PointArena: Probing Multimodal Grounding Through Language-Guided Pointing

Long Cheng[1*] **Jiafei Duan**[1,2*] **Yi Ru Wang**[1†] **Haoquan Fang**[1,2†] **Boyang Li**[1†]
Yushan Huang[1] **Elvis Wang**[3] **Ainaz Eftekhar**[1,2] **Jason Lee**[1,2] **Wentao Yuan**[1]
Rose Hendrix[2] **Noah A. Smith**[1,2] **Fei Xia**[1] **Dieter Fox**[1] **Ranjay Krishna**[1,2]

[1]University of Washington [2]Allen Institute for Artificial Intelligence
[3]Anderson Collegiate Vocational Institute
https://pointarena.github.io

Abstract

Pointing serves as a fundamental and intuitive mechanism for grounding language within visual contexts, with applications spanning robotics, assistive technologies, and interactive AI systems. While recent multimodal models have begun supporting pointing capabilities, existing benchmarks typically focus only on referential object localization. We introduce **PointArena**, a comprehensive platform for evaluating multimodal pointing across diverse reasoning scenarios. PointArena comprises three components: (1) **Point-Bench**, a curated dataset of approximately 1,000 pointing tasks across five reasoning categories; (2) **Point-Battle**, an interactive web-based arena facilitating blind, pairwise model comparisons, which has collected over 4,500 anonymized votes; and (3) **Point-Act**, a real-world robotic manipulation system allowing users to directly evaluate model pointing in practical settings. We conducted extensive evaluations of both state-of-the-art open-source and proprietary models. Results indicate that `Molmo-72B` consistently outperforms others, though proprietary models increasingly demonstrate comparable performance. Additionally, we find that supervised training targeting pointing tasks significantly improves performance. Across our multi-stage evaluation pipeline, we observe strong correlations, underscoring the critical role of precise pointing in enabling multimodal models to bridge abstract reasoning with real-world actions.

1 Introduction

Pointing focuses our attention. It is one of the earliest and most universal non-verbal methods we use to communicate intent; in fact, children learn to point as a prelinguistic form of communication [29]. Precise spatial grounding—*pointing*—enables a wide range of practical and high-impact applications across robotics, assistive technology, human-computer interaction, and vision-language interfaces. In robotics, a pointing-capable model can interpret language commands like "pick up the red cup next to the bowl" and translate them into precise spatial actions [37], enabling fine-grained object manipulation in cluttered environments [13]. In assistive technologies, systems can help visually impaired users by answering spatial queries such as "where is the handle on this door?" [6] or 'which one is the garlic?' In education or creative tools, pointing allows for interactive visual tutoring, such as identifying components in scientific diagrams or guiding a learner through a painting [16]. Even in everyday virtual assistants or search engines, the ability to refer to specific image regions via pointing could make multimodal interactions more intuitive and expressive [10]. Across these

[*]Co-first authors.
[†]Co-second authors.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

domains, pointing provides a low-bandwidth yet powerful spatial interface for grounding language in vision—precise enough for manipulation, intuitive enough for communication, and general enough to scale with modern multimodal models.

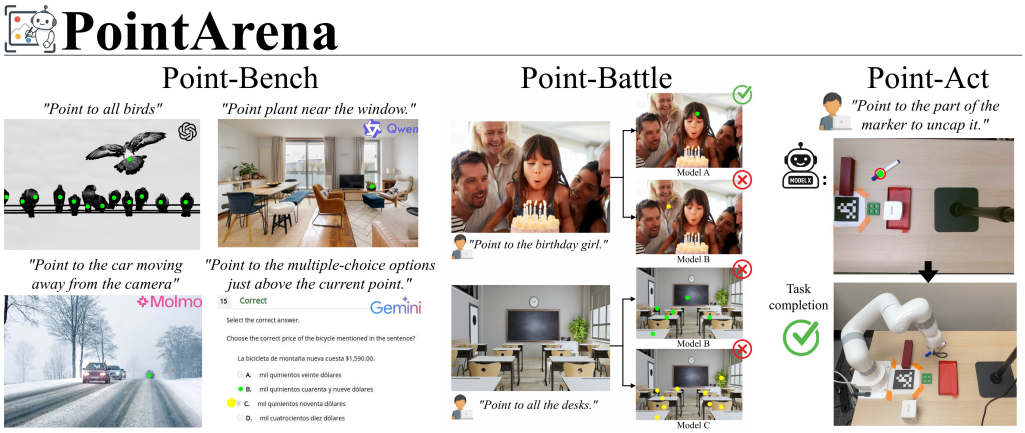

Figure 1: **Overview of *PointArena*.** PointArena consists of three components: **Point-Bench**, a curated dataset for evaluating grounded pointing across five reasoning types; **Point-Battle**, a live platform for blind, pairwise model comparisons with user voting; and **Point-Act**, real-world task involving manipulation via pointing-based language commands.

Recent advances in multimodal models have begun to incorporate more dynamic and spatially expressive forms of interaction. The Segment Anything Model (SAM) [19] enables segmentation from sparse visual prompts such as points or boxes, revealing the potential of fine-grained spatial control. Google's Gemini models [14] push the boundaries of long-context visual reasoning, incorporating multiple modalities over extended sequences. In parallel, new datasets have emerged to support explicit spatial referencing. Molmo's PixMo dataset [10] introduces 2D pointing as a form of multimodal alignment between images and instructions, while RoboPoint [37] focuses on spatial affordance prediction by linking instructions to interaction-relevant keypoints in robotic contexts. These setups bias evaluations toward pixel-level accuracy rather than conceptual reasoning and often lack diversity or scalability.

