

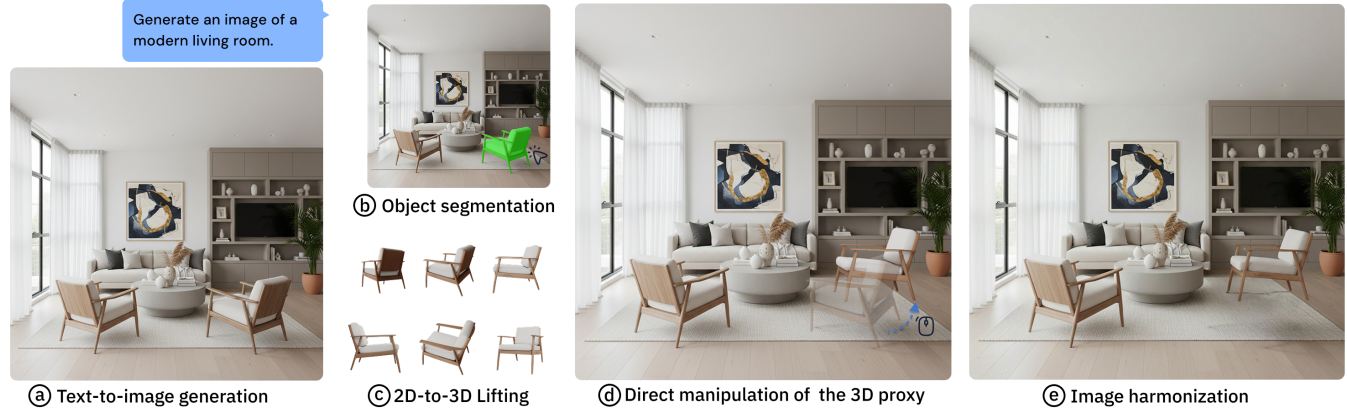
# ImageDirector: Expressing Spatial Intent via 3D Direct Manipulation for Generative Image Editing

Ricardo Mejia Araujo

Department of Computer Science  
New Jersey Institute of Technology  
Newark, New Jersey, USA  
ram227@njit.edu

Lei Zhang

Department of Informatics  
New Jersey Institute of Technology  
Newark, New Jersey, USA  
lei.zhang@njit.edu



**Figure 1: ImageDirector Workflow.** (a) **Text-to-image Generation:** An initial image is synthesized from a text prompt. (b) **Object Segmentation:** The user selects and segments an object of interest. (c) **2D-to-3D Lifting:** The system turns the 2D segment into a 3D proxy using a generative model. (d) **Direct Manipulation:** The 3D proxy enables interactive repositioning, rotation, and scaling while the background is automatically inpainted. (e) **Harmonization:** The system harmonizes the manipulated proxy back into the scene for a seamless final result.

## Abstract

Generative image editing through text prompting suffers from two fundamental limitations: text is imprecise for specifying spatial changes, and iterative edits often introduce unwanted visual drift. In contrast, people naturally approach image editing with a compositor’s intuition: manipulating individual objects while expecting the rest of the scene to remain intact. We present ImageDirector, a generative image editing system that supports this intuition by enabling users to manipulate objects in a 2D generated image as if they exist in 3D space. Given a generated image, our system segments a selected object, turns it into a manipulable high-fidelity 3D proxy via image-to-3D lifting, and lets users reposition, rotate, and scale it through direct manipulation. The system then re-composites and harmonizes the result with background regions preserved. In a within-subjects study ( $N = 12$ ), ImageDirector significantly outperformed a text-only prompting baseline in spatial accuracy and visual consistency across iterative edits; participants also reported substantially greater perceived control and expressiveness over the editing process. Finally, we discuss how this approach extends to

broader content creation workflows, opening new possibilities for intuitive, precise, and controllable human-AI co-creation.

## CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**; **Interaction design theory, concepts and paradigms**.

## Keywords

Generative AI, Human-AI Interaction, Image Generation, Large Language Models

### ACM Reference Format:

Ricardo Mejia Araujo and Lei Zhang. 2026. ImageDirector: Expressing Spatial Intent via 3D Direct Manipulation for Generative Image Editing. In *The 39th Annual ACM Symposium on User Interface Software and Technology (UIST '26)*, November 02–05, 2026, Detroit, MI, USA. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3830398.3830709>

## 1 Introduction

Generative text-to-image models have fundamentally changed how people create and iterate on visual content. What once required hours of manual compositing can now be accomplished in seconds: describe a scene, receive an image, refine through conversation. This fluency has made multi-turn generative editing a compelling paradigm for creative workflows, enabling users to incrementally build toward a visual intention without requiring expertise in traditional image editing tools [6, 51].



This work is licensed under a Creative Commons Attribution 4.0 International License. *UIST '26, Detroit, MI, USA*

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2856-3/26/11

<https://doi.org/10.1145/3830398.3830709>

Yet as users move from initial generation to the more demanding work of iterative refinement, two structural limitations become apparent. The first is *spatial imprecision*. Natural language is poorly suited to express spatial intent. Phrases like “rotate the lamp a little to the left” leave critical degrees of freedom unconstrained, and the model fills them stochastically. This imprecision manifests as two distinct forms of ambiguity: users lack a shared coordinate system to specify positions and orientations (*spatial reference ambiguity*), and struggle to identify specific objects when scenes contain multiple similar elements (*object reference ambiguity*). The resulting feedback loop of prompting via trial-and-errors is thus cognitively expensive, unpredictable, and frequently unsuccessful [25, 60]. The second limitation is *visual drift*. Because generative pipelines treat each edit as a new sampling event, regions that the user did not intend to change are subject to re-generation. Over iterative edits, unedited areas gradually diverge from the original (e.g., background elements move or disappear). Both limitations share a common root: existing systems treat the image as a flat linguistic artifact with no intuitive representation of objects in the image. The objects cannot be manipulated; they can only be regenerated.

We present ImageDirector, a generative image editing system that addresses these limitations by enabling users to manipulate objects in a 2D generated image as if they exist in 3D space. Given a generated image, our pipeline uses an image segmentation model to extract a selected object, lifts it into a manipulable 3D proxy via a pre-trained image-to-3D model, and enables direct manipulation of the proxy including translation, rotation, and scaling. The system then harmonizes the manipulated proxy back into the scene, producing a seamless composite. In this paradigm, direct selection and segmentation resolve object reference ambiguity; direct manipulation of the 3D proxy resolves spatial reference ambiguity. In addition, because the background is composited rather than regenerated, visual drift is eliminated at its source.

We evaluated our system in a within-subjects study (N=12) comparing ImageDirector against a state-of-the-art text-only prompting baseline (i.e., Nano Banana Pro) across structured spatial editing tasks for generated images. Participants using ImageDirector achieved significantly *higher spatial intent accuracy*, produced *lower visual drift* in unedited regions across iterative edits, and reported reduced cognitive workload and greater perceived control and expressiveness over the image editing process. At the same time, participants preferred the immediacy of the raw 3D proxy over the latency of harmonization. We also found that participants converged on a hybrid, multimodal workflow that combines text prompting for semantic contexts with direct manipulation for precise spatial edits. In summary, this paper makes the following contributions:

- An open-sourced<sup>1</sup>, 3D-grounded generative image editing system that enables users to express spatial intent via direct manipulation by integrating segmentation, inpainting, 2D-to-3D lifting, and harmonization models.
- Empirical evidence from a within-subjects user study comparing ImageDirector and a text-only prompting baseline.
- Design insights into how users construct and communicate spatial intent in generative image editing and multimodal human-AI interaction.

<sup>1</sup>Code available at: <https://github.com/Immersive-Creation-Lab/image-director>

## 2 Related Work

### 2.1 AI Image Generation and Editing

The rapid advancement of text-to-image generative models has fundamentally transformed how users create and edit visual content. Early approaches leveraged Generative Adversarial Networks (GANs), with Reed et al. demonstrating that convolutional GANs could synthesize images directly from natural language descriptions [40]. The introduction of transformer architectures and contrastive language-image pre-training (CLIP) [37] substantially improved semantic alignment between text and generated imagery, enabling models such as DALL-E [39] to achieve unprecedented image quality and fidelity. Latent diffusion models further advanced the field: Rombach et al.’s Stable Diffusion [41] performs the denoising process in a compressed latent space, enabling efficient synthesis at high resolution and providing the open-source foundation on which subsequent research including the work presented here builds.

**2.1.1 Limits of Spatial Language and Prompt Engineering.** Crafting effective text prompts for generative models has been a long studied challenge. Liu and Chilton’s foundational study [31] systematically characterized how different prompt formulations affect output quality, showing that style modifiers, rather than subject descriptions, exert disproportionate influence on aesthetic outcomes. Oppenlaender’s taxonomy of prompt modifiers documented the breadth of keyword categories practitioners use to steer image generation [35]. Despite this body of knowledge, prompting for spatial relationships presents a distinct and underexplored challenge. Subramonyam et al. theorize an instruction gap inherent to generative model interaction: models are highly sensitive to language precision, yet human language naturally tolerates variant expressions of the same intent—a mismatch that becomes most acute when the intended information is spatial and continuous rather than categorical [45]. Ko et al.’s interview study with 28 visual artists found that participants consistently identified the inability to specify precise spatial configurations as a critical unmet need, and that textual descriptions of orientation, depth placement, and compound transformations are inherently ambiguous [25]. Zamfirescu-Pereira et al. similarly documented that non-expert users frequently do not know where to begin when spatial intent must be expressed linguistically, and that model outputs regularly fail to reflect spatial instructions even when those instructions appear unambiguous to the author [60].

