** The 20 tasks span four capability families: semantic attributes, camera awareness, part structure, and surface properties.

often rests on appearance: under randomized colors, the single-view part-grounding accuracy of Gemini-3.1 Pro and GPT-5.6 Sol falls from above 82% to below 36%. Third, specialized 3D object models trail general-purpose VLMs on most tasks despite consuming 3D input directly, reaching at most 45.0% on top-level category classification. These findings reveal two key limitations: insufficient viewpoint awareness and over-reliance on appearance for structural reasoning. Addressing them requires models to reason over geometry explicitly aligned with RGB observations across views, for which existing VLMs lack an explicit mechanism.

To answer the second question, we propose ***Objective3D***, an object-centric fine-tuning framework for pretrained VLMs built on these observations. It targets three factors that govern adaptation: which views of an asset the model observes, how complementary rendering modalities interact, and how often each task contributes supervision. *View augmentation* samples different subsets of pre-rendered views during training while keeping all modalities aligned; *intra-view attention* exchanges information between aligned modalities, such as RGB and surface normals, at corresponding image patches; and a *fixed multitask data mixture* sets each task's share of the training examples. The attention module is inserted after the visual merger and initialized as an identity mapping, so the backbone's pretrained representations and autoregressive answer interface are preserved. Applied to Qwen3-VL-8B, Objective3D raises the average ObjectArena3D score from 43.0 to 70.3, surpassing the strongest proprietary models, GPT-5.6 Sol and Gemini-3.1 Pro (both 65.9). The gains are largest on view correspondence, 2D–3D projection, and multi-view part grounding, which improve by 56.0, 68.6, and 51.4 percentage points, respectively, whereas orientation analysis remains the weakest metric at 56.4% (Table 2). In summary, our contributions are threefold:

- We formulate object-level 3D evaluation as a *property-resolved* problem and introduce ***ObjectArena3D***, which evaluates 20 tasks across four capability families, together with randomized-orientation and randomized-color settings that test robustness to changes in pose and appearance.
- We design a unified evaluation protocol of general-purpose proprietary and open-source VLMs and specialized 3D object models, revealing weak camera awareness, a heavy reliance on appearance cues for part grounding, and a wide gap between specialized 3D models and general-purpose VLMs.
- We present ***Objective3D***, an object-centric fine-tuning framework that combines view augmentation, intra-view attention across aligned rendering modalities, and a fixed multitask data mixture, enabling an 8B open-source VLM to surpass leading proprietary VLMs on ObjectArena3D.

2 OBJECTARENA3D BENCHMARK

ObjectArena3D evaluates fine-grained understanding of 3D objects. The revised standard suite contains 16,319 questions in 20 tasks across four capability families (Figure 2): semantic and attribute understanding (5 tasks, 4,398 questions), camera awareness (7 tasks, 4,351 questions), part-structure understanding (3 tasks, 2,639 questions), and surface understanding (5 tasks, 4,931 questions). Each question is posed over rendered observations of one asset and scored deterministically against its reference answer, without a language-model judge. Two randomized conditions further test robustness to changes in orientation and appearance (Section 2.3). Detail task definitions, statistics, and metrics are detailed in Figure 1 and Appendx C.

2.1 TASK DESIGN

Figure 2 illustrates representative tasks from each of the four families, which we describe below.

Semantic and attribute understanding. This family evaluates whether recognition extends from an object's category to the characteristics of the particular instance. Category recognition assigns a two-level label from 11 top-level categories and 51 subcategories, and serves as a reference for semantic competence. In detail, modeling-style recognition identifies how an asset was created or acquired, choosing among seven labels such as scanned, procedural, voxel, and AI-generated. Attribute analysis comprises three tasks: symmetry distinguishes reflective from rotational symmetry and strict from approximate reflection; posture concerns the articulation of a subject's body rather than the rigid orientation of the asset; and object counting separates independent objects from their constituent parts and from repeated observations of the same object across views.

Camera awareness. This family evaluates whether a model can relate image observations to the underlying 3D object and the cameras that observe it. Orientation analysis estimates an object's pose from a single render without a canonical reference image: the model must infer the object's intrinsic front or top and then report its facing direction, its full three-axis pose, or the direction in which its top points. View correspondence relates two views of the same object: point correspondence locates a marked surface point in the second view or determines that it is not visible, and relative rotation estimates the change in camera azimuth between the views. 2D–3D projection links image locations to coordinates in an object-centered frame, in the forward direction from a 3D point to its image location and in the inverse direction from a point marked in two views to its 3D coordinates.

Part-structure understanding. This family evaluates whether a model can identify and localize an object's components. Part counting asks for the number of distinct instances of a named component across three views, in which the same instance may appear more than once. In part grounding, a query refers to a component by its name, its appearance, or its function, and the model returns a bounding box. Single-view grounding requires a box in one image, whereas multi-view grounding requires boxes for the same component in each of three views, tying the query to a physical part rather than to a region of one image.

Surface understanding. This family evaluates how well a model interprets an asset's surface appearance and material representation. Three texture tasks grade texture richness, detect and classify texture abnormalities such as untextured surfaces, chaotic mappings, and local texture loss, and identify lighting or shadows baked into the base-color texture. Material identity recognition asks for the dominant materials of the asset, and metallic/roughness assessment asks whether the corresponding parameter maps are plausible for the depicted materials. These tasks require distinguishing an asset's intended surface properties from baked-in rendering effects and authoring defects.

2.2 ANNOTATION AND VALIDATION

References come from human annotation, model-assisted labeling with human refinement, and known rendering geometry, and each family is validated according to how its references are produced. For semantic categories, we sample both assets on which Gemma-4-31B-it (Team, 2026) and Gemini-3.1 Pro (Google DeepMind, 2026) agree and assets on which they disagree; these predictions guide sampling only, and annotators then refine the labels and remove semantically ambiguous assets. Modeling-style labels are initialized from asset provenance, such as scanned or generated sources, and refined by annotators; symmetry, posture, and object counts are annotated manually.

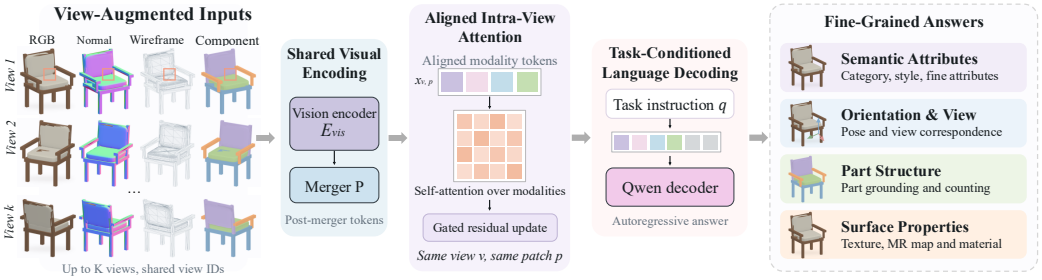


Figure 3: **Overview of Objective3D.** A shared encoder processes sampled views, and intra-view attention fuses aligned modalities. The language model generates task-conditioned answers.

Camera targets are computed from rendering parameters and known geometry, with orientation targets anchored to manually annotated canonical fronts. Candidate questions vary the viewpoint configuration and localization granularity, and are balanced over selected pose bins and answer options. Human reviewers screen each candidate for interpretability and for consistency between the renders and the reference answer, retaining 4,351 of 5,000 candidates (Table A3).

Part-counting targets are derived from PartNeXt (Wang et al., 2025b) and 3DCoMPaT200 (Ahmed et al., 2024) annotations. Grounding queries use either annotated part names or descriptions of a part's appearance or function generated by Gemini-3.1 Pro (Google DeepMind, 2026). For multi-view grounding, we retain the three views in which the target part has the largest visible area. Reviewers verify that the target is visible, the query is unambiguous, and each box localizes the target; failing examples are discarded.

Surface annotations are primarily human judgments under the protocol in Appendix B.4. To broaden the coverage of texture defects, we synthesize a subset of texture-abnormality examples by replacing materials, randomly perturbing texture mappings and adjusting lighting, which produces localized texture removal, mismatched textures and overexposed samples (Appendix B.5).

2.3 Robustness Evaluation

Strong performance under standard viewing conditions does not guarantee robustness to changes in orientation or appearance. We therefore re-render assets under two randomized conditions that test reliance on canonical poses and on familiar surface cues.

Randomized orientation. Assets are rotated away from their canonical upright configuration by nonzero multiples of $45°$ about one or two horizontal axes, optionally combined with a rotation about the vertical axis, and the task inputs are re-rendered. This breaks the usual alignment between an object's intrinsic up direction and the rendering frame. The condition contains 8,617 questions from the semantic and surface tasks whose references remain valid under rigid rotation.

Randomized color. We replace the original appearance with random solid colors while preserving mesh geometry; for semantic and structure inputs, roughness and metallicity are also randomized. The condition thus removes texture cues and alters material appearance, testing how much semantic, spatial, and structural judgments depend on them. It contains 10,675 questions from the semantic, camera, and part-structure families; surface tasks, whose references describe the original appearance, are excluded. Grounding boxes are recomputed for the re-rendered views, which may differ from the standard ones.

Both conditions reuse the standard task-specific metrics, and robustness is measured as the score change on matched question subsets, with each score computed against the reference for its own rendering. Task coverage, rendering details, and the matching protocol are provided in Appendix C.8.

3 Object-Centric Fine-Tuning

We introduce Objective3D, a supervised fine-tuning framework that adapts a pretrained VLM to fine-grained object understanding (Figure 3). It includes three components: *view augmentation* varies input views for model observation (Section 3.1); *intra-view attention* exchanges evidence between aligned rendering modalities (Section 3.2); and a *fixed multitask data mixture* sets how often each

task contributes supervision (Section 3.3). These components are designed to improve viewpoint awareness and to give structural reasoning access to geometric evidence beyond appearance.

Setup. Let $\mathcal{A}$ denote a 3D asset, q a task instruction, and y the target answer. Each asset has a set of pre-rendered views $\mathcal{V}_\mathcal{A}$ and a set of rendering modalities $\mathcal{M}_\mathcal{A}$, such as RGB and surface normals; I_v^m denotes the image of view v in modality m. The backbone embeds each image with its shared vision encoder E_{vis} and merger P as $X_v^m = P(E_{\text{vis}}(I_v^m)) \in \mathbb{R}^{L_v \times d}$, where L_v is the number of merged patches and d is the hidden size of the language model. All modalities of a view share one camera and image grid, so the p-th tokens $x_{v,p}^m$ of all modalities describe the same image region.

3.1 View Augmentation

View augmentation varies the observations of an asset across training examples while keeping the number of input images bounded. For each example, we select a subset $\mathcal{V}_{\mathcal{A},s} \subseteq \mathcal{V}_\mathcal{A}$ of at most K views using a seeded rule keyed by a sampling key s, so that the selection is reproducible yet varies with s. Assets with at most K views keep all of them. All modalities use the same selected views, which keep their original order. The model input is then

$$\mathcal{I}_{\mathcal{A},s} = \{I_v^m : v \in \mathcal{V}_{\mathcal{A},s},\ m \in \mathcal{M}_\mathcal{A}\}, \qquad |\mathcal{V}_{\mathcal{A},s}| \le K. \tag{1}$$

Because every modality is sampled at the same views, the augmentation changes which parts of the asset are observed without breaking cross-modal correspondence. The selected images and the instruction q form a single instruction-tuning example.

3.2 Intra-View Attention

Aligned tokens from different modalities carry complementary evidence about the same region: for example, RGB renderings convey color and texture, whereas surface normals encode local surface orientation independently of appearance. Because the backbone encodes each image independently, such tokens interact only inside the language model, which must first recover their correspondence. We therefore insert an attention module after the merger that lets them interact directly. For view v and patch p, we stack the tokens of the $M = |\mathcal{M}_\mathcal{A}|$ modalities as $X_{v,p} = [x_{v,p}^1; \dots; x_{v,p}^M] \in \mathbb{R}^{M \times d}$ and update them as

$$\begin{aligned} N_{v,p}^m &= \mathrm{LN}(x_{v,p}^m) + e_m, \\ A_{v,p} &= \mathrm{MHA}(N_{v,p}), \\ Y_{v,p} &= X_{v,p} + \tanh(\gamma_{\text{iv}})(A_{v,p} - X_{v,p}), \end{aligned} \tag{2}$$

where LN is layer normalization, e_m is a learned embedding of modality m, MHA is multi-head self-attention over the M tokens, and γ_{iv} is a learned scalar gate. Its initialization and training settings are given in Section A.1.

Attention runs only along the modality axis: each view–patch group is processed independently, and groups with a single modality are left unchanged. The updated tokens return to their original sequence positions before being combined with the text embeddings, so the number and order of image tokens are preserved. The module acts only on the final post-merger features, leaving the backbone's existing multi-scale visual pathways intact.

3.3 Data and Training

We train Objective3D in two stages from Qwen3-VL-8B-Instruct. Mid-training uses a 300K-example mixture emphasizing part structure, followed by supervised fine-tuning on a 416K-example mixture emphasizing camera awareness and surface properties. The second stage initializes from the mid-training checkpoint and enables view augmentation and intra-view attention. Both mixtures cover all four capability families, with task proportions fixed within each stage (Figures A1 and A2).

Both stages use LoRA adaptation and minimize autoregressive cross-entropy over answer tokens:

$$\mathcal{L}_{\text{SFT}} = -\frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \log p_\theta(y_{i,j} \mid y_{i,<j}, \mathcal{I}_{\mathcal{A}_i,s_i}, q_i), \tag{3}$$

where Ω contains answer-token positions in the minibatch, excluding prompt and image tokens, and θ denotes the trainable parameters. Stage-specific schedules and implementation details are provided in Section A.1.

Table 2: **Evaluation on ObjectArena3D**. Scores are percentages. Attribute, camera, and texture aggregates use question-count-weighted primary scores. Avg. averages the displayed columns over available measurements. Ranks compare fully updated rows, and boldface compares updated measurements.