There is a need for a holistic evaluation platform to make progress towards language-guided pointing. Although datasets for referring expressions exist (e.g.RefCOCO, RefCOCO+, and RefCOCOg [17, 34]), they are focused on a subset of pointing tasks: object location. They lack the ambiguity and contextual variability that users expect from modern interactive models, limiting their utility for studying pragmatic or interactive applications. As a result, they offer only partial insight into the full spectrum of grounding required for embodied or assistive agents.

We, therefore, propose **PointArena**, a platform to probe and evaluate grounded visual reasoning with pointing. PointArena presents a suite of tasks where a multimodal model must answer questions or resolve instructions by combining language and pointing gestures to identify specific image regions. These tasks go beyond traditional VQA by requiring spatial outputs (e.g., selecting a location or region) rather than purely textual ones. PointArena allows for both unambiguous and ambiguous scenarios, supporting studies of disambiguation, spatial commonsense, and pragmatic inference. Unlike bounding boxes, segmentation masks, or free-form text responses, pointing offers high-precision signal that avoids reliance on object contours or dense annotations and is directly compatible with human evaluation.

PointArena decomposes pointing into three stages of evaluation: 1) **Point-Bench** is a curated dataset of 982 manually selected, annotated, and verified image-question pairs across five high-level categories (Spatial, Affordance, Counting, Steerable, and Reasoning), an interactive, online platform for blind, pairwise comparison between models based on user instructions. Users select from curated or custom-uploaded images. Voting is anonymized, and we have collected over $4,500$ votes from more than 100 participants. 3) **Point-Act** is a real-world benchmark that evaluates the utility of pointing in for a downstream application. The system directs a robotic arm to manipulate objects

through pointing-based language commands. All three evaluation stages require minimal human effort; each is self-contained and can run live to evaluate any model.

Through our evaluation of both open-source and proprietary models across the three stages of the PointArena benchmark, we find that `Molmo-72B` achieves the highest performance on Point-Bench, with proprietary models such as `Gemini-2.5-Pro` performing comparably. Models trained with explicit pointing supervision consistently outperform those without. We also observe a strong correlation between static benchmark accuracy and human preference in Point-Battle. Notably, we find that adding language reasoning (e.g., Chain-of-Thought [31]) does not improve visual grounding for pointing tasks. Our study further reveals several other actionable insights into model behavior and evaluation design. We see **PointArena** as a missing component necessary as we develop general-purpose vision-language models that can reason about and interact with the world.

2 Related work

Grounding benchmarks. There are many benchmarks for visual grounding and spatial reasoning capabilities of multimodal large language models (MLLM). The RefCOCO, RefCOCO+, and RefCOCOg datasets focus on 2D visual grounding, with RefCOCOg emphasizing long-form referring expressions and fine-grained object distinctions [35]. In 3D, ScanRefer provides 51,583 descriptions across 800 RGB-D scans, supporting joint language-geometry models for 3D object localization [7]. ReferIt3D and CityRefer extend this to fine-grained and outdoor settings, with CityRefer incorporating geographic features from OpenStreetMap [2, 23]. Interactive benchmarks such as GuessWhat?! evaluate multi-turn object grounding through binary dialog across 150K games [9]. Flickr30K Entities supports phrase-region grounding through 276K bounding boxes and cross-caption coreference annotations [26]. These datasets primarily focus on bounding box grounding, object retrieval, or dialog-based localization. They do not explicitly evaluate pointing behavior.

Arena style evaluation. Arena-style evaluations have emerged as a popular method for comparing large language models (LLMs) through anonymized pairwise comparisons and user voting, initially popularized by *Chatbot Arena*, which amassed over 240K votes across diverse models and languages [8]. Extensions like *MT-Bench* added multi-turn dialogues with AI-driven scoring [39], while *am-ELO* stabilized rankings through maximum likelihood estimation and annotator reliability modeling [21]. Fully automated frameworks such as *Auto-Arena*, leveraging LLM-generated content and voting, closely mirrored human judgments [38]. *BenchBuilder* automated benchmark creation from crowdsourced data, enhancing statistical robustness. Specialized variants like *Werewolf Arena* examined social reasoning [5], and *OpenArena* facilitated offline evaluations [27]. Despite their scalability and user alignment, arena-style methods lack ground-truth validation and are vulnerable to manipulation [22], prompting alternative approaches such as *Tournament Evaluation* [18]. *PointArena* uniquely addresses these issues by combining ground-truth evaluation (Point-Bench) with human-preference arena comparisons (Point-Battle).

