More Views, More Problems? A Critical Analysis of Multi-View Aggregation for Agent Perception

Anonymous Author(s)
Submission Id: 1517

ABSTRACT

For an autonomous agent to interact effectively with its environment, it must construct accurate and robust representations from visual data. Multi-view perception is generally expected to enhance understanding, but the question of how to best integrate information across views remains unresolved. In this work, we examine an LLM-based synthesis strategy in which captions generated from multiple viewpoints by a VLM are aggregated into a single, unified description. We evaluate this approach against three conditions: a canonical single-view baseline, a naive average, and an oracle that selects the most informative viewpoint. Experiments are conducted on two datasets: a collection of in-domain, real-world objects and a domain-shifted set of 3D-printed objects. Our results demonstrate that synthesis can successfully combine complementary information when per-view captions are reliable, yielding descriptions superior to those from a static view. However, when domain shift reduces caption quality, the same strategy degrades substantially, often performing worse than the baseline. These findings highlight the brittleness of current multi-view aggregation methods and underscore the need for more robust information-fusion mechanisms to ensure reliable perception in autonomous agents, particularly in robotic settings where accurate scene understanding directly impacts control, manipulation, and safety.

KEYWORDS

Autonomous Robots, Active Perception, Vision-Language Models, Multi-view Synthesis

ACM Reference Format:

Anonymous Author(s). 2026. More Views, More Problems? A Critical Analysis of Multi-View Aggregation for Agent Perception. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages.

1 INTRODUCTION

Understanding and describing visual scenes in natural language is a fundamental capability for autonomous agents, particularly in Human-Robot Collaboration [2, 14, 48] scenarios where shared understanding is essential for success. The emergence of powerful Vision-Language Models (VLMs) offers a promising pathway for equipping robots with this ability [3, 35], allowing them to ground high-level instructions in the physical world and translate complex visual data into human-understandable descriptions of task-relevant objects and actions [1, 12, 15, 29, 50].

However, a critical gap exists between the static, single-image benchmarks on which these models are typically trained and the dynamic, multi-perspective reality of an *embodied agent* [18, 25, 39]. In the physical world, a single viewpoint is often insufficient due to occlusion, ambiguous angles, or poor lighting [7, 19, 24]. For instance, a wrench viewed from the side may appear indistinguishable from a metallic rod, a screwdriver's handle might obscure its tip, and reflective or transparent surfaces can distort appearance under certain lighting conditions. Such cases highlight how relying on a single perspective can lead to incomplete or misleading representations. While an agent's embodiment (i.e., its physical presence in the world) grants it the ability to actively gather information by moving, it remains unclear how this capability can be best leveraged for scene description. This raises fundamental questions about which are the most effective strategies for active perception and information integration.

In this paper, we reframe the problem of object description from a passive captioning task to an active, agent-driven process. We systematically evaluate distinct perceptual strategies, representing a gradient of agent complexity, to determine how an agent should best use its mobility and reasoning capabilities. These strategies range from relying on a fixed, canonical view, analogous to a static sensor; to actively seeking an optimal viewpoint, a simple but effective active perception strategy; and finally to intelligently synthesizing information from all available views into a single, holistic description generated by a Large Language Model (LLM). To stress-test these strategies, we evaluate them on two object sets: real-world tools and 3D-printed tools, which introduce a significant domain shift in texture and material properties. Our experimental pipeline is illustrated in Figure 1. This investigation allows us to answer the following research questions:

(1) To what extent can a mobile agent improve its descriptive accuracy by actively seeking an optimal viewpoint, compared to relying on a fixed, canonical perspective?
(2) Can an agent achieve superior descriptive performance by intelligently synthesizing information from multiple viewpoints, compared to simply identifying and using the single most informative view?
(3) How do these different perceptual strategies affect an agent's robustness when faced with a significant domain shift (i.e., from real-world to 3D-printed objects)?

Our work provides clear, evidence-based answers to these questions, offering valuable insights into the design of more capable and robust autonomous systems that can effectively perceive and describe the world around them.

2 RELATED WORK

2.1 Vision-Language Models for Captioning

Vision-language models (VLMs) have made significant advances in generating image captions and grounded language. Early systems

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). This work is licenced under the Creative Commons Attribution 4.0 International (CC-BY 4.0) licence.

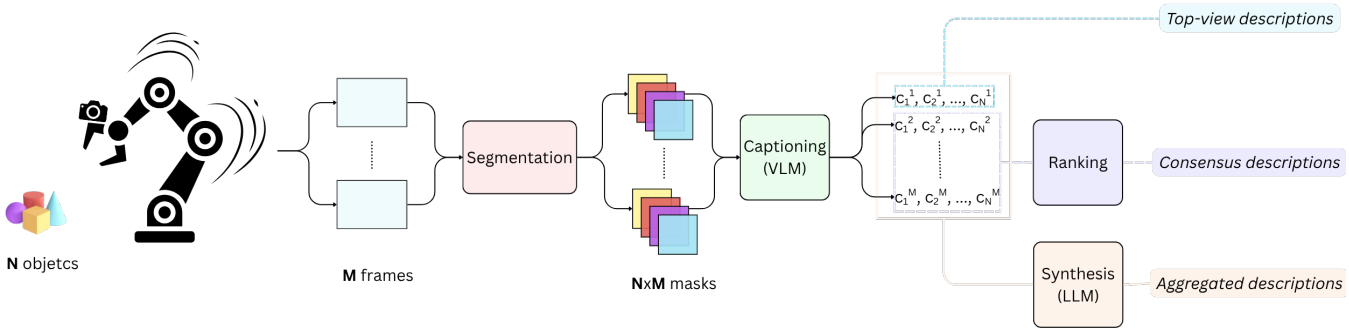


Figure 1: System architecture used to evaluate object descriptions from multiple viewpoints. From left to right, the pipeline begins by segmenting the objects in every single frame, which isolates the objects from each input view. Each segmented object is then passed to a VLM that produces candidate textual descriptions for every view. These per-view captions are evaluated from a single point of view, by ranking the different points of view, and parallelly processed by a LLM that synthesizes a coherent unique description. The outputs include top-view, consensus (i.e., ranking), and aggregated descriptions, reflecting alternative strategies for active perception. Together, these components enable an autonomous agent to interpret physical objects through vision–language reasoning and to potentially derive more robust descriptions.

like Show and Tell [43] and Show, Attend and Tell [44] used CNN-RNN architectures trained end-to-end. More recent transformer-based models such as BLIP/BLIP2 [16, 17] and Flamingo [3] achieve state-of-the-art performance by leveraging large-scale vision-language pretraining. CLIP [33] learns image-text alignment via contrastive learning and has been used in zero-shot recognition tasks. GPT-4V [30] and similar multimodal large language models (MLLMs) now enable complex reasoning and naturalistic captioning from images, even in few-shot or zero-shot settings. However, despite their impressive performance on benchmark datasets, the robustness of these models to real-world domain shifts (such as variations in texture, material, and lighting) remains a significant challenge.