Methods	Rank	Avg.	*Category L1*	*Category L2*	*Style*	*Attr. Analysis*	*View Corr.*	*Orient. Analysis*	*2D–3D Proj.*	*Part Ground*	*Part Ground MV*	*Counting*	*Texture*	*MR Quality*	*Mat. Pred.*
			Semantic Perception				Camera Awareness			Structure Perception			Texture and Materials		
Proprietary Models															
Gemini-3.1 Pro (Google DeepMind, 2026)	–	65.90	83.86	68.76	73.17	64.67	32.33	52.37	48.04	82.96	74.02	**73.67**	66.17	72.92	63.72
GPT-5.6 Sol (OpenAI, 2026)	–	65.92	79.13	62.93	65.49	60.47	52.74	47.21	60.62	**85.46**	**76.76**	73.26	52.74	78.65	61.44
Claude-5.0 Opus (Anthropic, 2026)	–	65.28	83.24	65.85	69.31	62.64	33.72	44.75	55.39	83.31	70.92	70.03	**69.32**	**75.93**	**64.25**
Open-source Models															
Qwen3.8-Flash-Next (Qwen Team, 2026b)	2	53.83	77.14	56.80	54.05	52.44	24.73	36.92	37.36	58.05	55.54	67.33	49.59	68.19	61.71
Qwen3.6-35B-A3B (Qwen Team, 2026a)	4	46.18	74.65	49.83	45.19	45.76	15.87	28.06	29.00	64.84	10.97	64.41	54.43	59.03	58.25
Qwen3-VL-30B-A3B (Bai et al., 2025)	5	46.12	66.35	40.62	48.85	46.51	13.53	22.50	24.57	65.08	50.18	62.85	49.94	53.15	55.42
Qwen3-VL-8B (Bai et al., 2025)	6	42.95	70.29	41.41	52.21	42.09	13.97	20.59	19.64	58.88	19.07	60.04	45.41	59.74	54.97
Intern-VL3.5-30B-A3B (Zhu et al., 2025)	–	35.43	43.43	25.22	44.07	36.95	14.89	25.10	21.19	7.87	–	50.47	60.48	50.86	44.59
Gemma-4-26B-A4B-it (Team, 2026)	3	50.94	77.47	56.80	54.66	47.26	17.63	32.03	23.06	79.74	43.86	61.71	47.32	58.17	62.53
Specialist Models															
PointLLM-R-7B (Chen et al., 2026)	–	24.15	44.98	–	23.64	22.56	–	8.39	–	–	–	–	33.80	–	11.54
ShapeLLM-Omni (Ye et al., 2025)	–	27.86	33.58	–	13.65	42.30	–	–	–	–	–	–	43.54	30.36	3.70
EVA01 (Yang et al., 2026)	–	28.48	22.61	–	32.98	32.59	13.46	11.40	14.10	–	–	–	43.78	52.58	32.82
SFT Models															
Objective3D (Qwen3-VL-8B)	1	70.26	**81.20**	**61.62**	**82.60**	**60.69**	**70.01**	**56.35**	**88.22**	79.98	70.44	68.57	58.48	74.36	60.89

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We implement Objective3D using the ms-swift training framework[1] and evaluate the open-source VLM baselines with the same codebase for a fair comparison. We fine-tune Qwen3-VL-8B-Instruct (Bai et al., 2025) using LoRA with rank 32, scaling factor 2, and dropout 0.05. The optimizer is fused AdamW (Loshchilov & Hutter, 2019) with a cosine learning-rate schedule and stage-dependent learning rates from 10^{-5} to 10^{-4}. All training is conducted on Nvidia A800 GPUs, with effective batch sizes of 128 for mid-training and 256 for supervised fine-tuning. Training details are provided in Appendix A.1; evaluation protocols and metrics are described in Appendix C.

4.2 MAIN RESULTS

Overall performance. As shown in Table 2, (i) proprietary models lead the baselines at 65.28–65.92% on average, yet none dominates across capability families; (ii) open-source models trail by about 12 points, led by Qwen3.8-Flash-Next at 53.83%; and (iii) specialized 3D object models, despite consuming 3D input directly, score below Qwen3-VL-8B on nearly every metric they support.

Capability-wise analysis. (i) Camera awareness is the weakest family for every general-purpose baseline: the proprietary models reach 79.13–83.86% on level-1 category recognition, but no baseline exceeds 53% on view correspondence or orientation analysis. (ii) On part structure, the gap between the best proprietary and open-source models widens from 5.72 points on single-view grounding to 21.22 points when a correct box is required in all three views. (iii) Surface understanding tracks overall capability only weakly: two open-source models fall within the proprietary range of 61.44–64.25% on material prediction.

Effect of object-centric fine-tuning. Objective3D ranks first at 70.26%, 27.31 points above its Qwen3-VL-8B backbone and 4.34 points above the strongest proprietary model. It improves on all 13 metrics and outperforms every baseline on all three camera-awareness tasks, raising 2D–3D projection from 19.64% to 88.22%. This lead comes mainly from camera awareness: it still trails the best proprietary model on 9 of 13 metrics, and orientation analysis is its weakest metric (56.35%).

Robustness. Objective3D achieves the highest average among the updated models in Table 3 (57.61%), 3.60 points above the strongest proprietary baseline. Under random texture/color, camera-awareness scores change by at most 6.43 points across these models (2.62 for Objective3D), whereas single-view grounding drops by 41.37–51.38 points. For Objective3D, single-view and all-view

[1] https://github.com/modelscope/ms-swift

Table 3: **Robustness of camera, structure, and texture understanding.** Camera and structure are evaluated under random texture/color; texture and material tasks are evaluated under random orientation. Scores are percentages. Avg. is the arithmetic mean of available displayed metric columns across the two conditions. Boldface compares updated models; ◇ marks retained earlier results, and – indicates unavailable measurements.

		Random texture/color						*Random orientation*			
		Camera Awareness			**Structure**			**Texture**			
Model	Avg.	View corr.	Orient.	2D–3D proj.	Ground. SV	Ground. MV all	Part count	Texture quality	MR quality	Material Top-1	Material Top-3 F1
Proprietary Models											
Gemini-3.1 Pro	53.39	29.85	52.10	48.64	35.57	30.93	**72.74**	60.81	73.21	64.50	**65.60**
GPT-5.6 Sol	54.01	46.31	43.80	58.81	35.33	31.91	69.93	48.81	**79.23**	64.04	62.00
Claude-5.0 Opus	53.89	32.63	40.88	56.90	**35.82**	**32.64**	69.51	**65.87**	76.50	**65.69**	62.41
Open-source Models											
Qwen3.8-27B◇	58.51	38.92	42.93	65.26	32.89	–	70.24	73.96	76.36	63.39	62.64
Qwen3.6-35B-A3B	38.75	15.87	27.29	29.91	23.47	5.26	55.67	51.70	59.89	58.90	59.52
Qwen3-VL-30B-A3B	36.90	13.09	23.36	24.07	21.03	15.65	55.78	43.44	55.87	57.52	59.23
Qwen3-VL-8B	35.07	13.75	22.02	22.26	15.89	11.49	49.64	43.88	59.17	55.41	57.23
Intern-VL3.5-30B-A3B◇	34.83	13.90	23.64	21.95	3.30	–	46.83	63.09	50.14	44.22	46.43
Gemma-4-26B-A4B-it	37.65	17.04	29.98	24.47	28.36	7.70	52.86	36.64	55.01	62.66	61.77
Objective3D	**57.61**	**68.76**	**53.89**	**90.84**	33.50	26.41	53.80	56.59	70.34	61.83	60.10

Camera and texture-quality aggregates use question-count-weighted primary scores. For ◇ rows, Avg. uses the retained scores rounded to two decimal places. SV grounding reports IoU $\geq$ 0.5 accuracy; MV all requires IoU $\geq$ 0.5 in every reference view. MR quality is metallic/roughness quality-label accuracy; material Top-3 reports set F1.

Table 4: **Semantic robustness on tasks shared by both perturbations.** Each cell reports randomized accuracy (%) followed by its change from standard accuracy in parentheses (percentage points). Changes use matched question IDs within each task and condition. Boldface marks the best randomized score in each row.

Condition	Gemini-3.1 Pro	GPT-5.6 Sol	Claude-5.0 Opus	Qwen3.6-35B-A3B	Qwen3-VL-30B-A3B	Qwen3-VL-8B	Gemma-4-26B-A4B-it	Objective3D
Category: level 1								
Random color	**81.64** (−2.25)	78.18 (−1.04)	79.68 (−3.54)	70.02 (−4.66)	64.99 (−1.33)	68.69 (−1.67)	74.56 (−2.91)	78.14 (−3.04)
Random orientation	**78.27** (−5.62)	73.40 (−5.83)	75.40 (−7.83)	66.53 (−8.16)	59.03 (−7.29)	61.62 (−8.74)	68.86 (−8.62)	70.19 (−10.99)
Symmetry								
Random color	**66.20** (−0.70)	63.87 (+3.26)	60.61 (+0.23)	55.94 (−0.47)	52.91 (−2.10)	42.66 (−11.19)	61.77 (+0.23)	60.84 (−1.40)
Random orientation	**59.67** (−7.23)	55.94 (−4.66)	57.34 (−3.03)	47.55 (−8.86)	51.05 (−3.96)	39.39 (−14.45)	54.78 (−6.76)	51.52 (−10.72)
Object counting								
Random color	**60.10** (−2.99)	53.87 (−3.49)	57.61 (−5.49)	21.45 (−0.50)	23.94 (−2.00)	17.71 (−1.75)	17.71 (−1.50)	55.11 (−0.75)
Random orientation	**56.22** (−6.97)	49.75 (−7.71)	47.76 (−15.42)	15.92 (−5.97)	20.90 (−4.98)	14.43 (−4.98)	15.92 (−3.23)	45.52 (−10.45)

Parentheses show $\Delta = \text{Random} - \text{Std.}$, computed before rounding. The same question subset is used across models within each task and perturbation. Object counting uses exact accuracy.

grounding fall from 79.98% and 70.44% to 33.50% and 26.41%, respectively. Random orientation is more disruptive than random color for all eight evaluated models on category recognition, symmetry, and object counting (Table 4). Level-1 category accuracy drops by 5.62–8.74 points across the seven baselines and 10.99 points for Objective3D, compared with 1.04–4.66 and 3.04 points under random color, respectively. Appearance-independent part understanding and orientation-invariant recognition therefore remain open challenges, including for our model.

4.3 Discussion and Exploration

Ablation studies. We examine the contributions of data and visual adaptation on Qwen3-VL-8B, using 120K training examples and a fixed training budget for the core ablations. The plain fine-tuning baseline achieves a task-macro score of 58.24%, while the this configuration combining view augmentation and intra-view attention reaches 59.63%, a gain of 1.39 percentage points. This joint improvement supports combining diverse views with interaction across aligned visual modalities, although it does not establish their individual contributions. In a further experiment, adding 10K training examples with component and wireframe renderings improves the score by another 0.73 points. These renderings expose part boundaries and geometric structure, providing complementary

cues to RGB appearance. The additional gain suggests that expanding visual supervision is useful, but it reflects both rendering diversity and increased data volume.

Generalization beyond ObjectArena3D. To assess whether object-centric fine-tuning preserves general-purpose 2D understanding, we additionally evaluate Objective3D and its Qwen3-VL-8B backbone on MMBench (Liu et al., 2024c). Objective3D achieves 88.35%, compared with 88.58% for the backbone, a decrease of only 0.23 percentage points. This small gap suggests that Objective3D largely retains the backbone's general-purpose 2D capabilities on this benchmark.

5 Related Work

3D object understanding with MLLMs. Following the success of 2D MLLMs (Liu et al., 2023; Bai et al., 2023; Liu et al., 2024a;b; Clark et al., 2026), recent work increasingly injects 3D representations into LLMs. At the scene level, LLM-driven generalists replace task-specific 3D vision–language models, unifying captioning, question answering, grounding, and planning in indoor environments (Hong et al., 2023; Wang et al., 2023; Chen et al., 2024; Huang et al., 2024; Zhu et al., 2024; Mei et al., 2025; 2026; Huang et al., 2026a). At the object level, point–language models align 3D-specific tokenizers with LLMs for open-ended shape understanding (Xu et al., 2024; Qi et al., 2024b;a; 2025); follow-ups explore encoder-free designs, alternative tokenizations, and explicit reasoning traces (Tang et al., 2025; Paul et al., 2026; Chen et al., 2026; Zhang et al., 2026), while native and unified 3D MLLMs further connect object understanding with generation (Ye et al., 2025; Wang et al., 2024; Ye et al., 2026a;b; Huang et al., 2026b; Yang et al., 2026). Yet most models rely on sparse point inputs and caption-style supervision and thus mainly capture holistic object semantics. Part-aware and locally aligned variants (Wang et al., 2025a; Hadgi et al., 2026; Huang et al., 2026c) move toward finer granularity, but fine-grained asset properties, such as appearance and geometric attributes, materials, orientation, and part structure, remain underexplored. We target this gap with a comprehensive benchmark and a multi-representation MLLM.

3D object perception, understanding, and benchmarking. Object-centric 3D vision–language research spans 3D object captioning (Luo et al., 2023; 2024; Yeshwanth et al., 2025), text–3D alignment and pre-training (Sahoo et al., 2025; Su et al., 2026), orientation estimation and canonicalization (Wang et al., 2025c; Lu et al., 2025; Jin et al., 2025), and part segmentation (Liu et al., 2025; Wang et al., 2025b), while 3D object benchmarks provide increasingly rich annotations (Fu et al., 2021; Li et al., 2022; Slim et al., 2025; Wu et al., 2023; Deitke et al., 2023; Ahmed et al., 2024; Lin et al., 2025). Evaluation, however, has not kept pace: 3D MLLMs are still often assessed on scene-level grounding and QA benchmarks with per-case accuracy. Recent diagnostic studies question this setup: object-centric chain-based evaluation exposes poor per-object coherence (Huang et al., 2025), and blind LLMs can match 3D MLLMs on standard 3D QA by exploiting language priors (Ma et al., 2026), mirroring spatial-cognition probes of 2D MLLMs (Yang et al., 2025; Jia et al., 2025). Existing object-level evaluation also remains narrow, often limited to coarse captioning and classification (Xu et al., 2024). These gaps motivate a benchmark for fine-grained understanding of individual 3D assets that measures attributes, materials, orientation and camera awareness, and structural parsing.