Models that point or sketch. Multimodal large language models (MLLMs) have made notable progress in vision-language reasoning, with GPT-4V showcasing tasks like story generation and OCR-free math [33]. Most use encoder-decoder architectures with visual-language fusion to enhance cross-modal integration. MiniGPT-4 aligns ViT features with language models via Q-Formers [30], while Molmo improves 2D spatial grounding by regressing normalized coordinates, reaching 92% precision in icon localization [11]. RoboPoint [36] instruction-tunes VLMs on synthetic data for robotic affordance prediction, outperforming GPT-4o and PIVOT [24] by over 20% in spatial accuracy. Other models, like VisCPM and Qwen-VL, add region-level control and multilingual support, while NExT-GPT integrates 3D, audio, and video inputs [33]. Yet, challenges in precise spatial localization persist, raising questions about what drives effective pointing in MLLMs.

3 PointArena

Evaluating the ability of MLLMs to localize language-referred entities in images requires benchmarks that are both precise and diagnostic. Existing benchmarks often emphasize classification or captioning, but fall short when it comes to assessing fine-grained spatial grounding—the ability to resolve natural language instructions into specific image coordinates. This capability is critical not only for understanding model alignment with human intent, but also for enabling downstream applications in

robotics [36], augmented reality [12], and interactive web agents [15], and potentially contributing to explainability [25]. As both specialized pointing models and general-purpose MLLMs improve, standardized evaluation across open-source and proprietary systems becomes essential.

We introduce **PointArena**, an evaluation suite for language-conditioned pointing, comprising three stages: (i) **Point-Bench**, a curated dataset for controlled measurement of spatial localization accuracy; (ii) **Point-Battle**, a live, blinded human-preference arena for pairwise model comparison; and (iii) **Point-Act**, a real-world robotic setting that evaluates pointing precision through physical execution. Together, these components form a unified framework to quantify and analyze how well MLLMs ground language into visual and physical space.

3.1 Task formulation

We formalize pointing as a language-conditioned fine-grained localization task. The input consists of an RGB image $I \in \mathbb{R}^{H\times W\times 3}$ and a natural-language instruction prompt $q = \{w_t\}_{t=1}^{T}$. A multimodal large language model (MLLM) $\mathcal{F}_\theta$ takes (I, q) as input and predicts a set of image-space coordinate points $P = \{(x_i, y_i)\}_{i=1}^{K}$, where each point lies within the image bounds: $x_i \in [0, W-1]$, $y_i \in [0, H-1]$.

Ground-truth supervision is provided as a set of binary masks $\{M_j\}_{j=1}^{K^*}$, with each mask $M_j \in \{0,1\}^{H\times W}$ denoting the valid region for one of K^* annotated targets. A predicted point (x_i, y_i) is considered correct if it lies within the spatial support of some mask M_j, i.e., $M_j[y_i, x_i] = 1$.

A prediction is considered *successful* if:

1. The number of predicted points matches the number of target regions: $K = K^*$,
2. Each target region M_j is covered by at least one predicted point: $\exists (x_i, y_i) \in P$ such that $M_j[y_i, x_i] = 1$.

This formulation enables fully automated evaluation given access to the ground-truth masks, with no need for human annotators at test time.

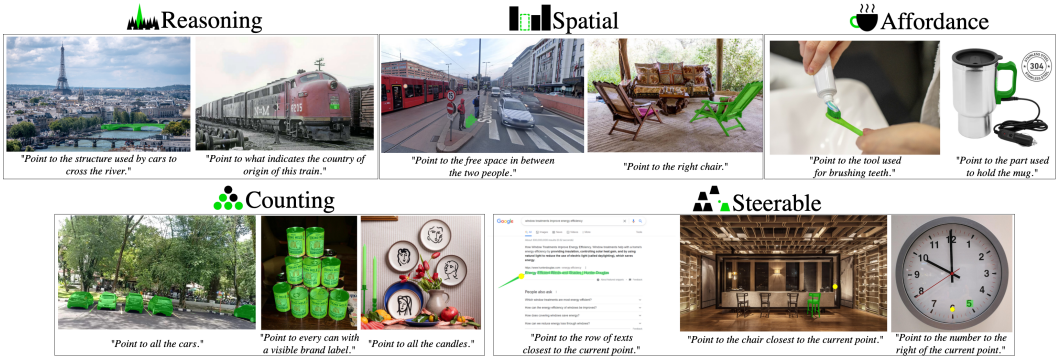

Figure 2: **Overview of the five Point-Bench categories and the annotation UI.** Point-Bench includes 982 image-query pairs grouped into five categories: **Spatial** (positional references), **Affordance** (functional part identification), **Counting** (attribute-based grouping), **Steerable** (relative pointing), and **Reasoning** (open-ended visual inference). Each example shows a representative query and the corresponding target. On the right, we show the Gradio-based annotation interface used to collect and refine segmentation masks. Initial masks are generated using SAM and refined by annotators, followed by manual verification.

3.2 Point-Bench

Point-Bench is the largest benchmark for evaluating language-guided pointing, comprising 982 text-image pairs with pixel-level target masks collected from public sources after April 20, 2025. The dataset is evenly divided into five task-driven categories—Spatial, Affordance, Counting, Steerable, and Reasoning—derived from a survey of question types frequently tackled by open-source MLLMs [11, 36, 28]. Each category targets a distinct capability: 1) *Spatial* focuses on positional

queries within scenes rich in spatial relationships or repeated objects (e.g., "Point to the leftmost tree in the image"); 2) *Affordance* emphasizes functional parts of objects, typically in tabletop scenes, prompting queries like "Point to the handle used for pouring"; 3) *Counting* features multiple similar items and supports queries about subsets based on number or attributes, such as "Point to all the blue cars in the image"; 4) *Steerable* leverages images from the PixMo dataset that include a reference point, guiding annotators to ask relative-position questions like "Point to the item closest to the marked point"; and 5) *Reasoning* presents event-rich or abstract scenes, inviting open-ended queries that require inference, such as "Point to the tallest man-made object in the image." Annotators, recruited via crowdsourcing, were free to ask any question, but carefully selected category-specific images naturally guided them toward prompts aligned with each reasoning axis. These curated splits together support systematic evaluation of an MLLM's ability to recognize, reason, and precisely ground language in visual space.

