

# Diffusion-Based Immersive Visual Reasoning

Yifeng Zhang<sup>✉</sup>, Ming Jiang<sup>✉</sup>, and Qi Zhao<sup>✉</sup>

University of Minnesota, Minneapolis MN 55455, USA  
{zhan6987, mjiang}@umn.edu, qzhao@cs.umn.edu

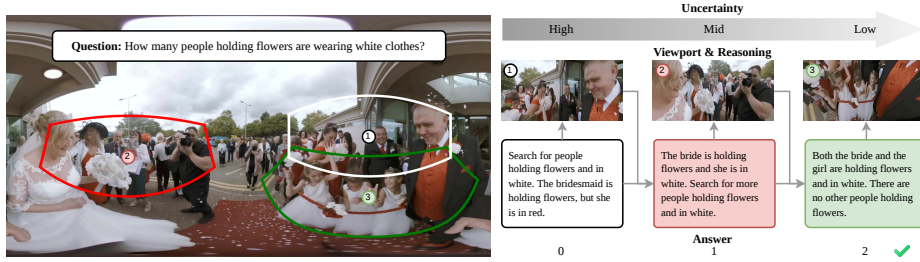
**Abstract.** Visual reasoning in immersive environments is critical for VR, AR, and robotics, yet is fundamentally challenged by partial observability. Existing Vision-Language Models (VLMs) are ill-suited for this task, as their rigid “look first, think later” paradigm decouples perception from reasoning. They either process distorted, low signal-to-noise panoramas or reason over a fixed set of pre-determined views, preventing dynamic, task-driven exploration. We reframe immersive visual reasoning as an iterative process of probabilistic belief updating. We introduce Diffusion-Based Immersive Visual Reasoning (DIVR), a novel framework that models reasoning as a denoising diffusion process, progressively refining a latent reasoning hypothesis from a state of high uncertainty to a denoised low-uncertainty state. DIVR comprises two tightly coupled modules: a Latent Hypothesis Refiner (LHR) that uses observed visual evidence to denoise the current hypothesis, and a Reasoning-Guided Observer (RGO) that uses the denoised hypothesis to actively select the next most informative viewport. This creates a closed-loop system where the evolving reasoning hypothesis continuously guides perception, distinct from the rigid, open-loop paradigm of prior work. Extensive experiments on multiple 360° VQA benchmarks demonstrate that DIVR establishes a new state-of-the-art, significantly outperforming strong baselines, particularly on complex questions that require dynamic, multi-step evidence gathering.

**Keywords:** Immersive Visual Reasoning · VLM · Diffusion Model

## 1 Introduction

Immersive 360° visual content is becoming central to domains like virtual reality (VR), augmented reality (AR), and robotics [8, 11, 14, 29, 42, 47]. Unlike traditional narrow field-of-view (FoV) imagery, 360° environments present a profound challenge of partial observability: an agent must reason over a vast scene where crucial evidence may be incomplete or lie outside its current view. This setting isolates the core challenge of perception-driven reasoning, where an agent must decide *where to look* to iteratively build a coherent understanding.

Existing Vision-Language Models (VLMs) [4, 30, 53], while powerful, are constrained by a naive “look first, think later” paradigm that fails in immersive contexts. This open-loop process, which decouples perception from reasoning, relies on either processing a full, geometrically distorted equirectangular image [24, 32, 48], which forces the model to ground details amidst severe projection



**Fig. 1:** Our DIVR framework performs iterative, active reasoning in a partially observable 360° immersive environment. Faced with a complex question, a diffusion model continuously refines its reasoning hypothesis under the guidance of selected viewports, and feeds back the refined hypothesis to guide the next round of viewport selection. As relevant information accumulates over iterations, the model progressively reduces uncertainty, converging toward a clearer reasoning hypothesis and more precise viewport selections, ultimately enabling more reliable answers.

artifacts, or reasoning over a fixed set of pre-determined viewports [12, 22, 36, 48]. Both approaches possess inherent limitations because they prevent the model from actively seeking new evidence if its initial perception or reasoning proves insufficient.

