GLANCE: FREE-FORM 3D OBJECT IDENTIFICATION THROUGH ACTIVE GEOMETRIC INSPECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Identifying an arbitrary 3D object from geometry alone is challenging: a single viewpoint may be insufficient, and geometric identity need not admit a unique textual ground truth. We study this problem as *free-form 3D object identification*: inferring an object's identity without color, texture, physical scale, canonical orientation, scene context, or a candidate vocabulary.

We introduce GLANCE, an active recognition approach in which an off-the-shelf multimodal large language model (MLLM) iteratively inspects rendered geometric observations. Rather than committing to an initial guess, GLANCE maintains competing free-form identity hypotheses, predicts geometric evidence that would distinguish them, and selects subsequent viewpoints to test these predictions. No task-specific 3D training or candidate vocabulary is required.

To evaluate free-form identification without imposing a predefined taxonomy, we construct a benchmark of $1{,}118$ 3D shapes, each independently identified by 4 human annotators together with their confidence score. We retain these annotations as multiple reference identities rather than collapsing them into a single ground truth, and evaluate predictions through semantic agreement with the references.

In controlled experiments, GLANCE improves recognition over passive multi-view baselines while requiring fewer observations. At larger scale, GLANCE achieves 83.2% primary accuracy on 570 benchmark objects. On ShapeNet, GLANCE achieves 72.0% accuracy while keeping the category vocabulary hidden from the recognition model.

1 INTRODUCTION

Recognizing what an object is from its geometry alone is surprisingly difficult. Without color, texture, physical scale, canonical orientation, or scene context, even familiar objects can become ambiguous: a cup may resemble a trash can, while a particular viewpoint may hide the feature that disambiguates them, such as the handle of a mug. Moreover, even when the geometry is understood, there may be no unique identity: different observers may describe the same shape at different levels of specificity or using different names. Humans address geometric ambiguity by inspecting an object from different viewpoints, searching for features that confirm or reject competing interpretations.

Supervised 3D recognition methods classify shapes among predefined categories Qi et al. (2017); Wang & Solomon (2021). Zero-shot approaches such as PointCLIP transfer pretrained vision-language representations to 3D recognition Zhang et al. (2022); Zhu et al. (2023), but still rely on candidate category names or text queries at inference time. Recent point-language models can instead generate object identities directly Xu et al. (2024), but require specialized 3D representations and substantial training, and assume point clouds with color and orientation.

At the same time, multimodal large language models (MLLMs) can recognize and reason about visual content and generate free-form textual responses Liu et al. (2023a). Reason–act–observe frameworks such as ReAct demonstrate how language models can interleave reasoning with actions to acquire additional information from external tools and environments Yao et al. (2022). These capabilities suggest an active approach to 3D identification: an MLLM could interrogate an object's geometry, form competing hypotheses, predict geometric evidence that would distinguish them, and acquire observations to test those predictions.

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.

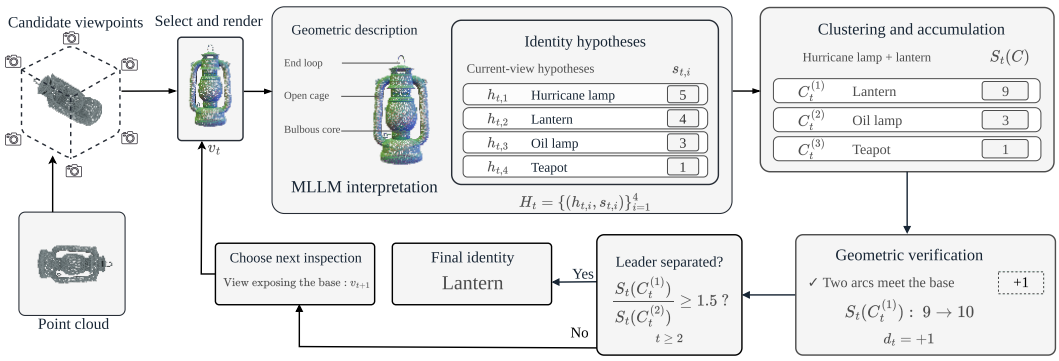


Figure 1: Overview of GLANCE. The MLLM iteratively observes an arbitrarily oriented 3D shape, maintains competing identity hypotheses, and selects viewpoints to resolve ambiguity.

Existing 3D recognition methods generally rely on predefined categories, candidate text queries, specialized 3D representations, or 3D-specific training Qi et al. (2017); Wang & Solomon (2021). In contrast, MLLMs can recognize visual content and generate free-form textual responses, suggesting a different approach: actively interrogating an object's geometry to form and test identity hypotheses Xu et al. (2024).

This raises a question: *can a general-purpose MLLM identify an arbitrary 3D object from geometry alone, without being given a vocabulary of possible answers?*

We investigate this question with GLANCE (*Guided Looking and Active Naming via Continued Exploration*), an active recognition approach in which an off-the-shelf MLLM iteratively inspects rendered geometric observations (Fig. 1). The model maintains competing free-form identity hypotheses, predicts geometric evidence that would distinguish them, and selects unseen viewpoints to test these predictions. An external controller aggregates evidence across observations and determines when the leading identity is sufficiently separated from its alternatives. Recognition therefore becomes an active process in which hypotheses guide which geometric evidence is inspected next.

Evaluating free-form identification is also challenging because geometry need not admit a unique textual identity. We therefore construct a benchmark of 1,118 Fusion/ABS shapes Willis et al. (2022); Izadyar et al. (2025), each independently identified with a confidence score by four human annotators (Sec. 4). We retain these annotations as multiple reference identities and evaluate predictions through semantic agreement, without imposing a predefined taxonomy.