2.2 Captioning in Robotics

In robotic applications, language grounding enables agents to interpret instructions, describe their environment, and communicate intent. Benchmarks like ALFRED [37] evaluate language understanding in interactive settings. Research has also integrated language models into manipulation tasks, such as SayCan [1], which connects language to robotic affordances, or LLM-Grounder [45], which maps natural language to object references in the scene. However, most captioning efforts assume a single fixed camera view or do not investigate the variations introduced by observing and captioning from multiple angles.

2.3 Multi-View Perception

Multi-view approaches are common in 3D reconstruction and object recognition. Su et al. [38] introduced Multi-View CNNs for object classification, showing that combining images from different viewpoints improves accuracy. NeRF-based methods [27] reconstruct 3D scenes from multiple views, but are not optimised for natural language output. In captioning, multi-image input has been explored in contexts such as photo albums or video summarization [46], but less so in robotics, where multiple works have focused on grasping [8, 21] and manipulation [32]. Despite success in recognition and reconstruction, multi-view methods have rarely been coupled with language output, leaving open questions about how viewpoint aggregation affects semantic grounding.

2.4 Summary

Our work addresses the under-explored problem of multi-view captioning for robotic tabletop perception. Unlike prior work that assumes single fixed viewpoints or focuses on 3D reconstruction without language output, we systematically compare single- and multi-view pipelines across multiple VLMs and domains, providing the first comprehensive analysis of how viewpoint diversity affects caption quality and grounding accuracy for robotic agents.

3 METHODOLOGY

In this section, we formalize the four agent-based strategies evaluated in our study. Our methodology is designed to create a clear, comparative analysis of how an agent's perceptual strategy impacts its ability to describe objects. We begin by defining the core components of our experimental setup, then detail the calculation of per-object scores for each strategy, and conclude by specifying how these scores are aggregated into the final dataset-level metrics used for our analysis.

3.1 Preliminaries and Notation

Let O be the set of all objects used in our evaluation with cardinality $|O| = N$. This set is partitioned into two disjoint subsets, O_{real} and O_{3D}, containing real-world tools and 3D-printed items, respectively. This partitioning allows us to assess the robustness of each strategy to the domain shift between real and synthetic-style objects in the real world. For each set of objects O_x, we capture a set of images V_o ($|V_o| = M$) to simulate different agent capabilities: (1) A single image from a fixed, top-down perspective, denoted v_{top} or v_0. This represents the data available to a static agent. (2) A

set of $M-1$ images acquired from moving through a 180 degrees hemisphere, denoted $\{v_1, v_2, \ldots, v_{M-1}\}$. This represents the richer visual data available to a mobile agent with active perception capabilities. Our framework utilizes two distinct foundation models. A Vision-Language Model, $\mathcal{M}_{VLM}$, serves as the agent's core perception module, generating a caption for a given image. A Large Language Model, $\mathcal{M}_{LLM}$, serves as an aggregation module for our synthesis strategy. The quality of any generated caption c is measured against a ground-truth description c_o^* using an evaluation metric $S(c, c_o^*)$, which returns a scalar score.

3.2 Per-Object Evaluation Strategies

We evaluate four strategies that represent a gradient of increasing agent complexity, from a static baseline to an intelligent synthesizer. For each object $o \in O$, we compute a score for each strategy.

(1) **Static Canonical View ($S_{\text{top}}(o)$):** This strategy serves as our most fundamental baseline, modelling an agent with a fixed sensor perpendicular to the table, a common setup in industrial robotics. The performance is the score of the caption $c_{top} = \mathcal{M}_{VLM}(v_{top})$ generated from the single top-down view:

$$S_{\text{top}}(o) = S(c_{top}, c_o^*). \tag{1}$$

(2) **Best Consensus View ($S_{\text{best}}(o)$):** This strategy models a sophisticated form of active perception, where an agent seeks the single most informative alternative viewpoint to the top-view. Simply taking the maximum score for each metric is problematic, as the "best" view can be ambiguous, as one view might yield a high score on a specific metric while another excels on a different one. Therefore, simply selecting the view with the maximum score for each metric independently would result in an inconsistent and moving baseline, making a fair comparison with the single aggregated caption impossible. Our consensus approach establishes a single, stable 'best-view' caption against which all metrics can be fairly evaluated. To create a robust and fair baseline, we first identify a single best consensus caption ($c_{\text{consensus}}$) for each object. This is achieved through a ranking-and-summation procedure: for each of our scores S based on different evaluation metrics, all $M-1$ candidate captions (one per view excluding v_{top}) are ranked from 1 to $M-1$ based on their scores. These ranks are then summed for each caption, yielding a total consensus score. The caption with the lowest total score (indicating the most consistently high performance across all metrics) is selected as $c_{\text{consensus}}(o)$. The performance for this baseline is then the score of this specific caption:

$$S_{\text{best}}(o) = S(c_{\text{consensus}}(o), c_o^*). \tag{2}$$

where $c_{\text{consensus}}(o)$ is the pre-selected consensus caption for object o. This ensures we compare our aggregated caption against a strong, stable, and unambiguously chosen baseline across metrics that evaluate different properties. This scenario assumes the existence of an oracle that can identify the optimal viewpoint - or equivalently, the most accurate caption. However, such ground-truth supervision is not always available, necessitating alternative strategies for selecting or generating captions that maximize informational content and align with human annotations.