To construct a Point-Bench, we developed an intuitive Gradio-based annotation interface. Annotators were shown images sampled from each category and asked to write natural language queries aligned with the category theme. These queries were then evaluated using predictions from three anonymized MLLMs. If one or fewer models produced a correct prediction as judged by human evaluators, the query was considered sufficiently challenging and accepted for inclusion in the dataset. Following this, the annotators used the same interface to annotate the target points directly on the image. A SAM model was used to generate initial masks based on the selected point, and users could refine these masks by editing or removing portions before submission. Finally, a separate group of annotators manually verified the masks to ensure they accurately reflected the user-generated queries.

3.3 Point-Battle

As MLLMs increasingly incorporate visually grounded reasoning and pointing capabilities, static benchmarks become inadequate for evaluating performance in open-ended, real-world scenarios—particularly with respect to human preferences. To address this limitation, we introduce **Point-Battle**, a dynamic platform for pairwise evaluation of MLLMs' pointing abilities based on user-provided language instructions. Point-Battle adopts a head-to-head evaluation format inspired by Chatbot Arena [8], implemented via a Gradio-based web interface. In each round, two anonymized models are randomly sampled from the top performers in **Point-Bench**—including `GPT-4o`, `Gemini 2.5 Flash`, `Molmo-7B-D`, `Qwen2.5-VL-7B`, and `Grok-2 Vision`. Users submit a natural language instruction and select an image from a curated dataset (post-April 20, 2025) or upload their own. The two models return point predictions, which are displayed side by side. Participants vote for the better output or select both good" or both bad" if applicable. No preset prompts are provided, encouraging diverse and unbiased instructions. Model identities are kept anonymous to prevent bias. Since launch, Point-Battle has collected over 4,500 votes from approximately 100 participants worldwide. Unlike the static **Point-Bench**, which may be subject to overfitting during model development, Point-Battle serves as a continuously updated benchmark that captures real-time human preferences and tracks progress in visually grounded reasoning across MLLMs. As Point-Battle scales, it also serves as a platform for collecting pointing data.

3.4 Point-Act

The first two stages of Point Arena evaluate MLLMs' pointing capabilities through quantitative metrics and human preference assessments. However, pointing is only meaningful insofar as it enables real-world utility. To evaluate such support, we introduce *Point-Act*—an interactive system where users issue natural-language instructions via a GUI to a double-blind MLLM. The model generates one or more predicted points, which are translated into actionable commands for an `xArm 6 Lite` robot. The robot executes a pick-or-place action at the indicated location using depth sensing for spatial reasoning. This setup operationalizes pointing into end-to-end physical manipulation, bridging language grounding with robotic control. Point-Act highlights the downstream consequences of grounding precision: even small localization errors can cause execution failures, whereas accurate predictions enable consistent real-world success.

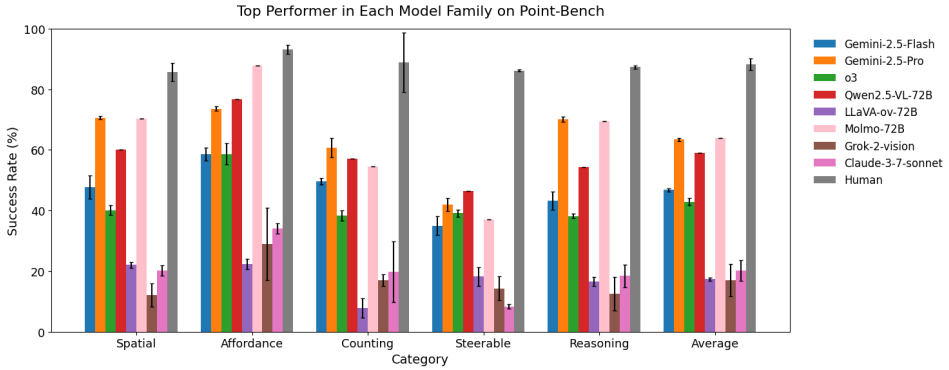

Figure 3: **Success rates of MLLMs on Point-Bench** across six task categories: *Spatial*, *Affordance*, *Counting*, *Steerable*, *Reasoning*, and *Average*. Each bar represents the mean success rate (%) for a given model, with error bars indicating standard deviation across three evaluation runs. The "Human" bar serves as an upper-bound reference. The results demonstrate substantial performance disparities, with top models (e.g., GPT-4o, Gemini-2.5-Pro, Molmo-72B) achieving near-human accuracy in select categories, while others (e.g., LLaVA, Grok, and Claude) consistently underperform.