Our experiments first separate the effects of multiview evidence and active inspection. On a controlled 100-object comparison (Sec. 5.2), primary accuracy increases from 53% with a single view to 73% with random multiview inspection and 76% with GLANCE. At the same time, GLANCE uses 15% fewer observations than random inspection and improves near-match accuracy from 75% to 82%, reflecting fewer severe semantic errors.

At larger scale (Sec. 5.3), GLANCE achieves 83.2% primary accuracy on the first five batches of our benchmark, 570 objects, while recognition accuracy consistently tracks human reference consensus. Finally, on ShapeNet (Sec. 5.4), the resulting free-form identities support conventional categorical recognition at 72.0% primary accuracy despite the taxonomy being withheld during recognition.

2 RELATED WORK

3D object recognition and open-vocabulary understanding. Classical 3D recognition typically classifies shapes within a predefined set of categories, including multi-view methods that aggregate rendered observations or learn informative viewpoints Su et al. (2015); Hamdi et al. (2021). Most additionally assume a canonical orientation; relaxing this assumption substantially degrades recognition Zhao et al. (2025). Recent vision-language methods such as PointCLIP, PointCLIP V2, and OpenShape enable zero-shot or open-vocabulary 3D recognition Zhang et al. (2022); Zhu et al. (2023); Liu et al. (2023b), but generally compare 3D representations against candidate category

names supplied at inference time. In contrast, we provide neither a candidate vocabulary nor canonical orientation, appearance, physical scale, or scene context: identities must be generated directly from geometric observations.

Language models for 3D understanding. PointLLM, PointLLM-R, and GPT4Point couple language models with learned 3D representations for tasks including free-form object classification and reasoning Xu et al. (2024); Chen et al. (2026); Qi et al. (2024). These methods require specialized 3D encoders and point-language training, whereas we study whether an off-the-shelf MLLM can identify an appearance-free shape through sequentially acquired 2D observations. Other approaches apply pretrained vision-language models to rendered 3D assets: Cap3D and Kabra et al. (2024) aggregate semantic information across predetermined views for captioning or annotation Luo et al. (2023); Kabra et al. (2024). We instead make observation acquisition part of recognition, using evolving identity hypotheses to determine which geometry to inspect next.

Active object recognition and viewpoint selection. Viewpoint selection has long been studied for active recognition. Recent methods learn shape-dependent viewpoints Hamdi et al. (2021), select views using CLIP confidence Song et al. (2025), or characterize the strong dependence of VLM recognition on viewing angle Rubinstein et al. (2024). Most closely related, Fan et al. (2024) actively acquires observations for open-vocabulary recognition, but ultimately classifies against candidate concepts supplied at inference time. GLANCE instead generates its own free-form identity hypotheses and uses them both as recognition outputs and to guide subsequent geometric inspection. Our experiments separate the effects of multiview evidence, model-directed viewpoint selection, and explicit prediction and verification of discriminating geometric evidence.

3 METHOD

Given an arbitrarily oriented 3D object represented as a boundary representation (B-Rep), GLANCE produces a single free-form identity prediction p that describes the object. We formulate recognition as a sequential process of hypothesis generation and testing (Fig. 1). We first sample the B-Rep into a point cloud and expose its geometry to a MLLM through rendered observations from a discrete set of viewpoints (Sec. 3.1). Because these identities are free-form, we use a separate language model to cluster semantically equivalent hypotheses and accumulate their scores (Sec. 3.2). The MLLM then predicts discriminating geometric evidence and selects an unseen viewpoint (Sec. 3.3). We repeat this process until the accumulated evidence sufficiently separates one identity from the alternatives or we reach the maximum observation budget (Sec. 3.4).

Importantly, the MLLM neither accumulates numerical confidence nor decides when recognition is complete. It receives each newly rendered view and the remaining viewpoints, while previous observations, descriptions, hypotheses, and decisions remain available through the history. Semantic clustering, evidence accumulation, and termination are handled separately by the controller.

3.1 GEOMETRY OBSERVATIONS

We represent each input B-Rep by uniformly sampling its surfaces to produce a point cloud. We render the points as spheres and color them according to their surface normals, which are evaluated analytically on the B-Rep faces from which the points are sampled. This exposes local surface orientation without introducing appearance information. The source B-Rep contains no color, texture, or material information. Since it also contains no physical scale, we normalize each object to fit the rendering volume. Absolute size is nevertheless often a discriminator between objects of similar shape. We therefore provide the MLLM, before the first observation, with the only dimensional information that is unambiguous without physical units: the ratios of the object's bounding-box extents; the MLLM is instructed that the real-world size is unknown.

We define a discrete set $\mathcal{V} = \{v_i\}_{i=1}^{14}$ of camera locations, consisting of six face views and eight corner views of the rendering volume. The input objects are not canonically oriented, and consequently these camera names do not correspond to semantic directions of the object. We use the nominal top camera as the initial viewpoint v_1. We select subsequent viewpoints from previously unseen elements of $\mathcal{V}$.

For each observation, the MLLM first produces a geometry-only description of the rendered point cloud. The description captures geometric properties such as proportions, symmetry, silhouette, distinctive features, and connectivity, without assigning semantic meaning to colors or assuming a canonical object orientation. These descriptions become part of the persistent interaction history and can therefore be used when reasoning about subsequent observations.

