Appendix

A Full Algorithmic Workflow

A.1 System Architecture and Agent Toolchain

Figure 1 shows a real-world execution example of PhotoAgent, which follows the modular structure outlined in our method. The agent operates in an Observe–Think–Act loop, invoking tool-style modules at each stage to process perception, perform spatial reasoning, and execute camera control.

The perception module extracts semantic and geometric information from raw images. We use GroundingDINO for open-vocabulary object detection. In portrait scenarios, we incorporate MediaPipe FaceMesh to extract facial landmarks such as head orientation, gaze, and key contour points, which serve as prior cues for viewpoint planning.To simulate view changes before real execution, we reconstruct the environment using AnySplat to produce a 3D Gaussian scene, aligned with camera trajectory via VINS-Fusion odometry. The 6-DoF pose solver is a custom module developed in-house. It analytically maps the aesthetic and spatial constraints predicted by the language model into an interpretable $SE(3)$ pose.

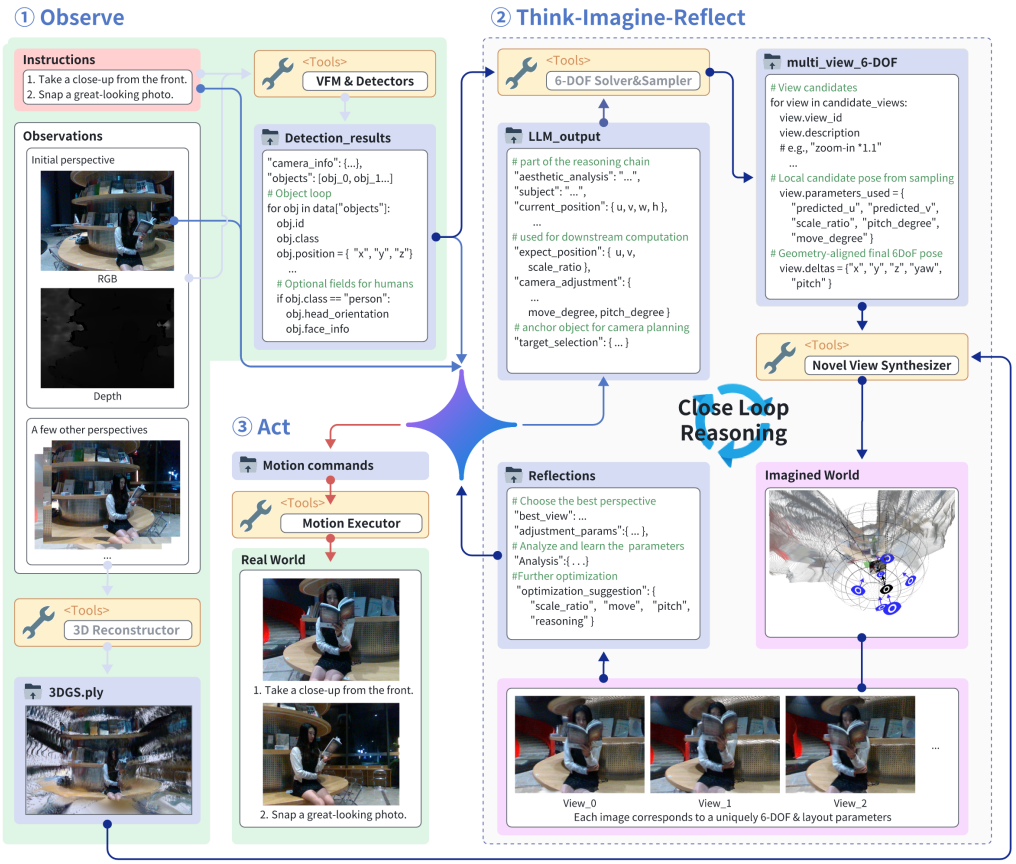


Figure 1: Real-world execution example of PhotoAgent.

A.2 Algorithmic Workflow

To provide a clear overview of the system logic, we include the complete pseudocode for the PhotoAgent execution pipeline in algorithm 1. The algorithm integrates three core modules: language-

guided intent parsing, geometric pose solving, and reflective viewpoint optimization. The output is the optimized camera pose $\mathbf{x}^*$ that satisfies both aesthetic and semantic intent.

