Google DeepMind UCL

Leveraging VLM-Based Pipelines to Annotate 3D Objects

Rishabh Kabra, Loic Matthey, Alexander Lerchner, Niloy J. Mitra

Contributions

1. A visually grounded aggregation (***ScoreAgg***) of VLM responses across multiple queries.
2. We use ScoreAgg to summarize VLM captions/annotations generated across different views of a 3D object.
3. We compare ScoreAgg with a leading approach which uses GPT4 for text-based summarization (CAP3D, Luo et al. 2023). We use *caption blow-up ratios* as a **measure of hallucination** to show our method is reliable where CAP3D is not.
4. We establish a SoTA on type and material inference w.r.t. given and collected human labels respectively.
5. We propose an **unsupervised *visual sensitivity metric*** that is predictive of VLM accuracy.

Background

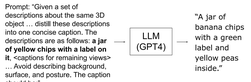
VLMs are text-generating systems which model the joint likelihood of image-text data. Pretrained VLMs have been successful at various zero-shot inference tasks. We rely on two families of pretrained VLMs to generate captions/annotations: 55B PaLI-X and 3B BLIP-2. We show, however, that they are inconsistent in recognizing different views of the same object.

Objaverse is a collection of 800k diverse but poorly annotated 3D models created by 100k artists. It provides a rich testing ground for VLM-based annotation pipelines. A subset of 47k objects called Objaverse-LVIS is accompanied by human-verified categories. We rely on it to validate our type annotations. We also introduce a subset with material labels.

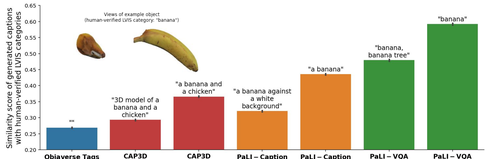
Baseline. A three-module pipeline (CAP3D) was recently proposed to generate captions for Objaverse. Their pipeline is as follows: a VLM (BLIP-2) first produces 5 candidate captions for 8 object views; CLIP filters all but one caption per view, and GPT4 performs a flawed detail-preserving but hallucination-prone aggregation. Our procedure is similar up to CAP3D's first stage, but we don't use any further neural modules for filtering or summarization.

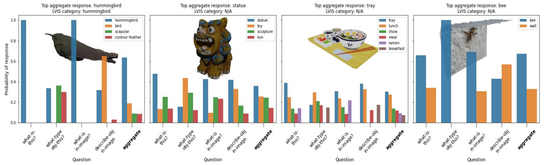
Multi-view differences can produce varying object descriptions

A. Aggregation in text space using an LLM and engineered prompt (CAP3D)



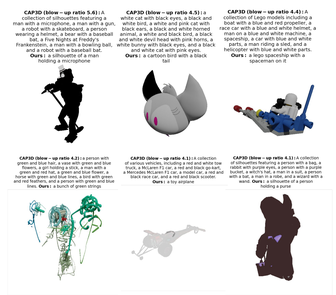
Task 1: Inferring object semantics/type





Hallucination measure

$$\text{blow-up ratio}(r) = \frac{\text{wordcount}(r)}{\max_{i,j} \text{wordcount}(r_{i,j})}$$



Task 2: Inferring object material

Table 1. **Material inference with two VLMs: PaLI-X and BLIP-2.**. The models are provided either an object type annotation or image as inputs or both. We report the top-3 accuracy as well as the soft accuracy averaged over our material test set of 860 objects. Whenever we use appearance as an input (i.e., VLM mode), we aggregate responses across object views. Thus the predicted distributions contain up to J=5 alternatives in LLM mode or up to IJ=40 in VLM mode.

		From Type (LLM mode)		From Appearance (VLM mode)	From Type and Appearance (VLM mode)	
		CAP3D captions	PaLI-VQA types	No caption/type information	CAP3D captions	PaLI-VQA types
PaLI-X 55B VQA	Top-3 acc.	0.73 ± 0.44	0.58 ± 0.49	0.83 ± 0.37	$\mathbf{0.87 \pm 0.34}$	0.86 ± 0.35
	Soft acc.	0.36 ± 0.29	0.25 ± 0.28	0.41 ± 0.28	$\mathbf{0.44 \pm 0.27}$	$\mathbf{0.44 \pm 0.29}$
BLIP-2 T5 XL	Top-3 acc.	0.24 ± 0.43	0.22 ± 0.41	0.68 ± 0.47	0.59 ± 0.49	$\mathbf{0.69 \pm 0.46}$
	Soft acc.	0.19 ± 0.35	0.16 ± 0.33	0.50 ± 0.41	0.42 ± 0.42	$\mathbf{0.51 \pm 0.42}$