(3) **Average Single-View ($S_{\text{avg}}(o)$):** This strategy quantifies the expected performance of an agent that gathers multi-view data but processes it naively. It answers whether, on average, more data is better than a single canonical view. The score is the mean performance across all N viewpoints from the set V_o:

$$S_{\text{avg}}(o) = \frac{1}{M} \sum_{i=0}^{M-1} S(\mathcal{M}_{VLM}(v_i), c_o^*). \tag{3}$$

(4) **Multi-View Synthesis ($S_{\text{agg}}(o)$):** Our proposed strategy, representing an intelligent agent that can reason over and synthesize information. This tests for synergy, i.e., whether a holistic approach through an LLM can produce captions that are as accurate as the best single perspective. A set of captions $C_o = \{\mathcal{M}_{VLM}(v_i)\}_{i=0}^{M-1}$ is generated from all views. The LLM then synthesizes these into a single caption, $c_{\text{agg}} = \mathcal{M}_{LLM}(C_o)$. The score is:

$$S_{\text{agg}}(o) = S(c_{\text{agg}}, c_o^*). \tag{4}$$

3.3 Dataset-Level Metrics

To provide a robust comparison of the strategies' overall effectiveness, we average the per-object scores across all objects within a given category. This aggregation mitigates the influence of outliers and provides a clear summary of performance. For the set of real objects O_{x}, the final metrics are:

$$S_{\text{y}}^{\text{x}} = \frac{1}{|O_{\text{x}}|} \sum_{o \in O_{\text{x}}} S_{\text{y}}(o), \tag{5}$$

where $x \in \{\text{real}, \text{3D}\}$ corresponds to our set of objects and $y \in \{\text{top}, \text{avg}, \text{best}, \text{agg}\}$ is our condition.

4 EXPERIMENTAL SETUP

4.1 Objects

For our experiments, the objects were grouped into two distinct sets, (O_{real} and O_{3D}). The first set comprises real, functional tools such as a digital multimeter, wire stripper, various wrenches, tweezers, pliers with insulated handles, and a hammer with a wooden handle. These items, commonly found in industrial and domestic contexts, are constructed from authentic materials including metal, wood, and rubber. The second set, by contrast, consists of 3D-printed plastic replicas of everyday household and workshop items, such as different types of bolts (hexagonal, wing, flange), a screwdriver, a hammer, a pan, and an adjustable wrench. While simplified and fabricated from uniform plastic, these replicas preserve the overall shape of their real-world counterparts and represent typical tools and utensils encountered in practice. Figure 2 shows all the objects used in our experiments. A ground-truth caption for each object was produced manually by a human annotator, and the lists of captions is provided in the supplemental material.

Keeping these two distinct object sets provides important advantages. It allows us to test whether VLMs - typically trained on internet-scale image-text data and biased toward realistic photographs - can generalize across variations in material properties,

(a) Real objects

(b) 3D printed objects

Figure 2: Our test objects. Both sets (real and 3D) are composed by 10 objects that differ over at least one property (e.g., color, shape).

textures, lighting conditions, and domains. This includes recognizing familiar object categories when presented in atypical forms. Although the exact composition of foundation model training datasets is not publicly available, it is reasonable to assume that textureless, 3D-printed objects are rarely, if ever, included. Consequently, this second set introduces a domain shift and can serve as a proxy for Out-Of-Distribution (OOD) objects, offering a valuable benchmark for evaluating model robustness and real-world transferability.

4.2 Data collection

To acquire multi-view images of tabletop scenes, we employ a Franka Emika Research 3 (FR3) robotic arm fitted with an Intel RealSense D435i RGB-D camera mounted on its end effector. The camera records RGB video data, while the robot's pose is controlled and logged via the `panda-py` library. Our experimental setup is illustrated in Figure 3.

Figure 3: Our experimental setup: the Franka Emika Research 3 is equipped with a Intel RealSense D435i RGB-D camera on its gripper. It collects pictures from several points of view by moving over a hemispherical trajectory.

We define eight equally spaced viewpoints arranged in a hemispherical pattern above the scene, simulating a top-view circular inspection trajectory, as done in [13]. Each viewpoint is specified by a 4×4 transformation matrix. During data collection, the robot sequentially visits these poses. The recorded video is then downsampled from 30 frames per second (FPS) to 1 FPS to reduce processing overhead, yielding a total of 50 images per set.

We employ the Segment Anything Model 2 (SAM2) [34], a state-of-the-art promptable segmentation framework from Meta AI, for both image and video segmentation. In our setup, SAM2 is used in automatic mode to generate dense object-level masks of tabletop scenes without manual prompting. For video, we leverage its zero-shot transfer capability by initializing object masks on the first frame and propagating them across subsequent frames to maintain temporal consistency. These masks serve both for visualization and as a preprocessing step for vision-language captioning, ensuring that captions focus only on relevant object regions. After propagation, we further downsample the image set, yielding 10 representative views per object for downstream processing, which correspond to v_{top} and V_o.

4.3 Captioning

To generate natural language descriptions of the objects $C_o = \{\mathcal{M}_{VLM}(v_i)\}_{i=0}^{M-1}$ for each $o \in O$, we evaluated a range of state-of-the-art VLMs: BLIP-2 (2.7B) [17], SmolVLM 2 (2.2B) [26], Gemma 3n (4B) [10], Gemma 3 (27B) [40], Mistral Small 3.1 (24B) [28], LLaVA-1.6 (34B) [22, 23], Qwen 2.5 VL (72B) [5], LLaMA-Vision 3.2 (90B) [11], LLaMA 4 (16×17B), Gemini 2.5 Flash-Lite [41], and GPT-4.1 [30]. We used publicly available implementations whenever possible to allow reproducibility of our results. BLIP-2 and SmolVLM 2 were run using the Hugging Face Transformers library[1], while all other models were run using Ollama[2] or their respective API. Model selection was guided by resource constraints: all locally hosted models were required to fit on an NVIDIA A100 GPU (80 GB VRAM), while Gemini 2.5 Flash-Lite and GPT-4.1 were chosen for their limited free API access and strong performance among proprietary models.

All of the models have their prompt set as follows:

Describe the object in the image using only its name and key visual characteristics. Answer in single words only. Provide a direct, concise description. For example, if the image contains a football, say: "white football ball with black patches".

4.4 Aggregation

Our synthesis strategy leverages a powerful LLM (GPT 4.1) to act as an intelligent aggregator. For each object, the set of captions generated from multiple viewpoints is passed to the LLM. The model is then prompted to perform a selection and synthesis task to create the aggregated caption $c_{\text{agg}} = \mathcal{M}_{LLM}(C_o)$. Specifically, it is instructed to evaluate the varied, and often noisy or conflicting, descriptions to identify the most salient and accurate information. Based on this evaluation, the LLM generates a single, cohesive description by either synthesising details from multiple inputs or, as a practical heuristic, by adopting a single input caption if it is judged to be singularly comprehensive and superior to a synthesised alternative. The final text generated by the LLM represents the

[1] https://huggingface.co
[2] https://ollama.com

aggregated description, which is then evaluated against the ground truth. The prompt for the LLM is the following:

You are given multiple captions of the same object from different views. Some may be wrong, generic, or describe irrelevant objects. Prefer captions that are specific, plausible, and focused on the main object. Consistency across captions is useful, but do not follow the majority if they are vague or incorrect. If one caption clearly gives a better description of the main object, use it. Write one concise caption of the main object only. No quotes, line breaks, or commentary.