3.2 Hypothesis Generation and Evidence Accumulation

For each observation, the MLLM generates four free-form hypotheses, sorted by confidence

$$H_t = \{(h_{t,i}, s_{t,i})\}_{i=1}^{4},$$

where $h_{t,i}$ is an object identity and $s_{t,i} \in \{1, \ldots, 5\}$ is its confidence. The confidence is explicitly defined with respect to the *current observation only*; the MLLM is instructed not to incorporate evidence from previous observations into these scores.

Because identities are generated as free-form text, the same object may be described using different but semantically equivalent identities across observations. We therefore maintain a set of identity clusters $\mathcal{C}_t$. Each cluster $C \in \mathcal{C}_t$ consists of a representative identity ℓ_C and a set H_C of hypothesis/confidence pairs (h, s) assigned to that identity.

After each observation, a separate lightweight language model partitions all identities proposed so far into the clusters $\mathcal{C}_t$. The clustering model receives only identity strings, without confidence scores, images, or geometric descriptions, and therefore has no information about the numerical or geometric evidence associated with each identity. The model recomputes the partition over the full set of identities at each call rather than extending it incrementally, so a later partition may group identities differently from an earlier one. Clustering never modifies the recorded per-identity scores, so changes to the partition do not alter evidence accumulated in previous observations.

The score $S_t(C)$ of each cluster after observation t is the accumulated score of the hypothesis/confidence pairs currently assigned to it. Each hypothesis score consists of its initial confidence together with the outcomes of any geometric tests associated with that hypothesis (Sec. 3.3).

3.3 Active Hypothesis Testing

After clustering the hypotheses accumulated through observation t, we instruct the MLLM to first name the geometric feature that would most cheaply disprove the leading identity $h_{t,1}$ and state whether anything observed so far already speaks to it. It then identifies a concrete geometric feature whose presence is expected under $h_{t,1}$ and whose observation would help distinguish it from the competing identities. It selects an unseen viewpoint $v_{t+1} \in \mathcal{V} \setminus \{v_1, \ldots, v_t\}$ that it expects to expose this discriminator. In addition to selecting the viewpoint, the MLLM provides an explicit justification for why this view is preferable to the remaining candidates. We evaluate the importance of requiring this explicit prediction and justification in Sec. 5.2.

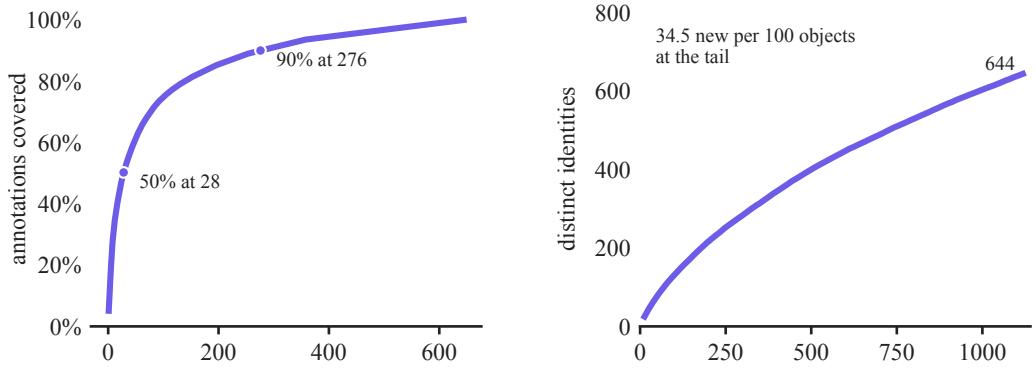
Upon receiving observation $t + 1$, the MLLM must first evaluate the prediction made at the preceding iteration. It reports whether the predicted geometric feature is present, absent, or cannot be determined from the new observation. We encode this verification as $d_{t+1} \in \{-1, 0, +1\}$, corresponding to absent, undetermined, and present, respectively. The verification result d_{t+1} is added to the score $s_{t,1}$ of $h_{t,1}$, the hypothesis whose prediction was under test, before the new hypotheses from observation $t + 1$ are accumulated. Thus, a new observation contributes both a new set of object hypotheses and the outcome of an explicit geometric test of the preceding interpretation.

3.4 Stopping Criterion

We evaluate the stopping criterion after at least two observations. Let

$$C_t^{(1)} = \arg\max_{C \in \mathcal{C}_t} S_t(C) \quad \text{and} \quad C_t^{(2)} = \arg\max_{C \in \mathcal{C}_t \setminus \{C_t^{(1)}\}} S_t(C)$$

denote the top two scores; recognition stops when $S_t(C_t^{(1)}) \geq \delta\, S_t(C_t^{(2)})$, where δ controls the required separation between the leading identity cluster and its strongest alternative. If only one



Cumulative coverage of free-form identities. Growth of the free-form identity vocabulary.

Figure 2: Vocabulary growth and cumulative coverage of free-form identities.

identity cluster exists, we consider the leading identity separated. If the criterion is not satisfied, the active-view procedure selects another unseen observation and the process repeats.

We impose a maximum budget of $T_{\max}$ observations; if this budget is reached before the stopping criterion is satisfied, the representative identity $\ell_{C_t^{(1)}}$ of the highest-scoring cluster is returned as p and the prediction is recorded as unseparated. Otherwise, $\ell_{C_t^{(1)}}$ is returned as p as soon as the stopping criterion is satisfied.

