

PhotoAgent: A Robotic Photographer with Spatial and Aesthetic Understanding

Lirong Che^{1*}, Zhenfeng Gan^{1*}, Yanbo Chen¹, Junbo Tan^{1†}, Xueqian Wang¹



Fig. 1. Overview of PhotoAgent and its capabilities. (a) illustrates the inefficiency of real-world trial-and-error, while PhotoAgent leverages internal simulation to achieve one-shot success. (b) highlights three key capabilities.

Abstract—Embodied agents for creative tasks like photography must bridge the semantic gap between high-level language commands and geometric control. We introduce PhotoAgent, an agent that achieves this by integrating Large Multimodal Models (LMMs) reasoning with a novel control paradigm. PhotoAgent first translates subjective aesthetic goals into solvable geometric constraints via LMM-driven, chain-of-thought (CoT) reasoning, allowing an analytical solver to compute a high-quality initial viewpoint. This initial pose is then iteratively refined through visual reflection within a photorealistic internal world model built with 3D Gaussian Splatting (3DGS). This “mental simulation” replaces costly and slow physical trial-and-error, enabling rapid

convergence to aesthetically superior results. Evaluations confirm that PhotoAgent excels in spatial reasoning and achieves superior final image quality.

I. INTRODUCTION

Endowing embodied agents with the ability to seamlessly collaborate with humans on creative tasks is a long-standing pursuit in robotics and artificial intelligence. Among creative domains, photography presents an ideal yet challenging testbed, as it deeply intertwines technical execution with subjective aesthetics. A successful photographer must comprehend not only the geometric properties of the world, such as occlusion and perspective, but also higher-level abstract intentions, like capturing “a dramatic photo”.

Early “robot photographers” hard-coded the rule-of-thirds yet falter outside curated scenes [1], [2]. Later methods split into two brittle camps. Reinforcement-learning treats viewpoint choice as black-box search [3], [4], but fusing geometric constraints and aesthetic semantics into one reward demands costly, environment-specific interaction data.

This work was supported by the Natural Science Foundation of Shenzhen (No. JCYJ20230807111604008, No. JCYJ20240813112007010), the Natural Science Foundation of Guangdong Province (No. 2024A1515010003) and Cross-disciplinary Fund for Research and Innovation (No. JC2024002) of Tsinghua SIGS.

*indicates equal contribution.

¹Center for Artificial Intelligence and Robotics, Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China, {clr24@mails., gzf24@mails., cyb23@mails., tjblql@sz., wang.xq@sz.}@tsinghua.edu.cn

†Corresponding author: Junbo Tan

Imitation systems such as PhotoBot simply retrieve and copy reference photos [5], amounting to template matching that cannot span the combinatorial diversity of novel scenes. Neither line bridges the fundamental semantic-to-geometric gap.

How can we unlock genuine creativity in robotic photography? Recent studies indicate that pretrained Large Multimodal Models (LMMs) already encode human-aligned aesthetic preferences [6], [7] and can be further tuned with only modest data [8]. Yet these models are not natively trained to translate language into camera motion; directly prompting an off-the-shelf LMM for a 6-DoF pose produced numerically erratic results.

To harness the LMM’s semantic power while restoring geometric soundness, we introduce PhotoAgent—an embodied photography agent whose entire decision loop is steered by the LMM. PhotoAgent carries out reflective reasoning in continuous geometric space: every “thought” emitted by the LMM is instantiated as a physically feasible pose, and every “reflection” is grounded in a view rendered on-the-fly by a real-time 3D Gaussian-splat world model. By internally simulating the visual consequences of candidate motions, the agent cuts real-world trial-and-error and converges rapidly to high-quality decisions. (see Figure 1 for an overview of the system and its key capabilities).

Our main contributions are:

- **An aesthetics-driven reasoning scheme.** We introduce an **anchor-point hypothesis** and craft an explicit chain-of-thought that maps subjective aesthetic goals into solvable spatial-geometric constraints.
- **An inverse viewpoint-solving paradigm.** Continuing the same chain-of-thought, we reformulate the resulting geometric constraints as explicit mathematical problems.
- **A closed-loop architecture based on 3DGS.** Leveraging 3DGS for real-time, photorealistic rendering, the agent performs visual reflection and iteratively refines its decisions.

PhotoAgent achieves state-of-the-art performance across two novel fronts, evaluated in both simulation and the real world through: (1) a language-conditioned spatial task that seeks the best viewpoint, and (2) a complete pipeline check of image aesthetics and instruction fidelity.

II. RELATED WORK

A. Robotic Photography

Early event-photography systems demonstrated end-to-end autonomy in crowds using composition heuristics and social interaction to capture portraits [1], [9]. Subsequent rule/score-based pipelines encoded guidelines or analytic scoring for repositioning and face-aware composition [2], [10], showing feasibility but limited adaptability in cluttered scenes. Learning-based approaches broadened this space, including template imitation with deep RL and direct optimization of learned aesthetic estimators on mobile platforms [3], [4]. Beyond fixed rules and generic scores, instruction-conditioned pipelines incorporate user intent by retrieving

a reference layout and imitating its pose [5], [11]; notably, a recent system (PhotoBot) introduces an LLM to reason over the user query and gallery captions before retrieval, then maps the query to the selected reference and solves a PnP-style pose to mimic composition [5]. Overall, prior systems either encode rules/scores or retrieve-and-mimic exemplars, whereas our method grounds free-form language directly in scene geometry to plan executable, novel viewpoints without dependence on a finite database.

B. Reasoning Architectures for Embodied Agents

Recent advances in large language models (LLMs) have driven a shift from reactive policies to reasoning-driven agents. Chain-of-Thought (CoT) prompting [12] enables step-by-step reasoning, but lacks grounding in real-world feedback. The ReAct framework [13] addresses this by interleaving thoughts and actions in a “thought-action-observation” loop, enabling reasoning to influence actions and vice versa.