**2.1.2 Generative Space Exploration.** PromptCharm provides a mixed-initiative loop: an RL-based agent suggests refined prompts, while users can tweak token-level attention or in-paint masked regions, exposing prompt, attention, and pixel masks [7]. In the domain of strictly textual tools, Sensecape allows users to decompose prompts into individual parts, arranged spatially, and compare variants in parallel, merging the parts they like [8]. Luminare asks writers to tag sentences along qualitative dimensions (e.g., formality, concreteness) and then recombines those dimensions to generate tailored drafts [9]. Wang et al.’s RePrompt approached this limitation through automated prompt refinement toward precise emotional expression, but spatial relationships remain outside the scope of current automated revision pipelines [53]. Our work is motivated directly by this gap: language is a poor medium for conveying

rotation axes, occlusion depth, and compound spatial transformations, and the resulting trial-and-error cycle imposes significant cognitive overhead on users. Promptify offers a lighter abstraction: each successive revision of a single prompt appears on a zoomable canvas, preserving visual history and encouraging lateral comparison, though genuine branching still requires manually copying the prompt [12]

**2.1.3 Spatially Targeted Editing for Images.** For spatially targeted editing, Hertz et al.’s Prompt-to-Prompt demonstrated that cross-attention maps within diffusion models can be manipulated to achieve layout-preserving edits by altering specific tokens in a prompt [16]. More recently, ControlNet [62] introduced auxiliary conditioning inputs—including depth maps, edge maps, and pose skeletons—to impose spatial structure on diffusion-based generation, while BrushNet [22] proposed a plug-and-play dual-branch inpainting architecture that provides content-aware per-pixel guidance within a masked region. While these methods expose structured conditioning channels, they remain difficult for end users to drive directly: the mapping from a desired 3D spatial transformation to a suitable conditioning representation is non-trivial and is not supported by any mainstream image generation interface. Recent systems have thus explored the conditioning channels that are more explicit in 3D. For instance, Canvas3D [32] proposes a 3D scene-editing canvas that allows users to spatially arrange elements in the scene for image generation. While such systems use a separate 3D editing interface for one-shot image generation, we aim to support users to specify spatial intent directly within the image during generation. Other systems such as ObjectMorpher [57] enable the deformation of single objects for one-shot edits; while there is some overlap in our technical pipeline, we explore the composition of multiple objects in iterative generative image editing workflows, including not only iteratively manipulating objects but also compositing their spatial relationships.

The closest concurrent related work to ours is Adobe Turnstyle [1] and Turntable [2], which enable users to lift *pre-made* vector and raster objects into 3D and manipulate them interactively. However, this project remains closed-source. Our work differs in several key respects. First, we propose a pipeline that incorporates image segmentation to enable the lifting of *any* objects within a generated image, as well as supporting editing effects such as occlusion that takes advantage of multi-object segmentation. Second, we contribute a fully open-sourced and working pipeline, allowing the broader research community to build upon it. Third, our user study provides empirical insight into the pipeline’s performance and how it can be integrated into image generation and editing workflows.

## 2.2 Direct Manipulation v.s. Natural Language Instruction

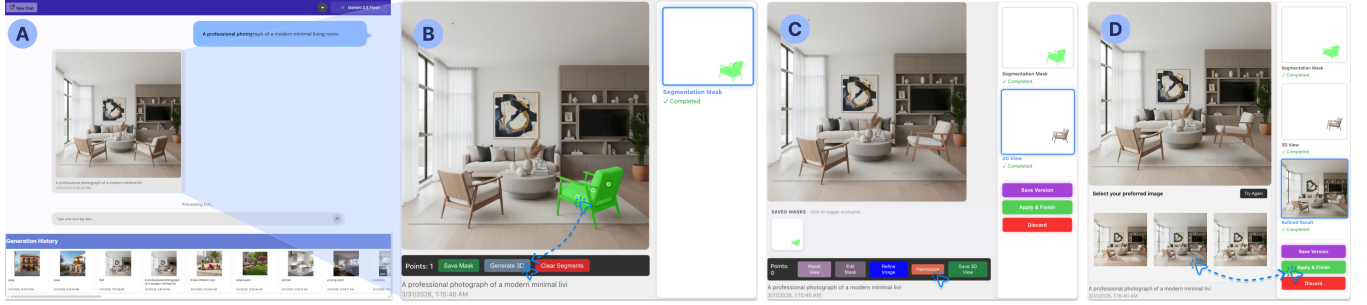
The theoretical foundation for our interaction design draws on Shneiderman’s principles of direct manipulation [44], which emphasize continuous representation of objects of interest, physical actions replacing complex syntax, and rapid incremental reversible operations. Standard 3D widgets (e.g., translate, rotate, scale handles) instantiate these principles in spatial editing contexts, making the intended transformation unambiguous in a way that verbal description cannot. Furthermore, research has explored *multimodal*

*cooperation* by blending natural language with direct manipulation to make interfaces more intuitive [42, 49]. Early systems such as Put-That-There [5] allowed users to issue vocal commands while pointing in mid-air. As language processing advanced, systems like PixelTone [26] further expanded command flexibility for photo editing, while other tools used graph mapping, vector representations, or interactive demonstrations (e.g., in PUMICE [29]) to resolve ambiguous natural language requests. Recent systems such as Stylette [24] and DirectGPT [33] blend speech and direct manipulation to customize websites and LLM prompts, respectively.

The mixed-initiative and multimodal design of our system is further informed by Amershi et al.’s 18 guidelines for human-AI interaction [3], which include making clear why the system did what it did, and enabling users to provide feedback during interaction. Brickify’s study adds a complementary finding: increased clarity in intent expression through direct manipulation helped users correctly attribute generation failures to model limitations rather than to inadequacies in their own instructions, reducing the unproductive back-and-forth that characterizes text-prompted iteration [43]. Liao et al.’s study of practitioner needs for explainable AI also found that users most urgently require support for explaining a prediction, i.e. understanding why a specific output looks as it does. [30]. Weisz et al.’s work on generative AI creative tools further established that version control and visualizing differences between iterations are important scaffolds when users may wish to return to earlier outputs [55]. Our system incorporates these principles by grounding spatial intent in an explicit 3D scene representation that makes each transformation reversible and inspectable, preserving scene history across the sequential edits that characterize real-world iterative composition tasks.

## 2.3 2D-to-3D Generative AI

Advances in neural representations, particularly NeRF [23], has significantly improved the fidelity of 3D reconstructions built from multi-view images. Simultaneously, generative AI has emerged as a central paradigm for 3D content generation. Early text-to-3D and image-to-3D pipelines, such as DreamFusion [36] relied on 2D diffusion via Score Distillation Sampling (SDS); however, they were frequently constrained by slow generation speeds and limited output diversity. Recent approaches such as Large Gaussian Models (LGMs) [47] bypass SDS altogether by converting 3D Gaussians into structured meshes. By leveraging multi-view diffusion priors and replacing basic text prompts with spatial and depth controls, these systems produce significantly more precise and detailed textured assets. Most recently, foundational models for 3D such as SAM 3D [12] now enable high-fidelity 3D asset generation directly from a single input image or text prompt. Beyond generation, prior work has also explored manipulating and editing these 3D assets. For instance, Cen et al. enable segmentation and editing of 3D NeRF representations [8]. HCI research has explored the benefits and challenges of 3D generative AI in scenarios such as group discussion [19, 34, 38], education [4, 14, 63], and AR/VR [20, 27, 50, 52, 58, 61]. We extend this line of research by repurposing 3D GenAI models for generative image editing workflows. Specifically, we view AI-generated 3D objects as interaction proxies that are directly manipulable for precise spatial intent in the context of 2D image editing.



**Figure 2: User Interface of ImageDirector.** (A) The overall generation interface with an initial text prompt, where the bottom shows a gallery of the generation history; (B) Users specify points of interests on the image, and the system generates a segmentation mask for the target object, highlighted in green; (C) Once the system lifts the segmentation into a spatial proxy, users directly manipulate the object in 3D space; (D) Users select from the harmonized options, and save the modified image.

### 3 ImageDirector System

In this section, we detail the design decisions and rationale behind ImageDirector, a generative image editing tool that enables users to express spatial editing intent via 3D direct manipulation. We built ImageDirector with the following design goals:

- D1: Enable direct manipulation for expressing spatial editing intent.** Prior work has shown that natural language serves as a poor medium for conveying precise spatial intent [25]. ImageDirector addresses this limitation by enabling users to manipulate 3D proxies of objects within generated images through direct dragging, rotating, and scaling, reducing reliance on linguistic approximation.
- D2: Support accurate and efficient iterative edits.** Effective editing workflows are iterative, involving more than manipulating a single object in isolation [9, 10]. ImageDirector supports iterative editing across multiple objects, including free exploration of spatial placement, management of multi-object selection, and composition of spatial relationships such as occlusion between objects within a scene.
- D3: Integrate spatial editing into generative image workflows.** Rather than positioning ImageDirector as a standalone editing tool, our goal is to embed spatial manipulation within broader image creation workflows. This requires producing output quality comparable to fully generated images, as well as supporting fluid transitions between natural language prompting and direct 3D manipulation as complementary input modalities.