4 Dataset

Evaluating free-form object identification requires reference identities that do not assume a predefined category vocabulary. We draw candidate shapes from Fusion/ABS Willis et al. (2022); Izadyar et al. (2025). During manual curation, we exclude scenes, multi-object assemblies, and obvious repeated instances encountered during inspection, while retaining geometrically distinct instances even when they share the same semantic identity. The resulting annotation pool contains 1,540 shapes. For annotation, each shape is uniformly sampled with 100,000 points to obtain a point cloud, which is displayed without color, texture, material, physical scale, or scene context.

Human annotation. We ask four human annotators to independently identify every retained shape using an interactive 3D viewer. Annotators observe only the point-cloud geometry and may freely rotate and zoom the object. For each shape s_i, annotator $a \in \{1, 2, 3, 4\}$ provides a free-form object identity $\ell_{i,a}$ of up to three words in their preferred language, together with a confidence score $c_{i,a} \in \{1, \ldots, 5\}$ from 1 (lowest) to 5 (highest). Annotators may instead skip an object when they cannot provide an identity. The annotation interface provides no candidate vocabulary, model predictions, or annotations from the other annotators. We collect multiple independent identities because geometry alone may admit several reasonable identities for the same object. Moreover, an annotator may recognize an object without knowing its most specific name, and equivalent concepts may be expressed differently across languages. We therefore treat the human annotations as independent reference identities rather than as votes toward a single categorical ground truth.

Construction. We retain only shapes for which all four annotators provide an identity and confidence score. The resulting benchmark contains 1,118 objects, each with four independently collected identity/confidence pairs. We preserve these annotations individually rather than merging them into a single reference identity. Non-English annotations are translated to English, and spelling and capitalization errors are corrected without otherwise changing their semantic content.

Statistics. Of the 1,540 shapes in the annotation pool, we retain the 1,118 objects with four independent reference identities (App. A.2). These comprise 4,472 identities and 644 distinct identity strings. The vocabulary is strongly long-tailed: 290 identities (45.0%) occur only once, the 28 most frequent identities account for half of all annotations, and 276 are required to cover 90% (Fig. 2, left).

Moreover, the number of distinct identities continues to grow throughout the benchmark (Fig. 2, right), illustrating why a fixed vocabulary is poorly suited to this setting.

5 EVALUATION

We evaluate GLANCE using GPT-6 Astra as the recognition MLLM and GPT-5.6 Luna for semantic clustering and evaluation (Sec. 5.1). Unless otherwise stated, we use a stopping threshold of $\delta = 1.5$ and a maximum of $T_{\max} = 5$ observations. Controlled experiments use a fixed subset of 100 benchmark objects, while the complete method is evaluated on a fixed subset of 570 objects comprising the first five batches, selected independently of model performance. Full implementation, rendering, API, and inference details are provided in App. A.2.

5.1 SEMANTIC EVALUATION PROTOCOL

Free-form identities cannot be evaluated using exact string equality. Moreover, the human annotations provide reference identities rather than canonical category labels. Different annotators may use synonyms or different levels of specificity, and an annotator may recognize an object without knowing its most specific name. Conversely, the recognition model may provide a more specific name than an annotator. We therefore evaluate semantic agreement between the predicted identity and each of the four independently collected human identities.

For every prediction p_i and human identity $\ell_{i,a}$, we use a separate GPT-5.6 Luna call as the semantic judge J. The judge receives only the two identities and assigns $J(p_i, \ell_{i,a}) \in \{r_s, r_b, r_r, r_d\}$ according to the following relations:

- SAME r_s: the identities denote the same object type, including synonyms or cases where the prediction is more specific than the reference;
- BROADER r_b: the prediction is a correct but less specific name for the object;
- RELATED r_r: the prediction denotes a closely similar but distinct object type, or has a part/whole relationship with the reference;
- DIFFERENT r_d: none of the above.

For example, *cup* is BROADER than *coffee cup*, whereas *cup* and *mug* are considered RELATED. The judge receives neither the object geometry nor information about which system generated the prediction. Identities that are identical after case and punctuation normalization are assigned SAME without a model call. Every verdict is cached on disk keyed by the model, prompt, and two normalized strings, so a reported evaluation is exactly reproducible.

The four relations induce three increasingly permissive evaluation criteria. The strict criterion $\mathcal{C}_s = \{r_s\}$ accepts only SAME; our primary criterion $\mathcal{C}_p = \{r_s, r_b\}$ additionally accepts BROADER, allowing a valid identity at a coarser level of specificity; and the near-match criterion $\mathcal{C}_n = \{r_s, r_b, r_r\}$ additionally accepts RELATED. For a given criterion $\mathcal{C}$, we define

$$c(p_i, \mathcal{C}) = \mathbf{1}[\exists a \in \{1, 2, 3, 4\} : J(p_i, \ell_{i,a}) \in \mathcal{C}],$$

and report the accuracy $\mathrm{A}(\mathcal{C}) = 1/n \sum_{i=1}^{n} c(p_i, \mathcal{C})$, where n is the number of objects. Thus, a prediction is considered correct if it agrees with at least one independently provided human reference under the chosen criterion. Each object contributes equally to the final accuracy; human confidence scores are not used to weight correctness. We instead use confidence independently to analyze the relationship between human reference consensus and model behavior in Sec. 5.3.

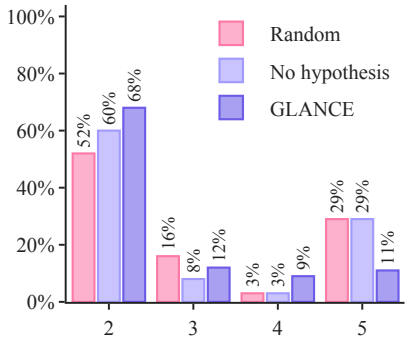
To validate the semantic judge, we manually annotate a random sample of 200 predictions against their reference identities, using the same four relations and without access to the judge's output. The judge agrees on 88.0% of items (Cohen's $\kappa = 0.78$); under the primary binary criterion $\mathcal{C}_p$ agreement is 94.5% ($\kappa = 0.86$). The judge is slightly more permissive, accepting 76.5% of predictions against the annotator's 71.0%.