Building on ReAct, Reflexion [13], [14] introduces a self-improvement layer, where the agent summarizes and critiques its own past behaviors using language. This “verbal reinforcement” loop enables iterative skill refinement through feedback-driven reflection.

Despite their power, these reasoning agents are typically deployed in symbolic domains (e.g., text games, API calls). How to ground such reflection into real-world visual consequence remains an open problem—particularly for tasks like photography that demand geometric precision.

C. World Models for Visual Foresight

Bridging the gap between symbolic reasoning and physical execution requires internal world models. Approaches like World Models [15] and Dreamer-style latent planners [16]–[18] learn compact latent dynamics to imagine action outcomes. While efficient, they often distort geometry, which is critical for image composition. Explicit 3D models address this. NeRF yields accurate appearance but is often too slow for online control, even with acceleration like Instant-NGP [19], [20]. 3D Gaussian Splatting (3DGS) attains real-time photorealistic rendering with explicit structure [21], and its uptake in robotics indicates promise for closed-loop use [22]–[24]. This combination makes 3DGS a practical backbone to couple language-level reasoning with controllable, view-accurate visual imagination in diverse scenes.

III. METHOD

Despite their powerful generalization capabilities, LMMs are limited in embodied control due to three factors: (1) an LMM lacks a dedicated spatial-reasoning mechanism and struggles to generalize in cluttered visual environments. (2) as token-based generators, even state-of-the-art models exhibit numerically ill-conditioned behavior when directly asked to output SE(3) poses and tend to conflate camera egomotion with object motion, likely due to limited egomotion-supervised pretraining; (3) Its inference is slow, especially for visual inputs, and each call must be followed by physical motion and reobservation, making naive closed-loop control impractically sluggish.

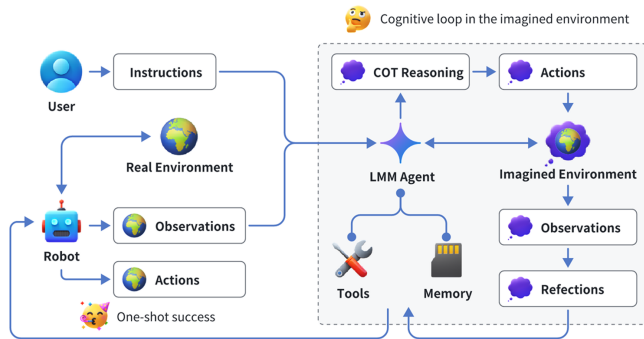


Fig. 2. Overall cognitive architecture of PhotoAgent.

A. System Overview

PhotoAgent couples high-level language reasoning with a geometry-faithful world model through a two-stage cognitive pipeline (Figure 2). At its core, PhotoAgent is powered by an LMM that functions as its central reasoning engine. This LMM-driven agent augments its CoT with a lightweight toolset and memory to bridge perception, reasoning, and actuation. The agent first fuses its multi-modal inputs \mathcal{O} into a metric 3D representation \mathcal{G} using real-time 3DGS [25], giving planning a geometry-faithful, photorealistic substrate. With \mathcal{G} in place, an LMM launches an internal counterfactual loop to determine the optimal action:

Intent parsing. The LMM decomposes the instruction \mathcal{L} and recent observations into compositional targets.

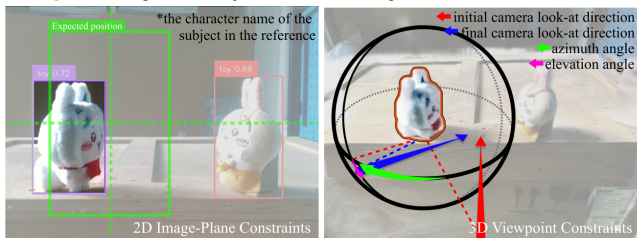
Pose proposal. It analytically solves candidate poses $\{x_i\}$ and renders predicted views $\mathcal{W}(x_i, \mathcal{G})$.

Reflective critique. Acting as a visual critic, the LMM scores and verbalizes how geometric changes affect aesthetics, then refines its hypothesis. This iterative process is inspired by the “reason-act” and “self-reflection” paradigms from recent work on language agents [13], [14].

B. Intention Parsing

To address the first critical gap of LMMs—their tendency to be overwhelmed by cluttered scenes—the ‘Intention Parsing’ module simplifies the problem space through our **Anchor-Point Hypothesis**. Instead of attempting to reason about all scene elements simultaneously, this strategy directs the LMM to emulate human cognition by selecting a single principal subject to serve as a compositional anchor point. This cognitive simplification is crucial: it reframes an ill-posed global optimization problem into a well-defined and tractable one of relative positioning: *how should the agent move relative to this anchor point to achieve the desired photographic outcome for the entire frame?* To reason meaningfully about this anchor point and its spatial context, the LMM must be grounded in the physical world. We achieve this via a structured input representation \mathcal{Z} derived from raw robot observations \mathcal{O} . Since LMMs are not natively trained for 3D geometric understanding, we explicitly supply the relevant cues through a modular tool-use paradigm [26], [27]. The structured perceptual inputs include camera intrinsics, and for each detected object: its semantic label, 2D bounding box (and its center (u, v)), and 3D world coordinates. The modular

Analysis: ... The center point of DIO* is located in the left center area of the screen, so we place the subject at ‘ $u=213$ ’ (about the left third line) and ‘ $v=240$ ’ (vertically centered) ... The camera needs to rotate significantly clockwise around the subject ... Lower the camera height and tilt it upwards to capture the subject from bottom to top...