4 Experiments

We evaluate a range of multimodal large language models (MLLMs)—both proprietary and open-source—using three components: **Point-Bench** (static benchmark evaluation), **Point-Battle** (human preference comparison), and **Point-Act** (real-world robotic execution). Section 4.1 describes the evaluation protocols, including model selection, prompting, and success metrics. Section 4.2 presents results on model performance and the impact of pointing supervision. Section 4.3 presents results demonstrating the correlation between benchmark accuracy, human preference judgments, and real-world task performance in pointing tasks. Section 4.4 includes ablations on prompt structure and output formats using GPT-4o to analyze factors affecting pointing accuracy.

4.1 Evaluation Setup

All evaluations were performed under zero-shot prompting conditions. To ensure consistent outputs across models with differing internal coordinate systems—particularly proprietary ones—we adopted a standardized output format: $[x, y]$, where x and y denote horizontal and vertical pixel coordinates, respectively. This format was used across all models, except for those like `Molmo`, `Qwen2.5-VL`, and `Gemini`, which provide explicit coordinate outputs or prompting instructions.

Success was measured using a binary metric: a prediction was considered correct if the point lay within the target mask. For non-counting tasks, models were prompted to predict a single point; if multiple were returned, only the first point was evaluated, assuming it reflected the highest-confidence prediction due to the autoregressive generation process.

Point-Bench. We benchmarked 16 MLLMs (spanning open-source and proprietary models, including key variants). Each model was evaluated on the same 982 image-instruction pairs, three times independently, to compute means and standard deviations. Open-source models were executed locally on NVIDIA A100 GPUs, while proprietary models were accessed via public APIs.

Point-Battle. To measure alignment with human preferences, we released a live evaluation platform and promoted it via social media and mailing lists. Users voted on head-to-head comparisons between anonymous model outputs. Elo ratings were computed from pairwise comparisons excluding ambiguous votes ("both good" or "both bad").

Point-Act. We recruited 10 remote participants to interact with our real-world robot setup. For a fixed scene, participants evaluated three agents—`Molmo-7B-D`, `GPT-4o`, and a human reference—across three trials. After each condition, they completed a System Usability Scale (SUS) survey.

Models. We evaluate variants from Molmo [10], Gemini [28], OpenAI [1], Claude [3], Grok [32], LLaVA [20], and Qwen [4]. See appendix for details.

Figure 4: **Qualitative predictions across Point-Bench categories.** Example model predictions are shown for each of the five Point-Bench categories: **Spatial**, **Affordance**, **Counting**, **Steerable**, and **Reasoning**. Each colored dot corresponds to a prediction from a different MLLM, labeled by model name in the legend. These examples highlight the diversity of pointing behaviors and the variation in performance across models.

4.2 Main Results

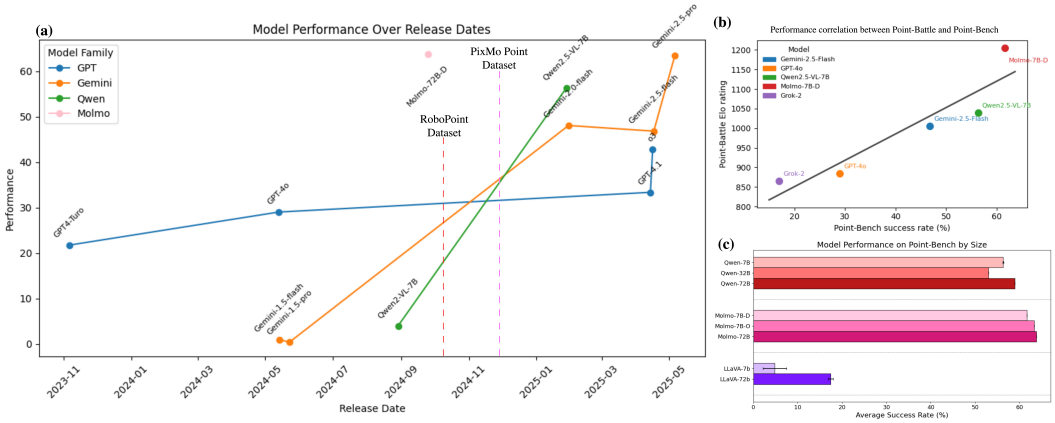

Figure 5: **Insights from Point-Battle and Point-Bench.** (a) Point-Bench accuracy over time by model family. Performance jumps notably post-PixMo (dashed line, Dec 2024), with `GPT-4.1` and `Gemini-2.0-Flash` showing large gains over predecessors, hinting at pointing supervision. (b) Point-Battle and Point-Bench are strongly correlated ($R^2 = 0.85$), validating consistency. (c) Open-source model performance vs. model size shows only minor gains, indicating diminishing returns with scale.

Open-source models perform comparably to proprietary models in pointing accuracy. Point-Bench results show that open-source MLLMs explicitly trained on pointing data often match or outperform proprietary models. For example, `Molmo-72B` outperformed `Gemini-2.5-Pro` by 0.43 percentage points—a statistically insignificant margin ($p \approx 0.29$). In affordance reasoning, open-source models like `Molmo-72B` and `Qwen2.5-VL` consistently exceed proprietary baselines. Overall, Molmo-72B achieves the highest performance on the Point-Bench benchmark as shown in Table 4.