4.5 Evaluation

To assess the quality of captions generated by the VLMs, we use several existing evaluation metrics, which represents the scores indicated in Equations (1) to (5). BLEU [31] measures n-gram precision against reference texts, while ROUGE [20] focuses on recall of overlapping n-grams, but both may miss semantic nuances. METEOR [6] combines precision, recall, and synonym matching, making it more robust to linguistic variations in machine translation and captioning. CIDEr [42] weights n-grams by consensus across multiple references, capturing commonly agreed-upon descriptions. Standard CIDEr assumes multiple references per image, and its TF-IDF weighting uses the logarithm of the reference count; for single-reference images, this leads to zero weights and meaningless scores. To address this, we replace the reference count with $max(2, |\text{refs}|)$, ensuring positive TF-IDF weights while leaving multi-reference cases unchanged. Additionally, we modify the unigram length calculation by setting $n = 0$ when counting words for the Gaussian length penalty, so that CIDEr-1 correctly reflects differences in short captions; without this fix, missing words in short captions would be ignored, inflating scores. Finally, all other steps – Gaussian length penalty parameter, including cosine similarity, n-gram averaging, and CIDEr-M for varying-length candidates – remain identical to standard CIDEr. SPICE [4] evaluates scene-graph similarity, focusing on objects, attributes, and relations, and aligns closely with human judgment for detailed descriptions. BERTScore [49] uses contextual embeddings to measure semantic similarity beyond surface-level word overlap, implemented here with `deberta-xlarge-mnli`. BLEURT [36] (implemented with `BLEURT-20-D12`) leverages pre-trained language models fine-tuned on human judgments to predict text quality directly, BARTScore [47] uses the likelihood of text under a BART model (`bart-large-cnn`) for reference-based and reference-free evaluation, and GPTScore [9] employs a large language model (`gpt2-xl`) to assign quality scores by assessing relevance, coherence, and alignment with references.

By combining multiple metrics, we get a more complete view of model performance (balancing lexical accuracy, semantic relevance, fluency, and diversity) and reduce the risk of overfitting to the characteristics of any one metric. This approach is particularly important when evaluating open-ended tasks like captioning or text generation, where there are many valid outputs. It is worth noting that all these metrics range from 0 to 1, except for CIDEr in {0, 10}, and BARTScore and GPTScore have negative values. For simplicity, we will refer to BERTScore and BARTScore as BERT and BART respectively, but we will be maintaing the nomenclature GPTScore to avoid confusion with the GPT-4.1 model.

5 EXPERIMENTS & RESULTS

5.1 Top-view

Our empirical investigation was designed to quantify the robustness of modern VLMs to a physical domain shift, a common challenge for embodied robotic agents, and to create a baseline to evaluate alternative strategies. By comparing the model performance on a set of real-world, multi-material tools against a visually distinct set of single-material, 3D-printed objects, we can precisely measure the impact of this shift. The evaluation, combining quantitative metrics with a qualitative analysis, reveals critical limitations in current state-of-the-art models and metrics.

Given $S_{\text{top}}^{\text{real}}$ and $S_{\text{top}}^{\text{3D}}$, the scores calculated from the top-view over all of the objects in the respective sets, we calculate ΔS as an indicator of the difference in performance between the two sets as follows:

$$\Delta S_{\text{top}} = S_{\text{top}}^{\text{real}} - S_{\text{top}}^{\text{3D}} \tag{6}$$

Table 1 illustrates a summary of the findings of our first experiment. On the real objects, we can see some examples of how n-gram metrics such as BLEU and ROUGE capture lexical overlap. BLIP-2 generates “a metal object with a long, curved handle” for “metal tweezers”, yielding BLEU = 0.125, reflecting low word overlap despite accurate semantics. Gemma3n outputs “Blue and gray wire stripper” for “grey and blue wire stripper”, achieving BLEU = 0.8, since “gray” and “grey” do not overlap. LLaVA produces “Pair of scissors with yellow and red handles” for “Combination pliers with red and yellow handles”, giving ROUGE-L=0.67, demonstrating how metrics that focus on lexical mismatch does not capture semantics. Semantic metrics better capture meaning. Llama 3.2 generates “The object in the image is a pair of pliers with yellow handles and silver metal jaws, positioned” for “Needle-nose pliers with red and yellow handles”. BLEU is low (0.2778), but METEOR, which reflects also semantic similarity, achieves a modest score (0.5506). Similarly, SPICE score is 0.8571 for “blue wire stripper” vs “grey and blue wire stripper” generated by Qwen, while the caption “blue and gray pliers” generated by SmolVLM2 scores 0.0 , highlighting semantic mislabeling. Embedding-based metrics such as BERTScore, BLEURT, and BARTScore are more robust to paraphrasing but occasionally inconsistent. For example, LLaVA receives BERTScore = 0.6391 for “scissors” vs “pliers”, overestimating similarity, while BARTScore ranges from -3.346 for Llama 3.2’s long description of a screwdriver to -8.107 for Gemini’s concise “Hex key”, showing sensitivity to brevity and lack of fluency. For 3D-printed objects, lexical metrics decline due to simplified visual cues and reduced texture detail captured by VLMs. SmolVLM2 produces “green plastic handle” for the object “plastic green hammer” (BLEU = 0.6667) and “black metal bolt“ for “plastic black hexagonal bolt” (BLEU = 0.4777), demonstrating moderate overlap but low semantic accuracy. BLIP2 scores ROUGE-L = 0.1667 for “plastic green hexagonal bolt” vs “a green threaded screw on a black background”, while Gemma3n achieves ROUGE-L = 0.6667 for “plastic black cast iron pan” vs “Black cast iron skillet”, capturing lexical similarity but not precise semantics. METEOR and SPICE scores remain moderately informative: BLIP2 achieves METEOR = 0.2941 for “a green plastic tool with a handle” vs “plastic green hammer”, SmolVLM-2 produces a CIDEr = 6.689 for “plastic green wing nut”, and SPICE = 0.8 for Qwen’s “green