Expect_position:
`{"u": 213, "v": 240, "scale_ratio": 1.7 }`
Expect_degree:
`{"azimuth_angle": -90, "elevation_angle": -5}`

Fig. 3. Intention Parsing workflow demonstration. “Take a photo of the toys with visual tension, like the reference.”

design allows for task-specific extensions, such as including facial orientation in portrait photography, which supports reasoning over composition rules like “looking room.” Having simplified the perceptual problem using an anchor point, we now address the second LMM limitation: its inability to reliably generate stable SE(3) poses for physical execution. Instead of tasking the LMM with direct pose regression, we guide it to produce a set of well-defined geometric constraints through a structured CoT reasoning process [12]. This process mirrors a human photographer’s workflow, breaking down the decision into iterative workflow:

- **Intent–Scene Alignment and Aesthetic Diagnosis.** The LMM maps the user’s goal onto specific scene elements, selects a single subject to serve as the anchor point, identifies occlusions, distractions, or layout flaws, and verbally proposes an aesthetic correction (e.g., “the subject should move slightly to the left”).
- **2D Image-Plane Constraints.** These determine where and how large the subject appears within the frame:
 - (u^*, v^*) : the target coordinates of the anchor point in the image plane, specifying horizontal and vertical layout.
 - s : the ratio between desired and current subject scale, controlling visual size.
- **3D Viewpoint Constraints.** These define the camera’s ideal spatial configuration relative to the anchor point:
 - θ : azimuth angle, determining the orbital direction around the subject.
 - φ : elevation angle, controlling vertical camera height.
 - ρ : camera-to-subject distance derived analytically from s , naturally coupling 2D visual scale with 3D spatial positioning without entanglement.

The output of this reasoning process is a structured vector of geometric constraints:

$$\mathbf{g} = (u^*, v^*, s, \theta, \varphi).$$

Figure 3 shows the subsequent reasoning-to-pose pipeline (using the input scene and user command already introduced in Figure 1). Although camera roll around the optical axis remains mathematically free with a single anchor point, we freeze it at 0° to avoid unstable horizon tilt—small rolls can

enhance aesthetics, but large rolls often ruin the frame. We deliberately adopt this spherical parameterization instead of directly regressing a 6-DoF pose. This design decouples distance control (via s) from directional control (via θ , φ), avoiding logical entanglement where a similar visual outcome could be produced by either translating or rotating the camera. Such disentanglement aligns with principles in language-conditioned robotics [28], enabling more stable and interpretable inference.

C. Geometric Solving

Given the geometric constraint vector \mathbf{g} , we recover a valid 6-DoF pose $\mathbf{T} \in \text{SE}(3)$ in two closed-form steps. This analytic mapping keeps every geometric term explicit, cleanly separates distance from direction, and remains numerically well-conditioned. Moreover, the explicit structure of this pipeline exposes interpretable intermediate steps, facilitating pattern discovery and causal reasoning by the LMM in the reflective optimization loop.

We begin by modeling the subject as an upright cylinder. This assumption ensures consistent projections: the subject appears rectangular regardless of azimuth, with a fixed aspect ratio and distance-dependent size. Let h_0 be the subject’s height in pixels, ρ_0 its original depth and H its true physical height. Given the focal length f and the desired-to-current scale ratio s , we solve for the new camera-to-subject distance ρ :

$$H = \frac{h_0 \rho_0}{f}, \quad h^* = s h_0, \quad (1a)$$

$$\rho = \frac{fH}{h^*} = \frac{fH}{s h_0} = \frac{\rho_0}{s}. \quad (1b)$$

Intuitively, $s > 1$ (a larger on-screen subject) implies moving *closer* (ρ decreases), while $s < 1$ implies moving back.

We then compute the camera’s 3D position \mathbf{p}_c in the subject-centric coordinate frame using the predicted global azimuth θ and elevation φ :

$$\mathbf{p}_c = \begin{bmatrix} \rho \cos \varphi \sin \theta \\ \rho \sin \varphi \\ \rho \cos \varphi \cos \theta \end{bmatrix}. \quad (2)$$

Next, we determine the initial 6-DoF camera pose \mathbf{T}_0 using a `look-at` function, which orients the camera from position \mathbf{p}_c to face the subject’s location $\mathbf{p}_{\text{subject}}$:

$$\mathbf{T}_0 = \text{look-at}(\mathbf{p}_c, \mathbf{p}_{\text{subject}}). \quad (3)$$

We adopt a z -up world frame and keep the camera’s roll angle fixed at $\psi = 0^\circ$ to avoid horizon tilt. The `look-at` function constructs the rotation matrix \mathbf{R} from the viewing direction and the world-up vector $(0, 0, 1)$, yielding the initial pose $\mathbf{T}_0 = [\mathbf{R} | \mathbf{t}]$.

This initial pose is then refined using a visual servoing loop. We first project the subject’s center under \mathbf{T}_0 and compute the pixel error vector:

$$\mathbf{e} = [u - u^*, v - v^*]^T. \quad (4)$$

Under small-angle assumptions, horizontal and vertical errors in the image plane are corrected by adjusting the camera’s

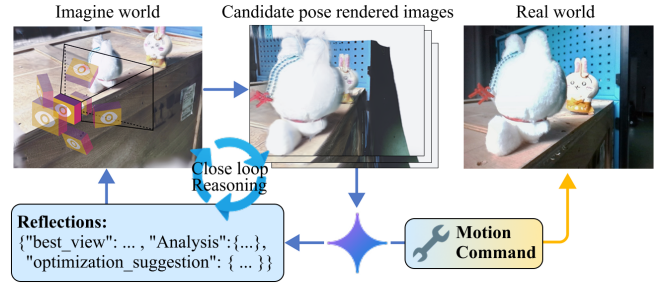


Fig. 4. Closed-loop reflective reasoning. Starting from an internal 3DGS “imagined world”, the agent renders candidate views, critiques them via the LMM, and issues optimized motion commands.

local yaw (θ_{yaw}) and pitch (φ_{pitch}) angles, respectively. This yields a near-diagonal image Jacobian:

$$\mathbf{J} \approx \begin{bmatrix} \frac{\partial u}{\partial \theta_{\text{yaw}}} & 0 \\ 0 & \frac{\partial v}{\partial \varphi_{\text{pitch}}} \end{bmatrix}. \quad (5)$$

This approximation holds for small angular errors (e.g., $|\Delta\theta_{\text{yaw}}|, |\Delta\varphi_{\text{pitch}}| \lesssim 5^\circ$) around the optical axis. We set the gain λ so that each update induces at most 5% of the image width in pixel shift, preventing overshoot and oscillation.