6 Conclusion

We introduced ObjectArena3D, a property-resolved benchmark that evaluates fine-grained 3D asset understanding with 20 tasks in four capability families on shared assets, together with randomized-orientation and randomized-color conditions that test reliance on canonical poses and appearance cues. Our evaluation shows that recognizing an asset's category does not imply understanding its properties: camera awareness is the weakest family, part grounding relies heavily on appearance, and specialized 3D object models trail general-purpose VLMs. We further proposed Objective3D, an object-centric fine-tuning framework that combines view augmentation, intra-view attention across aligned rendering modalities, and a fixed multitask data mixture, enabling an 8B open-source VLM to surpass leading proprietary models on ObjectArena3D. Orientation analysis and appearance-independent part grounding nevertheless remain open challenges, including for our model (Section F). We hope ObjectArena3D and Objective3D will help move 3D object understanding beyond recognition.

AI USE STATEMENT

We used Gemini-3.1 Pro to generate part-grounding queries and Gemma-4-31B-it and Gemini-3.1 Pro to guide category-candidate sampling (Section 2.2); all resulting labels were verified by humans. We also used AI tools for polishing writing, writing visualization code, and other auxiliary tasks, but not for research ideation, experiment design, or result interpretation. The authors reviewed all AI-assisted content and take full responsibility for it.

ETHICS STATEMENT

This work involves no human subjects or personal data. ObjectArena3D is built from existing 3D asset collections used under their licenses; broader impacts are discussed in Section G.

REPRODUCIBILITY STATEMENT

Benchmark construction and evaluation protocols are detailed in Section 2 and Section C, and Objective3D and its training setup in Section 3, Section 4.1, and Section A. All answers are scored deterministically without an LLM judge. We will release the benchmark, code, and model checkpoints.

REFERENCES

Mahmoud Ahmed, Xiang Li, Arpit Prajapati, and Mohamed Elhoseiny. 3dcompat200: language-grounded compositional understanding of parts and materials of 3d shapes. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pp. 135770–135782, 2024.

Anthropic. Claude Opus 4.7 system card. `https://www.anthropic.com/claude-opus-4-7-system-card`, April 2026. Dated 16 April 2026.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Chaoqi Chen, Qile Xu, Wenjun Zhou, and Hui Huang. Pointllm-r: Enhancing 3d point cloud reasoning via chain-of-thought. *arXiv preprint arXiv:2605.22013*, 2026.

Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26428–26438, 2024.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28652–28668, 2026.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36:35799–35813, 2023.

Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129 (12):3313–3337, 2021.

Google DeepMind. Gemini 3.1 Pro model card. `https://deepmind.google/models/model-cards/gemini-3-1-pro/`, February 2026. Published 19 February 2026.

Souhail Hadgi, Bingchen Gong, Ramana Sundararaman, Emery Pierson, Lei Li, Peter Wonka, and Maks Ovsjanikov. Patchalign3d: Local feature alignment for dense 3d shape understanding. *arXiv preprint arXiv:2601.02457*, 2026.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023.

Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. *Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada*, 2024.

Haifeng Huang, Yilun Chen, Zehan Wang, Jiangmiao Pang, and Zhou Zhao. Chat-scene++: Exploiting context-rich object identification for 3d llm. *arXiv preprint arXiv:2603.27507*, 2026a.

Jiangyong Huang, Baoxiong Jia, Yan Wang, Ziyu Zhu, Xiongkun Linghu, Qing Li, Song-Chun Zhu, and Siyuan Huang. Unveiling the mist over 3d vision-language understanding: Object-centric evaluation with chain-of-analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24570–24581, 2025.

Junming Huang, Chi Wang, Letian Li, Guangkai Xu, Donglin Huang, Hao Chen, Qiang Dai, and Weiwei Xu. Cg-mllm: Captioning and generating 3d content via multi-modal large language models. *arXiv preprint arXiv:2601.21798*, 2026b.

Peng Huang, Yifeng Chen, Zeyu Zhang, and Hao Tang. Unimesh: Unifying 3d mesh understanding and generation. *arXiv preprint arXiv:2604.17472*, 2026c.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. *arXiv preprint arXiv:2506.03135*, 2025.

Li Jin, Yujie Wang, Wenzheng Chen, Qiyu Dai, Qingzhe Gao, Xueying Qin, and Baoquan Chen. One-shot 3d object canonicalization based on geometric and semantic consistency. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16850–16859, 2025.

Yuchen Li, Ujjwal Upadhyay, Habib Slim, Ahmed Abdelreheem, Arpit Prajapati, Suhail Pothigara, Peter Wonka, and Mohamed Elhoseiny. 3d compat: Composition of materials on parts of 3d things. In *European conference on computer vision*, pp. 110–127. Springer, 2022.

Chendi Lin, Heshan Liu, Qunshu Lin, Zachary Bright, Shitao Tang, Yihui He, Minghao Liu, Ling Zhu, and Cindy Le. Objaverse++: Curated 3d object dataset with quality annotations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 6813–6822, October 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Minghua Liu, Mikaela Angelina Uy, Donglai Xiang, Hao Su, Sanja Fidler, Nicholas Sharp, and Jun Gao. Partfield: Learning 3d feature fields for part segmentation and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9704–9715, 2025.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, 2024c. URL https://arxiv.org/abs/2307.06281.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Yichong Lu, Yuzhuo Tian, Zijin Jiang, Yikun Zhao, Yuanbo Yang, Hao Ouyang, Haoji Hu, Huimin Yu, Yujun Shen, and Yiyi Liao. Orientation matters: Making 3d generative models orientation-aligned. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=YVOCwQF9k2.

Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pre-trained models. *Advances in Neural Information Processing Systems*, 36:75307–75337, 2023.

Tiange Luo, Justin Johnson, and Honglak Lee. View selection for 3d captioning via diffusion ranking. In *European Conference on Computer Vision*, pp. 180–197. Springer, 2024.

Xianzheng Ma, Tao Sun, Shuai Chen, Yash Bhalgat, Jindong Gu, Angel X Chang, Iro Armeni, Iro Laina, Songyou Peng, and Victor Adrian Prisacariu. Do 3d large language models really understand 3d spatial relationships? *arXiv preprint arXiv:2603.23523*, 2026.

Guofeng Mei, Wei Lin, Luigi Riz, Yujiao Wu, Fabio Poiesi, and Yiming Wang. Perla: Perceptive 3d language assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14369–14379, 2025.

Guofeng Mei, Wei Lin, Luigi Riz, Yujiao Wu, Yiming Wang, and Fabio Poiesi. Efficient encoder-free fourier-based 3d large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23785–23794, 2026.

OpenAI. GPT-5.6 system card. https://deploymentsafety.openai.com/gpt-5-6/gpt-5-6.pdf, July 2026. Dated 9 July 2026.

Sneha Paul, Zachary Patterson, and Nizar Bouguila. Point cloud as a foreign language for multi-modal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16676–16687, 2026.

Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. Shapellm: Universal 3d object understanding for embodied interaction. In *European Conference on Computer Vision*, pp. 214–238. Springer, 2024a.

Zhangyang Qi, Ye Fang, Zeyi Sun, Xiaoyang Wu, Tong Wu, Jiaqi Wang, Dahua Lin, and Hengshuang Zhao. Gpt4point: A unified framework for point-language understanding and generation. In *CVPR*, 2024b.

Zhangyang Qi, Ye Fang, Zeyi Sun, Xiaoyang Wu, Tong Wu, Jiaqi Wang, Dahua Lin, and Hengshuang Zhao. Gpt4point++: Advancing unified point-language understanding and generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026a. URL https://qwen.ai/blog?id=qwen3.6-27b.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026b. URL https://qwen.ai/blog?id=qwen3.8.

Aadarsh Sahoo, Vansh Tibrewal, and Georgia Gkioxari. Aligning text, images, and 3d structure token-by-token, 2025. URL https://arxiv.org/abs/2506.08002.

Habib Slim, Xiang Li, Yuchen Li, Mahmoud Ahmed, Mohamed Ayman, Ujjwal Upadhyay, Ahmed Abdelreheem, Arpit Prajapati, Suhail Pothigara, Peter Wonka, et al. 3dcompat++: An improved large-scale 3d vision dataset for compositional recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Yuanhao Su, Shaofeng Zhang, Xiaosong Jia, and Qi Fan. Pointalign: Feature-level alignment regularization for 3d vision-language models. In *CVPR*, 2026.

Yiwen Tang, Zoey Guo, Zhuhao Wang, Ray Zhang, Qizhi Chen, Junli Liu, Delin Qu, Zhigang Wang, Dong Wang, Bin Zhao, et al. Exploring the potential of encoder-free architectures in 3d lmms. *arXiv preprint arXiv:2502.09620*, 2025.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Chunshi Wang, Junliang Ye, Yunhan Yang, Yang Li, Zizhuo Lin, Jun Zhu, Zhuo Chen, Yawei Luo, and Chunchao Guo. Part-x-mllm: Part-aware 3d multimodal large language model. *arXiv preprint arXiv:2511.13647*, 2025a.

Penghao Wang, Yiyang He, Xin Lv, Yukai Zhou, Lan Xu, Jingyi Yu, and Jiayuan Gu. Partnext: A next-generation dataset for fine-grained and hierarchical 3d part understanding. In *Neurips 2025 Dataset and Benchmark Track*, 2025b.

Penghao Wang, Siyuan Xie, Hongyu Yan, Xianghui Yang, Jingwei Huang, Chunchao Guo, and Jiayuan Gu. Artllm: Generating articulated assets via 3d llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34281–34291, 2026.

Zehan Wang, Haifeng Huang, Yang Zhao, Ziang Zhang, and Zhou Zhao. Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. *arXiv preprint arXiv:2308.08769*, 2023.

Zehan Wang, Ziang Zhang, Jiayang Xu, Jialei Wang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Orient anything v2: Unifying orientation and rotation understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025c.

Zhengyi Wang, Jonathan Lorraine, Yikai Wang, Hang Su, Jun Zhu, Sanja Fidler, and Xiaohui Zeng. Llama-mesh: Unifying 3d mesh generation with language models, 2024. URL https://arxiv.org/abs/2411.09595.

Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 803–814, 2023.

Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11097–11107, 2020.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *ECCV*, 2024.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025.

Zongyuan Yang, Mingjing Yi, Wanli Ma, Chenzhuo Fan, Bocheng Li, Baolin Liu, Yuke Lou, Yingde Song, Yongping Xiong, Zhengdong Guo, et al. Eva01: Unified native 3d understanding and generation via mixture-of-transformers. *arXiv preprint arXiv:2605.16745*, 2026.

Chongjie Ye, Cheng Cao, Chuanyu Pan, Yiming Hao, Yihao Zhi, Yuanming Hu, and Xiaoguang Han. Omni123: Exploring 3d native foundation models with limited 3d data by unifying text to 2d and 3d generation. *arXiv preprint arXiv:2604.02289*, 2026a.

Junliang Ye, Zhengyi Wang, Ruowen Zhao, Shenghao Xie, and Jun Zhu. Shapellm-omni: A native multimodal llm for 3d generation and understanding. *arXiv preprint arXiv:2506.01853*, 2025.

Junliang Ye, Zehuan Huang, Yansong Qu, Chunshi Wang, Yunhan Yang, Yang Li, Yawei Luo, Zhuo Chen, Sheng Lu, Jun Zhu, et al. Universe3d: Emerging properties of unified multimodal models in 3d understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 613–623, 2026b.

Chandan Yeshwanth, David Rozenberszki, and Angela Dai. Excap3d: Expressive 3d scene understanding via object captioning with varying detail. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21699–21709, 2025.

Dongxu Zhang, Yiding Sun, Pengcheng Li, Yumou Liu, Hongqiang Lin, Haoran Xu, Xiaoxuan Mu, Liang Lin, Wenbiao Yan, Ning Yang, et al. Pointcot: A multi-modal benchmark for explicit 3d geometric reasoning. *arXiv preprint arXiv:2602.23945*, 2026.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Beyond Recognition: Benchmarking and Advancing Fine-Grained Object Understanding

Supplementary Material

CONTENTS

A More Implementation Details **A1**

A.1 Two-Stage Qwen3-VL-8B Training A2

A.2 Qwen3.8-Flash-Next Configuration A2

B Objective3D-Data **A4**

B.1 Detailed Semantic Tasks A5

B.2 Detailed Camera Tasks A6

B.3 Part-structure tasks A7

B.4 Surface tasks and annotation criteria A7

B.5 Synthetic texture anomalies A8

B.6 Annotation provenance and verification scope A9

C Evaluation Protocol **A9**

C.1 Evaluated Models A9

C.2 Response Parsing and Evaluation Scope A10

C.3 Semantic and Attribute Metrics A10

C.4 Camera Metrics A11

C.5 Part-Grounding Metrics A11

C.6 Surface Metrics A12

C.7 Aggregation and Reporting A12

C.8 Standard and randomized observation conditions A13

D Additional Visualization **A14**

E Additional Experiments **A14**

E.1 Fine-Grained Results under Random Texture Color A22

F Limitations and Future Work **A23**

G Broader Impacts **A23**

A MORE IMPLEMENTATION DETAILS

We detail the two-stage Qwen3-VL-8B training configuration, followed by a separate Qwen3.8-Flash-Next configuration that has completed only a one-step implementation check.

A.1 TWO-STAGE QWEN3-VL-8B TRAINING

Training stages and data. Mid-training starts from Qwen3-VL-8B-Instruct and uses the 300K A1 mixture for two epochs. Supervised fine-tuning initializes from its last checkpoint and uses the 416K C mixture for one epoch, restoring model weights without continuing the optimizer or scheduler state. The semantic, camera, part-structure, and surface proportions are 20%, 25%, 50%, and 5% in A1, and 10.82%, 44.23%, 7.21%, and 37.74% in C, respectively. These are fixed mixtures of sampled examples, with 1% held out for validation in each stage; their distributions are shown in Figures A1 and A2.