#### 3.1 User Interface

ImageDirector is designed as a continuous control loop that supports iterative spatial editing across three distinct operational phases: target specification and scene decoupling, apatial proxy manipulation, and generative harmonization and refinement.

**3.1.1 Target Specification and Scene Decoupling.** Each iteration begins with a base 2D generated image. The user identifies a target object by clicking over the region of interest, upon which ImageDirector utilizes an existing segmentation model to generate a corresponding instance mask. In parallel, the system initiates a scene decoupling process by utilizing an existing image inpainting model to reconstruct the background to preserve visual continuity.

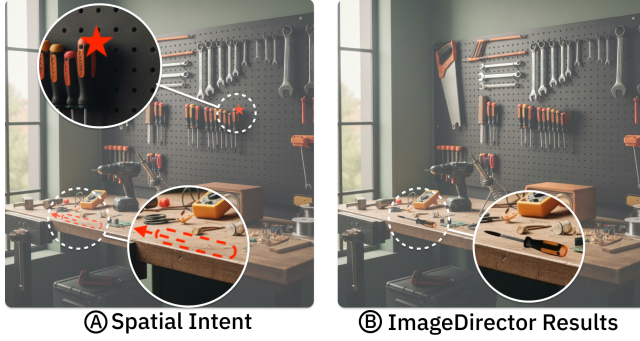
**3.1.2 Spatial Proxy Manipulation.** Once the user chooses to generate a 3D representation of the object, our system employs an existing 2D-to-3D generative model that lifts the isolated 2D segment into a manipulable 3D proxy. Once generated, the 3D proxy is overlaid onto the original inpainted image plane at its original position and orientation. The user interacts with the proxy through direct manipulation that offers direct control over translation (X, Y, Z), rotation (yaw, pitch), and uniform scaling. Unlike traditional 2D warping, this phase provides real-time visual feedback on the proxy’s depth and orientation, enabling the user to direct the scene’s composition with geometric precision.

**3.1.3 Generative Harmonization and Refinement.** Once the user is satisfied with the spatial arrangement, the system captures the applied transformation and passes it to the harmonization pipeline, which re-composites the manipulated proxy seamlessly into the inpainted background to produce a new 2D output image. The user is presented with three rendered variations for selection. They may return to any of the previous phase (e.g., segmentation or 3D manipulation) via the thumbnails list on the right (see Fig. 2). They may also save the current rendered image in the gallery of generation history. Once they apply the changes and save the image, they can re-enter the control loop for further iterations.

#### 3.2 Disambiguating Spatial Intent via Direct Manipulation

A central design decision in ImageDirector is replacing ambiguous linguistic approximation with direct manipulation (D1). As mentioned previously, We have established two sources of ambiguity that make text-based spatial editing unreliable: *spatial reference ambiguity*, where users lack a shared coordinate system to specify positions and orientations (e.g., “move it to the left” is ambiguous without a reference point), and *object reference ambiguity*, where scenes contain multiple similar elements that are difficult to disambiguate linguistically (e.g., “rotate the third screwdriver from the left”). ImageDirector addresses both forms of ambiguity (see Fig. 3). Rather than asking users to describe a desired spatial intent in words, the system externalizes objects’ spatial attributes as a manipulable 3D proxies that users can manipulate directly. Spatial reference ambiguity is eliminated because the transformation is the





**Figure 3: Disambiguating spatial intent via direct manipulation:** (A) The user wants to create a scenario where the screwdriver (starred in red) should be placed at the annotated circle with the tip aligned along the arrow; (B) Using ImageDirector, users can disambiguate the object and spatial reference via direct segmentation and manipulation.

input: users specify positions by placing objects there, not by describing where “there” is. Object reference ambiguity is eliminated because users select targets by pointing and clicking.

Our manipulation interface supports three types of direct spatial control, each mapped to a distinct input: right-click drag for translation across the image plane, mouse scroll wheel for depth scaling that simulates perspective-consistent movement along the Z-axis, and left-click drag for rotation about the proxy’s axes. Notably, because the proxy is a full 3D representation, rotation can reveal previously occluded surfaces rather than warping the original 2D pixels. The user sees what the object looks like from the new viewpoint, not a distorted version of the old one.

Together, these controls support rapid exploration of spatial placement (D2), where users can try multiple configurations in quick succession, compare results across the generation history gallery, and converge on a composition through direct spatial reasoning rather than iterative prompt revision. The “directness” and real-time feedback afforded by the 3D proxy makes each manipulation immediately legible, thus reducing the unpredictability that characterizes prompt-based editing.

### 3.3 Managing Multi-Object Occlusion

Realistic scene composition requires more than accurate object placement; it requires compositing *depth relationships* between objects, such as determining which elements occlude which. Such spatial dynamics pose unique challenges that automated monocular depth estimation alone cannot reliably resolve, particularly in scenes with multiple similar or overlapping objects (D2). ImageDirector addresses this through multi-mask management that puts explicit occlusion control in the hands of the user (see Fig. 4).

**3.3.1 Multi-Object Segmentation.** During the target Specification phase (Section 3.1.1), users are encouraged to segment all objects of interest before beginning manipulation. Each selection is stored as a named instance mask in a persistent *Saved Masks* panel. While only one mask is lifted into an active 3D proxy at a time, the remaining masks are retained as potential occlusion layers that can be activated at any point during editing.



**Figure 4: Iterative editing involving multiple objects:** (A) The initial generated image; (B) The user creates two segmentation masks to handle occlusion between the knife and the avocado; (C) The user toggles the avocado mask to the front, hiding the 3D knife proxy behind the avocado; (D) The user iteratively edits the image by modifying the blue mug.

**3.3.2 Interactive Occlusion Toggling.** When the user determines that the active 3D proxy should appear *behind* another object in the scene, they can toggle the corresponding saved mask to the foreground. The system renders the toggled mask as a static 2D layer with higher depth priority than the 3D proxy, effectively “tucking” the proxy behind it. This allows users to resolve complex depth conflicts — for instance, positioning a knife between an avocado and a cutting board — through a direct, reversible action rather than relying on the generative model to infer the correct depth hierarchy. This approach aligns with D1 and D2: it makes depth relationships an explicit, user-controlled parameter rather than an implicit model decision, and it scales naturally to scenes with multiple interacting objects, supporting the kind of precise multi-object composition that text prompting cannot reliably achieve.

### 3.4 Multimodal Cooperation: Blending Natural Language and Direct Manipulation

Multimodal cooperation in creative tools has a rich history by combining natural language instructions with direct manipulation to reduce interaction friction [5, 42, 49]. Building on this paradigm, ImageDirector enables fluid transitions between natural language prompting and direct 3D manipulation into generative image workflows (D3), as seen in Fig. 2A. Instead of relying on a single input channel, our system employs a complementary division of labor, where natural language input establishes the semantic context and high-level creative intent (e.g., scene generation, material changes, or style adjustments), and direct manipulation provides precise,

unambiguous selection and fine-grained spatial action (e.g., precise translation, rotation, depth scaling, and multi-object occlusion control) within that established semantic context.

This cooperative design is also tailored specifically for iterative editing workflows rather than single-shot generation. For example, a user might first use a text prompt to generate an image of a modern kitchen by describing the overall context (see Fig. 4A). They can then select specific items via direct manipulation to precisely adjust their position, orientation, and occlusion (see Fig. 4B–C). Finally, users can proceed to the next iteration by repositioning the mugs to opposite sides of the counter to convey a warm, conversational atmosphere (see Fig. 4D), or use text prompts to modify the overall image style or aesthetics.

### 3.5 Architecture and Implementation

ImageDirector is implemented as a web application that orchestrates a suite of inference models hosted on a cloud server with an A100 GPU (80GB). As seen in Fig. 5, the core pipeline comprises three modules: (1) *Object Segmentation*, which uses SAM3 [7] to isolate the target element via instance segmentation, followed by the LaMa inpainting model [46] to reconstruct the background in disoccluded regions; (2) *2D-to-3D Lifting*, which utilizes SAM3D [12] to convert the 2D segment into a high-fidelity textured 3D proxy; and (3) *Harmonization*, which uses a combination of IP-Adapter [59] and Nano Banana Pro to diverse results that harmonize the manipulated proxy back into the scene.

**3.5.1 Intent Classification.** Our system supports two intent categories: spatial *EDIT* intents (e.g., repositioning or rotating an object) and non-spatial *GENERATE* intents (e.g., changing an object’s material or synthesizing new elements). By default, non-spatial intents are initiated via text prompt, and spatial intents are initiated via direct point and click on the image. To demonstrate our system’s compatibility with existing text-prompting workflows, we implemented a routing rule that converts spatial text commands (e.g., “tilt it up” or “make it a bit larger”) into discrete 3D proxy transformations (e.g., 15 degree rotation increments, 1.5× scale multipliers) and applied them directly to the proxy, rather than sending them through the generation model. This feature was described to demonstrate compatibility and thus was excluded from the user study.