Using the classic image-based visual servoing framework [29], [30], the required angular correction is calculated as:

$$\begin{bmatrix} \Delta\theta_{\text{yaw}} \\ \Delta\varphi_{\text{pitch}} \end{bmatrix} = -\lambda \mathbf{J}^{-1} \mathbf{e}. \quad (6)$$

Since \mathbf{J} is diagonal, this results in two decoupled first-order control loops. Applying the correction $(\Delta\theta_{\text{yaw}}, \Delta\varphi_{\text{pitch}})$ to the initial pose \mathbf{T}_0 yields the final, refined pose \mathbf{T}^* .

D. Reflective Reasoning

In complex or multi-subject scenes (with distractors or subtle compositional requirements), the single-anchor prior can be brittle; hence the analytical solution may require further refinement. To address this, we introduce a reflective optimization module that iteratively improves the camera pose by performing *visual reflection* within a geometrically faithful internal simulator. Inspired by the “reflection” loop from language agents [13], [14], we adapt this paradigm from symbolic text space into 3D visual reasoning, leveraging a photorealistic and geometry-accurate world model based on 3D Gaussian Splatting [25]. By probing small viewpoint changes, the LMM exploits motion-induced regularities, enabling robust composition in multi-object scenes, see Sec. IV-A.

Unlike latent imagination methods prone to hallucinations [16], our 3DGS model offers precise control over rendered views, ensuring that each “reflection” corresponds to a true visual consequence.

As illustrated in Figure 4, the agent closes the perception–action loop entirely inside the 3DGS world model before a single physical motion command is issued. Algorithm 1 concisely summarizes this Observe–Think–Act reflective loop.

Algorithm 1 Reflective Reasoning

Require: Natural-language instruction \mathcal{L} , observations \mathcal{O} (including the current RGB image \mathcal{I}_0), 3DGS world model \mathcal{G}

Ensure: Final camera pose $\mathbf{T}^* \in \text{SE}(3)$

1. $(\mathcal{I}_0, \mathcal{Z}) \leftarrow \text{EXTRACTINPUTS}(\mathcal{O})$
 2. $\mathbf{g} = (u^*, v^*, s, \theta, \varphi) \leftarrow \text{INTENTPARSING}(\mathcal{L}, \mathcal{I}_0, \mathcal{Z})$
 3. $\mathbf{T}_0 \leftarrow \text{GEOMETRICSOLVE}(\mathbf{g})$
 4. $\mathbf{x}^* \leftarrow \text{POSETOSPHERICAL}(\mathbf{T}_0)$
 5. **for** $t = 0$ **to** $K - 1$ **do**
 6. $\mathcal{C}_t \leftarrow \{\mathbf{x}^* \oplus \pm\delta\rho, \mathbf{x}^* \oplus \pm\delta\theta, \mathbf{x}^* \oplus \pm\delta\varphi\}$
 7. **for all** $\mathbf{x}_i \in \bar{\mathcal{C}}_t$ **do**
 8. $\tilde{\mathcal{I}}_i \leftarrow \mathcal{W}(\mathbf{x}_i, \mathcal{G})$
 9. $a_i \leftarrow A(\tilde{\mathcal{I}}_i, \mathcal{L})$
 10. **end for**
 11. $\mathbf{x}' \leftarrow \arg \max_{\mathbf{x}_i \in \bar{\mathcal{C}}_t} a_i$
 12. **if** $a(\mathbf{x}') - a(\mathbf{x}^*) < \epsilon$ **then**
 13. **break**
 14. **end if**
 15. $\mathbf{x}^* \leftarrow \mathbf{x}'$
 16. **end for**
 17. $\mathbf{T}^* \leftarrow \text{SPHERICALTOPOSE}(\mathbf{x}^*)$
 18. **return** \mathbf{T}^*
-

At each iteration t , we begin from the current best pose. To enable fine-grained and interpretable optimization, we operate not on the full $\text{SE}(3)$ pose T directly, but on its spherical coordinate parameterization, $\mathbf{x}_t^* = (\rho_t, \theta_t, \varphi_t)$. This three-parameter vector defines the camera’s position, while its orientation is implicitly determined by two constraints: the camera always points towards the subject, and its roll angle is fixed at zero (as detailed in Section III-C). This representation decomposes the complex 6-DoF exploration problem into independent adjustments of distance, azimuth, and elevation.

A set of candidate poses is generated by applying a single-axis perturbation to the current best parameters \mathbf{x}_t^* :

$$\mathcal{C}_t = \{\mathbf{x}_t^* \oplus [\pm\delta\rho, 0, 0], \mathbf{x}_t^* \oplus [0, \pm\delta\theta, 0], \mathbf{x}_t^* \oplus [0, 0, \pm\delta\varphi]\}, \quad (7a)$$

$$\bar{\mathcal{C}}_t = \mathcal{C}_t \cup \{\mathbf{x}_t^*\}, \quad |\bar{\mathcal{C}}_t| = 7. \quad (7b)$$

To evaluate each candidate $\mathbf{x}_i \in \bar{\mathcal{C}}_t$, it is first converted from its spherical parameterization back into a full $\text{SE}(3)$ pose, which is then rendered into an image $\tilde{\mathcal{I}}_i$ via the world model \mathcal{G} . This ensures both simplicity in optimization and fidelity in physical representation.