Pointing supervision significantly boosts performance. Access to explicit pointing data is a key driver of model accuracy as shown in Figure 5a. Within the Qwen family, incorporating the PixMo corpus into `Qwen2.5-VL-7B` increased performance to 52.3%, a substantial gain over the 17.4% achieved by `Qwen2-VL-7B`, which did not use such data. In contrast, LLaVA variants—also trained without explicit pointing supervision—achieved only 4.8–17.4% on average.

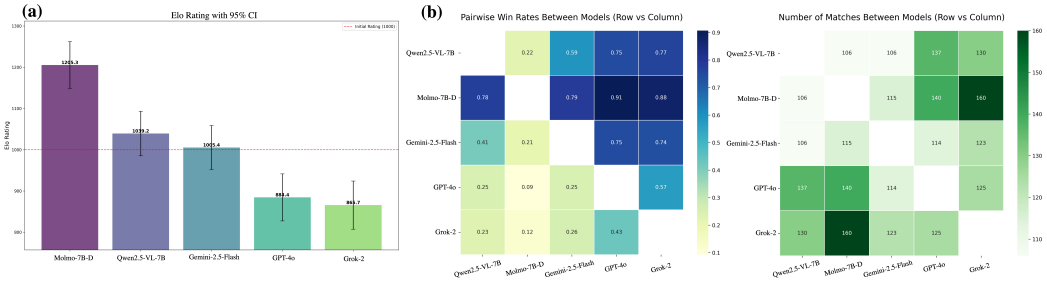

Figure 6: **Performance on Human Preference Evaluation with Point-Battle.** We collected over 4,500 votes from more than 100 global participants. Based on the Elo ratings derived from these votes, we observed a clear preference for outputs from open-source models such as `Molmo-7B-D` and `Qwen2.5-VL-7B`, which consistently outperformed proprietary models in terms of human preference.

Proprietary models likely benefit from open-source pointing datasets. While proprietary training data is opaque, we observe large performance jumps in models released shortly after the PixMo [11] and RoboPoint dataset [36]. For instance, `GPT-o3` improved by 21.1 percentage points over `GPT-4-Turbo`, and `Gemini-2.5-Flash` improved by 45.9 points over `Gemini-1.5-Flash` (Figure 5a). These results suggest that recent proprietary models may have incorporated PixMo or a similar corpus.

Open-source models align more closely with human preferences. In Point-Battle, `Molmo-7B-D` outperformed `Gemini-2.5-Flash` by 196 Elo points. Their 95% confidence intervals do not overlap, and `Molmo-7B-D` won 79% of the 115 direct head-to-head comparisons, as shown in Figure 6. Both `Qwen2.5-VL-7B` and `Molmo-7B-D` surpass proprietary models in human preference evaluations and exceed the 1000-point baseline, indicating a statistically significant advantage over random guessing. However, in terms of preference-aligned pointing performance, `Molmo-7B-D` remains clearly superior to `Qwen2.5-VL-7B`.

Molmo excels on Point-Act evaluation. User study results shown in Figure 7 that `Molmo-7B-D` outperforms the proprietary `GPT-4o` model by a substantial margin, achieving 65% higher performance and approaching human (oracle) baseline levels. This superiority is also reflected in user preference, with `Molmo-7B-D` scoring 60.3 points higher in SUS than `GPT-4o`.

Model size does not impact pointing performance. As shown in Figure 5c, the performance of open-source models (LLaVA-OV, Molmo, and Qwen-VL) on Point-Bench remains largely unchanged with increased model size. For example, `Qwen2.5-VL-7B` performs within 3% of `Qwen2.5-VL-72B`, and `Molmo-7B-O` differs by less than 1% from `Molmo-72B`. These results suggest that scaling model size does not significantly improve pointing accuracy.

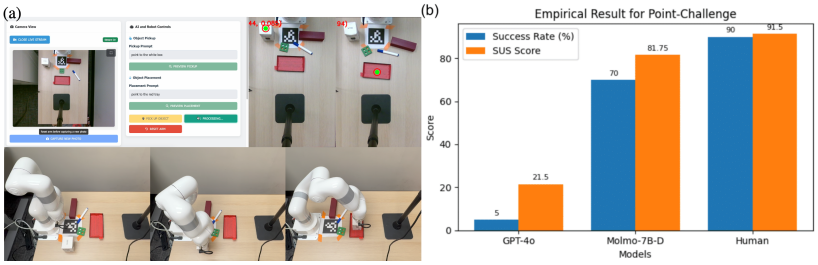

Figure 7: **Overview of the Point-Act system.** (a) The Point-Act manipulation setup enables remote control of a real-world xArm 6 Lite robot via language instructions, allowing users to evaluate pointing MLLMs. (b) User-blind evaluations and SUS preference scores collected for each model.

4.3 Results between three evaluation frameworks.

The three-stage evaluation of MLLMs' pointing capabilities should not be viewed as isolated components, but as complementary steps in a progressive pipeline. As MLLMs improve, they are expected to advance through these stages. Therefore, understanding the correlation and agreement between stages is crucial for assessing consistent performance gains.

Human-preference and static dataset evaluations are highly consistent. Point-Bench's static dataset will inevitably plateau as MLLMs improve by training on ever-larger, real or synthetic pointing corpora (e.g., RoboPoint [36]). To stay ahead, we introduce **Point-Battle**, a live arena that updates continuously and enables open-ended model comparison in real time. Validating this setup, we re-evaluated the models tested on Point-Bench and observed strong alignment: Point-Battle scores correlate with Point-Bench results at $R^2 = 0.85$ (Figure 5b).