**3.5.2 Object Segmentation and 2D-to-3D Lifting.** Our technical pipeline transforms a user’s selection into a manipulable 3D proxy through three sequential stages:

- **Segmentation.** We employ SAM3 [7] for high-precision object extraction. The system supports both point-click and text-based prompts to specify target objects and generate instance masks
- **Background Restoration.** To ensure scene integrity after an object is lifted, we apply the LaMa inpainting model [46] to reconstruct the vacated region. A  $15 \times 15$  pixel dilation kernel is applied to the object’s mask prior to inpainting to eliminate edge artifacts, producing a seamless “clean plate” for subsequent repositioning.
- **3D Proxy Generation.** The isolated 2D segment is processed by SAM3D [12] to produce a textured 3D proxy, stored

as a .ply mesh. This proxy is rendered in the frontend using Three.js, enabling the real-time cursor-based manipulation.

**3.5.3 Re-rendering and Harmonization.** Once the user finalizes a spatial transformation, the system captures the proxy’s updated pose—position, orientation, and scale—for the synthesis stage. We employ two models to harmonize the re-rendered object with the original inpainted background, producing a coherent composite image. First, we use IP-Adapter [59] that takes a style reference and a structural reference to generate the harmonized (style-transferred) image as an output. Second, we send the 3D overlaid image directly to Nano Banana Pro and prompt it to harmonize the shadow, lighting, etc. Our interface then shows a blend of results from both options for the user choose from.

## 4 Evaluation

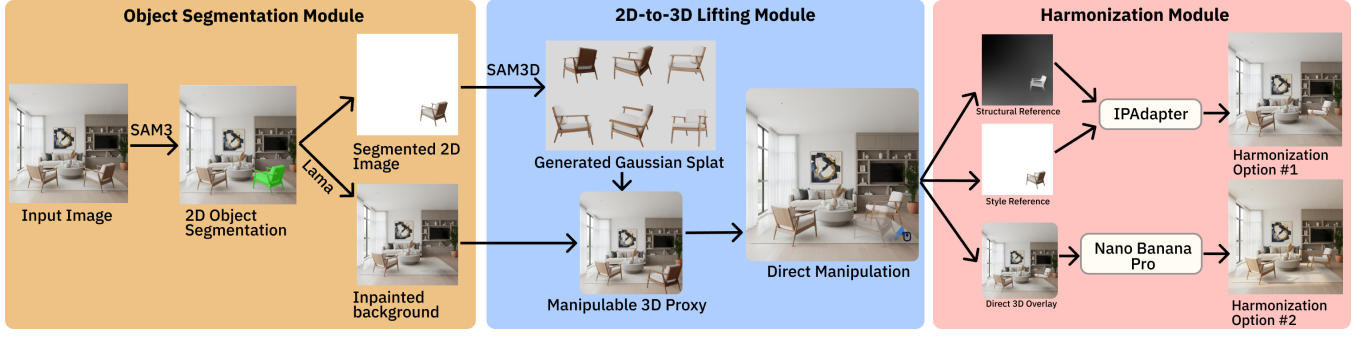
To understand the benefits and challenges of ImageDirector, we conducted a user study comparing our 3D-grounded direct manipulation system against a text-only prompting baseline. Specifically, we have the following research questions:

- **RQ1: Spatial Accuracy and Efficiency.** To what extent does ImageDirector improve user speed and accuracy in achieving intended spatial transformations compared to the baseline?
- **RQ2: Visual Consistency.** To what extent does ImageDirector preserve visual consistency in unedited regions across iterative edits compared to the baseline?
- **RQ3: Perceived Control, Workload, and Expressiveness.** How do users perceive using ImageDirector compared to the baseline in terms of cognitive workload, control over spatial outcomes, and expressiveness of spatial intent?
- **RQ4: User Perception and Adoption.** How do users perceive the overall output quality of ImageDirector compared to the baseline, what interaction strategies emerge, and how might the tool integrate into broader image creation workflows?

### 4.1 Participants

We recruited 12 participants (10 male, 2 female) through a combination of physical flyers on a university campus and word of mouth. Participants ranged in age from 19 to 33 years ( $M = 22.5$ ,  $SD = 3.62$ ). All participants had at least beginner level of image editing experiences such as cropping and applying filters to images. Most participants are novices in AI image generation: three participants reported having no experience with AI image generation tools (e.g., DALL-E and Midjourney) at all and seven participants had tried them only once or twice. We also excluded individuals with professional 3D software experience (e.g., Blender, Unity) to avoid ceiling effects in the ImageDirector condition.

Our study used a within-subjects design with two conditions. In the **Baseline** condition, participants performed spatial edits using text-only prompting through Google Nano Banana Pro [15], a state-of-the-art commercial image generation system developed by Google DeepMind. We selected Nano Banana Pro as the baseline because it represents the current frontier of conversational image editing, with demonstrated strength in multi-turn editing [11] and



**Figure 5: System Pipeline of ImageDirector.** The pipeline shows the transition from the initial input image and interactive object segmentation (left), to 3D proxy generation and direct manipulation (middle), and to harmonization (right).

reference-based image reasoning [54]—making it a strong comparison point for evaluating whether 3D direct manipulation offers benefits beyond what the best available text-only workflow can achieve. In the **ImageDirector** condition, participants used our proposed 3D-grounded pipeline, which supports directly manipulating translation, rotation, and scaling of object proxies. Although ImageDirector’s full design supports both text-based and direct manipulation editing, we disabled the natural language editing channel in this condition to isolate the effect of 3D direct manipulation as an input modality, enabling a controlled comparison between the two interaction paradigms. Condition order and task set assignment were counterbalanced using an AB/BA scheme across two groups of six participants each, to control for learning and fatigue effects.

## 4.2 Procedure

Each session lasted approximately 75 minutes and was conducted in person. Upon arrival, participants provided informed consent, completed the demographic and screening instrument, and received a brief orientation to generative image editing. Participants were then trained on the two experimental conditions in counterbalanced order, including the text-only prompting baseline (i.e., Nano Banana Pro), which supports iterative editing through natural language instructions alone, and ImageDirector, which supports direct 3D object manipulation. Training for each condition consisted of a researcher-led walkthrough followed by two practice tasks not included in the analysis. Then, participants completed all experimental tasks in both conditions, with condition order counterbalanced using a Latin square design to control for order and learning effects. After each condition, participants completed the NASA Task Load Index (NASA-TLX) and a perceived control and expressiveness questionnaire. After completing both conditions, participants responded to a forced-choice overall preference question and a per-task preference survey, then participated in a 15-minute semi-structured interview. Sessions were screen-recorded and audio-recorded with participant consent.

## 4.3 Tasks

During the study, participants completed a series of spatial image editing tasks framed around an art director scenario: they were asked to imagine they were preparing a product scene or instructional photograph and needed to arrange objects to match a given

target composition. Each task presented a generated scene image alongside an annotation specifying the spatial goal. Annotations took one of two forms, reflecting the natural structure of spatial instructions: (1) a red circle or oval indicating the target position for an object, (2) a red circle with a directional arrow indicating both target position and intended orientation. Tasks were designed to span a range of spatial manipulation types including object translation, rotation, and scaling while remaining interpretable without domain expertise. Stimuli were drawn from everyday object and scene categories (e.g., tools, produce, household items) to ensure all participants could engage with tasks without specialized knowledge. Task details can be seen in Appendix A. Each participant completed three iterative steps per condition. Condition and task order was counterbalanced to avoid learning effects.

## 4.4 Measure

We evaluated our system using a combination of objective performance metrics, behavioral logs, surveys, qualitative interviews, and expert evaluation.

**4.4.1 Spatial Accuracy.** We measured spatial accuracy using position error, computed as the Euclidean distance in pixels between the centroid of the edited object in the participant’s output and the center of the target annotation (circle or oval). Object centroids were extracted from segmentation masks generated by SAM3 [7], computed as the mean  $(x, y)$  coordinate of all mask pixels. The image size was 1024 x 1024 px across all tasks.

**4.4.2 Visual Drift.** Visual drift was measured as the change in background regions between the pre-edit and post-edit image at each iteration, using the Structural Similarity Index (SSIM) [17]. To isolate background stability from object placement, we computed a union mask combining the object’s original position (A) and target position (B) for each task, derived from segmentation masks of both the original and edited objects. This union mask was applied identically to both the pre-edit and post-edit images before computing SSIM, ensuring that the object regions were excluded from the comparison in both conditions equally. Notably, this masking strategy is conservative: it excludes object replication artifacts — cases in the text baseline where the system created a copy of the object at the target position rather than moving it, leaving the original in place — from the SSIM computation. By excluding these artifacts

from SSIM, we ensure that any SSIM advantage for ImageDirector reflects pure background stability rather than penalizing the baseline for its most severe failure mode.

**4.4.3 Surveys and Interviews.** After completing each condition, participants completed the NASA Task Load Index (NASA-TLX) to assess workload across six dimensions. We also measured perceived control and expressiveness using 7-point Likert scales (1–7) adapted from [48]. Following both conditions, participants responded to a forced-choice overall preference question ("Which system did you prefer overall?") and a per-operation preference instrument in which they indicated, for each task type, which condition better supported their intent. Each session concluded with a 15-minute semi-structured interview exploring how participants approached spatial intent in each condition, what strategies they developed, and where they experienced breakdowns. Interviews were audio-recorded and transcribed verbatim.