The LMM serves as a vision-language critic $A(\cdot, \mathcal{L})$, assigning a 5-point scalar score $a_i = A(\tilde{\mathcal{I}}_i, \mathcal{L})$ to each image. After identifying the highest-scoring candidate, the model performs causal reasoning to explain the success (e.g., “Increasing azimuth θ improved composition by creating more looking room for the subject”). The single-axis sampling strategy is critical to isolating such causal effects. We update the best pose \mathbf{x}_{t+1}^* based on the scoring results. The LMM’s causal explanation from the previous step informs the next sampling direction. The process repeats until either the score

gain falls below a threshold ϵ or the iteration count reaches a maximum K . The final pose \mathbf{x}^* is selected globally from the union of all candidates:

$$\mathbf{x}^* = \arg \max_{\mathbf{x}_i \in \bigcup_{t=0}^{K-1} \bar{\mathcal{C}}_t} A(\mathcal{W}(\mathbf{x}_i, \mathcal{G}), \mathcal{L}). \quad (8)$$

This “propose-simulate-critique-learn” loop transforms a blind exploration problem into a guided optimization process. Enabled by the high-quality initial solution and millisecond-level 3DGS rendering speeds [21], our reflective loop typically converges within few iterations, achieving real-time performance. We use $K=3$, $\epsilon=0.2$, $\delta\theta=\delta\varphi=8^\circ$, and a radial step $\delta\rho = \alpha \rho_t$ with $\alpha=0.1$ (10% of the current distance).

IV. EXPERIMENTS

A. Spatial Reasoning Evaluation

We designed several scenarios to separately evaluate the agent’s spatial imagination and instruction-following capabilities, demonstrated through horizontal movements and pitch adjustments, as well as its spatial composition skills when capturing images involving multiple objects.

1) *Experimental Setup:* All experiments were conducted on a workstation equipped with an NVIDIA RTX 4090D GPU. Physics-based simulation and rendering were performed in Isaac Sim [31], with all scene assets sourced exclusively from its built-in content library to ensure consistency and reproducibility. All methods use the same multimodal model (GPT-4.1).

We define one iteration (step) as predicting one camera pose, executing it by directly setting the camera in Isaac Sim, and rendering the next RGB observation as input for the subsequent step. As this process is fully simulated, it introduces no physical motion time and requires no 3D reconstruction overhead.

2) *Baseline:* We adopt two direct-pose baselines following the ReAct/Reflexion-style prompting paradigm [13], [14], while keeping the input interface consistent with our method.

Direct-6-DoF. We use chain-of-thought (CoT) prompting [12] to directly predict a 6-DoF camera pose at each step.

Direct-6-DoF w/ Reflection. This baseline follows the same direct 6-DoF execution loop, but after observing the outcome of the previous step, it performs an explicit reflection to analyze the consequence of the executed motion and decide how to adjust the next move, before predicting the subsequent 6-DoF pose.

Both methods operate under the same three-step environment-interaction budget; the reflection stage is an additional reasoning call and does not increase the number of environment interactions.

3) *Comparative Experiments:* To evaluate the performance of different methods, we designed three gradient tasks with increasing difficulty, as visually illustrated in Figure 5.

- **Easy—Isolated Banana.** Center-frame a single banana in a low-clutter scene to test the reliability of object localization and basic pan-tilt control.

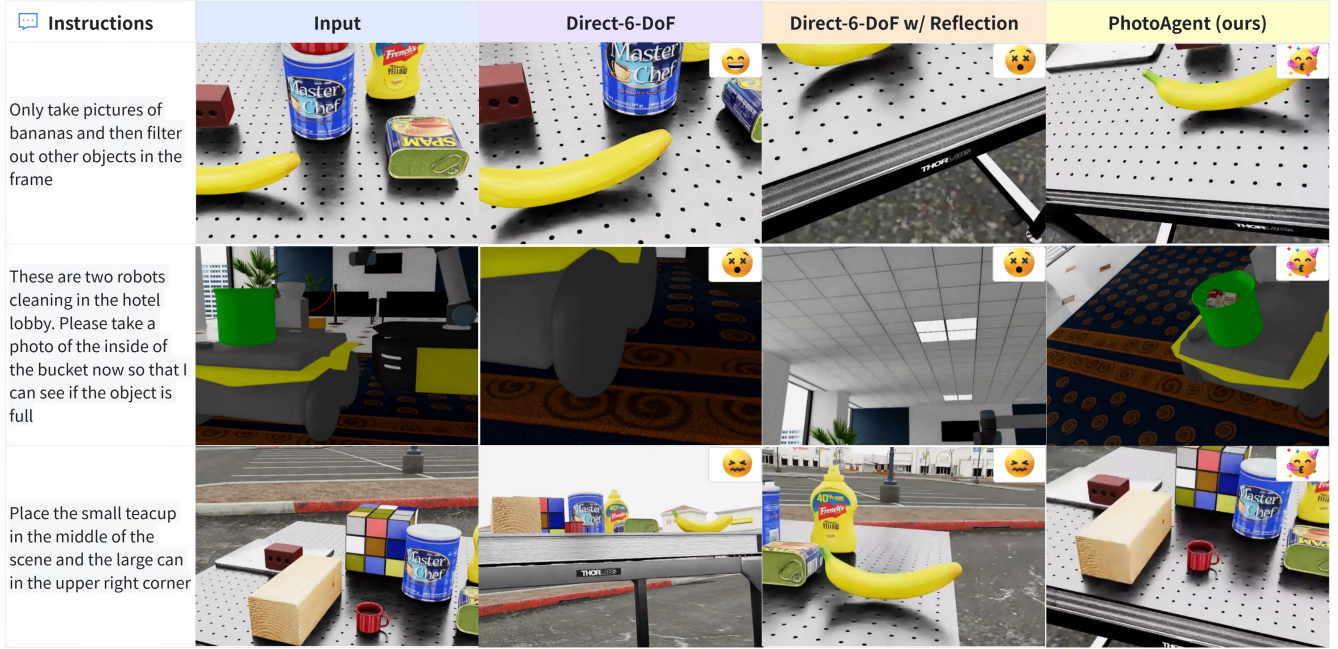


Fig. 5. Performances of our method and baselines on three tasks of different levels.