Algorithm 1 PhotoAgent: Inverse Viewpoint-Solving with Reflective Optimization

Require: Natural-language instruction L, observations $\mathcal{O}$, 3DGS world model $\mathcal{G}$
Ensure: Optimal camera pose $\mathbf{x}^* \in SE(3)$

$\mathbf{Z} \leftarrow \textsc{ExtractInputs}(\mathcal{O})$ ▷ 2D/3D detections and semantic cues
$\mathbf{g} = (u^*, v^*, s, \theta, \phi) \leftarrow \textsc{IntentParsing}(L, \mathbf{Z})$ ▷ Geometric constraint vector
$\mathbf{x}_0 \leftarrow \textsc{GeometricSolve}(\mathbf{g})$ ▷ Closed-form mapping to $SE(3)$
$\mathbf{x}^* \leftarrow \mathbf{x}_0$
for $t = 0$ **to** $K - 1$ **do**
  $C_t \leftarrow \{\mathbf{x}^* \oplus \delta\rho,\ \mathbf{x}^* \oplus \delta\theta,\ \mathbf{x}^* \oplus \delta\phi\}$ ▷ Perturbations in distance/angle
  for all $\mathbf{x}_i \in C_t$ **do**
    $I_i \leftarrow W(\mathbf{x}_i, \mathcal{G})$ ▷ Render via 3D Gaussian Splatting
    $a_i \leftarrow A(I_i, L)$ ▷ LMM critic score
  end for
  $\mathbf{x}' \leftarrow \arg\max_{\mathbf{x}_i \in C_t} a_i$
  if $a(\mathbf{x}') - a(\mathbf{x}^*) < \epsilon$ **then**
    break
  end if
  $\mathbf{x}^* \leftarrow \mathbf{x}'$
end for
return $\mathbf{x}^*$

B Space Reasoning

As shown in Figure 2, it presents the results of each iterative step of different methods under various spatial reasoning tasks. Different spatial reasoning scenarios are included, such as taking pictures of specific objects (like bananas) and filtering out others, photographing the interior of a bucket in a scene with robots, and placing objects (small teacup and large can) in designated positions. Through these visual results of iterative steps, we can intuitively observe the performance differences of methods like ReACT, Reflexion, and OURS in handling spatial reasoning problems.

C Prompt Templates

To support language-driven spatial reasoning, we design two task-specific prompt templates that guide the LMM to perform different roles within the PhotoAgent system. Both prompts elicit structured JSON outputs to ensure interpretability and facilitate downstream execution.

The first prompt (Figure 3) is used in scenarios where the model is given a single image with detected objects and a user instruction. The model is asked to identify the most relevant subject, analyze the visual composition, and recommend how the framing can be improved. The output includes the chosen subject ID, semantic justification, aesthetic reasoning, expected position and scale, and a set of camera adjustment parameters (e.g., azimuth and pitch).

The second prompt (Figure 4) is used in cases where multiple candidate viewpoints are rendered from the same scene under varied camera parameters (e.g., zoom, azimuth, pitch). The model is asked to evaluate these options and select the best view, providing detailed reasoning for its choice. Additionally, it may suggest further refinements to improve composition. The response includes both comparative analysis of parameter changes and a final set of optimized camera parameters for execution.

Figure 2: The detailed procedures in the process of spatial reasoning.