**4.4.4 Expert Evaluation.** To complement automated metrics with perceptual judgment, we recruited two expert raters with professional image editing experience to evaluate participant outputs. One expert was a ui/ux designer with 3-5 years of professional experience in image editing and graphics design, and the other was a senior graphics designer with more than 6 years of experience. Raters were blind to condition throughout the evaluation. For each trial, raters were shown the task instruction and annotation, the pre-edit image, and the post-edit image, and rated four metrics using a structured rubric: (1) Spatial Intent Achievement – whether the edit fulfilled the task instruction; (2) Background Preservation – whether unedited regions remained stable; (3) Visual Cohesion – whether the edited object appeared to belong in the scene, assessed across lighting, shadow, scale, and edge blending; and (4) Overall Edit Quality – a holistic professional judgment of the output. Prior to rating, both raters participated in a joint calibration session in which they scored five practice trials together and discussed their reasoning until reaching agreement. Raters then independently scored all trials. Inter-rater reliability was assessed using Krippendorff's alpha computed separately for each metric.

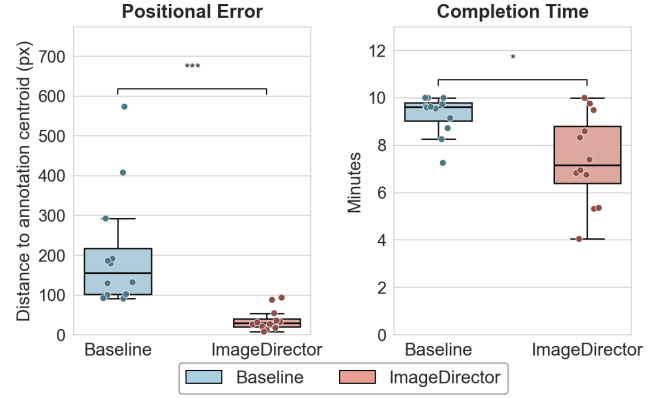
## 4.5 Analysis

For each quantitative dependent variable, we applied Wilcoxon signed-rank tests to compare C1 and C2 directly. Overall preference and per-operation preference proportions were tested using chi-square. Interview transcripts were analyzed using reflexive thematic analysis [13]. One researcher developed a codebook drawing deductively from the research questions and inductively from emergent themes in the data. The full dataset was then coded and similar codes were grouped into themes.

## 5 Results

### 5.1 Accuracy and Efficiency of Spatial Edits (RQ1)

RQ1 examined how accurately and efficiently participants achieved intended spatial transformations. We report position error, expert-rated Spatial Intent Achievement, and task completion time.



**Figure 6: Baseline vs. ImageDirector across positional position error (↓) and task completion time (↓). \*  $p < .05$ , \*\*  $p < .01$ , \*\*\*  $p < .001$ .  $N = 12$ .**

**5.1.1 Position Error.** Participants using ImageDirector achieved significantly lower position error than the text-only baseline (ImageDirector:  $M = 37.4$  px,  $SD = 27.8$ ; Baseline:  $M = 206.6$  px,  $SD = 149.0$ ;  $p < 0.001$ ), representing an **81.9%** reduction in centroid-to-target distance (see Fig. 6). The text-only baseline also exhibited substantially higher variance, reflecting the unpredictability of spatial placement via natural language.

**5.1.2 Expert-Rated Spatial Intent Achievement.** Two expert raters assessed Spatial Intent Achievement on a 3-point scale (0 = not achieved, 1 = partially achieved, 2 = fully achieved; Krippendorff's  $\alpha = 0.69$ ). ImageDirector outputs received significantly higher scores (ImageDirector:  $M = 1.57$ ,  $SD = 0.37$ ; Baseline:  $M = 0.80$ ,  $SD = 0.48$ ;  $p < 0.01$ ), as seen in Fig. 7.

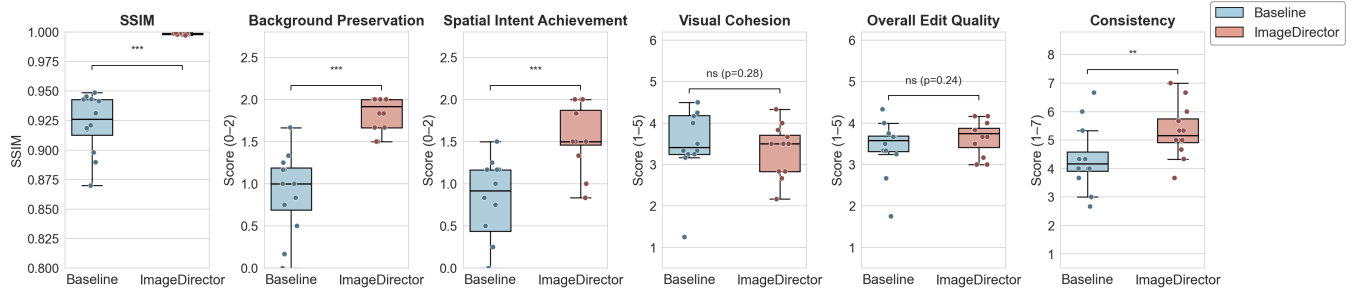
**5.1.3 Task Completion Time.** We also examined whether the two conditions differed in the time participants needed to complete spatial editing tasks. Total session time—the cumulative duration across all tasks within each condition—was significantly lower for ImageDirector (ImageDirector:  $M = 7.40$  min,  $SD = 109.1$  s; Baseline:  $M = 9.29$  min,  $SD = 47.9$  s;  $p = 0.02$ ), representing a **20.4%** reduction in total editing time (see Fig. 6). We found that completion time for ImageDirector has a larger SD partly because the 3D manipulation learning curve experienced by some participants, as indicated in Section 5.3.1.

### 5.2 Visual Consistency Across Iterative Edits (RQ2)

RQ2 examined how the two systems differ in preserving background regions across iterative edits. We report SSIM-based background drift and expert-rated Background Preservation (see Fig. 7).

**5.2.1 Background Drift (SSIM).** Background SSIM, computed over regions outside a union mask covering both original and target object positions, was significantly higher for ImageDirector (ImageDirector:  $M = 0.9982$ ,  $SD = 0.0007$ ; Baseline:  $M = 0.9222$ ,  $SD = 0.0251$ ;  $p < 0.001$ ), indicating substantially less background drift.





**Figure 7: Baseline vs. ImageDirector across image editing accuracy (SSIM), expert-rated metrics in image quality. ImageDirector achieves significantly stronger background preservation and greater rated consistency. \*  $p < .05$ , \*\*  $p < .01$ , \*\*\*  $p < .001$ .  $N = 12$ .**

**5.2.2 Expert-Rated Background Preservation.** Expert raters assessed Background Preservation on a 3-point scale (0 = not preserved, 1 = partially preserved, 2 = fully preserved; Krippendorff’s  $\alpha = 0.83$ ). ImageDirector received significantly higher scores (ImageDirector:  $M = 1.83$ ,  $SD = 0.19$ ; Baseline:  $M = 0.90$ ,  $SD = 0.47$ ;  $p < 0.01$ ), corroborating the automated SSIM findings.

**5.2.3 Expert-Rated Overall Quality.** Visual Cohesion (5-point scale assessing lighting, shadow, scale, and edge blending;  $\alpha = 0.74$ ) and Overall Edit Quality (5-point;  $\alpha = 0.60$ ) did not differ significantly between conditions. Experts noted that while ImageDirector more accurately fulfilled spatial intent, the text baseline occasionally produced more naturalistic lighting and shadow—an artifact of full scene regeneration that, despite introducing drift, can incidentally improve illumination consistency. This suggests that the two approaches present a trade-off between spatial precision and photorealistic integration, a tension we return to in Section 6.

### 5.3 Users’ Perceptions of Control, Workload, and Expressiveness (RQ3)

RQ3 examined users’ subjective experience including NASA-TLX workload, perceived control, expressiveness, and overall preference. Complete results can be seen in Appendix B.

**5.3.1 Perceived Workload and Performance (NASA-TLX).** Overall cognitive workload did not differ significantly between conditions ( $p = 0.84$ ). We attribute this to the learning curve associated with ImageDirector’s 3D manipulation pipeline, which introduced unfamiliar interaction demands that offset the reduced iteration burden, a pattern consistent with participants’ interview feedback (Section 5.4). However, the NASA-TLX *Performance* subscale—which captures how successful participants felt in accomplishing the task—was significantly more favorable for ImageDirector (ImageDirector:  $M = 2.75$ ,  $SD = 1.54$ ; Baseline:  $M = 4.50$ ,  $SD = 1.31$ ;  $p = 0.02$ ). Note that lower NASA-TLX scores indicate better perceived performance.

**5.3.2 Perceived Control.** Participants reported significantly greater perceived control over spatial editing outcomes when using ImageDirector (ImageDirector:  $M = 5.89$ ,  $SD = 0.87$ ; Baseline:  $M = 4.39$ ,  $SD = 1.47$ ;  $p = 0.02$ ).

**5.3.3 Perceived Expressiveness.** Participants rated ImageDirector as significantly more expressive of their spatial intent (ImageDirector:  $M = 5.56$ ,  $SD = 1.00$ ; Baseline:  $M = 4.33$ ,  $SD = 1.31$ ;  $p = 0.02$ ).