VLM	METEOR			CIDEr			SPICE			GPTScore		
	Real	3D	ΔS	Real	3D	ΔS	Real	3D	ΔS	Real	3D	ΔS
SmolVLM2 (2.2B)	0.21 ± 0.21	0.19 ± 0.09	0.02	4.19 ± 2.78	4.64 ± 2.00	-0.45	0.23 ± 0.25	0.19 ± 0.25	0.04	-8.77 ± 2.75	-8.60 ± 1.39	-0.17
BLIP2 (2.7B)	**0.39 ± 0.24**	0.22 ± 0.13	0.17	3.27 ± 2.50	2.12 ± 0.94	1.15	**0.35 ± 0.32**	0.07 ± 0.13	0.28	-3.50 ± 0.71	-4.44 ± 0.80	0.94
Gemma 3n (4B)	0.37 ± 0.32	0.25 ± 0.16	0.12	3.88 ± 2.74	3.51 ± 1.92	0.37	0.35 ± 0.33	0.08 ± 0.18	0.27	-4.99 ± 1.57	-5.16 ± 0.83	0.17
Mistral 3.1 (24B)	0.31 ± 0.20	0.23 ± 0.17	0.08	2.37 ± 1.59	3.42 ± 2.19	-1.05	0.17 ± 0.24	0.12 ± 0.25	0.05	-5.24 ± 2.10	-5.67 ± 1.27	0.43
Gemma 3 (27B)	0.35 ± 0.33	0.21 ± 0.14	0.14	4.40 ± 3.62	3.73 ± 1.90	0.67	0.30 ± 0.38	0.13 ± 0.27	0.17	-5.61 ± 2.28	-6.33 ± 0.96	0.72
LLaVA 1.6 (34B)	0.18 ± 0.13	0.12 ± 0.08	0.06	2.08 ± 1.91	2.03 ± 1.74	0.05	0.10 ± 0.12	0.00 ± 0.00	0.1	-4.76 ± 1.20	-5.92 ± 1.72	1.16
Qwen 2.5 VL (72B)	0.39 ± 0.27	**0.32 ± 0.20**	0.07	**5.57 ± 2.69**	**5.15 ± 2.26**	0.42	0.33 ± 0.40	**0.33 ± 0.35**	0	-4.82 ± 1.77	-5.79 ± 0.99	0.97
Llama 3.2 (90B)	0.26 ± 0.26	0.24 ± 0.11	0.02	1.15 ± 2.06	0.69 ± 0.78	0.46	0.21 ± 0.24	0.18 ± 0.18	0.03	**-3.50 ± 0.65**	**-3.61 ± 0.41**	0.11
Llama 4 (17B x 16)	0.20 ± 0.27	0.23 ± 0.14	-0.03	3.24 ± 2.71	4.43 ± 2.06	-1.19	0.21 ± 0.31	0.13 ± 0.27	0.08	-6.37 ± 1.70	-7.22 ± 1.57	0.85
Gemini 2.5 FL	0.35 ± 0.30	0.28 ± 0.19	0.07	4.62 ± 2.51	4.36 ± 2.49	0.26	**0.46 ± 0.30**	0.27 ± 0.34	0.19	-4.13 ± 1.16	-6.28 ± 1.66	2.15
GPT-4.1	**0.58 ± 0.31**	**0.40 ± 0.19**	0.18	**6.58 ± 2.79**	**6.54 ± 2.15**	0.04	0.45 ± 0.38	**0.38 ± 0.41**	0.07	-4.19 ± 2.01	-5.58 ± 1.09	1.39

Table 1: Summary of captioning performance across multiple VLMs on the real and 3D object datasets. Overall, the results indicate that generating human-like captions is generally easier for real objects than for 3D-printed objects across most models over multiple metrics.

screwdriver" vs "plastic green screwdriver". GPTScore evaluations of other VLMs reveal fluency biases: SmolVLM-2 scores -11.0 for the terse caption "hammer", and LLaVA -3.897 for a long 3D object description, reflecting the metric's preference for detailed, fluent phrasing.

Among closed models, Gemini produces concise captions that are lexically precise but can be penalized by surface-based metrics. For instance, "Allen hexagonal wrench" generated as "Hex key" yields BARTScore = -8.1067, demonstrating brevity-driven penalties. In contrast, OpenAI GPT VLM produces longer, fluent captions with high semantic fidelity. For example, "Needle-nose pliers with red and yellow handles" is generated as "needle-nose pliers with yellow and red handles", achieving CIDEr = 10, indicating near-perfect alignment in unigrams, while SPICE and METEOR reflect also strong semantic coverage. These observations highlight a key trade-off: Gemini favors brevity and lexical precision, whereas GPT prioritizes semantic richness and fluent, descriptive phrasing.

Overall, lexical metrics like BLEU and ROUGE underestimate the quality of paraphrased or fluent captions, while semantic metrics (METEOR, SPICE, embedding-based scores) better capture meaning but can overvalue verbosity. GPTScore evaluations further illustrate the relevance of fluency, penalizing overly short captions even when semantically correct. From these examples, no single metrics manages to capture similarity on their own.

The gap between real and 3D-printed objects underscores a persistent visual bias: models correctly identify object categories but often omit material or functional cues in simplified 3D representations. A more comprehensive evaluation should therefore combine lexical precision, semantic fidelity, and fluency metrics to capture both descriptive accuracy and linguistic quality. For the following experiments, we select the two best performing models from each category (open- vs closed- source), Qwen 2.5 VL and GPT 4.1.

5.2 Multiple views

In our second experiment, we examine whether leveraging multiple viewpoints leads to more accurate object descriptions. For each object $o \in O_x$, where $x \in \{\text{real}, \text{3D}\}$, we compute our scores from the top view $S_{\text{top}}(o)$ (from Equation (1)) and compare it with the consensus score $S_{\text{best}}(o)$ (from Equation (2)), assessing whether a particular viewpoint provides more informative descriptions that better align with human annotations.

Figure 4 reports the percentage of objects for which the consensus caption scores higher (>) or equal (≥) to the top-view caption ($S_{best} \geq S_{top}$) across all evaluation metrics. On the real set, Qwen 2.5 VL shows consistent improvements, with strictly higher scores for 30-70% of the objects and equal or better scores in nearly all cases (up to 100%). GPT-4.1 follows a similar trend but with slightly smaller margins, reaching 40-60% of higher scores and 80-100% when including ties. On the 3D set, both models exhibit a mild drop in strictly higher cases, particularly for GPT-4.1, which "All" column decreases from 30% (real) to 10% (3D). Qwen 2.5 VL remains more robust, maintaining 50-60% improvements across most metrics. Overall, these results suggest that considering multiple points of view provides broader benefits for Qwen 2.5 VL, while GPT-4.1, which already achieves higher single-view scores, shows more limited potential for further gains.