D Qualitative Examples Across Scenarios

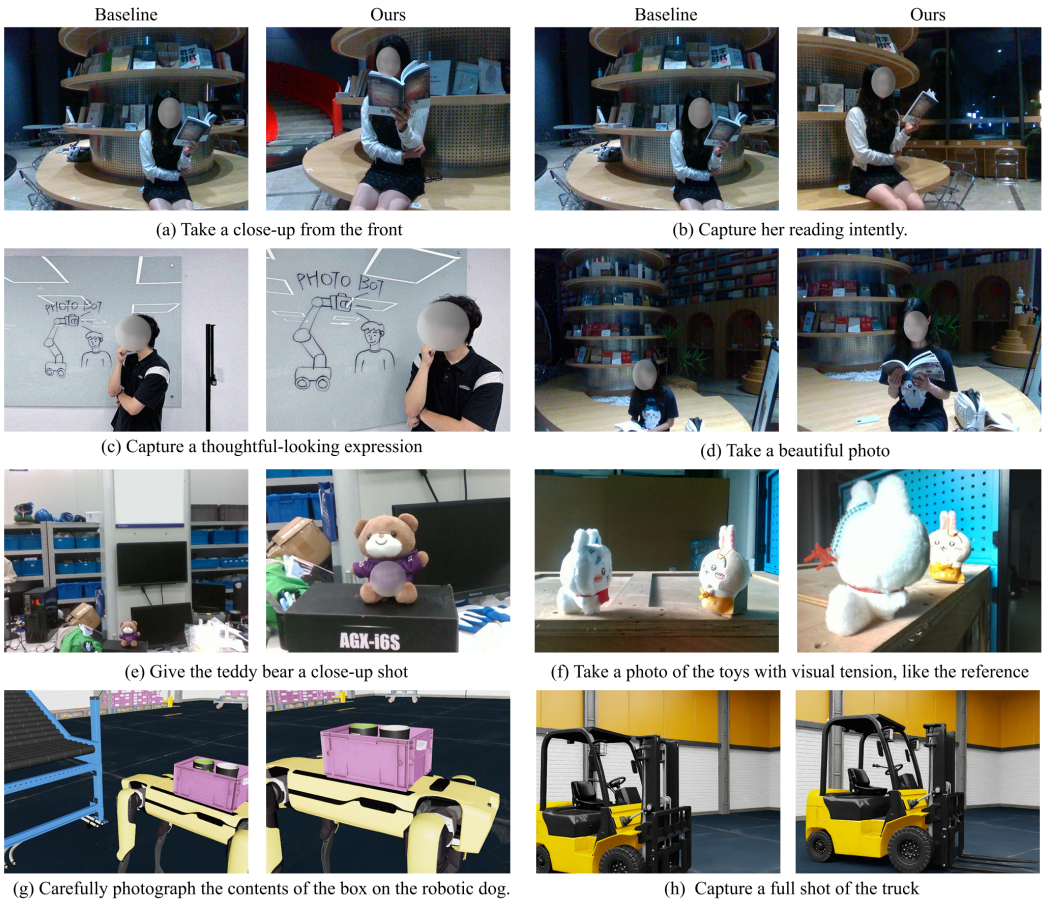
Figure 5 presents visual examples from eight different test scenarios used in our experiments. These tasks span both real-world and simulated environments, and involve a range of compositional goals—such as emphasizing a subject's expression, creating visual tension between objects, or capturing the contents of a robotic container.

```
System: You are a professional photographic-composition expert with deep knowledge
    of camera parameters and 3D spatial imaging principles.

User: Please analyze this photo with a bounding box and provide detailed aesthetic
    analysis based on the following instruction:
        User_instruction: {User_instruction}
        Camera_info: {camera_info}
        Detected_objects: {detected_objects}
        Image_size: {width}x{height}
        [Input Image]

Assistant:
{
  "aesthetic_analysis": "A detailed critique covering: spatial relationships;
    subject-scene context; and alignment with the user's intent.",
  "subject": {
    "target_object_id": <integer>,         // ID of the selected object
    "reasoning": "<why this object?>"       // Rationale for selection
  },
  "valuable_elements": ["<element1>", ...],   // List of other elements to preserve
  "recommended_composition": {
    "style": "<e.g. 'right-third composition'>",
    "justification": "Why this rule, from photographic, spatial, and aesthetic
    perspectives"
  },
  "current_position": {
    "u": <pixels>,   // Bounding-box center u-coordinate
    "v": <pixels>,   // Bounding-box center v-coordinate
    "w": <pixels>,   // Bounding-box width
    "h": <pixels>    // Bounding-box height
  },
  "expect_position": {
    "u": <pixels>,         // Desired u after adjustment
    "v": <pixels>,         // Desired v after adjustment
    "scale_ratio": <float> // >1 to zoom in, <1 to zoom out
  },
  "camera_adjustment": {
    "azimuth_constraint": -1 | 0 | 1,  // -1: CW; 0: none; 1: CCW around subject
    "pitch_constraint":   -1 | 0 | 1,  // -1: tilt up; 0: none; 1: tilt down
    "azimuth_explanation":  "Why this azimuth move? (relation to gaze & composition)
    ",
    "elevation_explanation":"Why this pitch move? (narrative & environmental context
    )",
    "expect_angle": {
      "azimuth":   <degrees>, // Absolute degrees to rotate around subject
      "elevation": <degrees>  // +: tilt down; -: tilt up
    }
  },
  "coordinate_system":     {your-coordinate-config},
  "face_orientation_spec": {your-face-orientation-config}
}
```