- **Medium—Cabin Inspection.** Capture a top-down image of a cleaning robot’s cabin to determine whether it is full. This task introduces moderate visual clutter, requires nontrivial viewpoint selection, and includes a simple semantic verification.
- **Hard—Multi-Object Composition.** Reframe via camera motion to center the *cup* and place the *can* in the upper-right quadrant without altering the scene; this requires precise horizontal translation, pitch control, and multi-object spatial reasoning.

4) *Evaluation Metrics:* Evaluation is based on success rate under a uniform three-step interaction budget. A trial is counted as successful if the agent achieves the objective within this budget. Table I reports the mean number of interaction steps over successful trials, with success counts (out of three trials) in parentheses; failed trials are excluded. Lower step counts indicate higher efficiency.

TABLE I: Performance Comparison on Gradient Tasks

Method	Easy	Medium	Hard
Direct-6-DoF	3.00 (2/3)	N/A (0/3)	N/A (0/3)
Direct-6-DoF w/ Reflection	3.00 (1/3)	N/A (0/3)	N/A (0/3)
PhotoAgent (ours)	2.33 (3/3)	2.00 (3/3)	2.00 (2/3)

5) *Results and Analysis:* As shown in Table I, our method achieves higher success rates and requires fewer interaction steps across all task difficulties. Empirically, Direct-6-DoF and Direct-6-DoF w/ Reflection produce coherent actions on simple tasks but degrade on complex scenarios. The key failure mode is direct 6-DoF pose regression, which is highly sensitive to initial deviations: a large first step pushes inference outside a trust region, lacks contractivity, and compounds errors. In contrast, our method exploits

spatial structure and adopts an azimuth-based incremental parameterization. This preserves spatial coherence, promotes contractive updates, reduces sensitivity to the first step, and curbs error accumulation, yielding superior stability, convergence, and success rates on complex tasks.

B. Aesthetic and Instruction Adherence Evaluation

We conducted a human-centered evaluation study to evaluate the aesthetic quality and instruction adherence of the photographs generated by **PhotoAgent**. Drawing on prior work in robotic photography user studies [1], [5], we aimed to assess: (1) aesthetic improvement, (2) instruction alignment, and (3) statistical significance of results.

1) *Experimental Setup.* : We deployed PhotoAgent on a custom mobile manipulator composed of an Agilex RangeMini2 mobile base and a TechRobots TB6-R3 6-DoF arm. An Intel RealSense D435i was mounted as the end-effector camera. Onboard computation was handled by a Thunderobot mini PC with an NVIDIA RTX 4070 Laptop GPU (8GB).

The system architecture followed our method design. We used GroundingDINO [32] for open-vocabulary detection. In portrait scenarios, we employed MediaPipe FaceMesh [33] to extract facial landmarks as prior cues. The 3D scene was constructed using AnySplat [25], aligned via VINS-Fusion odometry [34]. We capture 5-7 views around the subject and reconstruct a 3DGS scene via a single feed-forward AnySplat pass; reconstruction is seconds-level [25].

For each scenario, we compare a baseline photo and an optimized photo. The baseline is the initial (unoptimized) view, deterministically chosen as the first captured view on our predefined initialization trajectory, while the optimized photo is the final output produced by our full pipeline.

TABLE II: Experimental scenarios and user instructions. \S denotes simulated scenes; \ddagger indicates same subject under different instructions.

ID	Scene Name	User Instruction
(a)	Girl_Portrait \ddagger	Take a close-up from the front.
(b)	Girl_Reading \ddagger	Capture her reading intently.
(c)	Man_Whiteboard	Capture a thoughtful-looking expression.
(d)	Girl_Library \ddagger	Take a beautiful photo.
(e)	Teddy_Lab	Give the teddy bear a close-up shot.
(f)	Dolls_Confrontation	Take a photo of the toys with visual tension, like the reference.
(g)	RoboDog_Factory \S	Carefully photograph the contents of the box on the robotic dog.
(h)	Truck_Warehouse \S	Capture a full shot of the truck.

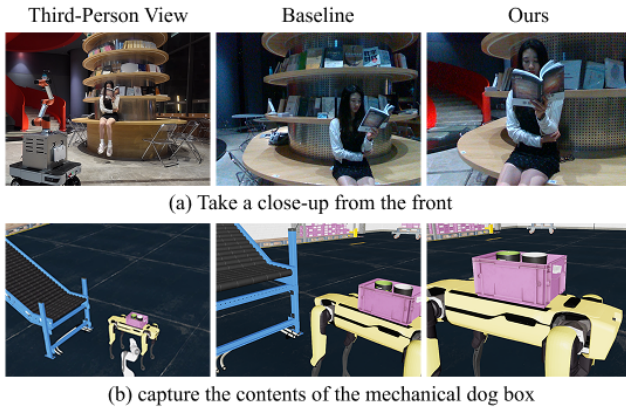


Fig. 6. Qualitative examples from real (a) and simulated (b) environments. Each row shows: third-person view, baseline, and PhotoAgent’s output.

A total of 100 volunteers participated in a two-phase online study.