When comparing top-view captions to consensus captions, the improvements are strongly object-dependent and can be quantified across metrics. For instance, for the metal tweezers, the reference caption is "metal tweezers". Qwen's top-view caption, "metal tongs", achieved BLEU-1=0.50, ROUGE-1=0.50, METEOR=0.25, and SPICE=0, reflecting only partial lexical alignment. The consensus caption, "metal tweezers", significantly improved all these metrics: BLEU-1 rose to 1.0, ROUGE-1 to 1.0, METEOR to 0.94, and SPICE to 1.0, while semantic metrics like BERT and BLEURT also increased to 1.0 and 0.95 respectively. Similarly, GPT-4.1's top-view caption "metal yarn needle" initially had BLEU 0.33, ROUGE-L 0.40, and METEOR 0.24, but the consensus caption "metal tweezers" brought BLEU to 1.0, ROUGE-L to 1.0, METEOR to 0.94, and SPICE to 1.0. For objects where the top-view caption is partially descriptive, we observe moderate but meaningful gains. For example, for the object "orange and black digital multimeter", GPT-4.1's top-view caption, "digital multimeter", had BLEU=0.22, ROUGE-1=0.57, METEOR=0.40, and SPICE=0.67. The consensus caption, "orange digital multimeter", increased BLEU to 0.51, ROUGE-1 to 0.75, METEOR to 0.53, and SPICE to 0.86, showing clear lexical and semantic improvements even though the top-view already captured part of the object.

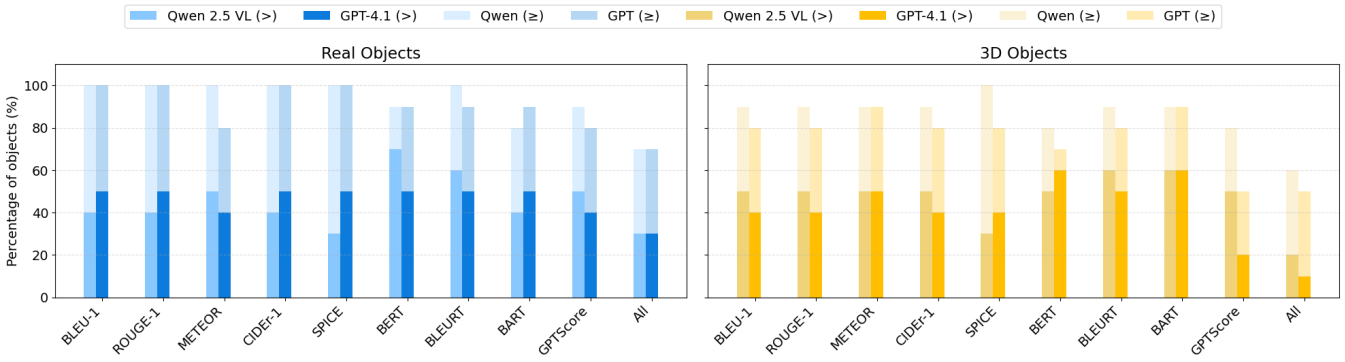


Figure 4: Percentage of objects for which the consensus caption scores higher (>) or equal (≥) than the top-view caption over each single metric. The last group (All) indicates the percentage of objects for which the consensus caption improves simultaneously over all the scores.

In contrast, objects with top-view captions that are highly mismatched to the reference show smaller improvements. For "Allen hexagonal wrench", Qwen's produced top-view caption "blue L-shaped tool" had BLEU, ROUGE-1, and METEOR all equal to 0, while the consensus caption "blue hex key" increased semantic metrics such as BERT (from 0.08 to 0.33) and BLEURT (0.10 to 0.19), showing a better semantic alignment, while lexical metrics remained 0, reflecting minimal lexical alignment. For 3D objects, consensus captions also consistently improved metrics over top-view captions. Qwen's "plastic white adjustable wrench" had a top-view caption "adjustable wrench" with BLEU-1=0.37, ROUGE-1=0.67, METEOR=0.49, and SPICE=0.67. The consensus caption "white adjustable wrench" increased BLEU-1 to 0.72, ROUGE-1 to 0.86, METEOR to 0.75, and SPICE to 0.86, showing improvements in both lexical and semantic alignment. Similarly, GPT-4.1's "plastic black wing nut" had a top-view caption "black plastic knob" with BLEU-1=0.48, ROUGE-1=0.57, METEOR=0.26, and SPICE=0. The consensus caption "black wing nut" raised BLEU-1 to 0.72, ROUGE-1 to 0.86, METEOR to 0.75, and SPICE to 0.61, improving semantic and n-gram alignment.

Overall, our results shows that consensus captions can improve metrics over top-view captions, especially when the top-view is partially accurate. BLEU-1 and ROUGE-1 gains reflect better lexical overlap with the reference, while METEOR and SPICE capture semantic improvements. When top-view captions are largely inaccurate, improvements are modest and mostly seen in semantic metrics, like BERT and BLEURT, rather than lexical scores.

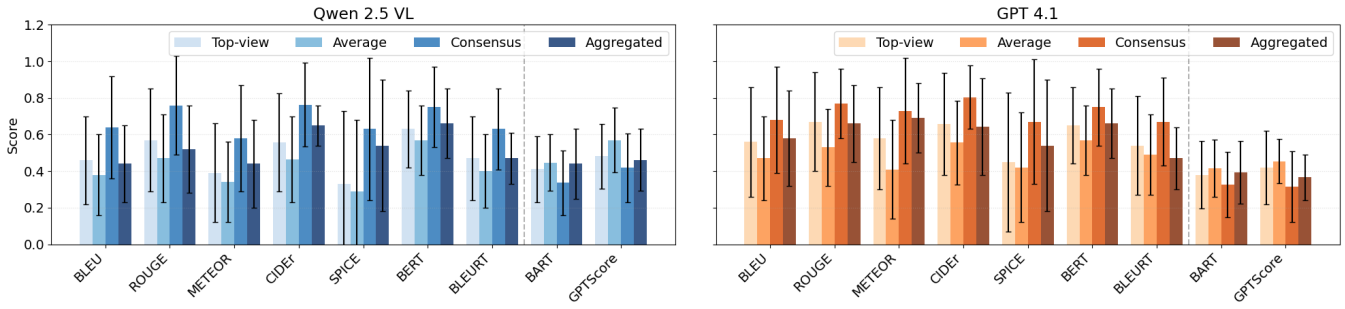
To conclude, Figure 5 illustrates the comparison between the three conditions: top-view S_{top}, average S_{avg}, and consensus S_{cons}. It is evident that simply averaging the scores across all views is the least effective strategy, as it often performs worse than using a single top-view caption, particularly for GPT-4.1 on real objects. When models already describe the object accurately from the top-view, incorporating alternative views that produce different captions appears to introduce noise, which in turn reduces the evaluation scores. This suggests that careful aggregation or consensus mechanisms are preferable to naive averaging when combining multiple perspectives.