m "glass"
t_{cap3d} "hat and a jar, both with ropes tied around them"
t_{pali} "potion"
$p_{pali}(\hat{m}|t_{cap3d})$ "cotton" (0.64), "can't tell" (0.36)
$p_{pali}(\hat{m}|t_{pali})$ "potion" (0.35), "glass" (0.27)
$p_{pali}(\hat{m}|A)$ "cork" (0.45), "glass" (0.19)
$p_{pali}(\hat{m}|t_{cap3d}, A)$ "burlap" (0.44), "canvas" (0.30)
$p_{pali}(\hat{m}|t_{pali}, A)$ "glass" (0.67), "cork" (0.17)
$p_{blip}(\hat{m}|t_{cap3d})$ "straw" (0.49), "plastic" (0.33)
$p_{blip}(\hat{m}|t_{pali})$ "a tainted potion made of a tainted potion and a tainted potion" (0.77), "a tainted potion made of a tainted potion, and a tainted pot" (0.14)
$p_{blip}(\hat{m}|A)$ "wood" (0.83), "rope" (0.10)
$p_{blip}(\hat{m}|t_{cap3d}, A)$ "wood" (0.68), "leather" (0.13)
$p_{blip}(\hat{m}|t_{pali}, A)$ "wood" (0.95), "stone" (0.04)

m "marble"
t_{cap3d} "white marble skull"
t_{pali} "skull"
$p_{pali}(\hat{m}|t_{cap3d})$ "marble" (0.79), "porcelain" (0.09)
$p_{pali}(\hat{m}|t_{pali})$ "bone" (0.75), "bones" (0.09)
$p_{pali}(\hat{m}|A)$ "clay" (0.35), "marble" (0.22)
$p_{pali}(\hat{m}|t_{cap3d}, A)$ "marble" (0.55), "clay" (0.20)
$p_{pali}(\hat{m}|t_{pali}, A)$ "clay" (0.33), "marble" (0.27)
$p_{blip}(\hat{m}|t_{cap3d})$ "limestone" (0.68), "marble" (0.32)
$p_{blip}(\hat{m}|t_{pali})$ "calcium phosphate" (0.83), "calcareous limestone" (0.08)
$p_{blip}(\hat{m}|A)$ "marble" (0.81), "white marble" (0.10)
$p_{blip}(\hat{m}|t_{cap3d}, A)$ "white marble" (0.85), "marble" (0.07)
$p_{blip}(\hat{m}|t_{pali}, A)$ "marble" (0.43), "limestone" (0.36)

Score-Based Aggregation (ScoreAgg)

B. Aggregation using available VLM scores of each description (**ours**)

During VLM sampling (e.g., beam search), the likelihood of any sampled text can be computed without any additional cost. When VLM queries are correlated (e.g. views of the same object), we can expect recurring responses across queries. Say we run I queries to get J (response, score) pairs per query, for a total of IJ pairs $\{(r_{i,j}, s_{i,j})\}$. Let f be a map to postprocess strings and reduce them to a canonical form. The following aggregation helps identify responses r which occur frequently while accounting for the model's confidence in each occurrence.

$$\forall r \in \{r_{i,j}\}:$$

$$s_i(r) := \sup\{s_{i,j} \mid f(r_{i,j}) = r \text{ and } j = 1, 2, ..., J\} \quad (1)$$

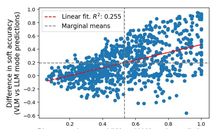
$$s_{agg}(r) := \log \sum_i \exp(s_i(r)) \quad (2)$$

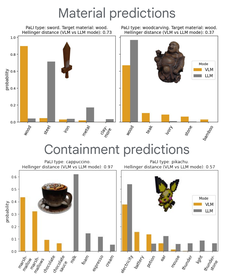
$$p(r|\{r_{i,j}, s_{i,j}\}) := \exp(s_{agg}(r)) / \sum_{r'} \exp(s_{agg}(r')) \quad (3)$$

Unsupervised evaluation of VLM annotations

Table 2. **Object properties assessed without validation, via visual ablation of PaLI-X responses.** Placeholders such as T or M are filled in with a prior type or material inference. The indefinite article "a" is replaced with "an" if the next word requires it.

Question type	LLM mode / VLM mode question	Hellinger distance (b/w LLM & VLM mode outputs)
Fragility	Is a/this T fragile?	0.110 ± 0.058
	Is a/this M T fragile?	0.103 ± 0.054
Lift-ability	Can a human lift a/this T?	0.133 ± 0.063
	Can a human lift a/this M T?	0.124 ± 0.055
Afford-ance	How is a/the T typically used?	0.575 ± 0.223
Contain-ment	What might a/the T contain?	0.690 ± 0.260
	What is something that typically goes into a/the T?	0.570 ± 0.235
	What items or substance might a/the T contain?	0.731 ± 0.236
Color	What color is a/this T?	0.894 ± 0.163
	What color is a/this M T?	0.875 ± 0.173





Takeaways

1. Running multiple VLM probes/aggregating across them is a useful technique to uncover/deal with VLM uncertainty.
2. VLM evaluation can be scaled to cases where we don't have validation data.
3. We are releasing our annotations for Objaverse to help downstream applications such as retrieval, 3D generation, and physical simulation.