Shared optimization settings. We use LoRA with rank 32, alpha 64, and dropout 0.05, with linear layers selected subject to the backbone-freezing configuration. Both stages use fused AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$, weight decay 0.05, a cosine learning-rate schedule, and gradient clipping at 1.0. The per-device batch size is two with one gradient-accumulation step; the stage-specific settings are listed in Table A1. Training uses bfloat16, scaled dot-product attention, and gradient checkpointing for both language and visual modules. Packing, padding-free execution, and DeepSpeed are disabled; training and data seeds are 42. Checkpoints are saved every 500 steps, retaining the latest five.

Table A1: **Two-stage training configuration.** Mixture sizes precede the 1% validation split. Visual learning rates apply to trainable parameters in the vision group. Warmup is a fraction of the stage's update steps.

Setting	Mid-training	Supervised fine-tuning
Mixture	A1 (300K)	C (416K)
Initialization	Pretrained backbone	Mid-training checkpoint
Epochs	2	1
Main learning rate	10^{-4}	10^{-5}
Visual learning rate	5×10^{-6}	10^{-4}
Warmup	5%	2%
Data-parallel workers	64	128
Effective global batch	128	256
Evaluation interval	250 steps	500 steps

Visual inputs and attention. Both stages allow 16,384 tokens per example, truncate longer sequences from the right, and cap each image at 262,144 pixels. View augmentation and intra-view attention are not explicitly enabled in the first-stage arguments. The second stage enables view sampling with at most six views and seed 42, without a modality-subsampling cap. Its intra-view module uses eight attention heads, learned embeddings for up to eight modalities, and a configured gate initialization of 0.1; a restored gate weight takes precedence over initialization. The first stage explicitly sets non-reentrant gradient checkpointing, while the second leaves this choice to the runtime default.

A.2 QWEN3.8-FLASH-NEXT CONFIGURATION

Model adaptation and distributed training. The C-mixture configuration uses Megatron with Qwen3.8-Flash-Next, a model with approximately 8.95B parameters and 512 experts. It starts with a new LoRA adapter, using rank 8, alpha 32, and dropout 0.05 on `in_proj`, `out_proj`, `linear_proj`, and `linear_qkv`. The visual backbone and aligner are frozen. The intra-view attention module is fully trainable, with eight heads, learned modality embeddings for up to eight modalities, and a scalar gate initialized to zero. The recorded trainable parameter count is 27.79M (0.3106%): 26.25M for this module and 1.54M for LoRA. The inspected checkpoint contains all eight attention-module tensors. The configuration uses 256 GPUs (32 nodes with eight GPUs each), with tensor, pipeline, expert, expert-tensor, and data parallelism degrees of 2, 2, 4, 1, and 64, respectively. Sequence parallelism and a distributed optimizer with parameter, gradient, and optimizer-state sharding are enabled.

Optimization and planned duration. Adam uses a learning rate of 10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$, weight decay 0.1, and gradient clipping at 1.0. The cosine schedule decays to zero after a 3% warmup (24 steps). The micro-batch size is one; 64 data-parallel workers and eight accumulation steps give a global batch size of 512. The configuration targets one epoch over the 416,000-example SFT data mixture, corresponding to 812 planned updates. This mixture contains 392,627 standard, 5,003 random-color, 5,228 random-gravity, and 13,142 additional-channel examples.

Visual inputs and modality sampling. The sequence limit is 32,768 tokens, and examples exceeding it are discarded rather than truncated. The per-image limit is 262,144 pixels, corresponding to at most 256 merged visual tokens, with up to six views per modality. An example containing RGB, normals, wireframe, and component renderings can therefore expand to 24 images (6,144 visual tokens). The C configuration uniformly samples at most three modalities, reducing this bound to 18 images (4,608 visual tokens); RGB is not forcibly retained. Sampling is deterministic for a fixed seed and sample key, independent of rank and iteration order, and removed images and their prompt placeholders are dropped together. It applies only to training inputs and leaves validation and benchmark images unchanged. The inspected validation inputs contain 12 images for semantic tasks, 10 for texture abnormalities and material identity, nine for metallic/roughness quality, six for baked lighting and texture richness, three for part counting and multi-view grounding, and one or two for multi-view camera tasks.

B OBJECTIVE3D-DATA

Figures A1 and A2 summarize the task allocation and rendering variants of the 300K mid-training mixture and the 416K supervised fine-tuning mixture, respectively.

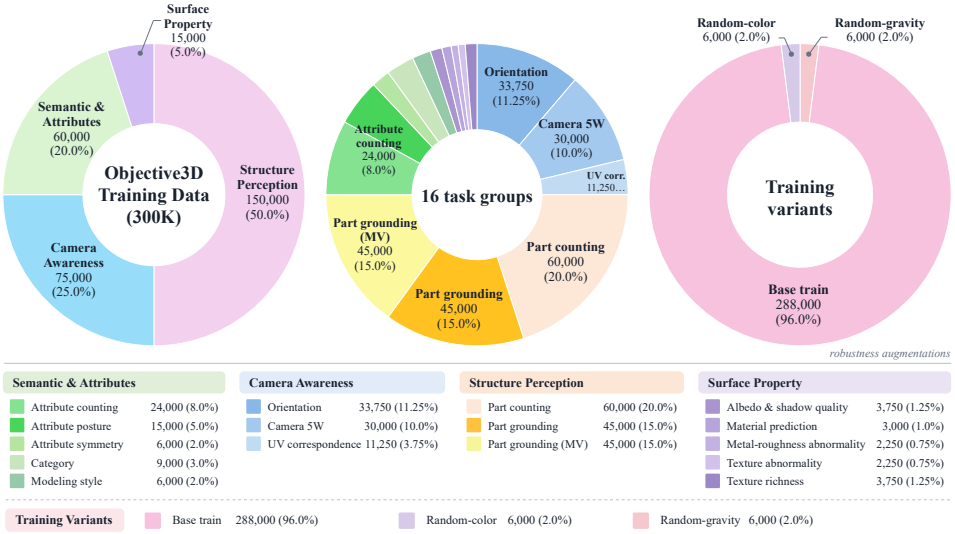


Figure A1: **Mid-training data (300K).** Task-family allocation and rendering variants of the mid-training mixture.

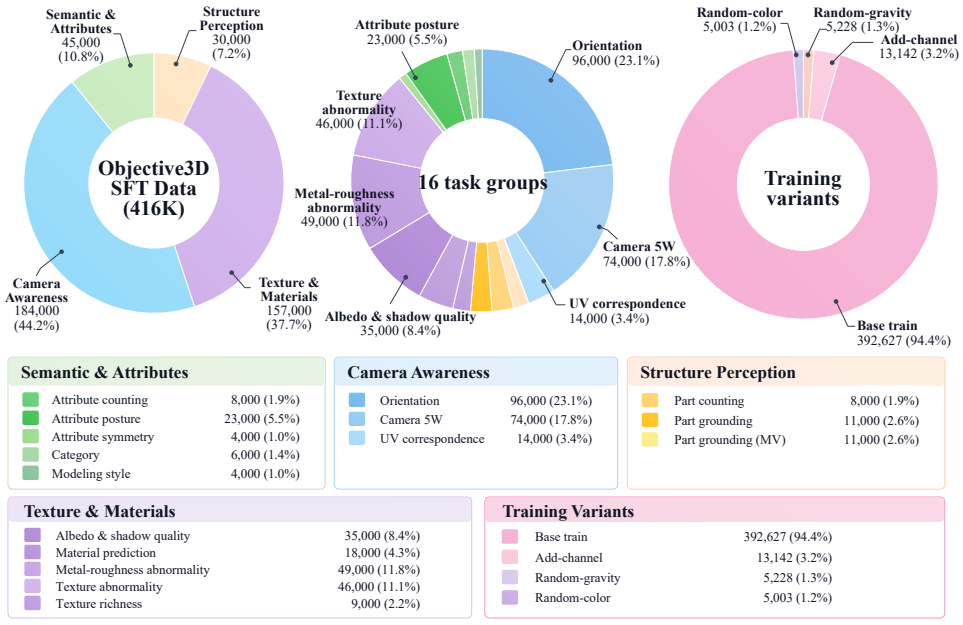


Figure A2: **Supervised fine-tuning data (416K).** Task-family allocation and rendering variants of the SFT mixture.

B.1 Detailed Semantic Tasks

The semantic family contains 4,398 questions. Each question provides six views of the asset, with a corresponding RGB render and normal map for each view (12 images in total). Table A2 summarizes the question formats. Category, style, and attribute judgments refer to the asset as a whole rather than to each rendered image independently.

Table A2: **Semantic question formats.** The two symmetry formats belong to the same symmetry task.

Task	Questions	Answer space
Category	2,410	Hierarchical category pair
Modeling style	655	Four choices from seven labels
Symmetry existence	158	Boolean
Symmetry type	271	Five choices
Posture	499	Six choices
Object counting	405	$1, \ldots, 10$, or $11+$

Category hierarchy. The category vocabulary has 11 substantive top-level categories and 51 substantive subcategories, all represented in the evaluation set. The prompt additionally permits an `other`/`na` fallback, yielding 12 and 52 entries in the respective answer vocabularies; the fallback is not a reference label in the retained set. The top-level categories are prop (738), character (490), built structure (276), animal (177), scene composition (154), robot (132), combat equipment (114), vehicle (109), natural element (95), wearable (87), and scientific asset (38). Each answer identifies both levels. The supplied hierarchy distinguishes complete entities from recognizable components and coherent scenes from bounded collections of independent entities. For example, a building is distinct from an architectural component, while a room layout is distinct from an isolated furniture set. The prompt also requests a visual description and confidence, but the categorical targets are the two hierarchy levels.

Modeling style. The seven labels are hybrid (104), scanned (99), procedural (97), AI-generated (96), voxel (93), hard surface (83), and organic (83). Each question presents four candidate labels and asks which best describes the asset's modeling style. Labels are initialized from asset provenance, including scanned and generated sources, and manually refined. They describe the annotated visual and production characteristics of the asset; a model is not asked to infer the identity of a particular generation system.

Symmetry. The binary format asks whether any global structural symmetry is present and contains 79 positive and 79 negative examples. The five-way format distinguishes strict reflection (79), approximate reflection (79), no symmetry (79), finite rotational symmetry (26), and continuous rotational symmetry (8). Continuous rotational symmetry refers to approximate preservation of the overall geometry under arbitrary rotation about an axis; finite rotational symmetry holds only at discrete angles. Strict reflection requires matching geometry and sufficiently visible base color or texture across a reflection plane, discounting illumination effects. Approximate reflection retains a global bilateral organization but permits local differences in pose, parts, or appearance. Repetition of isolated local elements is insufficient to establish global symmetry.

Posture. The six evaluation labels are neutral (89), locomotion (89), low support (89), lying (89), action pose (83), and sitting (60). Annotation concerns the subject's articulated body configuration rather than camera pose or rigid rotation of the asset. Overlapping cases follow the priority lying, sitting, low support, locomotion, action pose, then neutral. In particular, low support requires substantial lowering relative to the subject's normal stance, whereas locomotion requires a recognizable phase of whole-body travel. Species-specific neutral configurations avoid interpreting ordinary quadrupedal standing or natural ground contact as a human crouching or lying pose. Although the annotation instructions define fallback categories for inapplicable or indeterminate examples, the retained questions use the six concrete labels above.

Object counting. Questions ask for the number of independently identifiable objects or semantically independent subjects represented by an asset. Counts one through ten have 40, 40, 77, 55, 31, 27, 16, 18, 15, and 9 examples, respectively; 77 examples use the overflow label $11+$. Repeated instances and distinct subjects, such as a rider and a horse, count separately. Constituent parts of one semantic object, display stands, shadows, reflections, and objects depicted only in textures do not.

The same instance appearing in multiple views is counted once. The $11+$ label denotes more than ten instances and is not an exact count of eleven.

B.2 Detailed Camera Tasks

The camera family contains 4,351 questions across seven tasks. Orientation questions provide one render without a canonical-front reference image. Correspondence and relative-rotation questions provide two views; forward projection uses one view, and inverse projection uses two. Geometric targets are derived from known rendering configurations, with the canonical fronts used for orientation anchored to human annotations. The task-specific coordinate conventions are given in the prompts.

Canonical orientation. Facing-direction estimation contains 851 questions asking for the azimuth correction that brings the object's front toward the camera. Its 12 values are multiples of 30° in $(-180^\circ, 180^\circ]$, with counterclockwise rotation positive when viewed from above. The three-axis task contains 399 questions and asks for the current pose relative to a front-facing, upright canonical configuration: azimuth belongs to $\{0^\circ, 30^\circ, \ldots, 330^\circ\}$, while elevation and roll each belong to $\{-30^\circ, 0^\circ, 30^\circ\}$. Elevation is positive when more of the object's top is visible, and positive roll is counterclockwise in the image. The prompt provides category-dependent definitions of the front, such as the face of a character, the heading end of a vehicle, or the main entrance of a building. A fixed compass indicates the coordinate convention but does not encode the object's pose.

Top-direction estimation. The 741 questions depict tilted or inverted objects and ask which way their intrinsic top points. The classes are forward (toward the camera; 141), back (away from the camera; 107), left (164), right (167), and inverted (162). The object's horizontal heading is irrelevant to this judgment. The model must first identify the surface or end that normally faces upward, such as a bottle opening, a car roof, or a character's head crown, and then determine its observed direction.

Cross-view point correspondence. The 908 questions mark a physical surface point in the source view and ask whether it remains visible in a target view. If visible, its target location is expressed as an integer grid cell (u, v), with the origin at the top left, u increasing rightward, and v downward. These are screen cells rather than texture UV coordinates. Grid sizes 4×4, 6×6, 8×8, and 10×10 have 222, 225, 233, and 228 examples, respectively. There are 782 visible and 126 non-visible targets; self-occlusion, rotation to the back, and leaving the frame all count as non-visibility. The grid variation changes the required localization precision while retaining the same correspondence formulation.