**5.3.4 Overall Preference.** When asked which system they preferred overall, 8 of 12 participants chose ImageDirector, 3 chose the text baseline, and 1 expressed no preference ( $\chi^2(2) = 6.50$ ,  $p = 0.04$ ).

### 5.4 User Perception and Adoption (RQ4)

**5.4.1 Perceived Control and Agency.** The most prominent finding across all participants was the heightened sense of control afforded by the 3D direct manipulation system compared to text-only prompting. Participant identified ImageDirector as the one that made them feel more in control. P04 stated: “I can move an object around anywhere in 3D manipulation because I know exactly where I want it to be.” P06 similarly noted the system gave “a lot more precision and control over where exactly I needed to be.”

On the contrary, participants explicitly valued not needing to translate spatial intent into detailed and tiring prompts. P02 preferred the direct manipulation over crafting detailed prompts: “I don’t have to give a specific direction and when I’m giving the direction I have to be concise with top left or top right corner... but here [with ImageDirector] I just take that job.” P09 echoed that “[With Baseline] you have to write a lot to be able to tell it what to pick up, read, how to rotate and how to resize.”

The frustration of crafting detailed prompt was heightened due to AI hallucination. P07 explained the cognitive mismatch: “the AI generator would do not much as you want it, because it’s not [what’s] in your head. It’s just going by what the words are used in the prompt.” Similarly, P13 noted that “[In Baseline] you are mentally playing with the LLM. You are trying to make sure your prompts are being interpreted the way you want to be, but it’s not always the case.”

**5.4.2 The Communication Gap in Text Prompting.** Participants described situations where natural language proved insufficient for spatial editing due to **spatial reference ambiguity** and **object reference ambiguity**. For spatial reference ambiguity, P03 mentioned that “Rotation was a lot easier because in 2D it’s hard to really quantify angles and I think that even if you said like any kind of angle you’d have to say in what it’s in relation to.” P04 provided an extended example: “I wanted to move the banana over here. It has little to no references... It could be here. It could be here. It could be anywhere.”

Object disambiguation was another distinct challenge. For example, P03 noted that “[In Baseline] I described a little bit too much when there were multiple objects that all looked the same... there

were three mugs that all looked the same. If I was able to select that specific mug and then rotate that, I think that would have been a lot easier.” P11 described the workshop scene as having “so many screwdrivers, so not many ways to reference the position.”

**5.4.3 Reduced Iteration Overhead.** Participants described text-based prompting as fundamentally unpredictable, requiring costly iteration cycles. P13 characterized it as “more of kind of luck that you are playing with Gemini... you kind of give prompts again and again.” P06 recounted that Gemini “just thinking oh he probably means this and then we’re moving half of the stuff off the desk.” P07 similarly found that “no matter how many times I did it with different AI systems, it still makes mistakes.” P04 contrasted the two systems: “In AI, I said, hey, let’s try to do it like this. The AI might or might not do what I wanted to do... But in the 3D manipulation, I know what I wanted to do because I tried it.”

**5.4.4 Visual Fidelity and Realism Concerns.** Despite the strong preference for the control offered by 3D manipulation, participants identified visual quality as the system’s primary weakness. P07 stated: “It’s not 100% looks real. It looks like it’s over the image, which it looks like a montage.” P13 added that: “When we create the 3D model and get back to the picture, it is like not as natural as it could be.” P06 noted lighting discrepancies, saying, “the lighting changes a little bit. And then you can sort of tell that it’s been added in post.” P09 found the issue of realism particularly for small objects: “Sometimes it’s not very accurate for particularly for the small objects in the image when it creates the 3D model out of it.” Notably, all participants did not use the harmonization feature for at least one sub task, partly due to the need for task completion, which primarily focused on spatial manipulation, and efficiency in the timed study.

**5.4.5 Complementary Modalities.** 8 of 12 participants believed that spatial manipulation tasks could not be replicated through text prompting alone. P03 mentioned: “I don’t think that I could have moved this wrench without it moving any of the other tools on the board.” Conversely, 7 participants believed text-based tasks were achievable via 3D manipulation, despite with lower visual quality. Participants also envisioned combining both modalities in their workflows. For example, P06 articulated the clearest cooperation: “If I wanted to angle the sofa... I would use 3D manipulation to just rotate it... But if I wanted to say make it like a reclining sofa or a plush sofa instead of a leather one I’d probably ask it to do it via text.” P08 proposed a sequential workflow: “Move the mug on exact position that I want... and then use the natural language processing to maybe make it look more ideal.”

## 6 Discussion

Overall, our results demonstrate that ImageDirector significantly outperformed the text-prompting baseline in achieving precise spatial edits, preserving background stability, and offering users control and expressiveness in the image editing process. Eight of twelve participants judged that such spatial manipulation tasks could not be replicated through text prompting alone, while seven independently converged on hybrid workflows combining both modalities. Below, we discuss our implications for image generation workflows, broader applications areas, and the design of multimodal human-AI interaction.

## 6.1 Improving Text-to-Image Editing via Direct Manipulation

Our findings empirically reveal a *spatial intent bottleneck*: the systematic failure of natural language to encode the spatial mental models users hold when editing images. This corresponds to Hutchins et al.’s analysis of the gulf of execution [21]. In text-based editing, users face a double gulf: a *semantic gulf* in translating spatial intentions into verbal descriptions, and an *articulatory gulf* because natural language lacks primitives for precise geometric specification. This aligns with findings from prior work such as [25, 60] that spatial intent can be inherently ambiguous and difficult to express linguistically. Our results demonstrate that, by applying Shneiderman’s classic principles of direct manipulation [44] to generative image editing contexts, we could bridge these gulfs, where the spatial transformation itself becomes the input.

The system also shifts editing from holistic scene description to object-level decomposition, aligning with how participants naturally conceptualized their edits: not as global transformations, but as operations on specific objects. This decomposition additionally resolves the object disambiguation problem identified in our interviews—pointing is unambiguous in a way that language fundamentally is not [5]. However, nine participants identified visual fidelity as the system’s primary limitation. P04’s observation that the 3D proxy is “really fantastic as a mind map” even when visually imperfect suggests a design opportunity for *progressive fidelity*—rapid, low-fidelity spatial exploration followed by higher-fidelity rendering once the composition is finalized.

## 6.2 Spatial Intent Beyond Image Editing

Our core design insight for future human-AI interaction is that systems should manipulable intermediate proxies on latent structure when language is insufficient. This generalizes beyond image editing; generative interfaces in general can externalize difficult-to-prompt intent dimensions through controllable representations, including but not limited to 3D object proxies, spatial layouts, paths and trajectories. For instance, generative video systems inherit the same spatial intent bottleneck, compounded by the temporal dimension. A 3D proxy paradigm could allow users to keyframe object positions through direct manipulation, creating a “spatial storyboard” for video generation [18]. Similarly, camera path specification could benefit from manipulating a virtual camera in a 3D proxy scene rather than prompting “slowly pan left while zooming in.” Participants spontaneously identified further applications including interior design and photography post-production.

## 6.3 Multimodal Prompting & Workflows

Our findings show that 7 of 12 participants preferred a hybrid workflow model that combines text prompting for semantic changes with 3D manipulation for spatial precision. This behavior highlights the strength of framing human-AI creative interaction through multimodal cooperation rather than relying on a single input channel. In this cooperative paradigm, the two input modalities serve distinct yet complementary roles: when text prompting sets the high-level semantic context by setting up global styles, themes, objects, direct manipulation operates within that semantic frame by providing fine-grained, deterministic control over 3D transforms where verbal

descriptions naturally fail. New interaction challenges could emerge for future systems to realize the full potential of this multimodal paradigm, such as minimizing modality switching costs, resolving conflicts when two inputs produce contradictory instructions, and implementing progressive disclosure to support varying levels of user expertise. Beyond structured tasks in our study, multimodal workflows also have distinct opportunities and constraints for professional, open-ended creative tasks. In professional settings, such as commercial layout design, visual storyboarding, or art creation, where creators author with a spatial mental model, direct manipulation could reduce the time spent battling model unpredictability and hallucination. The ability to deterministically anchor, rotate, and layer assets enables artists to rapidly explore compositions while maintaining exact visual control [56]. However, this multimodal workflow assumes the user has a concrete editing intent. In highly open-ended or early-stage ideation scenarios, where users lack specific spatial targets or explicit manipulation goals, the requirement to manually manipulate 3D proxies may introduce frictions compared to the rapid, serendipitous discovery afforded by text prompts. Thus, future work could explore how direct manipulation can be best deployed in open-ended creative systems that balance precision spatial control and generative exploration.