5.2 ACTIVE RECOGNITION

We evaluate active recognition on the same 100-object subset using two baselines and one ablation of our complete method. All variants use the recognition MLLM and semantic clustering (Sec. 3), the

Method	$A(\mathcal{C}_p) \uparrow$	$A(\mathcal{C}_n) \uparrow$	Views $\downarrow$
Single view	53%	54%	1.00
Random views	73%	75%	3.09
No hypothesis test	69%	73%	3.01
GLANCE	**76%**	**82%**	**2.63**

Controlled comparison on 100 benchmark objects.



Distribution of the observations.

Figure 3: Controlled evaluation of active recognition.

same candidate viewpoints and rendering procedure (Sec. 3.1), and the semantic evaluation protocol (Sec. 5.1). We change only the observation strategy and the geometric verification.

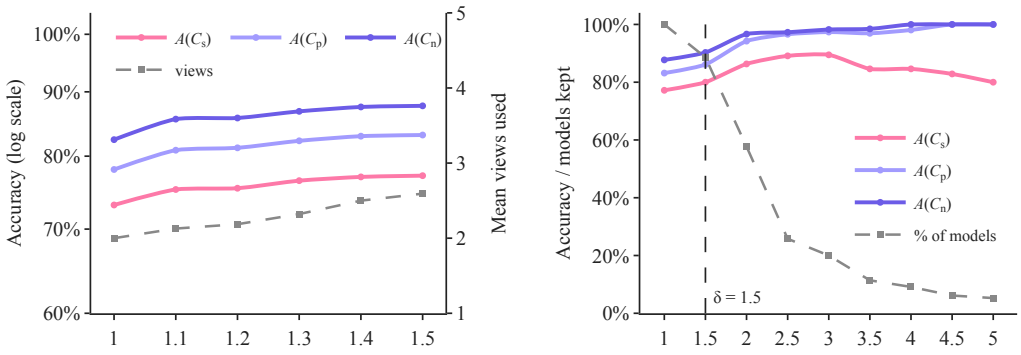
Single view. We use only the initial observation (Sec. 3.1) and perform no additional inspection. We obtain the final identity from the hypotheses generated from this observation using the same semantic clustering and identity scoring as our complete method (Sec. 3.2). This baseline measures recognition from a single geometric observation without accumulating evidence from additional viewpoints.

Random views. We replace active viewpoint selection (Sec. 3.3) with uniform random selection among the remaining unseen viewpoints. We process each observation using the same recognition, semantic clustering, identity scoring, and stopping criterion as our method. Because random selection makes no geometric prediction about what the next observation should reveal, we omit the prediction and verification term d_t (Sec. 3.3). This baseline measures the benefit of accumulating evidence from multiple geometric observations without actively choosing which geometry to inspect. This variant therefore differs from GLANCE in two respects at once, since dropping active selection also removes the verification term.

No hypothesis testing. We allow the recognition MLLM to select the next unseen viewpoint, but remove the explicit geometric reasoning used by our complete method (Sec. 3.3). In particular, the model selects a viewpoint without specifying the geometric discriminator it intends to inspect or the feature it expects to observe. Consequently, we also remove the subsequent verification and its $+1/0/-1$ contribution d_t to the identity score. The model is still required to name the feature that would most cheaply disprove its leading hypothesis, so this ablation removes the forward-looking prediction and its test, not self-criticism as such. We otherwise retain the same recognition, semantic clustering, identity scoring, candidate viewpoints, and stopping criterion. This ablation tests whether explicitly formulating and subsequently testing a geometric hypothesis contributes beyond allowing the MLLM to select its own viewpoint.

Single versus multiple views. The single-view baseline achieves only 53% primary accuracy, while accumulating randomly selected observations increases accuracy to 73%. Near-match accuracy similarly increases from 54% to 75%. Thus, additional geometric observations provide a substantial benefit when objects are presented in arbitrary orientations.

Observation efficiency. GLANCE achieves 76% primary accuracy using 2.63 observations on average, compared with 73% using 3.09 randomly selected observations, requiring approximately 15% fewer observations. The stopping-time distribution (Fig. 3, right) shows that this difference is not due merely to small reductions distributed uniformly across objects. GLANCE terminates after only two observations for 68% of objects, compared with 52% for random observation. Conversely, only 11% of GLANCE runs reach the maximum five-view budget, compared with 29% for random



Accuracy and observation cost vs δ. Accuracy and coverage vs δ.

Figure 4: Tradeoffs induced by hypothesis separation.

observation. Thus, the active procedure substantially reduces the long tail of objects that remain unresolved until the observation budget is exhausted.

Explicit hypothesis testing. Allowing the MLLM to control the camera is not by itself sufficient. Without explicit forward prediction and verification, primary accuracy decreases to 69%, despite requiring 3.01 observations on average. Moreover, 29% of these runs reach the maximum observation budget, the same fraction as random observation and substantially more than GLANCE's 11%. Near-match accuracy also increases from 73% without the forward test to 82% with the complete method (Fig 3, left). App. A.3 illustrates cases where GLANCE substantially reduces errors that persist under random multiview observation. These results suggest that the benefit of the active procedure does not arise simply from asking the MLLM where to look next, but from coupling viewpoint selection to an explicit geometric distinction between competing identities and subsequently testing the predicted evidence.