Relative camera rotation. The 459 questions show a fixed object from two camera positions and ask for the camera's azimuth change, defined as target minus source azimuth. The eight answer bins are $\{-135^\circ, -90^\circ, -45^\circ, 0^\circ, 45^\circ, 90^\circ, 135^\circ, 180^\circ\}$, with 56–59 examples per bin. Elevation changes of $-60^\circ, -30^\circ, 0^\circ, 30^\circ, 60^\circ$ occur in 61, 92, 111, 138, and 57 examples, respectively. This variation requires separating the requested horizontal rotation from a change in elevation and from the object's apparent rotation in the image.

Projection between image and object coordinates. Both projection tasks use a right-handed, Z-up object frame centered at the asset's bounding-sphere center. Colored arrows indicate the projected positive coordinate axes. Forward projection (495 questions) provides an integer 3D query point and one orthographic render containing four marked surface candidates. The model selects the candidate closest to the query's projection; answer positions A–D occur 125, 123, 123, and 124 times. Inverse projection (498 questions) marks the same query point in two views and provides reference surface anchors with known integer coordinates. The model recovers (x, y, z), with each component in $[-2, 2] \cap \mathbb{Z}$. The retained set contains 42 distinct coordinate triples. Showing the same point in two views provides the geometric evidence needed to constrain depth.

Sampling and human validation. Candidate selection varies pose bins, elevation changes, visibility, and grid resolution. The retained facing examples contain 62–79 instances per azimuth bin; three-axis azimuth bins contain 29–37 instances, and elevation and roll are also approximately balanced. This balancing targets selected answer and difficulty variables, rather than imposing a uniform distribution over every asset category or visibility class. Table A3 reports the candidate and retained counts after human quality review. Overall, 1,991 of 2,500 orientation candidates and 2,360 of 2,500 correspondence/projection candidates are retained. These percentages quantify retention after screening; they are not estimates of final annotation accuracy or inter-annotator agreement.

Table A3: **Camera question retention after human quality review.**

Task	Candidates	Retained	Retention (%)
Facing direction	1,000	851	85.1
Three-axis orientation	500	399	79.8
Top direction	1,000	741	74.1
Point correspondence	1,000	908	90.8
Relative rotation	500	459	91.8
Forward projection	500	495	99.0
Inverse projection	500	498	99.6
Total	**5,000**	**4,351**	**87.0**

B.3 Part-structure tasks

Part counting. The counting set contains 961 questions derived from PartNeXt annotations (Wang et al., 2025b). Each question supplies three views of one object and names a component type. The reference count is the number of distinct instances of that component on the object, rather than the sum of its appearances across images. The requested response contains a short evidence sentence and an integer count; the count is the evaluated field.

Language-conditioned grounding. The single-view and multi-view sets each contain 839 questions, paired by object, target part, and referring expression. The query types are part names (261), descriptive expressions (446), and functional descriptions (132). For example, a description can distinguish a particular blade component by its shape and position, while a functional expression refers to what a component is used for. The single-view image corresponds to the first image of its matched three-view example. For multi-view grounding, candidate renderings are ranked by the visible area of the target part and the top three are retained. Human reviewers check target visibility, the uniqueness of the reference, and bounding-box correctness; failed examples are excluded. The standard grounding sets contain visible targets in all supplied views.

Localization format. Each box is represented as $(y_{\min}, x_{\min}, y_{\max}, x_{\max})$ in normalized image coordinates, with the origin at the top-left corner. Multi-view responses associate one box with each zero-based image index. The prompt reserves $(0, 0, 0, 0)$ for an invisible or out-of-frame target, although this case does not occur in the retained standard grounding set. This output convention is separate from the IoU scoring rule in Section C. The structure inputs use the updated 1,024-pixel rendering resolution, consistently across the applicable standard and randomized conditions.

B.4 Surface tasks and annotation criteria

Surface evaluation contains 4,931 questions: 974 on baked lighting, 966 on texture richness, 1,196 on texture abnormalities, 1,097 on material identity, and 698 on metallic/roughness (MR) maps. The following criteria define the benchmark's annotation rubric; they are judgments of the supplied asset representations rather than measurements of physical material parameters.

Baked lighting and shadows. Six base-color observations are provided for each question. The three grades are *unqualified* (330), *qualified* (314), and *excellent* (330). Clear baked illumination boundaries, strong highlights or shadows, apparent emission, and severe ambient-occlusion patterns are unqualified. Qualified examples may contain subtle local shading without a discernible illumination direction. Excellent examples exhibit uniform base color without visible baked highlights, shadows, or ambient occlusion. The rubric permits clean, geometric eye highlights on stylized characters as a specific exception; such marks alone do not establish a lighting defect.

Texture richness. Each question contains six images and uses the same three grade names, with 329 unqualified, 330 qualified, and 307 excellent examples. The rubric treats absent textures, large monotonous color regions, error colors covering more than 1% of the surface, or blur covering more than one-third of the asset as unqualified. For color-block textures, fewer than four distinct colors over large regions is a cue for insufficient richness; this rule does not apply to fine-detail textures such as rust. Excellent examples combine sharp, layered local detail with coherent colors. Qualified examples satisfy neither the rejection criteria nor the excellent standard. Richness is judged separately from mapping correctness: visually detailed content can remain rich even when its mapping is distorted.

Texture abnormalities. Ten rendered images support a qualified/unqualified judgment and, for unqualified examples, an abnormality label. The set contains 496 qualified examples and 700 abnormal examples: 200 untextured or white models, 200 with texture chaos, 100 with local texture loss, 100 that are excessively dark, and 100 that are excessively light. Texture chaos includes semantically incoherent patterns, diagnostic checkerboards or coordinate textures, and mappings that disagree with the object's geometry. Local texture loss refers to missing regions exposing default or solid materials. The excessively dark and excessively light categories concern appearance that obscures the asset's details. Qualified examples have an empty abnormality label.

Material identity. Ten rendered images are provided. The answer vocabulary comprises wood, metal, plastic, clay or brick, leather, marble, fabric, stone, ceramics, plaster or cement, concrete, fur, skin, bark, leaf, others, and ground. The reference annotations use 16 of these 17 categories as the dominant material; ground is an available answer but does not occur as a dominant-material reference. The output specifies the dominant material and up to three materials ordered by apparent surface coverage, using an empty-material placeholder when fewer than three are present. The rubric draws on specularity, roughness cues, local patterns, and agreement across views. The dominant-material and material-set scores are defined separately in Section C.

Metallic/roughness assessment. Nine images are arranged in three view groups (front, back, and an oblique view), each containing a final render, a metallic map, and a roughness map. There are 350 qualified and 348 unqualified examples. The unqualified annotations distinguish overly uniform maps (100) from incoherent maps (248). The first category uses the benchmark's criterion of nearly constant MR maps without meaningful regional variation. The second covers implausible material assignments or abrupt, incoherent variation within a component. These are visual annotation categories, and do not imply that physical metallicity or roughness has been measured from the object.

B.5 Synthetic texture anomalies

Controlled synthesis supplements a subset of texture-abnormality examples. It modifies surface representations while retaining mesh geometry, and is distinct from the appearance-robustness evaluation in Section C.8.

Localized texture removal. The synthesis routine selects a subset of material slots or, when configured, connected face components, and assigns a neutral, untextured material to the affected faces. Selection retains an unaffected portion of the asset. The replacement material uses metallicity zero and roughness 0.85; the intervention therefore changes the local material representation as well as its base color.

Texture chaos and untextured appearance. Texture chaos is produced through randomized UV scaling, translation, and rotation, or by assigning a base-color texture from another material slot. Randomized noise can be overlaid with different blending modes and color treatments; noise also serves as a fallback when no mapping or texture-swap operation is applied. The untextured-model operation replaces all materials with white, nonmetallic materials of roughness 0.8. The synthesis routine records affected faces and transformation parameters for subsequent rendering. An applied material operation does not itself establish that a defect is perceptually apparent, so visible evidence and the task's annotation rubric determine its interpretation.

B.6 Annotation provenance and verification scope

The semantic category pool includes both agreement and disagreement cases from Gemma4-31B-it and Gemini-3.1-pro, followed by human refinement as described in Section 2.2. Their agreement is a sampling attribute, not an automatic acceptance criterion or a substitute for a reference annotation. Acquisition and creation provenance informs modeling-style labels, including scanned assets and assets generated by systems such as Tripo or Luma; human refinement determines the final labels. Symmetry, posture, and primary-object counts are manually annotated. For grounding, generated referring expressions are checked against the intended part and its visible extent. For camera tasks, an object's manually specified canonical orientation supplies the semantic reference, while rendering geometry determines the spatial targets.

Verification is task-specific. Semantically uninterpretable assets are removed from category evaluation, and ambiguous references or incorrect boxes are removed from grounding evaluation. By contrast, a clearly identifiable surface defect is a valid texture-assessment example. Camera candidate retention statistics in Section B.2 measure the proportion of questions retained after review, rather than annotator agreement or final label accuracy. Human validation and model-performance metrics serve different purposes; Section C specifies only the latter.

C Evaluation Protocol

Metric overview. Labels, counts, grid cells, and 3D coordinates require exact matches, angles allow a 3° tolerance, and part grounding requires IoU ≥ 0.5 in every reference view. Response parsing, task-specific definitions, and score aggregation are detailed below.

C.1 Evaluated Models

Table 2 compares Objective3D with twelve baselines in three categories: three proprietary VLMs, six open-weight VLMs, and three specialized 3D object models. The two VLM categories answer from the rendered observations of each question, supplied as images, whereas the specialized models take 3D input directly.

Proprietary models. We query Gemini-3.1 Pro (Google DeepMind, 2026), GPT-5.6 Sol (OpenAI, 2026), and Claude-5.0 Opus (Anthropic, 2026) through their official APIs. These frontier general-purpose multimodal models serve as the strongest image-based reference on ObjectArena3D. Gemini-3.1 Pro was also involved in benchmark construction: together with Gemma-4-31B-it, it guided the sampling of category candidates, and it generated the appearance and function descriptions used as part-grounding queries (Section 2.2). All resulting labels and queries were verified by humans, but this involvement may still affect the relative difficulty of these questions for Gemini-3.1 Pro (Section F).

Open-source models. We evaluate six open-weight VLMs from the Qwen, InternVL, and Gemma families: Qwen3.8-Flash-Next (Qwen Team, 2026b), Qwen3.6-35B-A3B (Qwen Team, 2026a), Qwen3-VL-30B-A3B and Qwen3-VL-8B (Bai et al., 2025), Intern-VL3.5-30B-A3B (Zhu et al., 2025), and Gemma-4-26B-A4B-it (Team, 2026). Qwen3-VL-8B is a dense model, whereas the models with an A3B or A4B suffix are mixture-of-experts models that activate about 3B or 4B parameters per token. Qwen3-VL-8B is also the backbone of Objective3D, so its row reports the same backbone before our object-centric fine-tuning (Section 4.1).

Specialized 3D object models. PointLLM-R-7B (Chen et al., 2026), ShapeLLM-Omni (Ye et al., 2025), and EVA01 (Yang et al., 2026) are object-level 3D multimodal language models that take 3D input directly. PointLLM-R-7B reasons over point clouds with chain-of-thought, whereas ShapeLLM-Omni and EVA01 are native 3D models that unify 3D understanding and generation, the latter through a mixture-of-transformers architecture. This category tests whether direct access to geometry, without multi-view rendering, suffices for fine-grained object questions.

Shared input and scoring protocol. All image-based models receive the same task-specific inputs and prompts. Semantic questions provide six views, each with an RGB render and a normal map

(Section B.1); camera questions provide one or two renders (Section B.2); part-structure questions provide one or three views (Section B.3); and surface questions provide six to ten images, including metallic and roughness maps where applicable (Section B.4). Every prompt requests a JSON answer, which is parsed and scored with the same deterministic rules for all models (Section C.2). Missing or unparseable answers remain in the denominator and count as incorrect.

Task coverage. All general-purpose VLMs are evaluated on the 13 metrics of Table 2, except that Intern-VL3.5-30B-A3B has no multi-view grounding result. The specialized models are evaluated only on the tasks their interfaces support, and the remaining entries are marked "–". All three are evaluated on level-1 category, modeling style, attribute analysis, texture, and material prediction. In addition, PointLLM-R-7B is evaluated on orientation analysis, ShapeLLM-Omni on metallic/roughness quality, and EVA01 on all three camera-awareness groups and metallic/roughness quality, which gives 6, 6, and 9 of the 13 metrics, respectively. None of them is evaluated on level-2 category or the part-structure tasks. Since Avg. is taken over the available columns and a "–" is not a zero score (Section C.7), these averages are not directly comparable with those of models evaluated on all 13 metrics.

Robustness evaluation. The randomized conditions are evaluated on a subset of these models, with model coverage reported separately for each condition in Tables 3 and 4.

C.2 Response Parsing and Evaluation Scope

Responses are parsed as JSON and scored deterministically against the references, without an LLM judge; unparseable answers count as incorrect. Before parsing, we remove an optional leading reasoning block and Markdown code fences. If the full response is not valid JSON, the parser attempts to recover an enclosed JSON object or array. String labels are compared after trimming surrounding whitespace and converting to lowercase. For multiple-choice tasks, either the option identifier or its corresponding label is accepted. Surface-quality labels additionally normalize spaces and hyphens to underscores and map equivalent forms, such as *normal* to *qualified* and *abnormal* to *unqualified*.

Evaluation is defined by the selected benchmark suite and its dataset versions. Predictions are matched to reference questions by task, version, and question identifier, and the reference answers are taken from the suite. Missing predictions remain in the evaluation denominator rather than being dropped; absent or unparseable answers cannot earn the primary score for a labeled question. Predictions outside the selected suite are excluded, and repeated predictions for the same question contribute only once. The reported parsing rate requires a complete answer with valid task fields; syntactically valid JSON alone is insufficient. Truncated responses and answers that fail these field checks receive zero primary scores.