Phase 1 (Independent Rating). All 16 images (8 scenarios \times 2 versions: baseline and ours) were shown in randomized order. Participants rated aesthetic appeal on a 5-point Likert scale.

Phase 2 (Paired Comparison). Participants compared baseline and optimized photos side-by-side with the original instruction and selected which better fulfilled the goal.

Evaluation Metrics. We used:

- **Mean Opinion Score (MOS):** average human rating;
- **GoB (%)**: “Good-or-Better” rate (score \geq 4);
- **Instruction Adherence Win Rate (IAWR) (%)**: instruction adherence preference in Phase 2.

Our metric choices follow established practice in *robotic photography* user studies: prior systems evaluate image quality with 5-point human ratings and report MOS and distributional summaries [1], [2], [9]–[11]; and instruction-following or aesthetic preference is routinely measured via *paired comparison* with per-scene win rates (voting-based preference), as in AutoPhoto and PhotoBot [3], [5].

We treat participants as the independent unit. Per scenario ($n=100$), MOS was tested with a paired Wilcoxon signed-rank test, and Phase-2 IAWR with a one-sided exact binomial test ($H_0: p=0.5$, $H_1: p>0.5$). Bonferroni correction was applied across 8 scenarios ($\alpha'=0.00625$), following common practice

TABLE III: Performance comparison between PhotoAgent (ours) and the baseline across categories.

Category (N)	Metric	Baseline	PhotoAgent (ours)	Δ
Portraits (4)	MOS	2.86	3.82	+0.96
	GoB	25.2%	68.5%	+43.2 pp
	IAWR	—	89.5%	—
Still Life (4)	MOS	2.88	3.94	+1.07
	GoB	28.2%	71.2%	+43.0 pp
	IAWR	—	96.2%	—
Overall (8)	MOS	2.87	3.88	+1.01
	GoB	26.8%	69.9%	+43.1 pp
	IAWR	—	92.9%	—

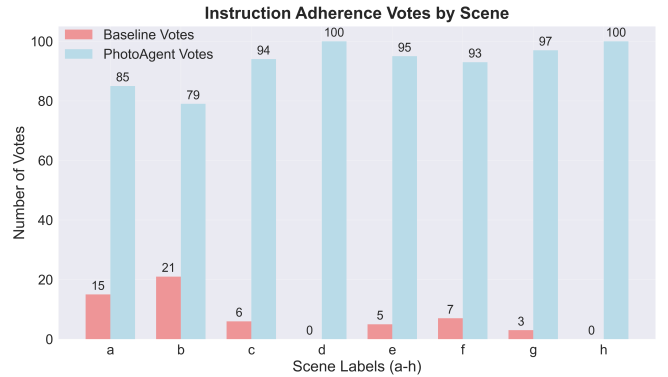


Fig. 7. Stage-2 instruction adherence. Per-scene win rates under paired comparison. Scene IDs (a–h) are defined in Table II.

in robot-photography user studies [3], [5], [11].

2) *Experimental Design:* We curated 8 scenarios (Table II) spanning portraits and still-life settings in both real and simulated environments. User instructions varied in abstraction—from direct composition commands (e.g., “close-up”) to affective intent (e.g., “thoughtful-looking expression”).

3) *Results and Analysis:* Figure 6 shows results from two representative scenarios. In *Girl_Portrait*, PhotoAgent produces a frontal close-up with subject-centered composition that fulfills the aesthetic intent. In *RoboDog_Factory*, it interprets spatially complex instructions, selects a novel viewpoint to reveal the box contents, and excludes visual clutter.

The quantitative data in Table III confirms our qualitative findings. PhotoAgent markedly improves aesthetic outcomes, boosting the overall MOS by **1.01** points and the GoB rate by **43.1** percentage points. This improvement is consistent across both portrait (**+0.96** MOS) and still-life (**+1.07** MOS) categories, demonstrating the robustness of our method while still achieving a **92.9%** instruction adherence rate.

Figure 7 reports Stage-2 instruction-adherence results by scene: win rates span **79–100%**. Object-centric scenes average \sim **96.2%**, slightly higher than portraits \sim **89.5%**. The lowest case (\sim **79%**, *Girl_Reading*) reflects a more abstract, mid-level instruction alongside limited permissible camera motion, which constrains attainable improvement despite correct intent understanding.

Across all 8 scenarios, MOS gains were significant (Wilcoxon; all raw $p<.001$, remaining significant under

Bonferroni), with Cohen’s d_z ranging from 0.55 to 0.88. IAWR values also exceeded chance in all scenarios (one-sided binomial; all raw $p < .001$, remaining significant under Bonferroni).

V. CONCLUSION

We introduced *PhotoAgent*, an embodied robotic-photography system powered by LMM-guided reasoning. By formulating camera control as an inverse *viewpoint-solving* problem, *PhotoAgent* interprets user instructions, solves geometric constraints analytically, and refines its decisions via visual reflection in a 3D Gaussian-splat world model. Experiments demonstrate superior spatial reasoning and aesthetic composition over baselines.

Beyond photography, the viewpoint-solving paradigm holds promise for broader embodied AI tasks. Most current systems decouple navigation and manipulation, with the latter often relying on a static camera view—making occlusions or ambiguity critical failure points. Active viewpoint exploration of *PhotoAgent*, similar to human perspective shifts, can bridge this gap, enabling stronger embodied intelligence that is aware of perception.