Semantic errors. The semantic error distribution provides a complementary view. A single observation produces an identity judged DIFFERENT from all human references for 46% of objects. Random multiview observation reduces this fraction to 25%, while GLANCE further reduces it to 18%. Interestingly, random observation and GLANCE have similar strict SAME accuracy (69% and 68%, respectively). Their predictions differ instead in the distribution of the remaining semantic outcomes: under GLANCE, fewer predictions are judged completely DIFFERENT from all human references, leading to near-match accuracy of 82%, compared with 75% for random observation.

5.3 LARGE-SCALE EVALUATION

We evaluate GLANCE on 570 objects from the fixed five-batch subset described above. Across these 570 objects, GLANCE achieves 77.2%, 83.2%, and 87.7% accuracy under the strict, primary, and near-match criteria, respectively. Most objects terminate after only two observations 70.2%, while 12.6% reach the maximum five-view budget.

Recognition difficulty. The stopping behavior is strongly associated with recognition difficulty. Objects resolved after two views have substantially higher recognition accuracy than objects requiring additional observations (App. A.4). The semantic outcomes follow the same trend: approximately 84% of objects resolved after two views receive a SAME verdict, while DIFFERENT predictions become increasingly frequent with additional observations (App. A.4).

Accuracy/observation tradeoff. The stopping threshold δ directly controls the tradeoff between recognition accuracy and observation cost. Replaying the recorded GLANCE trajectories with thresholds from 1.0 to 1.5, primary accuracy increases from approximately 78% to 83%, while the mean number of observations increases from approximately 2.0 to 2.6 (Fig. 4, left). Most of the accuracy gain occurs at small thresholds; beyond $\delta \approx 1.1$, accuracy improves gradually while observation cost

continues to increase. We use $\delta = 1.5$ throughout our experiments, favoring recognition accuracy over observation cost.

Accuracy/coverage tradeoff. The final separation between the leading hypotheses also provides a useful measure of prediction reliability. If we retain only predictions whose final score ratio exceeds a threshold δ, primary accuracy increases from 83.2% on the full set to approximately 94% at $\delta = 2$, while retaining approximately 58% of the objects (Fig. 4, right). Higher thresholds further increase primary and near-match accuracy, at the cost of rapidly decreasing coverage.

Human agreement. To relate model performance to human agreement, we define a reference-consensus score that combines semantic agreement between reference identities with annotator confidence. Two annotators agree when their identities are judged SAME or BROADER, treating this relation symmetrically. Let G_i denote the set of annotators agreeing with annotator i, including i itself, and let $k_{o,j} \in \{1, \ldots, 5\}$ denote annotator j's confidence for object o. We define $q_o = 1/4 \max_i \sum_{j \in G_i} k_{o,j}$. Thus, confidence contributes to the score only when supported by a semantically agreeing group of reference identities. Model performance closely tracks human reference consensus (App. A.5). Primary accuracy decreases from 91.2% for high-consensus objects to 76.4% and 68.3% for medium- and low-consensus objects, respectively. Thus, objects for which independent human annotators provide mutually consistent, high-confidence identities are also substantially easier for GLANCE to identify from geometry alone.

5.4 EVALUATION ON SHAPENET

We randomly sample 100 objects from ShapeNet across 21 randomly selected categories and run GLANCE unchanged, without exposing the category names to the recognition MLLM. To match our geometry-only setting, we remove color and texture and exclude real-world scale and canonical upright orientation as recognition cues. Thus, the model identifies objects from geometry alone without access to a predefined category vocabulary.

Using the evaluation described in Sec. 5.1 with ShapeNet categories replacing the human reference identities, GLANCE achieves $A(\mathcal{C}_s) = 63.0\%$, $A(\mathcal{C}_p) = 72.0\%$, and $A(\mathcal{C}_n) = 73.0\%$. Performance varies sharply across categories. Even under the strict criterion $\mathcal{C}_s$, all airplanes, benches, motorcycles, chairs, guitars, pillows and cars are correctly identified. In contrast, GLANCE reaches 20% on clocks and 40% on water tanks, which are often identified as silos, and 20% on bathtubs, identified as sinks. Such ambiguities illustrate the difficulty of inferring an object's intended function from shape alone (App. A.6). Color, texture, and physical size could help distinguish these objects, but those cues are unavailable under our protocol.

6 CONCLUSION

We introduced GLANCE, an active approach for free-form 3D object identification from geometry without a predefined candidate vocabulary. Our experiments show that multiple observations are crucial for this task and that actively predicting and testing discriminating geometric evidence improves recognition while reducing the number of observations required. To evaluate this setting, we introduced a benchmark with multiple independent human reference identities and a semantic evaluation protocol that accounts for the inherent ambiguity of geometric identity. GLANCE achieves 83.2% primary accuracy on our benchmark and 72.0% on ShapeNet without exposing its category vocabulary to the recognition model.

Our study is limited to a single recognition MLLM and a fixed set of rendered geometric observations. Future work could provide GLANCE with richer geometric inspection and recognition tools, allowing it to actively choose not only where to look, but also how to interrogate the geometry.

Together, these results show that object identities can be inferred directly from geometric observations and used to guide the acquisition of additional evidence without first specifying the set of possible answers.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, code development, and literature discovery. All AI-assisted content was reviewed and verified by the authors, and references were checked against the original publications. Generative AI tools were not used to produce human annotations. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work studies object identification from 3D geometry and does not involve personal or sensitive data. Human participants provided object identity annotations for geometric shapes; no personal information was collected.