Figure 3: Prompt template for the PhotoAgent composition task.

```
System:
You are a senior photography evaluation expert with professional experience in
    composition analysis, aesthetic judgment, and parameter-aware interpretation.

Your task is to analyze a set of candidate photos taken under different camera
    parameters and select the best viewpoint that aligns with both the user's
    intent and aesthetic standards.

Evaluation criteria:
1. Aesthetic quality: composition balance, visual guidance, and spatial
    relationships.
2. Instructional alignment: how well the view matches the user's intent and subject
    emphasis.
3. Parameter-specific effects: analyze how zoom ratio, azimuth shift, and pitch
    angle affect the actual visual composition (based on observable content, not
    prior knowledge).

Viewpoint definitions:
- View 0: Original parameters (reference)
- Views 1,2: Zoom variations
- Views 3,4: Azimuth shifts
- Views 5,6: Pitch angle shifts

User:
Please evaluate the following multi-view candidate images and select the best
    shooting angle.

    User instruction:{user_instruction}
    Initial aesthetic analysis:{extracted_analysis}
    Camera parameters for each viewpoint:{view_params_info}
    [Input Image]

Focus your analysis on the following parameter variations:
1. Zoom ratio (Views 1 & 2): Impact on subject size and image framing.
2. Azimuth angle (Views 3 & 4): Influence on perspective and spatial balance.
3. Pitch angle (Views 5 & 6): Effect on camera height, perspective, and narrative
    expression.

Assistant:
{
  "best_view_id": "<selected view index (0-6)>",
  "scale_ratio_analysis": "How zoom variation affects subject prominence and
    composition.",
  "move_angle_analysis": "How azimuth shift affects framing and spatial
    relationships.",
  "pitch_angle_analysis": "How pitch changes affect perspective and narrative
    emphasis.",
  "optimization_suggestion": {
    "scale_ratio_adjustment": "<suggested zoom ratio (e.g., 1.0 means unchanged)>",
    "move_degree_adjustment": "<suggested azimuth adjustment (in degrees)>",
    "pitch_degree_adjustment": "<suggested pitch adjustment (in degrees)>",
    "reasoning": "Justification for the above refinements. These are optional fine-
    tuning steps based on the selected best view and must not duplicate prior
    reasoning. Values should differ from existing settings unless reusing is
    strongly justified."
  }
}
```

Figure 4: Prompt template for multi-view composition evaluation.



(a) Take a close-up from the front

(b) Capture her reading intently.

(c) Capture a thoughtful-looking expression

(d) Take a beautiful photo

(e) Give the teddy bear a close-up shot

(f) Take a photo of the toys with visual tension, like the reference

(g) Carefully photograph the contents of the box on the robotic dog.

(h) Capture a full shot of the truck

Figure 5: Qualitative results across eight scenarios used in our evaluation. Each pair of images corresponds to a distinct user instruction.

To enable pairwise comparison in Stage 1, we collected user ratings on a total of 16 images, comprising 7 baseline views and 9 optimized views. Due to scenes *a* and *b* sharing the same baseline image, we introduced an additional optimized image from scene *d*—capturing the same environment but responding to a different instruction. As shown in Figure 6, baseline images concentrate around mid-range scores (2–3), while optimized results shift ratings rightward, with higher densities at Score 4 and 5. This distributional change complements the average MOS and GoB gains reported in the main text, and visually illustrates the perceptual margin between baseline and optimized outputs across both portrait and object-centric scenarios.

While the Stage 1 ratings yielded consistent and statistically significant improvements, we acknowledge several potential limitations inherent to subjective evaluation protocols. First, as images were presented in randomized but static order per participant, initial exposure to particularly weak or strong samples may have anchored subsequent ratings due to primacy or contrast effects. Second, the five-point Likert scale may induce centrality bias or generosity bias, wherein participants tend to avoid extreme scores (e.g., "1 – very poor") and favor mid-to-high ratings. This may explain the relatively low frequency of Score 1 even in baseline images . While these effects do not invalidate the observed trend—since our method still demonstrates a marked shift toward higher ratings—they highlight the value of incorporating complementary evaluation paradigms (e.g., forced pairwise comparisons, as in Stage 2) to validate aesthetic alignment from multiple perspectives.

As shown in Figure 7, PhotoAgent achieved dominant instruction adherence across all eight scenarios, with win rates ranging from 79% to 100%. Object-centric tasks (e.g., "capture the truck's full extent") generally yielded slightly higher agreement (avg. 96%) than person-centric ones (avg.