Point-Bench accuracy predicts real-world task success. We validated Point-Bench as a reliable proxy by testing three agents—`Molmo-7B-D`, `GPT-4o`, and a human reference—on Point-Act. Success rates closely aligned with Point-Bench scores, yielding a strong linear correlation ($R^2 = 0.92$). This high correlation indicates that Point-Bench is a reliable proxy for the pointing capability of multimodal LLMs in practical settings.

4.4 What Other Factors Drive Pointing Performance?

To understand the design choices that impact pointing, we conducted ablations on `GPT-4o` using variations in prompt structure and output representation, as shown in Table 1.

Targeted prompts outperform verbose reasoning. Incorporating Chain-of-Thought (CoT) reasoning reduced pointing accuracy by 2.9% for `GPT-4o` and by a substantial 16% for `Gemini-2.5-Flash`. Using raw, unfiltered user queries led to an additional drop of 2.6% and 3.7% for `GPT-4o` and `Gemini-2.5-Flash`, respectively. These results suggest that clear, targeted prompts with well-defined coordinate systems are crucial for effective pointing, while additional reasoning through language does not enhance MLLMs' pointing capabilities.

Table 1: Performance (%) of GPT-4o and Gemini-2.5-Flash variants across evaluation categories.

Method	Affordance	Spatial	Reasoning	Steerability	Counting	Average
GPT-4o + In-Context (2-shots)	46.0	26.7	23.9	22.5	33.2	30.4
GPT-4o + Chain-of-Thought (CoT) [31]	41.4	24.6	21.2	13.5	32.1	26.6
GPT-4o + Unparsed language instruction	37.8	23.7	19.7	21.9	31.1	26.9
GPT-4o (Default)	42.4	25.6	23.8	24.5	31.1	29.5
Gemini-2.5-Flash + In-Context (2-shots)	50.0	40.5	35.8	32.0	40.3	39.7
Gemini-2.5-Flash + Chain-of-Thought (CoT) [31]	44.4	25.6	24.9	26.0	33.7	30.9
Gemini-2.5-Flash + Unparsed language instruction	55.0	44.6	43.5	25.0	47.5	43.1
Gemini-2.5-Flash (Default)	58.6	47.7	43.2	35.0	49.7	46.8

5 Limitations, Discussion, and Conclusions

Discussions. PointArena benchmarks spatial reasoning in multimodal models via static datasets (Point-Bench) and human preference comparisons (Point-Battle). To improve annotation quality, we propose replacing grid-based tools with free-form contouring for finer object boundaries. We also augment Point-Bench with user-generated Point-Battle data to reduce staleness, and introduce adaptive sampling to compare similarly performing models for more informative evaluations.

Limitations. Current annotations rely on SAM [19] and grid refinement, which often yield coarse masks—especially for fine-grained objects—reducing quality. Static benchmarks also risk data leakage as models train on public data. Additionally, Point-Battle's uniform sampling wastes comparisons between mismatched models.

Conclusion. PointArena provides a scalable framework for evaluating grounded pointing, blending controlled benchmarks with live user feedback. While limitations remain in annotation quality, dataset freshness, and evaluation efficiency, our proposed updates improve precision, scalability, and signal strength for real-world spatial reasoning assessments

References

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

[2] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. ReferIt3D: Neural listeners for fine-grained 3d object identification in real-world scenes. In *16th European Conference on Computer Vision (ECCV)*, 2020.

[3] Anthropic. The claude 3 model family: Opus, sonnet, haiku. `https://www.anthropic.com/news/claude-3-family`, 2024. Claude-3 Model Card.

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[5] Suma Bailis, Jane Friedhoff, and Feiyang Chen. Werewolf arena: A case study in llm evaluation via social deduction, 2024.

[6] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342, 2010.

[7] Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language, 2020.

[8] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024.

[9] Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue, 2017.

[10] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Jen Dumas, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024.

[11] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models, 2024.

[12] Jiafei Duan, Yi Ru Wang, Mohit Shridhar, Dieter Fox, and Ranjay Krishna. Ar2-d2: Training a robot without a robot. *arXiv preprint arXiv:2306.13818*, 2023.

[13] Jiafei Duan, Wentao Yuan, Wilbert Pumacay, Yi Ru Wang, Kiana Ehsani, Dieter Fox, and Ranjay Krishna. Manipulate-anything: Automating real-world robots using vision-language models. *arXiv preprint arXiv:2406.18915*, 2024.

[14] Petko Georgiev et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[15] Izzeddin Gur, Hiroki Furuta, Austin Huang, Mustafa Safdari, Yutaka Matsuo, Douglas Eck, and Aleksandra Faust. A real-world webagent with planning, long context understanding, and program synthesis. *arXiv preprint arXiv:2307.12856*, 2023.

[16] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *arXiv preprint arXiv:2406.09403*, 2024.

[17] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

[18] Richard Kelley and Duncan Wilson. Tournament evaluation of large language models, 2025.

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

[20] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[21] Zirui Liu, Jiatong Li, Yan Zhuang, Qi Liu, Shuanghong Shen, Jie Ouyang, Mingyue Cheng, and Shijin Wang. am-elo: A stable framework for arena-based llm evaluation, 2025.