5.3 Aggregation

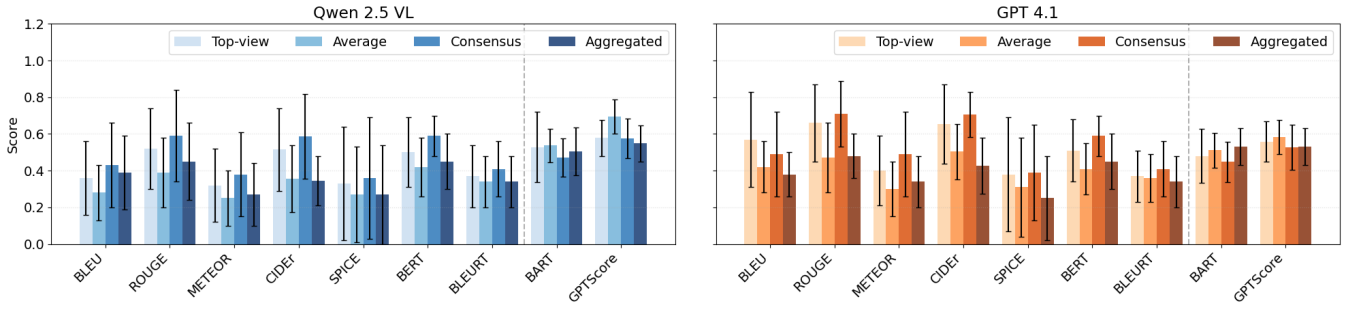
When comparing our three previous conditions to the aggregated scores S_{agg}, the picture shifts slightly. Aggregation generally produces lower values than the consensus, but often higher than the naive average, and on average on the same level as the top-view. For example, in the real dataset, GPT-4.1 achieves an average BLEU-1 of 0.58 on the aggregated captions, which is below the average consensus score of 0.73 but above the naive average of 0.47. Similarly, GPT-4.1 aggregates METEOR=0.69, higher than the top-view score (0.58) and the naive average score (0.48), but slightly below the consensus score (0.72). Aggregated scores reflect a summarization over multiple frames, which smooths out peaks and reduces variability (lower standard deviations in some cases), while still capturing stronger performance than naive averaging.

Qualitative examples illustrate how aggregation can either refine or distort the semantic content of the caption. On real objects, GPT-4.1 correctly refines "red and yellow pliers" into the more precise "Red and yellow insulated combination pliers," improving both lexical and semantic alignment with the ground truth ("Combination pliers with red and yellow handles"). Similarly, Qwen 2.5 VL transforms a noisy single-frame description ("yellow and red-handled wire cutters") into a more accurate "Diagonal wire cutters with yellow and red handles," capturing both shape and color more faithfully, but changing the main object. Complete failures also emerge: GPT-4.1's aggregation of "hex key" across views into "Black L-shaped hex key (Allen wrench)" slightly improves specificity but does not reach full alignment with the ground truth "Allen hexagonal wrench," suggesting that aggregation may add stylistic detail without semantic correction.

In the 3D dataset, the effect of aggregation is more pronounced. For example, Qwen 2.5 VL's "green screwdriver" becomes "Green or teal screwdriver," which introduces uncertainty rather than refinement, leading to reduced BLEU and ROUGE scores (0.5 vs. 0.61 and 0.57 vs. 0.8, respectively). GPT-4.1 shows a more severe case of semantic drift: its aggregation of across the different captions "black plastic knob, dark blue key, black flashlight, black ladle, black handheld rotary cutter, black handheld hair dryer, black three-blade boat propeller, dark three-bladed propeller, black circular object



(a) Real objects



(b) 3D objects

Figure 5: Caption. For visualization purposes, CIDEr has been rescaled by 10, and BARTScore and GPTScore have been rescaled by -10 (in this case lower is better). Raw values are reported in the supplemental material.

with two side handles, black wing nut" results in "Black three-blade boat propeller," completely diverging from the target concept and drastically lowering all metrics (e.g., BLEU 0.25 vs. 0.72). These examples aggregation can dilute strong single-frame cues, potentially merging inconsistent visual evidence across views.

Overall, aggregation provides a trade-off across frames, producing more stable, albeit occasionally less precise, descriptions. It mitigates variability and preserves salient features, yet it may fail to reach the peak accuracy attained by consensus captions, particularly in complex or ambiguous 3D scenarios where the object is not identified properly from multiple points of view. Nevertheless, aggregation constitutes a practical alternative in the absence of an oracle that can select the best scenario.

6 CONCLUSION

In this work, we conducted a systematic evaluation of modern VLMs, assessing their robustness under a physical domain shift and examining how active, multi-view perception can mitigate perceptual failures. Our analysis provides three main findings with direct implications for the design of autonomous agents.

First, we show that domain shifts (such as from real-world to 3D-printed objects) lead to a degradation in captioning performance across VLMs. This brittleness exposes a fundamental limitation: models trained predominantly on web-scale data remain poorly equipped to handle the visual novelty of unstructured environments.

Second, our results demonstrate the value of active perception. Moving from a fixed, top-down view to selecting an optimal viewpoint among multiple angles by using an oracle produces consistent and substantial gains. Conversely, naively averaging information across all views is detrimental, often underperforming a single view by allowing noisy or conflicting inputs to degrade the all the different scores. This highlights a key principle for agent design: robust perception requires selective and informed information fusion rather than indiscriminate data aggregation.

Finally, we evaluated LLM-based aggregation as one such fusion mechanism. While this approach improves over naive averaging and can refine object descriptions, it also introduces instability. Aggregation sometimes leads to severe semantic drift, misidentifying objects, particularly under domain shift. These findings suggest that while LLM-based reasoning over perceptual inputs is promising, current systems lack the grounding necessary for reliable deployment in safety-critical settings.

Our study is limited by the scale of the dataset, and future work should validate these findings across broader object categories and domain shifts. Nevertheless, these findings motivate the development of aggregation methods that explicitly account for uncertainty and selectively integrate reliable perceptual evidence. Based on our finding, future studies should focus on: finding the optimal strategy for selecting information across different viewpoints (via smarter prompts or reasoning models), integrating and developing robust metrics for similar scenarios and addressing the issue of domain-shift for VLMs.