REPRODUCIBILITY STATEMENT

We will release the code, dataset, annotations, prompts, and evaluation pipeline upon publication.

REFERENCES

Chaoqi Chen, Qile Xu, Wenjun Zhou, and Hui Huang. Pointllm-r: Enhancing 3d point cloud reasoning via chain-of-thought. In *ACM SIGGRAPH Conference Papers*, pp. 190:1–190:12, 2026. doi: 10.1145/3799902.3811081.

Lei Fan, Jianxiong Zhou, Xiaoying Xing, and Ying Wu. Active open-vocabulary recognition: Let intelligent moving mitigate CLIP limitations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16394–16403, 2024. doi: 10.1109/CVPR52733.2024.01551.

Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. MVTN: Multi-view transformation network for 3d shape recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. doi: 10.1109/ICCV48922.2021.00007.

Nafiseh Izadyar, Sai Chandra Madduri, and Teseo Schneider. Better step, a format and dataset for boundary representation. *arXiv preprint arXiv:2506.05417*, 2025.

Rishabh Kabra, Loic Matthey, Alexander Lerchner, and Niloy Mitra. Leveraging VLM-based pipelines to annotate 3d objects. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22710–22747. PMLR, 2024.

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

Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. OpenShape: Scaling up 3d shape representation towards open-world understanding. In *Advances in Neural Information Processing Systems*, volume 36, 2023b.

Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. doi: 10.1109/CVPR52733.2024.02495.

Jacob Rubinstein, Francis Ferraro, Cynthia Matuszek, and Don Engel. A large model's ability to identify 3d objects as a function of viewing angle. In *2024 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)*, pp. 281–288, 2024. doi: 10.1109/AIxVR59861.2024.00047.

Dan Song, Xinwei Fu, Ning Liu, Weizhi Nie, Wenhui Li, Lanjun Wang, You Yang, and Anan Liu. MV-CLIP: Multi-view CLIP for zero-shot 3d shape recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(9):8767–8779, 2025. doi: 10.1109/TCSVT.2025.3551084.

Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 945–953, 2015. doi: 10.1109/ICCV.2015.114.

Yue Wang and Justin M Solomon. Object dgcnn: 3d object detection using dynamic graphs. *Advances in Neural Information Processing Systems*, 34:20745–20758, 2021.

Karl DD Willis, Pradeep Kumar Jayaraman, Hang Chu, Yunsheng Tian, Yifei Li, Daniele Grandi, Aditya Sanghi, Linh Tran, Joseph G Lambourne, Armando Solar-Lezama, et al. Joinable: Learning bottom-up assembly of parametric cad joints. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15849–15860, 2022.

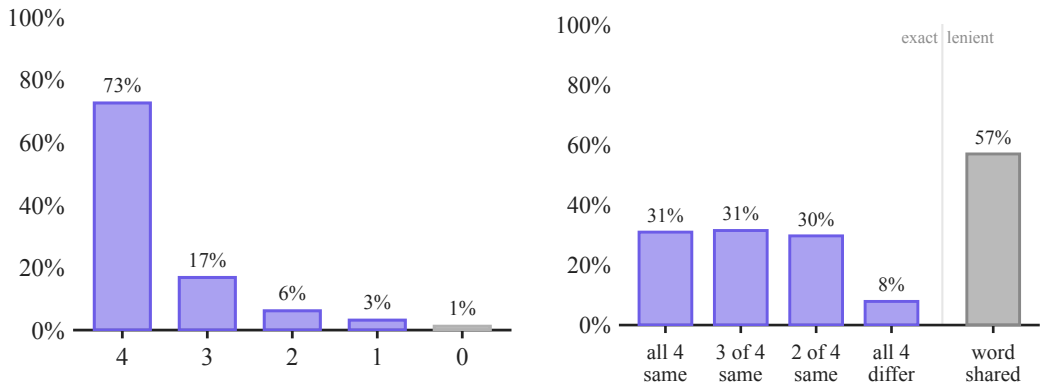
Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. PointLLM: Empowering large language models to understand point clouds. In *European Conference on Computer Vision (ECCV)*, 2024. doi: 10.1007/978-3-031-72698-9_8.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. PointCLIP: Point cloud understanding by CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8552–8562, 2022. doi: 10.1109/CVPR52688.2022.00836.

Weiguang Zhao, Guanyu Yang, Rui Zhang, Chenru Jiang, Chaolong Yang, Yuyao Yan, Amir Hussain, and Kaizhu Huang. Open-pose 3d zero-shot learning: Benchmark and challenges. *Neural Networks*, 181:106775, 2025. doi: 10.1016/j.neunet.2024.106775.

Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. PointCLIP V2: Prompting CLIP and GPT for powerful 3d open-world learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2639–2650, 2023. doi: 10.1109/ICCV51070.2023.00249.



Annotation coverage.

Exact agreement among reference identities.

Figure 5: Annotation coverage and exact agreement among reference identities.

A Appendix

A.1 Dataset Statistics

Of the 1,540 shapes in the annotation pool, 1,118 (72.6%) receive identities from all four annotators, 259 (16.8%) from three, 95 (6.2%) from two, 49 (3.2%) from one, and 19 (1.2%) from none (Fig. 5, left). Exact agreement is limited: all four annotators provide the same translated identity for only 346 objects (30.9%) (Fig. 5, right). The largest group of identical identities has size three for 352 objects (31.5%), size two for 332 (29.7%), and size one for 88 (7.9%). This diversity motivates retaining multiple independent reference identities rather than collapsing the annotations into a single category.