C.3 Semantic and Attribute Metrics

Categories and discrete labels. Category recognition reports exact-match accuracy at both hierarchy levels, together with joint accuracy requiring both levels to be correct. The fine-category (second-level) accuracy is the primary category metric. Posture, symmetry, and modeling style use exact-match accuracy over the task's answer choices.

Counting. Object counting and part counting use exact count accuracy. When both the prediction and reference can be interpreted numerically, equality is evaluated numerically. Mean absolute error and accuracy within one count are supplementary diagnostics. The mean absolute error is computed only when both counts are numerical, whereas an invalid count receives zero exact and within-one accuracy. The object-counting overflow category is compared as the discrete label `11_plus`; it is not converted to an exact count of eleven and does not contribute to the numerical-error diagnostic.

Class-balanced diagnostics. For multiple-choice and counting tasks, and for each category level, the evaluator additionally computes macro recall and macro F1 over the labels present in the evaluated references. These diagnostics complement the primary question-averaged accuracy when class frequencies differ.

C.4 CAMERA METRICS

Facing direction and relative azimuth. For angles in degrees, the circular error between a prediction $\hat{\theta}$ and its reference θ is

$$d_{\text{circ}}(\hat{\theta}, \theta) = \left| \left((\hat{\theta} - \theta + 180) \bmod 360 \right) - 180 \right|. \tag{4}$$

Facing-direction estimation and relative-camera-azimuth estimation use accuracy within 3° as the primary metric. Exact-angle accuracy, accuracy within 30°, and mean angular error are supplementary diagnostics. Missing or nonnumerical predictions receive zero threshold accuracy.

Three-axis pose and top direction. For azimuth, elevation, and roll, let the respective errors be

$$e_a = d_{\text{circ}}(\hat{a}, a), \qquad e_e = |\hat{e} - e|, \qquad e_r = d_{\text{circ}}(\hat{r}, r). \tag{5}$$

The primary score for a complete three-axis prediction is the mean of the three threshold indicators,

$$s_{\text{axis}} = \frac{1}{3} \sum_{k \in \{a,e,r\}} \mathbf{1}[e_k \leq 3^\circ]. \tag{6}$$

If any of the three predicted angles is missing or nonnumerical, the score is zero. This axis-averaged score differs from the stricter supplementary joint score,

$$s_{\text{joint}} = \prod_{k \in \{a,e,r\}} \mathbf{1}[e_k \leq 3^\circ], \tag{7}$$

which requires all three angles to satisfy the tolerance. We also retain joint exact accuracy, joint accuracy within 30°, and the mean error across the three axes as diagnostics. Top-direction prediction uses exact-match accuracy over the discrete direction labels.

Point correspondence. The primary score jointly evaluates visibility and image-grid localization. For a reference point that is visible in the target view, a prediction is correct only if it declares the point visible and returns the exact reference grid cell (u, v). For a reference point that is not visible, a prediction of nonvisibility is sufficient and its grid coordinates are not scored. In a parsed JSON object, the visibility field is interpreted as visible only when it is the Boolean value `true`; other values are interpreted as nonvisibility. Visibility accuracy and localization within one cell along both grid axes are supplementary diagnostics. The grid coordinates refer to image cells, rather than texture UV coordinates.

2D–3D projection. Forward projection uses exact accuracy over the candidate image locations. Inverse projection requires exact agreement with the reference three-dimensional coordinate triplet. Euclidean coordinate error and accuracy within a Euclidean distance of one coordinate unit are supplementary diagnostics. If any predicted coordinate is missing or nonnumerical, both exact and within-one accuracy are zero.

C.5 PART-GROUNDING METRICS

Single-view grounding. We measure the intersection over union (IoU) between the predicted box $\hat{B}$ and reference box B,

$$\text{IoU}(\hat{B}, B) = \frac{|\hat{B} \cap B|}{|\hat{B} \cup B|}. \tag{8}$$

The primary score is $\mathbf{1}[\text{IoU}(\hat{B}, B) \geq 0.5]$. Mean IoU and success rates at thresholds of 0.25 and 0.75 are supplementary diagnostics. The canonical box format is $[y_{\min}, x_{\min}, y_{\max}, x_{\max}]$ with coordinates in $[0, 1]$. For compatibility with alternative output scales, all four coordinates are divided by 1,000 when any coordinate exceeds one in magnitude. Axis order is handled separately: for each model run and grounding task, the evaluator fixes one convention by comparing aggregate IoU under the original and axis-swapped orders. Coordinates are clipped to the image bounds and the endpoints of each axis are ordered before computing overlap. An invalid box receives zero IoU; a pair of boxes with zero union also receives zero IoU, without a separate absent-target success rule.

Multi-view grounding. The all-view accuracy reported in Tables 2 and 3 requires correct localization in every reference view of a question. For V reference views, its per-question score is

$$s_{\text{all}} = \prod_{v=1}^{V} \mathbf{1}[\text{IoU}(\hat{B}_v, B_v) \geq 0.5]. \tag{9}$$

Predicted boxes are matched to their reference view indices; any missing or invalid box makes the question unsuccessful. The reported all-view accuracy is the mean of this binary score over questions, so it measures consistent part localization across the full set of observations. The supplementary view-averaged score is

$$s_{\text{view}} = \frac{1}{V} \sum_{v=1}^{V} \mathbf{1}[\text{IoU}(\hat{B}_v, B_v) \geq 0.5]. \tag{10}$$

For three-view questions, this diagnostic gives one third of a point for each correctly localized view; all-view accuracy gives no credit unless all three are correct. Missing or invalid boxes contribute zero for their reference views. Mean IoU over reference views and valid-box coverage are additional diagnostics.

C.6 Surface Metrics

Texture and metallic/roughness quality. Texture richness and baked-in lighting/shadow assessment use exact-match accuracy for their quality labels. Texture-abnormality assessment uses joint accuracy: both the quality label and the abnormality subtype must match the reference. For an optional subtype, a missing value is treated as an empty label in both the prediction and the reference. Metallic/roughness assessment uses quality-label accuracy as its primary metric, with subtype accuracy and joint quality–subtype accuracy retained as diagnostics. Consequently, the texture-abnormality and metallic/roughness primary scores impose different subtype requirements.

Material identity. Dominant-material recognition uses exact-match Top-1 accuracy as the primary metric. For the supplementary Top-3 metric, let P and G be the sets formed from the first three predicted and reference material names, respectively, after label normalization and removal of empty or *none* entries. We compute

$$F_{1,\text{set}}(P, G) = \frac{2|P \cap G|}{|P| + |G|}, \tag{11}$$

with a score of zero when both sets are empty. This score measures set overlap; the order within the retained material names does not affect it. Material Top-3 therefore denotes set F1, rather than Top-3 classification accuracy.

C.7 Aggregation and Reporting

Each task score is the arithmetic mean of its per-question primary scores. For a task group $\mathcal{T}$ whose subtasks contain N_t questions and have mean primary scores S_t, the aggregate is the question-count-weighted mean

$$S_{\mathcal{T}} = \frac{\sum_{t \in \mathcal{T}} N_t S_t}{\sum_{t \in \mathcal{T}} N_t}. \tag{12}$$

This weighting preserves the contribution of each evaluated question; it is not an unweighted average over subtasks. Avg. is the arithmetic mean of the available reported metric columns, with equal weight assigned to each column. Three-axis pose averages threshold indicators across axes within a question; multi-view grounding requires all reference views to pass the IoU threshold before averaging over questions.

Primary scores, accuracies, and set F1 are displayed as percentages, whereas angular, count, and coordinate errors retain their original units. Continuous-error diagnostics are averaged over the questions for which their required numerical predictions are available; their sample counts are tracked separately. An aggregate diagnostic uses the number of contributing questions as its weight. Joint pose accuracy, view-averaged grounding accuracy, and material Top-3 set F1 are supplementary diagnostics. Multi-view grounding contributes all-view accuracy to the reported aggregates, as defined in Section C.5. Missing model–task measurements are shown as "–" rather than interpreted as zero performance.

C.8 STANDARD AND RANDOMIZED OBSERVATION CONDITIONS

Scope and matching. The revised standard suite contains 16,319 questions across all four families. Random orientation contains 8,617 questions (3,887 semantic and 4,730 surface questions), while random texture/color contains 10,675 questions (3,727 semantic, 4,351 camera, and 2,597 structure questions). These conditions cover different task subsets, as detailed in Table A4. They are evaluated with the same task-specific scoring rules. To measure robustness, standard and randomized scores must use corresponding question subsets, matched by source question and task identity. If a geometric target changes with the rendering, the corresponding condition-specific reference is used. Structure rendering resolution is consistent across the applicable conditions.

Table A4: **Question counts by evaluation condition.** A dash indicates that a task is outside the scope of that randomized condition.

Task	Standard	Random orientation	Random texture/color
Category	2,410	2,402	2,402
Modeling style	655	654	–
Object counting	405	402	401
Posture	499	–	495
Symmetry	429	429	429
Facing direction	851	–	851
Three-axis orientation	399	–	399
Top direction	741	–	741
Point correspondence	908	–	908
Relative azimuth	459	–	459
Forward projection	495	–	495
Inverse projection	498	–	498
Part counting	961	–	961
Single-view grounding	839	–	818
Multi-view grounding	839	–	818
Baked lighting/shadows	974	974	–
Texture richness	966	966	–
Texture abnormalities	1,196	1,002	–
Material identity	1,097	1,090	–
Metallic/roughness quality	698	698	–
Total	**16,319**	**8,617**	**10,675**

Random orientation. Objects are re-rendered after rotations about one or two horizontal axes, optionally accompanied by a vertical-axis rotation. The transformation records use nonzero multiples of 45° from 45° to 315°. Per-example records retain the applied rotations, random seed, and view identifiers. This condition covers category, modeling style, symmetry, object counting, and the five surface tasks; posture, camera tasks, and part-structure tasks are outside its scope. The retained references concern properties intended to remain meaningful after rigid rotation. Scores characterize performance under the changed observations, whose visible surfaces can differ from the standard condition.

Random texture/color. This condition replaces the original surface appearance with randomized solid coloring while preserving the mesh geometry. For the semantic and structure renderings, the recorded procedure assigns a shared randomly sampled hue across an asset's mesh objects and also changes roughness and metallicity. The resulting perturbation therefore removes texture cues and alters material appearance, rather than changing hue alone. For camera tasks, the pose definitions and task overlays provide the reference for the re-rendered questions. Grounding boxes are recomputed for the supplied views, which can differ from the original views. Consequently, grounding results in this condition assess the combined change in appearance and selected observations; they do not isolate a color-only effect.

Interpretation of score changes. Let $\mathcal{I}_{t,v}$ denote the source questions for task t retained in variant v. Matched degradation in percentage points is defined as

$$\Delta_{t,v} = 100 \left(\frac{1}{|\mathcal{I}_{t,v}|} \sum_{i \in \mathcal{I}_{t,v}} s_{t,i}^{\mathrm{std}} - \frac{1}{|\mathcal{I}_{t,v}|} \sum_{i \in \mathcal{I}_{t,v}} s_{t,i}^{v} \right), \tag{13}$$

where each score uses the reference associated with its own observation condition. A positive value indicates lower performance under the perturbation. Task-level changes should not be inferred by subtracting aggregates over different task mixtures or unmatched subsets. The random-orientation and random-color suites are separate conditions rather than a full factorial evaluation of both interventions.

D Additional Visualization

E Additional Experiments

Result snapshots and coverage. Tables A5–A16 expand the five standard and four randomized full-suite baseline summaries, the standard and randomized summaries for our Qwen3-VL-8B model, and the supplied subset reports for Gemini-3.1 Pro, GPT-5.6 Sol, and Claude-5.0 Opus. We use the immediate `summary.md` files in the supplied directories, excluding nested repair and partial-report snapshots. The full standard summaries identify the suite as `benchmark_20260916`; the randomized summaries use `taiji_benchmark_random_20260918`, whose version field separates random color from random orientation. The full-suite baselines are Qwen3.8-Flash-Next, Qwen3.6-35B-A3B, Qwen3-VL-30B-A3B, Qwen3-VL-8B, and Gemma-4-26B-A4B-it; Qwen3.8-Flash-Next has no supplied randomized summary. The proprietary reports identify their models as `gemini-3.1-pro-preview`, `gpt-5.6-sol`, and `claude-opus-5`. All three include seven camera tasks (4,351 questions), three structure tasks (2,639), and three texture-quality tasks (3,136), totaling 10,126 questions across these 13 subtasks. Opus additionally reports material recognition and MR quality. These reports update the corresponding main-table columns; other measurements retain their earlier values where available, and Avg. is recomputed over available displayed columns. These partially updated rows retain the $\diamond$ marker and remain outside the ranking of fully updated models. Historical MV measurements without an available `all_iou50` score are shown as "–" and excluded from Avg. The detailed tables report only the supplied measurements. The revised standard suite retains all 16,127 questions from the original export and adds 192 texture-abnormality questions, bringing this task to 1,196 questions and the suite to 16,319. Reference labels in the supplied prediction records establish the revised subtype counts in Section B.4. Retained historical measurements are not assumed to cover these additional questions.

Answer validity and repair. The full-suite baseline snapshots include selective retries for malformed or truncated outputs, and the randomized Ours report records two repaired samples; the standard Ours summary does not specify a repair history. The reported results should therefore not be interpreted as uniformly single-pass measurements. Residual invalid or incomplete answers remain in the evaluated question counts and receive zero primary scores; Table A5 reports source error counts separately from parse success. In particular, standard Ours reports zero errors but one category answer fails complete parsing. The companion metadata confirms full expected-suite coverage for the eleven complete reports, with no missing, extra, or duplicate records. The proprietary Markdown reports combine results from separately evaluated task families and are authoritative for those subsets. Available JSON files corroborate only part of this coverage: Gemini and GPT provide texture-quality results, while the Opus root JSON retains an earlier missing-prediction evaluation explicitly superseded by its corrected Markdown report. Their subset record and parse statistics are reported separately in Table A5; no full-suite coverage or uniform retry policy is inferred. The randomized suite contains 19,292 questions, split into 10,675 random-color and 8,617 random-orientation questions. These conditions have different task coverage; the blocks below preserve that distinction rather than pooling their scores.