## 6.4 Limitations

Our results demonstrate that ImageDirector enabled users to express spatial intent via direct manipulation in generative image editing. However, our work is not without limitations. Regarding our system, first, our system presents a trade-off between interactivity and visual fidelity: the harmonization step that produces photorealistic composites introduces latency, and all participants skipped it for at least one sub-task, preferring the immediacy of the raw 3D proxy despite its artificial appearance. We made the harmonization step optional to grant users agency to prioritize speed or visual refinement, and it is not required for core spatial editing tasks in our study. Future work should thus reduce harmonization latency to near-real-time, allowing users to maintain both the speed of direct manipulation and the visual quality of the final output. Second, our system does not support perspective adjustment, as found in other tools (e.g., Adobe Photoshop, GIMP), which may reduce the object compositing realism and spatial consistency when harmonizing the object back into the image. Third, heavily occluded or tiny 2D objects can also be difficult to reconstruct, and such capabilities are attributed to the underlying 2D-to-3D model. Future work could thus explore fine-grained control over generative image elements across different scales and layers [28]. Regarding our study, our sample ( $N = 12$ ) was drawn from a university population with limited generative AI experience, which may not reflect the strategies of professional designers or 3D artists. Second, our evaluation tasks used predefined spatial goals with clear success criteria; while this enabled objective measurement, it may not capture the open-ended exploration of real-world creative workflows.

## 7 Conclusion

We presented ImageDirector, a generative image editing system that lifts selected objects from 2D generated images into manipulable 3D proxies and re-composites results with background preservation,

addressing both the spatial imprecision and visual drift inherent in text-only editing. Our within-subjects study ( $N = 12$ ) showed that ImageDirector reduced position error by 81.9%, maintained near-perfect background stability (SSIM = 0.998), and completed sessions 20.4% faster than a state-of-the-art text-only baseline. Participants reported greater perceived control and expressiveness and preferred ImageDirector overall, while also envisioning multimodal hybrid workflows. Our work demonstrates that language excels at specifying *what* to create, while spatial manipulation excels at specifying *where* and *how* to arrange it. We release ImageDirector as an open-source pipeline to support future research at this intersection.

## References

- [1] Adobe Inc. 2025. Project Turn Style: Edit 2D Objects as if They Were 3D. Adobe MAX 2025 Sneaks, Los Angeles, CA. Available at <https://www.adobe.com/max/2025/sessions/project-turn-style-gs3-10.html>.
- [2] Adobe Inc. 2025. View 2D Objects from New Angles (Turntable). Adobe Illustrator Desktop Help. Available at <https://helpx.adobe.com/illustrator/desktop/use-generative-ai/view-artwork-from-any-angle.html>.
- [3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.
- [4] Jad Bendarkawi, Ashley Ponce, Sean Chidozie Mata, Aminah Aliu, Yuhuan Liu, Lei Zhang, Amna Liaqat, Varun Nagaraj Rao, and Andrés Monroy-Hernández. 2025. Conversar: Exploring embodied llm-powered group conversations in augmented reality for second language learners. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–11.
- [5] Richard A Bolt. 1980. “Put-that-there” Voice and gesture at the graphics interface. In *Proceedings of the 7th annual conference on Computer graphics and interactive techniques*. 262–270.
- [6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18392–18402.
- [7] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. 2025. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719* (2025).
- [8] Jiazhong Cen, Zanwei Zhou, Jiemin Fang, Wei Shen, Lingxi Xie, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, et al. 2023. Segment anything in 3d with nerfs. *Advances in Neural Information Processing Systems* 36 (2023), 25971–25990.
- [9] Hsiang-Ting Chen, Li-Yi Wei, and Chun-Fa Chang. 2011. Nonlinear revision control for images. *ACM Transactions on Graphics (TOG)* 30, 4 (2011), 1–10.
- [10] Hsiang-Ting Chen, Li-Yi Wei, Björn Hartmann, and Maneesh Agrawala. 2016. Data-driven adaptive history for image editing. In *Proceedings of the 20th ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*. 103–111.
- [11] Tianyu Chen, Yasi Zhang, Zhi Zhang, Peiyu Yu, Shu Wang, Zhendong Wang, Kevin Lin, Xiaofei Wang, Zhengyuan Yang, Linjie Li, et al. 2025. EdiVal-Agent: An Object-Centric Framework for Automated, Fine-Grained Evaluation of Multi-Turn Editing. *arXiv preprint arXiv:2509.13399* (2025).
- [12] Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. 2025. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624* (2025).
- [13] Victoria Clarke and Virginia Braun. 2017. Thematic analysis. *The journal of positive psychology* 12, 3 (2017), 297–298.
- [14] Mariana Fernandez-Espinosa, Kai Zhang, Jad Bendarkawi, Ashley Ponce, Sean Chidozie Mata, Aminah Aliu, Lei Zhang, Francisco Fernandez Medina, Elena Mangione-Lora, Andres Monroy-Hernandez, et al. 2025. Practicing a Second Language Without Fear: Mixed Reality Agents for Interactive Group Conversation. *arXiv preprint arXiv:2510.08227* (2025).
- [15] Google. 2025. *Introducing Nano Banana Pro*. Retrieved March 28, 2026 from <https://blog.google/innovation-and-ai/products/nano-banana-pro/>
- [16] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626* (2022).
- [17] Alain Hore and Djemel Ziou. 2010. Image quality metrics: PSNR vs. SSIM. In *2010 20th international conference on pattern recognition*. IEEE, 2366–2369.
- [18] Erzhen Hu, Frederik Brudy, David Ledo, George Fitzmaurice, and Fraser Anderson. 2026. PrevizWhiz: Combining Rough 3D Scenes and 2D Video to Guide Generative Video Previsualization. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–21.
- [19] Erzhen Hu, Mingyi Li, Jungtaek Hong, Xun Qian, Alex Olwal, David Kim, Seongkook Heo, and Ruofei Du. 2025. Thing2Reality: Enabling Spontaneous

- Creation of 3D Objects from 2D Content using Generative AI in XR Meetings. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–16.
- [20] Xincheng Huang, Dieter Frehlich, Ziyi Xia, Peyman Gholami, and Robert Xiao. 2025. GaussianNexus: Room-Scale Real-Time AR/VR Telepresence with Gaussian Splatting. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–18.
- [21] Edwin L Hutchins, James D Hollan, and Donald A Norman. 1986. Direct manipulation interfaces. In *User centered system design*. CRC Press, 87–124.
- [22] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. 2024. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *European Conference on Computer Vision*. Springer, 150–168.
- [23] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* 42, 4 (2023), 139–1.
- [24] Tae Soo Kim, DaEun Choi, Yoonseo Choi, and Juho Kim. 2022. Stylette: Styling the web with natural language. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [25] Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. 2023. Large-scale text-to-image generation models for visual artists' creative works. In *Proceedings of the 28th international conference on intelligent user interfaces*. 919–933.
- [26] Gierad P Laput, Mira Dontcheva, Gregg Wilensky, Walter Chang, Aseem Agarwala, Jason Linder, and Eytan Adar. 2013. Pixeltone: A multimodal interface for image editing. In *proceedings of the SIGCHI conference on human factors in computing systems*. 2185–2194.
- [27] Jaewook Lee, Filippo Aleotti, Diego Mazala, Guillermo Garcia-Hernando, Sara Vicente, Oliver James Johnston, Isabel Kraus-Liang, Jakub Powierza, Donghoon Shin, Jon E Froehlich, et al. 2025. Imaginatear: AI-assisted in-situ authoring in augmented reality. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–21.
- [28] Jaewook Lee, Jaylin Herskovitz, Yi-Hao Peng, and Anhong Guo. 2022. ImageExplorer: Multi-layered touch exploration to encourage skepticism towards imperfect AI-generated image captions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [29] Toby Jia-Jun Li, Marissa Radensky, Justin Jia, Kirielle Singarajah, Tom M Mitchell, and Brad A Myers. 2019. Pumice: A multi-modal agent that learns concepts and conditionals from natural language and demonstrations. In *Proceedings of the 32nd annual ACM symposium on user interface software and technology*. 577–589.
- [30] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–15.
- [31] Vivian Liu and Lydia B Chilton. 2022. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–23.
- [32] Yubin Lu, Yihua Dai, Yutao Jiang, Tianyi Xiang, Chuhua Xian, and Shengfeng He. 2026. Canvas3D: Translating 2D Edits into Implicit Neural Instance Field. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 9, 1 (2026), 1–20.
- [33] Damien Masson, Sylvain Malacria, Géry Casiez, and Daniel Vogel. 2024. Directgpt: A direct manipulation interface to interact with large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [34] Nels Numan, Shwetha Rajaram, Balasaravanan Thoravi Kumaravel, Nicolai Marquardt, and Andrew D Wilson. 2024. Spaceblender: Creating context-rich collaborative spaces through generative 3d scene blending. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–25.
- [35] Jonas Oppenlaender. 2024. A taxonomy of prompt modifiers for text-to-image generation. *Behaviour & Information Technology* 43, 15 (2024), 3763–3776.
- [36] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988* (2022).
- [37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmlR, 8748–8763.
- [38] Shwetha Rajaram, Nels Numan, Balasaravanan Thoravi Kumaravel, Nicolai Marquardt, and Andrew D Wilson. 2024. Blendscape: Enabling end-user customization of video-conferencing environments through generative ai. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–19.
- [39] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*. Pmlr, 8821–8831.
- [40] Scott Reed, Zeynep Akata, Xinchun Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International conference on machine learning*. Pmlr, 1060–1069.
- [41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
- [42] Rajeev Sharma, Vladimir I Pavlovic, and Thomas S Huang. 1998. Toward multi-modal human-computer interface. *Proc. IEEE* 86, 5 (1998), 853–869.
- [43] Xinyu Shi, Yinghou Wang, Ryan Rossi, and Jian Zhao. 2025. Brickify: Enabling expressive design intent specification through direct manipulation on design tokens. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–20.
- [44] Shneiderman. 1983. Direct manipulation: A step beyond programming languages. *Computer* 16, 8 (1983), 57–69.
- [45] Hari Subramonyam, Roy Pea, Christopher Pondoc, Maneesh Agrawala, and Colleen Seifert. 2024. Bridging the gulf of envisioning: Cognitive challenges in prompt based interactions with LLMs. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [46] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2149–2159.
- [47] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. 2024. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*. Springer, 1–18.
- [48] Adam Tapal, Ela Oren, Reuven Dar, and Baruch Eitam. 2017. The sense of agency scale: A measure of consciously perceived control over one's mind, body, and the immediate environment. *Frontiers in psychology* 8 (2017), 1552.
- [49] Matthew Turk. 2014. Multimodal interaction: A review. *Pattern recognition letters* 36 (2014), 189–195.
- [50] Cyrus Vachha, Yixiao Kang, Zach Dive, Ashwat Chidambaram, Anik Gupta, Eunice Jun, and Björn Hartmann. 2025. Dreamcrafter: Immersive editing of 3d radiance fields through flexible, generative inputs and outputs. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [51] Peng Wang, Yichun Shi, Xiaochen Lian, Zhonghua Zhai, Xin Xia, Xuefeng Xiao, Weilin Huang, and Jianchao Yang. 2025. Seededit 3.0: Fast and high-quality generative image editing. *arXiv preprint arXiv:2506.05083* (2025).
- [52] Xueyang Wang, Qinxuan Cen, Weitao Bi, Yunxiang Ma, Xin Yi, Robert Xiao, Xinyi Fu, and Hewu Li. 2026. Roomify: Spatially-Grounded Style Transformation for Immersive Virtual Environments. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–24.
- [53] Yunlong Wang, Shuyuan Shen, and Brian Y Lim. 2023. Reprompt: Automatic prompt editing to refine ai-generative art towards precise expressions. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–29.
- [54] Hongyang Wei, Bin Wen, Yancheng Long, Yankai Yang, Yuhang Hu, Tianke Zhang, Wei Chen, Haonan Fan, Kaiyu Jiang, Jiankang Chen, et al. 2026. UniRef-Image-Edit: Towards Scalable and Consistent Multi-Reference Image Editing. *arXiv preprint arXiv:2602.14186* (2026).
- [55] Justin D Weisz, Jessica He, Michael Muller, Gabriela Hofer, Rachel Miles, and Werner Geyer. 2024. Design principles for generative AI applications. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–22.
- [56] Haijun Xia, Bruno Araujo, Tovi Grossman, and Daniel Wigdor. 2016. Object-oriented drawing. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 4610–4621.
- [57] Yuhuan Xie, Aoxuan Pan, Yi-Hua Huang, Chirui Chang, Peng Dai, Xin Yu, and Xiaojuan Qi. 2026. ObjectMorpher: 3D-Aware Image Editing via Deformable 3DGS. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5828–5838.
- [58] Zhihao Yao, Xiwen Yao, Haowei Xiong, Yuan-Ling Feng, Qirui Sun, Yijie Guo, and Haipeng Mi. 2026. GestuProp: 3D virtual reality prop generation with co-speech gestures. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [59] Hu Ye, Jun Zhang, Sibio Liu, Xiao Han, and Wei Yang. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721* (2023).
- [60] J Diego Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–21.
- [61] Lei Zhang, Jin Pan, Jacob Gettig, Steve Oney, and Anhong Guo. 2024. Vrcopilot: Authoring 3d layouts with generative ai models in vr. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–13.
- [62] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3836–3847.
- [63] Lei Zhang, Shuyao Zhou, Amna Liaquat, Tinney Mak, Brian Berengard, Emily Qian, and Andrés Monroy-Hernández. 2025. Empowering Children to Create AI-Enabled Augmented Reality Experiences. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–16.