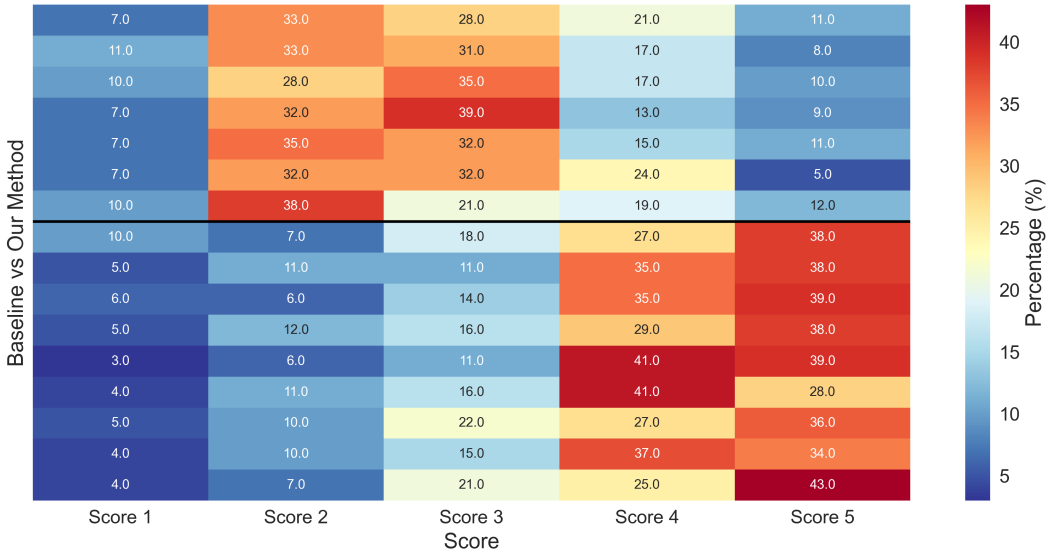


Figure 6: Normalized rating distribution heatmap for all 16 images in Stage 1. The top seven rows correspond to baseline images, while the bottom nine represent optimized outputs.

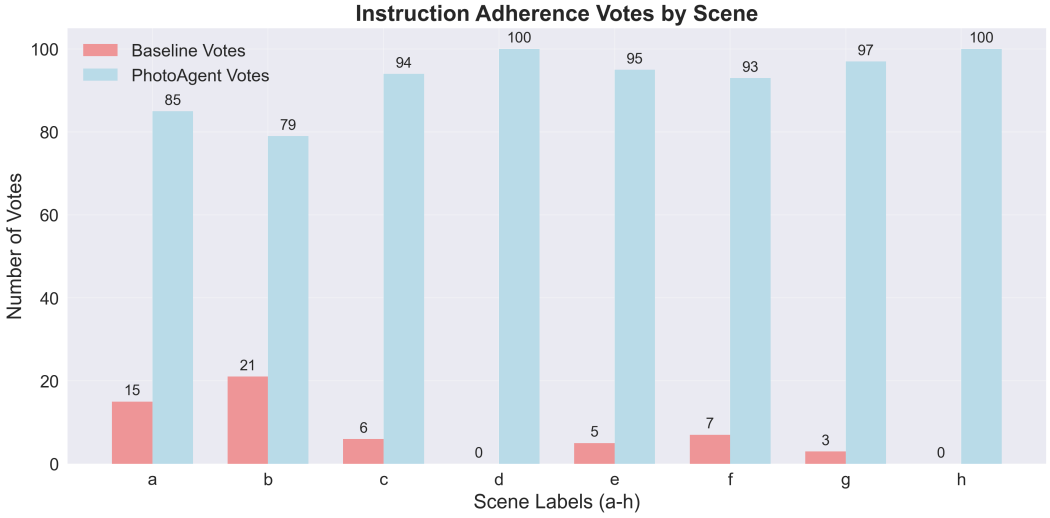


Figure 7: Instruction adherence votes per scene in the pairwise comparison study.

91%), likely due to reduced ambiguity and greater geometric clarity. Similarly, concrete instructions led to more consistent user preference than abstract ones (e.g., "take a beautiful photo"), though the system still performed strongly under both conditions. The only scenario with a comparatively lower win rate (79%) was the "focused reading" task, where the instruction's mid-level abstraction—emphasizing attentiveness without clear spatial constraints—combined with minimal pose variation between baseline and optimized views. As the subject remained largely static, user preference may have been influenced by secondary cues such as background composition or camera angle. This suggests that, for semantically diffuse goals with limited content variability, additional semantic reinforcement may be needed to amplify the optimization's perceived alignment.