[22] Rui Min, Tianyu Pang, Chao Du, Qian Liu, Minhao Cheng, and Min Lin. Improving your model ranking on chatbot arena by vote rigging, 2025.

[23] Taiki Miyanishi, Fumiya Kitamori, Shuhei Kurita, Jungdae Lee, Motoaki Kawanabe, and Nakamasa Inoue. Cityrefer: Geography-aware 3d visual grounding dataset on city-scale point cloud data, 2023.

[24] Soroush Nasiriany, Fei Xia, Wenhao Yu, Ted Xiao, Jacky Liang, Ishita Dasgupta, Annie Xie, Danny Driess, Ayzaan Wahid, Zhuo Xu, Quan Vuong, Tingnan Zhang, Tsang-Wei Edward Lee, Kuang-Huei Lee, Peng Xu, Sean Kirmani, Yuke Zhu, Andy Zeng, Karol Hausman, Nicolas Heess, Chelsea Finn, Sergey Levine, and Brian Ichter. Pivot: Iterative visual prompting elicits actionable knowledge for vlms, 2024.

[25] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence, 2018.

[26] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, 2016.

[27] SYV-AI. Openarena: An open platform for llm-as-a-judge evaluation. `https://github.com/syv-ai/OpenArena`, 2024. Accessed: 2025-05-09.

[28] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, Steven Bohez, Konstantinos Bousmalis, Anthony Brohan, Thomas Buschmann, Arunkumar Byravan, Serkan Cabi, Ken Caluwaerts, Federico Casarini, Oscar Chang, Jose Enrique Chen, Xi Chen, Hao-Tien Lewis Chiang, Krzysztof Choromanski, David D'Ambrosio, Sudeep Dasari, Todor Davchev, Coline Devin, Norman Di Palo, Tianli Ding, Adil Dostmohamed, Danny Driess, Yilun Du, Debidatta Dwibedi, Michael Elabd, Claudio Fantacci, Cody Fong, Erik Frey, Chuyuan Fu, Marissa Giustina, Keerthana Gopalakrishnan, Laura Graesser, Leonard Hasenclever, Nicolas Heess, Brandon Hernaez, Alexander Herzog, R. Alex Hofer, Jan Humplik, Atil Iscen, Mithun George Jacob, Deepali Jain, Ryan Julian, Dmitry Kalashnikov, M. Emre Karagozler, Stefani Karp, Chase Kew, Jerad Kirkland, Sean Kirmani, Yuheng Kuang,

Thomas Lampe, Antoine Laurens, Isabel Leal, Alex X. Lee, Tsang-Wei Edward Lee, Jacky Liang, Yixin Lin, Sharath Maddineni, Anirudha Majumdar, Assaf Hurwitz Michaely, Robert Moreno, Michael Neunert, Francesco Nori, Carolina Parada, Emilio Parisotto, Peter Pastor, Acorn Pooley, Kanishka Rao, Krista Reymann, Dorsa Sadigh, Stefano Saliceti, Pannag Sanketi, Pierre Sermanet, Dhruv Shah, Mohit Sharma, Kathryn Shea, Charles Shu, Vikas Sindhwani, Sumeet Singh, Radu Soricut, Jost Tobias Springenberg, Rachel Sterneck, Razvan Surdulescu, Jie Tan, Jonathan Tompson, Vincent Vanhoucke, Jake Varley, Grace Vesom, Giulia Vezzani, Oriol Vinyals, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Fei Xia, Ted Xiao, Annie Xie, Jinyu Xie, Peng Xu, Sichun Xu, Ying Xu, Zhuo Xu, Yuxiang Yang, Rui Yao, Sergey Yaroshenko, Wenhao Yu, Wentao Yuan, Jingwei Zhang, Tingnan Zhang, Allan Zhou, and Yuxiang Zhou. Gemini robotics: Bringing ai into the physical world, 2025.

[29] Michael Tomasello, Malinda Carpenter, and Ulf Liszkowski. A new look at infant pointing. *Child development*, 78(3):705–722, 2007.

[30] Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, Yutong Zhang, Zihao Wu, Zhengliang Liu, Tianyang Zhong, Bao Ge, Tuo Zhang, Ning Qiang, Xintao Hu, Xi Jiang, Xin Zhang, Wei Zhang, Dinggang Shen, Tianming Liu, and Shu Zhang. A comprehensive review of multimodal large language models: Performance and challenges across different tasks, 2024.

[31] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[32] xAI. Grok-2 model card. `https://x.ai/news/grok-2`, 2024. Large language model.

[33] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 11 2024.

[34] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. *arXiv preprint arXiv:1608.00272*, 2016.

[35] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions, 2016.

[36] Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics. *arXiv preprint arXiv:2406.10721*, 2024.

[37] Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction in robotics. In *Proc. Conference on Robot Learning (CoRL)*, volume 270, pages 4005–4020, 2025.

[38] Ruochen Zhao, Wenxuan Zhang, Yew Ken Chia, Weiwen Xu, Deli Zhao, and Lidong Bing. Auto-arena: Automating llm evaluations with agent peer battles and committee discussions, 2024.

[39] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification:[NA]

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer:[Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:[NA]

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: [NA]

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.