Metrics and units. We report the per-task metrics using the scoring criteria explained in Section C, including all-view accuracy for multi-view grounding. Category results distinguish the two

Table A5: **Record coverage and answer validity.** Companion JSON metadata reports zero missing, extra, and duplicate records for the full-suite sources. Reported error counts are retained in the evaluated rows and are distinct from parse failures. Parse success requires all task-required answer fields and is aggregated by question count. Proprietary-model subset rows cover the reported camera, structure, and surface tasks, with task errors and parse rates from the root Markdown summaries; whole-suite coverage is unavailable. Randomized conditions follow the source version field.

Model	Evaluated rows	Reported errors	Parse success (%)
Standard			
Qwen3.8-Flash-Next	16,319	0	100.00
Qwen3.6-35B-A3B	16,319	2	99.99
Qwen3-VL-30B-A3B	16,319	83	99.49
Qwen3-VL-8B	16,319	23	99.86
Gemma-4-26B-A4B-it	16,319	19	99.88
Ours (Qwen3-VL-8B)	16,319	0	99.99
Standard reported subset			
Gemini-3.1 Pro	10,126	0	99.01
GPT-5.6 Sol	10,126	1	99.99
Claude-5.0 Opus	11,750	0	99.99
Random color			
Qwen3.6-35B-A3B	10,675	0	100.00
Qwen3-VL-30B-A3B	10,675	89	99.17
Qwen3-VL-8B	10,675	0	100.00
Gemma-4-26B-A4B-it	10,675	3	99.97
Ours (Qwen3-VL-8B)	10,675	0	100.00
Random orientation			
Qwen3.6-35B-A3B	8,617	0	100.00
Qwen3-VL-30B-A3B	8,617	0	100.00
Qwen3-VL-8B	8,617	0	100.00
Gemma-4-26B-A4B-it	8,617	14	99.84
Ours (Qwen3-VL-8B)	8,617	0	100.00

hierarchy levels from joint correctness; camera results separate axis-wise, joint, thresholded, and continuous-error measurements; grounding reports all-view success as primary and view-averaged success as supplementary. The surface tables likewise separate texture quality from abnormality subtype, MR quality from joint correctness, and dominant-material accuracy from material-set F1. Accuracy, recall, F1, IoU, coverage, and parse-success values are shown as percentages. Count MAE, angular MAE, and coordinate L2 error retain count-label, degree, and coordinate units, respectively; smaller errors are better. Continuous-error diagnostics retain the source evaluator's eligible-prediction denominators rather than assigning a fictitious distance to an invalid answer. Macro recall and macro F1 follow the source's class-level calculation, while grouped primary scores use question-count weighting over the participating subtasks as in Section C.7. The tables below present individual task metrics rather than pooling unlike diagnostics; only the parse-success overview combines tasks, weighted by their question counts. Numerical values follow the precision of the supplied summaries, and a dash indicates an unavailable task or condition rather than a zero score. Standard and randomized blocks summarize their respective exported sets; direct subtraction does not establish a matched-subset robustness effect when their retained questions differ.

Table A6: **Hierarchical category recognition.** Acc., R, and F1 denote accuracy, macro recall, and macro F1. Joint accuracy requires both hierarchy levels to be correct. All values are percentages. There are 2,410 standard and 2,402 questions in each randomized condition.

Model	Level 1 Acc.	Level 1 R	Level 1 F1	Level 2 Acc.	Level 2 R	Level 2 F1	Both levels Joint
Standard							
Qwen3.8-Flash-Next	77.14	69.58	71.04	56.80	54.39	53.65	56.80
Qwen3.6-35B-A3B	74.65	67.92	70.12	49.83	49.15	48.58	49.63
Qwen3-VL-30B-A3B	66.35	58.25	61.98	40.62	39.98	40.73	38.09
Qwen3-VL-8B	70.29	59.76	61.72	41.41	35.50	35.32	41.16
Gemma-4-26B-A4B-it	77.47	74.88	75.64	56.80	54.04	52.24	56.68
Ours (Qwen3-VL-8B)	81.20	77.45	77.98	61.62	55.05	56.42	61.58
Random color							
Qwen3.6-35B-A3B	70.02	60.37	62.82	38.55	41.87	40.17	38.43
Qwen3-VL-30B-A3B	64.99	55.06	58.56	34.97	37.09	35.64	33.18
Qwen3-VL-8B	68.69	55.19	58.29	33.56	31.83	30.64	33.26
Gemma-4-26B-A4B-it	74.56	68.34	69.94	51.71	51.66	49.38	51.62
Ours (Qwen3-VL-8B)	78.14	72.31	73.46	56.74	48.92	49.36	56.70
Random orientation							
Qwen3.6-35B-A3B	66.53	58.73	59.53	41.51	40.66	37.85	41.42
Qwen3-VL-30B-A3B	59.03	51.23	52.89	33.51	33.51	31.99	31.47
Qwen3-VL-8B	61.62	49.58	49.94	33.10	29.15	26.98	32.93
Gemma-4-26B-A4B-it	68.86	62.14	62.63	47.09	44.01	41.63	46.92
Ours (Qwen3-VL-8B)	70.19	64.09	64.77	48.63	40.65	41.44	48.58

Table A7: **Modeling style, posture, and symmetry.** Each task reports accuracy, macro recall (R), and macro F1 in percent. Dashes identify tasks outside a condition's scope. Standard question counts are 655/499/429 (style/posture/symmetry), versus –/495/429 for random color and 654/–/429 for random orientation.

Model	Style Acc.	Style R	Style F1	Posture Acc.	Posture R	Posture F1	Symmetry Acc.	Symmetry R	Symmetry F1
Standard									
Qwen3.8-Flash-Next	54.05	55.35	48.18	60.92	60.74	59.06	66.20	53.70	53.61
Qwen3.6-35B-A3B	45.19	45.89	39.95	56.11	56.57	53.31	56.41	49.31	48.91
Qwen3-VL-30B-A3B	48.85	49.73	44.13	55.91	56.39	53.48	55.01	55.99	55.01
Qwen3-VL-8B	52.21	53.52	45.59	50.30	51.03	47.63	53.85	51.88	49.49
Gemma-4-26B-A4B-it	54.66	56.01	50.00	57.92	57.68	55.18	61.54	55.50	57.87
Ours (Qwen3-VL-8B)	82.60	82.50	81.10	63.53	64.00	61.67	62.24	61.47	56.30
Random color									
Qwen3.6-35B-A3B	–	–	–	54.14	54.48	52.17	55.94	50.06	50.08
Qwen3-VL-30B-A3B	–	–	–	56.57	56.82	54.91	52.91	55.97	54.03
Qwen3-VL-8B	–	–	–	56.57	56.73	54.24	42.66	42.83	39.25
Gemma-4-26B-A4B-it	–	–	–	58.59	58.48	57.02	61.77	56.92	57.81
Ours (Qwen3-VL-8B)	–	–	–	64.24	64.85	62.94	60.84	62.72	57.44
Random orientation									
Qwen3.6-35B-A3B	45.11	46.08	39.56	–	–	–	47.55	44.28	43.49
Qwen3-VL-30B-A3B	50.61	51.69	46.39	–	–	–	51.05	55.63	50.97
Qwen3-VL-8B	53.82	55.13	48.54	–	–	–	39.39	42.01	37.09
Gemma-4-26B-A4B-it	56.27	57.93	52.03	–	–	–	54.78	52.73	51.70
Ours (Qwen3-VL-8B)	80.89	80.79	79.74	–	–	–	51.52	52.78	47.43

Table A8: **Primary-object counting.** Accuracy, within-one accuracy, macro recall, and macro F1 are percentages; MAE is in count-label units under the evaluator's numerical treatment of the overflow class. Question counts are 405 (standard), 401 (random color), and 402 (random orientation).

Model	Exact	Within one	MAE ↓	Macro recall	Macro F1
Standard					
Qwen3.8-Flash-Next	27.41	42.22	1.913	22.12	18.69
Qwen3.6-35B-A3B	21.73	39.26	2.021	19.34	16.34
Qwen3-VL-30B-A3B	25.93	33.09	2.117	18.35	12.98
Qwen3-VL-8B	19.51	33.09	2.349	15.13	9.81
Gemma-4-26B-A4B-it	19.01	31.85	2.256	16.36	12.84
Ours (Qwen3-VL-8B)	55.56	63.46	0.695	46.96	45.94
Random color					
Qwen3.6-35B-A3B	21.45	38.90	2.043	18.96	15.82
Qwen3-VL-30B-A3B	23.94	30.92	2.177	17.23	12.03
Qwen3-VL-8B	17.71	30.67	2.372	13.97	8.43
Gemma-4-26B-A4B-it	17.71	32.42	2.247	13.97	8.90
Ours (Qwen3-VL-8B)	55.11	64.84	0.718	45.79	45.15
Random orientation					
Qwen3.6-35B-A3B	15.92	29.60	2.553	14.22	9.35
Qwen3-VL-30B-A3B	20.90	27.36	2.404	15.85	9.29
Qwen3-VL-8B	14.43	26.87	2.721	12.36	5.80
Gemma-4-26B-A4B-it	15.92	26.12	2.704	13.53	8.85
Ours (Qwen3-VL-8B)	45.52	58.21	1.007	36.23	34.38

Table A9: **Facing, relative camera rotation, and top direction.** Exact and threshold accuracies are percentages; angular MAE is in degrees. Exact and 3° accuracies are reported separately even where their values coincide. Both conditions contain 851 facing, 459 relative-rotation, and 741 top-direction questions.

Model	Facing Exact	Facing $\leq 3°$	Facing $\leq 30°$	Facing MAE ↓	Relative rot. Exact	Relative rot. $\leq 3°$	Relative rot. $\leq 30°$	Relative rot. MAE ↓	Top dir. Exact
Standard									
Gemini-3.1 Pro	29.26	29.26	58.64	51.847	25.71	25.71	25.71	75.590	81.24
GPT-5.6 Sol	27.85	27.85	56.29	54.642	35.95	35.95	35.95	57.353	70.18
Claude-5.0 Opus	30.90	30.90	55.70	53.478	19.61	19.61	19.61	79.706	63.29
Qwen3.8-Flash-Next	23.27	23.27	50.76	62.150	20.70	20.70	20.70	78.235	53.04
Qwen3.6-35B-A3B	15.86	15.86	43.48	72.832	18.30	18.30	18.30	83.628	41.16
Qwen3-VL-30B-A3B	13.51	13.51	35.49	81.469	12.64	12.64	12.85	90.495	32.66
Qwen3-VL-8B	13.63	13.63	37.13	76.569	13.73	13.73	13.73	86.863	26.05
Gemma-4-26B-A4B-it	20.09	20.09	40.31	75.881	13.51	13.51	13.51	83.628	47.91
Ours (Qwen3-VL-8B)	40.89	40.89	61.10	51.892	94.55	94.55	94.55	5.440	70.85
Random color									
Qwen3.6-35B-A3B	15.04	15.04	41.72	74.841	19.17	19.17	19.17	81.471	41.03
Qwen3-VL-30B-A3B	14.45	14.45	34.90	81.857	11.33	11.55	11.55	90.924	33.60
Qwen3-VL-8B	13.51	13.51	35.61	79.636	13.07	13.07	13.07	87.745	29.82
Gemma-4-26B-A4B-it	18.92	18.92	40.07	75.176	13.07	13.07	13.07	83.529	43.72
Ours (Qwen3-VL-8B)	37.72	37.72	56.99	55.911	94.34	94.34	94.34	5.490	69.50

Table A10: **Three-axis orientation.** Axis-averaged 3° accuracy is the primary score. Joint scores require all three axes to satisfy the condition. Scores are percentages and MAE is in degrees. Each condition contains 399 questions.

Model	Axis avg. $\leq 3°$	Joint exact	Joint $\leq 3°$	Joint $\leq 30°$	MAE ↓
Standard					
Gemini-3.1 Pro	48.04	9.52	9.52	34.34	31.646
GPT-5.6 Sol	45.86	9.27	9.27	39.10	31.479
Claude-5.0 Opus	39.85	4.26	4.26	23.81	40.752
Qwen3.8-Flash-Next	36.09	4.01	4.01	17.04	42.983
Qwen3.6-35B-A3B	29.74	1.00	1.00	27.32	41.830
Qwen3-VL-30B-A3B	22.81	0.50	0.50	25.06	43.759
Qwen3-VL-8B	25.31	0.00	0.00	25.81	42.531
Gemma-4-26B-A4B-it	27.99	1.00	1.00	21.55	42.181
Ours (Qwen3-VL-8B)	62.41	22.06	22.06	49.12	25.514
Random color					
Qwen3.6-35B-A3B	27.90	1.25	1.25	26.32	42.155
Qwen3-VL-30B-A3B	23.31	1.00	1.00	20.05	44.386
Qwen3-VL-8B	25.65	0.00	0.00	26.07	42.582
Gemma-4-26B-A4B-it	28.07	1.50	1.50	24.81	41.955
Ours (Qwen3-VL-8B)	59.40	17.79	17.79	41.60	28.045

Table A11: **Cross-view point correspondence.** Visibility accuracy evaluates the visibility judgment. Cell exact and within-one scores also account for nonvisible references, following Section C; they are not localization scores conditioned on visibility. All values are percentages. Each condition contains 908 questions.

Model	Visibility accuracy	Cell exact	Cell within one
Standard			
Gemini-3.1 Pro	80.84	35.68	73.35
GPT-5.6 Sol	84.14	61.23	79.30
Claude-5.0 Opus	82.82	40.86	74.56
Qwen3.8-Flash-Next	47.69	26.76	43.61
Qwen3.6-35B-A3B	18.61	14.65	17.29
Qwen3-VL-30B-A3B	19.71	13.99	16.30
Qwen3-VL-8B	19.27	14.10	16.85
Gemma-4-26B-A4B-it	70.70	19.71	50.88
Ours (Qwen3-VL-8B)	83.15	57.60	78.85
Random color			
Qwen3.6-35B-A3B	15.86	14.21	15.31
Qwen3-VL-30B-A3B	17.07	13.88	15.31
Qwen3-VL-8B	17.95	14.10	16.41
Gemma-4-26B-A4B-it	72.91	19.05	52.86
Ours (Qwen3-VL-8B)	82.93	55.84	78.08

Table A12: **2D–3D projection.** Forward projection reports accuracy, macro recall, and macro F1. Inverse projection reports exact accuracy, accuracy within one coordinate unit, and mean Euclidean error. Scores are percentages; L2 error retains coordinate units. Both conditions contain 495 forward and 498 inverse questions.