Ⓐ Original Image

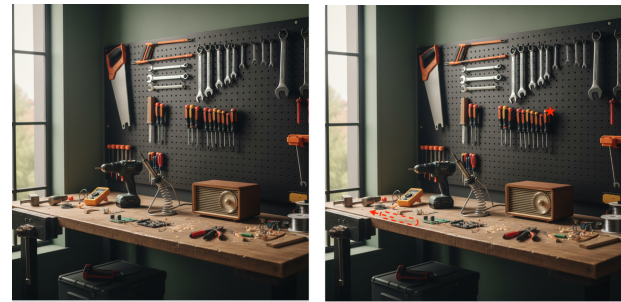
Ⓑ Step #1



Ⓒ Step #2

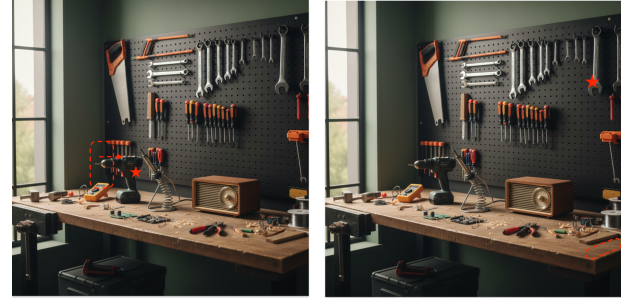
Ⓓ Step #3

Figure 8: Images for Study Task A.



Ⓐ Original Image

Ⓑ Step #1



Ⓒ Step #2

Ⓓ Step #3

Figure 9: Images for Study Task B.

## A Study Tasks

Participants were given the following framing throughout the study:

*Scenario: Imagine you are a junior art director for a visual design studio. Your client has approved the general aesthetic of several product shots but has sent back a list of critical spatial revisions. You must use the provided tool to execute each edit without changing the style or identity of the objects in the scene.*

### A.1 Task A: The Kitchen Scene

Below is the list of steps we asked participants (see Fig. 8):

- Step #1: Bring the blue mug (starred in red) to the circle and rotate it. The arrow indicates the direction of the handle. This change creates a feeling as if the photographer is using the mug.
- Step #2: The knife looks dangerous. Hide it between the avocado and the bread. The red arrow indicates the direction of the blade.
- Step #3: The two bananas should be spread out to make the scene more balanced. Specifically, the banana on the top should be moved to the red circle.

### A.2 Task B: The Workshop Scene

Below is the list of steps we asked participants (see Fig. 9):

- Step #1: To show a scene after someone just used the tools, move the screw driver (starred in red) to the red circle. The arrow indicates the direction of the tip

- Step #2: The power drill is too close to the user and looks dangerous! Move it to the red circle but behind the multimeter. The arrow indicates the direction of the drill.
- Step #3: To show a scene after someone just used the wrench, move the wrench (starred in red) to the red circle.

## B Additional Results

Figures 10 & 11 provide additional results described in Section 5.3.

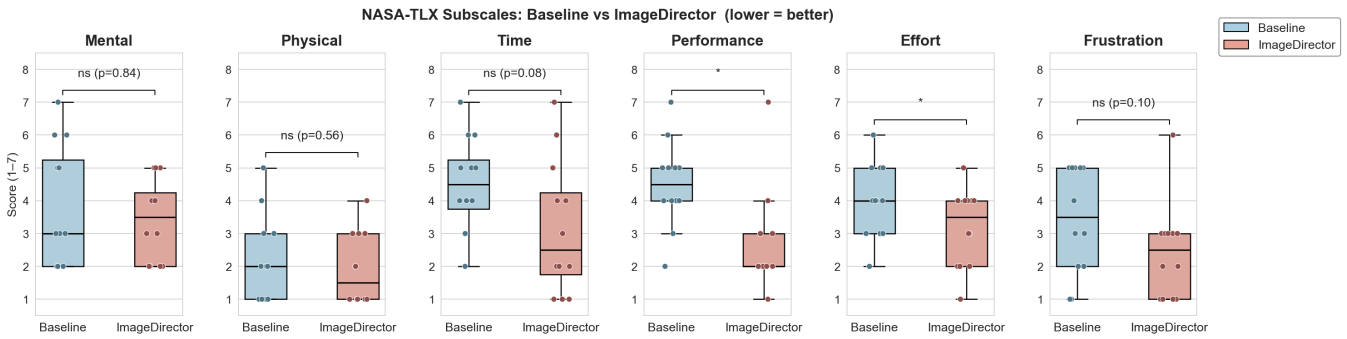


Figure 10: NASA-TLX Results. \*  $p < .05$ , \*\*  $p < .01$ , \*\*\*  $p < .001$ . N = 12.

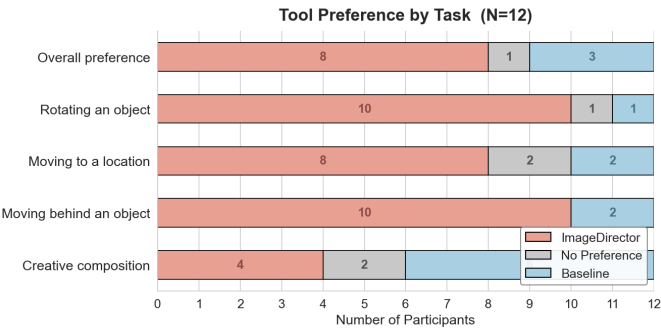


Figure 11: Participants' preferences comparing ImageDirector and Baseline.