A.2 Experimental Setup

We uniformly sample 100,000 points from each B-Rep and render them as spheres of radius 1.8 pixels at a resolution of 256×256. Surface normals are mapped to RGB as $\mathbf{c} = (\mathbf{n} + \mathbf{1})/2$, where $\mathbf{n}$ is the exact per-point surface normal evaluated analytically on the B-Rep face from which the point was sampled. We use a stopping threshold of $\delta = 1.5$ and a maximum observation budget of $T_{\max} = 5$.

All models are accessed through the OpenAI Responses API between 14 and 20 September 2026.[1] Each turn must emit exactly one tool call, and turns are chained server-side so that every previous image and report remains in context. Images are submitted at the API's low-detail setting. Each report contains four hypotheses, each limited to three content words. Sampling temperature and reasoning effort are left at the provider defaults.

Processing the large-scale evaluation requires 1,479 MLLM calls over 1,478 rendered observations, approximately $4.3M$ input and $0.63M$ output tokens, 8.6 hours of wall-clock time, and an API cost of approximately $52.

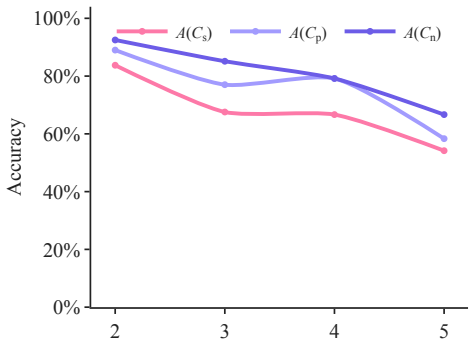
A.3 Qualitative Examples

Figure 6 shows four examples where GLANCE identifies the object correctly while random multi-view recognition does not. The top row shows the initial observation, and the bottom row shows an informative observation selected by GLANCE.

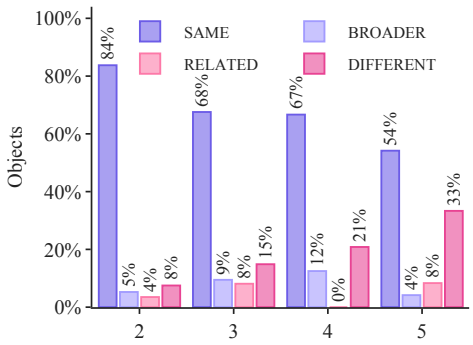
[1] We report the model aliases as called; providers may update the snapshot an alias resolves to, so results are tied to this access window.



Figure 6: Examples for which random multiview recognition remains incorrect while GLANCE identifies the object correctly. Random recognition predicts, from left to right, *climbing hold*, *iron*, *door knocker*, and *spur*; GLANCE identifies the objects as *mouse*, *bottle*, *sink*, and *chair*, respectively.



Recognition accuracy conditioned on the number of views.



Distribution of semantic verdicts conditioned on the number of views.

Figure 7: Recognition difficulty as a function of the number of observations.

A.4 RECOGNITION DIFFICULTY

Figure 7 provides the full breakdown of recognition performance as a function of the number of observations required before termination. We report both recognition accuracy and the distribution of semantic verdicts for objects terminating after two through five observations.

A.5 HUMAN AGREEMENT

Figure 8 reports recognition accuracy after grouping objects by the reference-consensus score defined in Sec. 5.3. We use high ($q_o \geq 4.5$), medium ($3 \leq q_o < 4.5$), and low ($q_o < 3$) consensus groups.

A.6 SHAPENET

Tab. 1 reports the complete per-category breakdown for the ShapeNet evaluation in Sec. 5.4.

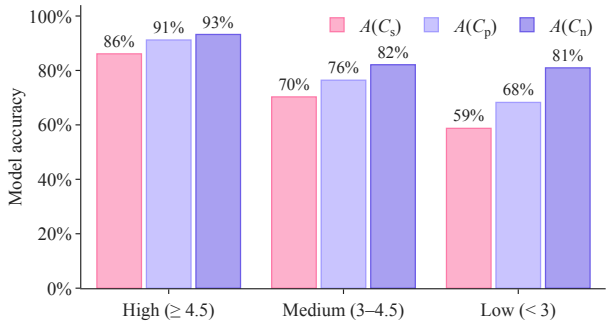


Figure 8: Recognition accuracy increases with human reference consensus.

Table 1: Per-category accuracy of GLANCE on ShapeNet, with the predefined category name as the single reference identity. Sample sizes are small, so these figures describe the sampled instances rather than performance over the full ShapeNet categories.

Category	n	$A(\mathcal{C}_s)$	$A(\mathcal{C}_p)$	$A(\mathcal{C}_n)$
Airplane	5	100	100	100
Bench	5	100	100	100
Motorcycle	4	100	100	100
Chair	5	100	100	100
Guitar	5	100	100	100
Pillow	5	100	100	100
Car	5	100	100	100
Keyboard	5	80	80	80
Mug	5	80	100	100
Sofa	5	60	100	100
Trash bin	6	60	80	80
Gun	5	60	60	60
Bookshelf	5	40	60	60
Camera	5	40	40	40
Laptop	5	40	40	40
Pistol	5	40	40	40
Water tank	5	40	60	60
Bathtub	5	20	40	60
Clock	5	20	20	20
Microphone stand	2	0	100	100
Microphone	3	0	0	0
All	100	63.0	72.0	73.0