Model	Forward Exact	Forward Macro R	Forward Macro F1	Inverse Exact	Inverse Within one	Inverse L2 ↓
Standard						
Gemini-3.1 Pro	77.78	77.77	78.74	18.47	58.84	1.052
GPT-5.6 Sol	90.30	90.31	90.31	31.12	70.08	0.864
Claude-5.0 Opus	83.43	83.42	83.40	27.51	70.28	0.910
Qwen3.8-Flash-Next	52.73	52.71	52.60	22.09	49.40	1.271
Qwen3.6-35B-A3B	42.83	42.86	42.17	15.26	44.78	1.401
Qwen3-VL-30B-A3B	35.76	35.73	34.66	13.45	45.18	1.397
Qwen3-VL-8B	31.52	31.61	29.02	7.83	39.76	1.390
Gemma-4-26B-A4B-it	36.16	36.16	35.73	10.04	35.94	1.596
Ours (Qwen3-VL-8B)	87.88	87.87	87.96	88.55	99.40	0.125
Random color						
Qwen3.6-35B-A3B	43.64	43.66	43.22	16.27	44.78	1.379
Qwen3-VL-30B-A3B	35.76	35.71	34.57	12.45	41.57	1.471
Qwen3-VL-8B	34.95	35.03	33.41	9.64	42.57	1.337
Gemma-4-26B-A4B-it	40.00	40.01	39.78	9.04	36.55	1.621
Ours (Qwen3-VL-8B)	91.31	91.30	91.30	90.36	99.40	0.105

Table A13: **Part counting.** Exact accuracy, within-one accuracy, macro recall, and macro F1 are percentages; MAE is in numbers of parts. Each condition contains 961 questions.

Model	Exact	Within one	MAE ↓	Macro recall	Macro F1
Standard					
Gemini-3.1 Pro	73.67	85.95	0.636	45.69	43.31
GPT-5.6 Sol	73.26	85.64	0.628	51.57	50.76
Claude-5.0 Opus	70.03	83.98	0.699	39.77	40.88
Qwen3.8-Flash-Next	67.33	82.83	0.727	33.57	34.46
Qwen3.6-35B-A3B	64.41	80.44	0.821	31.57	31.86
Qwen3-VL-30B-A3B	62.85	79.40	0.862	32.70	32.92
Qwen3-VL-8B	60.04	78.56	0.882	30.71	31.04
Gemma-4-26B-A4B-it	61.71	79.29	0.911	32.58	32.67
Ours (Qwen3-VL-8B)	68.57	81.79	1.025	43.08	41.06
Random color					
Qwen3.6-35B-A3B	55.67	74.61	1.096	27.56	29.35
Qwen3-VL-30B-A3B	55.78	75.03	1.076	25.80	26.35
Qwen3-VL-8B	49.64	70.45	1.262	20.57	19.54
Gemma-4-26B-A4B-it	52.86	69.41	1.324	27.15	28.07
Ours (Qwen3-VL-8B)	53.80	70.14	1.492	29.15	28.17

Table A14: **Single-view (SV) and multi-view (MV) part grounding.** IoU threshold scores, mean IoU, and valid-box coverage are percentages. The primary MV score in this paper is all-view accuracy (`all_iou50`): every reference view must reach IoU ≥ 0.5; a missing or invalid box in any reference view makes the question incorrect. Per-view accuracy (`view_iou50`), retained as an auxiliary metric, is the source evaluator's primary score. SV and MV each contain 839 standard and 818 random-color questions. The latter condition can also change the selected views.

Model	SV IoU≥ .25	SV IoU≥ .50	SV IoU≥ .75	SV Mean IoU	MV all IoU≥ .50 (primary)	MV view IoU≥ .50 (aux.)	MV Mean IoU	MV Coverage
Standard								
Gemini-3.1 Pro	89.87	82.96	66.75	76.00	74.02	78.90	71.37	93.44
GPT-5.6 Sol	91.90	85.46	71.28	78.37	76.76	83.75	76.40	100.00
Claude-5.0 Opus	92.25	83.31	58.76	70.73	70.92	81.01	67.72	100.00
Qwen3.8-Flash-Next	74.26	58.05	39.09	54.28	55.54	68.65	60.56	100.00
Qwen3.6-35B-A3B	81.76	64.84	40.76	59.19	10.97	24.43	29.67	100.00
Qwen3-VL-30B-A3B	76.76	65.08	46.72	60.23	50.18	61.22	57.51	99.88
Qwen3-VL-8B	76.52	58.88	29.68	52.84	19.07	34.68	39.51	99.52
Gemma-4-26B-A4B-it	89.63	79.74	57.81	70.80	43.86	65.28	58.11	99.52
Ours (Qwen3-VL-8B)	91.06	79.98	62.10	73.66	70.44	78.55	71.90	100.00
Random color								
Qwen3.6-35B-A3B	34.84	23.47	12.96	24.81	5.26	10.80	15.57	100.00
Qwen3-VL-30B-A3B	30.68	21.03	12.96	23.08	15.65	20.54	25.03	100.00
Qwen3-VL-8B	23.59	15.89	7.21	16.19	11.49	18.95	23.84	100.00
Gemma-4-26B-A4B-it	36.55	28.36	18.95	28.32	7.70	19.56	21.02	99.76
Ours (Qwen3-VL-8B)	42.79	33.50	23.11	33.56	26.41	31.87	31.85	100.00

Table A15: **Texture assessment.** Richness and baked-lighting columns report exact accuracy. Texture abnormality separates quality-label accuracy, subtype accuracy, and their joint accuracy; the joint score is primary. All values are percentages. Richness and lighting have 966 and 974 questions in both conditions; abnormalities have 1,196 standard and 1,002 random-orientation questions.

Model	Texture richness	Baked lighting	Abnormality Quality	Abnormality Subtype	Abnormality Joint
Standard					
Gemini-3.1 Pro	66.67	56.78	82.19	73.41	73.41
GPT-5.6 Sol	43.06	52.26	72.16	60.95	60.95
Claude-5.0 Opus	65.11	59.86	86.54	80.43	80.43
Qwen3.8-Flash-Next	40.48	54.83	69.06	52.68	52.68
Qwen3.6-35B-A3B	47.83	46.92	79.43	65.89	65.89
Qwen3-VL-30B-A3B	47.62	40.55	72.41	59.45	59.45
Qwen3-VL-8B	42.86	38.50	72.91	53.09	53.09
Gemma-4-26B-A4B-it	47.00	39.73	71.24	53.93	53.76
Ours (Qwen3-VL-8B)	62.01	58.83	71.32	55.35	55.35
Random orientation					
Qwen3.6-35B-A3B	48.24	45.48	73.15	61.08	61.08
Qwen3-VL-30B-A3B	42.86	36.55	67.37	50.70	50.70
Qwen3-VL-8B	40.06	35.22	68.36	55.99	55.99
Gemma-4-26B-A4B-it	39.23	36.86	54.99	33.93	33.93
Ours (Qwen3-VL-8B)	59.83	55.85	78.64	54.19	54.19

Table A16: **Material identity and metallic/roughness (MR) assessment.** Material Top-3 is set F1, not Top-3 classification accuracy. MR quality accuracy is primary; subtype and joint quality–subtype accuracy are supplementary. All values are percentages. Material identity reports 1,097 standard questions (926 for Claude-5.0 Opus) and 1,090 randomized questions; MR has 698 in both conditions.

Model	Material Top-1	Material Top-3 set F1	MR Quality	MR Subtype	MR Joint
Standard					
Claude-5.0 Opus	64.25	62.02	75.93	72.49	72.49
Qwen3.8-Flash-Next	61.71	63.95	68.19	63.47	63.47
Qwen3.6-35B-A3B	58.25	60.84	59.03	33.81	33.81
Qwen3-VL-30B-A3B	55.42	59.55	53.15	27.65	27.65
Qwen3-VL-8B	54.97	57.73	59.74	51.43	51.43
Gemma-4-26B-A4B-it	62.53	62.13	58.17	43.12	36.39
Ours (Qwen3-VL-8B)	60.89	61.37	74.36	50.14	42.69
Random orientation					
Qwen3.6-35B-A3B	58.90	59.52	59.89	32.81	32.81
Qwen3-VL-30B-A3B	57.52	59.23	55.87	28.94	28.94
Qwen3-VL-8B	55.41	57.23	59.17	51.86	51.86
Gemma-4-26B-A4B-it	62.66	61.77	55.01	42.41	34.96
Ours (Qwen3-VL-8B)	61.83	60.10	70.34	50.14	37.97

E.1 Fine-Grained Results under Random Texture Color

Table A17 reports attribute-analysis subtasks and expands the view-correspondence column of Table 3 for the eight models with current reports. Each task uses the same question subset across these models. UV denotes point correspondence on the image grid, and relative rotation is the camera-azimuth change (Section B.2).

Table A17: **Fine-grained results under random texture/color.** Scores are percentages. Boldface marks the best score in each column among the eight models shown.

	Attribute analysis			View correspondence	
Model	Posture	Symmetry	Object counting	Point corr. (UV)	Relative rotation
Questions	495	429	401	908	459
Proprietary models					
Gemini-3.1 Pro	**65.66**	**66.20**	**60.10**	32.05	25.49
GPT-5.6 Sol	63.43	63.87	53.87	52.31	34.42
Claude-5.0 Opus	63.64	60.61	57.61	39.76	18.52
Open-source baselines					
Qwen3.6-35B-A3B	54.14	55.94	21.45	14.21	19.17
Qwen3-VL-30B-A3B	56.57	52.91	23.94	13.88	11.55
Qwen3-VL-8B	56.57	42.66	17.71	14.10	13.07
Gemma-4-26B-A4B-it	58.59	61.77	17.71	19.05	13.07
SFT model					
Ours (Qwen3-VL-8B)	64.24	60.84	55.11	**55.84**	**94.34**

The same question subset is used across the eight models within each task; only models with current reports are included. Attribute tasks use exact accuracy. UV denotes image-grid point correspondence (`cell_exact`); relative rotation uses 3° accuracy (`within3`).

Attribute analysis. Among the eight models in Table A17, Gemini-3.1 Pro leads on posture, symmetry, and object counting, with scores of 65.66%, 66.20%, and 60.10%, respectively. The best of the four open-source baselines trails by 7.07, 4.43, and 36.16 percentage points on these tasks. Ours reaches 55.11% on object counting.

View correspondence. Among the seven baselines shown, GPT-5.6 Sol scores highest on both point correspondence (52.31%) and relative rotation (34.42%). Ours improves these scores to 55.84% and 94.34%, respectively, although point correspondence remains substantially more challenging than relative rotation for our model.

F Limitations and Future Work

Benchmark. ObjectArena3D is built from rendered views of digital 3D assets under controlled cameras and lighting. This setting makes reference answers verifiable, but it differs from real-world observations, where objects appear with clutter, occlusion, and sensor noise. The task set also covers a bounded part of object understanding and does not evaluate part mobility, affordance, or physical properties beyond the surface attributes considered here, such as mass or deformability. Several surface labels, such as texture richness and baked lighting, are rubric-based human judgments rather than physical measurements, and angles and coordinates are discretized into closed answer spaces to enable deterministic scoring. In addition, category candidates are partly sampled according to the predictions of two evaluated models; although the final labels are refined by humans, this selection may affect the relative difficulty of these questions for those models. Finally, the two randomized conditions cover different task subsets, and the random texture/color condition alters material appearance together with hue, so the observed drops do not isolate a single factor.

Method. Objective3D relies on aligned multi-view renderings of a 3D asset, which are not directly available from casual object photographs. Processing several rendering modalities per view also increases the number of visual tokens and the inference cost. Empirically, the gains are most pronounced on camera-awareness tasks, whereas orientation estimation and surface-quality assessment remain challenging, including for our model. Because Objective3D is fine-tuned on data that follow the task formats of ObjectArena3D, part of its improvement may reflect adaptation to these formats; generalization to other asset sources, rendering pipelines, and question styles remains to be verified.

Future work. On the benchmark side, natural extensions include real captures and objects in scene context, part mobility and functional properties, and open-ended responses that require models to explain the evidence behind their answers. On the modeling side, promising directions include reducing the reliance on privileged 3D inputs, for example by distilling geometry into image-only models, introducing finer patch-level geometric supervision, and reasoning explicitly in canonical object frames to improve orientation estimation. Since most tasks are scored deterministically against verifiable targets, they can also serve as reward signals for reinforcement learning.

G Broader Impacts

Positive impacts. Reliable fine-grained understanding of 3D assets can benefit applications that currently depend on manual inspection. In content creation, it can assist quality control of textures and material maps, and support the retrieval and curation of large asset collections, including data filtering for 3D generative models. In robotics and simulation, recognizing an object's orientation and localizing its parts are prerequisites for placement and manipulation. Beyond these applications, ObjectArena3D serves as a diagnostic tool that reveals where current models fail before they are deployed in such settings.

Potential risks. Automated quality assessment could be used to rate or filter creative work; its errors may unfairly penalize assets or their creators, and its adoption may reduce demand for manual review in 3D content production. Modeling-style predictions, including the AI-generated label, describe visual characteristics and should not be treated as evidence of an asset's provenance or authorship. Moreover, strong scores on rendered assets do not guarantee reliability in the physical world, and incorrect orientation or part predictions could lead to unsafe actions if used directly for robotic control; we therefore recommend human oversight and task-specific validation before such deployment. Finally, the asset distribution reflects its source collections and is imbalanced across categories, so some object types may be under-represented, and users of the benchmark should follow the licenses of the source assets.