REFERENCES

[1] Michael Ahn et al. 2022. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. *arXiv* (2022). arXiv:2204.01691 [cs.RO] https://arxiv.org/abs/2204.01691

[2] Arash Ajoudani, Andrea Maria Zanchettin, Serena Ivaldi, Alin Albu-Schäffer, Kazuhiro Kosuge, and Oussama Khatib. 2018. Progress and prospects of the human–robot collaboration. *Autonomous robots* 42, 5 (2018), 957–975.

[3] Jean-Baptiste Alayrac et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* (2022).

[4] Peter Anderson et al. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*.

[5] Shuai Bai et al. 2025. Qwen2.5-VL Technical Report. *arXiv* (2025). arXiv:2502.13923 [cs.CV] https://arxiv.org/abs/2502.13923

[6] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*.

[7] Shivam Chandhok and Pranav Tandon. 2024. Do Vision-Language Foundational models show Robust Visual Perception? *arXiv* (2024). arXiv:2408.06781 [cs.CV] https://arxiv.org/abs/2408.06781

[8] Daniele De Gregorio, Federico Tombari, and Luigi Di Stefano. 2016. RobotFusion: Grasping with a Robotic Manipulator via Multi-view Reconstruction. In *Computer Vision – ECCV 2016 Workshops*, Gang Hua and Hervé Jégou (Eds.). Springer International Publishing.

[9] Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as You Desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

[10] Google AI for Developers. [n.d.]. Gemma 3n model overview | Google AI for Developers - ai.google.dev. https://ai.google.dev/gemma/docs/gemma-3n. [Accessed 02-10-2025].

[11] Aaron Grattafiori et al. 2024. The Llama 3 Herd of Models. *arXiv* (2024). arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[12] Physical Intelligence et al. 2025. $\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization. arXiv:2504.16054 [cs.LG] https://arxiv.org/abs/2504.16054

[13] Ivan Kapelyukh, Yifei Ren, Ignacio Alzugaray, and Edward Johns. 2024. Dream2Real: Zero-Shot 3D Object Rearrangement with Vision-Language Models. In *IEEE International Conference on Robotics and Automation (ICRA)*.

[14] Ali Keshvarparast, Daria Battini, Olga Battaia, and Amir Pirayesh. 2024. Collaborative robots in manufacturing and assembly systems: literature review and future research agenda. *Journal of Intelligent Manufacturing* 35, 5 (2024), 2065–2118.

[15] Moo Jin Kim et al. 2024. OpenVLA: An Open-Source Vision-Language-Action Model. *arXiv preprint arXiv:2406.09246* (2024).

[16] Junnan Li et al. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR.

[17] Junnan Li et al. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR.

[18] Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Erran Li Li, Ruohan Zhang, et al. 2024. Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems* 37 (2024), 100428–100534.

[19] Yifan Li et al. 2023. Evaluating Object Hallucination in Large Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2023.emnlp-main.20

[20] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*.

[21] Huei-Yung Lin, Shih-Cheng Liang, and Yu-Kai Chen. 2021. Robotic Grasping With Multi-View Image Acquisition and Model-Based Pose Estimation. *IEEE Sensors Journal* (2021).

[22] Haotian Liu et al. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023).

[23] Haotian Liu et al. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[24] Hanqing Liu et al. 2025. When Lighting Deceives: Exposing Vision-Language Models' Illumination Vulnerability Through Illumination Transformation Attack. *arXiv* (2025). arXiv:2503.06903 [cs.CV] https://arxiv.org/abs/2503.06903

[25] Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. 2025. Aligning Cyber Space with Physical World: A Comprehensive Survey on Embodied AI. *IEEE/ASME Transactions on Mechatronics* (2025).

[26] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2025. SmolVLM: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299* (2025).

[27] Ben Mildenhall, , et al. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* (2021).

[28] Mistral AI. [n.d.]. Mistral Small 3.1 | Mistral AI - mistral.ai. https://mistral.ai/news/mistral-small-3-1. [Accessed 01-10-2025].

[29] NVIDIA et al. 2025. GR00T N1: An Open Foundation Model for Generalist Humanoid Robots. arXiv:2503.14734 [cs.RO] https://arxiv.org/abs/2503.14734

[30] OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

[31] Kishore Papineni et al. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics.

[32] Shengyi Qian et al. 2024. 3d-mvp: 3d multiview pretraining for robotic manipulation. *arXiv preprint arXiv:2406.18158* (2024).

[33] Alec Radford et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR.

[34] Nikhila Ravi et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714* (2024).

[35] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. 2022. A generalist agent. *arXiv preprint arXiv:2205.06175* (2022).

[36] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of ACL*.

[37] Mohit Shridhar et al. 2020. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

[38] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. 2015. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*. 945–953.

[39] Andrew Szot, Bogdan Mazoure, Omar Attia, Aleksei Timofeev, Harsh Agrawal, Devon Hjelm, Zhe Gan, Zsolt Kira, and Alexander Toshev. 2025. From Multimodal LLMs to Generalist Embodied Agents: Methods and Lessons. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10644–10655.

[40] Gemma Team, , et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786* (2025).

[41] Gemini Team et al. 2025. Gemini: A Family of Highly Capable Multimodal Models. *arXiv* (2025). arXiv:2312.11805 [cs.CL] https://arxiv.org/abs/2312.11805

[42] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-Based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[43] Oriol Vinyals et al. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[44] Kelvin Xu et al. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. PMLR.

[45] Jianing Yang et al. 2024. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE.

[46] Haonan Yu et al. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[47] Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating Generated Text as Text Generation. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 27263–27277. https://proceedings.neurips.cc/paper/2021/file/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Paper.pdf

[48] Muhammad Hamza Zafar, Even Falkenberg Langås, and Filippo Sanfilippo. 2024. Exploring the synergies between collaborative robotics, digital twins, augmentation, and industry 5.0 for smart manufacturing: A state-of-the-art review. *Robotics and Computer-Integrated Manufacturing* 89 (2024), 102769. https://doi.org/10.1016/j.rcim.2024.102769

[49] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SkeHuCVFDr

[50] Brianna Zitkovich et al. 2023. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *Proceedings of The 7th Conference on Robot Learning (Proceedings of Machine Learning Research, Vol. 229)*, Jie Tan, Marc Toussaint, and Kourosh Darvish (Eds.). PMLR, 2165–2183. https://proceedings.mlr.press/v229/zitkovich23a.html