REFERENCES

- [1] Z. Byers, M. Dixon, W. D. Smart, and C. M. Grimm, “Say cheese! experiences with a robot photographer,” *AI magazine*, vol. 25, no. 3, pp. 37–37, 2004.
- [2] K. Lan and K. Sekiyama, “Autonomous robot photographer with kl divergence optimization of image composition and human facial direction,” *Robotics and Autonomous Systems*, vol. 111, pp. 132–144, 2019.
- [3] H. AlZayer, H. Lin, and K. Bala, “Autophoto: Aesthetic photo capture using reinforcement learning,” in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 944–951.
- [4] H. Kang, J. Zhang, H. Li, Z. Lin, T. Rhodes, and B. Benes, “Lerop: A learning-based modular robot photography framework,” *arXiv preprint arXiv:1911.12470*, 2019.
- [5] O. Limoyo, J. Li, D. Rivkin, J. Kelly, and G. Dudek, “Photobot: Reference-guided interactive photography via natural language,” in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 2479–2486.
- [6] S. Hentschel, K. Kobs, and A. Hotho, “Clip knows image aesthetics,” *Frontiers in Artificial Intelligence*, vol. 5, p. 976235, 2022.
- [7] R. Jiang and C. W. Chen, “Multimodal llms can reason about aesthetics in zero-shot,” in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 6634–6643.
- [8] Z. Liao, X. Liu, W. Qin, Q. Li, Q. Wang, P. Wan, D. Zhang, L. Zeng, and P. Feng, “Humanaesexpert: Advancing a multi-modality foundation model for human image aesthetic assessment,” *arXiv preprint arXiv:2503.23907*, 2025.
- [9] M. Zabaruskas and S. Cameron, “Luke: An autonomous robot photographer,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 1809–1815.
- [10] R. Gadde and K. Karlapalem, “Aesthetic guideline driven photography by robots,” in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, vol. 22, no. 3, 2011, p. 2060.
- [11] R. Newbury, A. Cosgun, M. Koseoglu, and T. Drummond, “Learning to take good pictures of people with a robot photographer,” in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 11 268–11 275.
- [12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.
- [13] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, and Y. Cao, “React: Synergizing reasoning and acting in language models,” in *The Eleventh International Conference on Learning Representations*, 2023.
- [14] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, “Reflexion: Language agents with verbal reinforcement learning,” *Advances in neural information processing systems*, vol. 36, pp. 8634–8652, 2023.
- [15] D. Ha and J. Schmidhuber, “World models,” *arXiv preprint arXiv:1803.10122*, vol. 2, no. 3, p. 440, 2018.
- [16] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, “Dream to control: Learning behaviors by latent imagination,” *arXiv preprint arXiv:1912.01603*, 2019.
- [17] D. Hafner, T. Lillicrap, M. Norouzi, and J. Ba, “Mastering atari with discrete world models,” *arXiv preprint arXiv:2010.02193*, 2020.
- [18] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, “Mastering diverse control tasks through world models,” *Nature*, vol. 640, no. 8059, pp. 647–653, 2025.
- [19] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “Nerf: Representing scenes as neural radiance fields for view synthesis,” *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.
- [20] T. Müller, A. Evans, C. Schied, and A. Keller, “Instant neural graphics primitives with a multiresolution hash encoding,” *ACM transactions on graphics (TOG)*, vol. 41, no. 4, pp. 1–15, 2022.
- [21] B. Kerbl, G. Kopanas, T. Leimkühler, G. Drettakis *et al.*, “3d gaussian splatting for real-time radiance field rendering,” *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.
- [22] H. Matsuki, R. Murai, P. H. Kelly, and A. J. Davison, “Gaussian splatting slam,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 039–18 048.
- [23] N. Keetha, J. Karhade, K. M. Jatavallabhula, G. Yang, S. Scherer, D. Ramanan, and J. Luiten, “Splatam: Splat track & map 3d gaussians for dense rgb-d slam,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 21 357–21 366.
- [24] Z. Zhu, S. Peng, V. Larsson, W. Xu, H. Bao, Z. Cui, M. R. Oswald, and M. Pollefeys, “Nice-slam: Neural implicit scalable encoding for slam,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 786–12 796.
- [25] L. Jiang, Y. Mao, L. Xu, T. Lu, K. Ren, Y. Jin, X. Xu, M. Yu, J. Pang, F. Zhao *et al.*, “Anysplat: Feed-forward 3d gaussian splatting from unconstrained views,” *ACM Transactions on Graphics (TOG)*, vol. 44, no. 6, pp. 1–16, 2025.
- [26] S. G. Patil, T. Zhang, X. Wang, and J. E. Gonzalez, “Gorilla: Large language model connected with massive apis,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 126 544–126 565, 2024.
- [27] Y. Qin, S. Liang, Y. Ye, K. Zhu, L. Yan, Y. Lu, Y. Lin, X. Cong, X. Tang, B. Qian *et al.*, “Toolllm: Facilitating large language models to master 16000+ real-world apis,” *arXiv preprint arXiv:2307.16789*, 2023.
- [28] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman *et al.*, “Do as i can, not as i say: Grounding language in robotic affordances,” *arXiv preprint arXiv:2204.01691*, 2022.
- [29] S. Hutchinson, G. Hager, and P. Corke, “A tutorial on visual servo control,” *IEEE Transactions on Robotics and Automation*, vol. 12, no. 5, pp. 651–670, 1996.
- [30] F. Chaumette, “Visual servoing,” in *Computer vision: a reference guide*. Springer, 2021, pp. 1367–1374.
- [31] NVIDIA Corporation, “Isaac sim,” Online documentation, 2024. [Online]. Available: <https://developer.nvidia.com/isaac-sim>
- [32] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su *et al.*, “Grounding dino: Marrying dino with grounded pre-training for open-set object detection,” in *European conference on computer vision*. Springer, 2024, pp. 38–55.
- [33] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee *et al.*, “Mediapipe: A framework for building perception pipelines,” *arXiv preprint arXiv:1906.08172*, 2019.
- [34] T. Qin, P. Li, and S. Shen, “Vins-mono: A robust and versatile monocular visual-inertial state estimator,” *IEEE transactions on robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.