Consider the question posed in Fig. 1: “How many people holding flowers are wearing white clothes?”. A static model might initially focus on a salient region (Viewport 1), identify a person holding flowers but wearing red, and incorrectly conclude the answer. Such open-loop systems lack the ability to correct their initial, incomplete perception through subsequent, targeted exploration. To overcome these limitations, we reframe immersive visual reasoning as an iterative process of updating probabilistic beliefs. We introduce Diffusion-Based Immersive Visual Reasoning (DIVR), a framework that models reasoning through the lens of denoising diffusion. As shown in Fig. 1, DIVR begins with a reasoning hypothesis in a state of high uncertainty. It first inspects Viewport 1 and refines its hypothesis. Guided by this updated hypothesis, which now includes the need to find people in white, it actively selects a more informative view, observes the bride, and refines its hypothesis further (Viewport 2). This cycle of reasoning guiding perception, and perception informing reasoning, continues until the model has gathered sufficient evidence, reduced its uncertainty, and converged on the correct answer (Viewport 3). This closed-loop system, where reasoning actively steers perception to resolve uncertainty, forms the core principle of our DIVR framework.

Our framework comprises two novel modules: (1) A Latent Hypothesis Refiner (LHR) acts as the core reasoning module of our framework. This module operates on the model’s current, uncertain reasoning hypothesis on how to solve the task. At each step, it conditions on new visual evidence to progressively refine the hypothesis. (2) A Reasoning-Guided Observer (RGO) serves as the active perception component. It is guided by the current reasoning hypothesis

of the LHR to actively select the next most informative viewport. This ensures that the model’s perceptual focus is always aligned with its immediate reasoning needs. Through end-to-end training, these modules jointly learn to optimize the RGO’s perception policy and the LHR’s reasoning process for the final task.

In sum, our contributions are:

1. We introduce DIVR, a novel diffusion-based framework that models immersive visual reasoning as an iterative belief updating process. We implement this concept via a denoising diffusion model, providing a principled approach to resolving uncertainty through active perception.
2. We propose a new architecture composed of the LHR and RGO modules, which are jointly trained in a closed loop, coupling the model’s reasoning with its perception through iterative reasoning and active evidence acquisition.
3. Our DIVR establishes a new state-of-the-art on multiple 360° VQA benchmarks, providing evidence that our closed-loop paradigm is fundamentally superior to conventional open-loop methods for immersive reasoning.

## 2 Related Works

**Visual Reasoning in Immersive Scenes.** Traditional VLM reasoning assumes full observability, where all evidence is present in a static, FoV image [1, 3, 20, 31]. Using massive pre-training, large-scale VLMs excel in this paradigm [4, 28, 30, 35, 53]. To address visual reasoning in immersive scenes, early approaches attempted to adapt the static paradigm in two main methods. They first involve compressing the entire equirectangular image into a single spherical feature representation [12, 24, 25, 48]. While this approach preserves global context, it suffers from severe information overload and introduces significant geometric distortions, making it difficult for models to ground fine-grained details. The second strategy involves pre-sampling the 360° scene into a set of static viewports (*e.g.*, six cubemap faces) and processing them either in parallel or with a static attention mechanism [15, 50]. Both of these approaches share a critical flaw: they decouple the act of perception from the process of reasoning. The model is forced to reason over a pre-determined, static set of visual inputs, whether it’s a distorted global view or a fixed collection of crops. If the necessary evidence is not captured or is obscured in the initial representation, the reasoning process has no recourse to seek it out. In contrast, our work proposes a paradigm shift toward a tightly coupled, dynamic framework where reasoning actively guides perception in a continuous feedback loop.

**Viewport Selection and Active Perception.** Actively selecting informative viewpoints in 360° imagery has been explored via saliency or reinforcement learning (RL) to mimic human gaze or optimize for task rewards [10, 26, 27, 36, 54]. However, these methods typically treat viewport selection as a pre-processing step, detached from an evolving, multi-step reasoning task. Embodied AI, particularly embodied VQA [38, 43, 46], integrates perception and action but entangles visual reasoning with the complexities of physical navigation (*e.g.*, reward engineering, sim-to-real gaps). Our work deliberately isolates the core reasoning

process: we treat viewport selection as a lightweight perceptual behavior (where to look, not how to move), allowing us to study the fundamental feedback loop between belief updating (reasoning) and information acquisition (perception). By abstracting away the complexities of navigation, we can focus directly on the task of visual reasoning in partially observable settings.

**Uncertainty Modeling and Diffusion-Based Reasoning.** Classical reasoning systems often parse a question into a fixed, executable program [1–3, 16, 19, 41], a rigid approach that makes them vulnerable in partially observed environments: if an early step in the program fails, the entire reasoning chain can break. Diffusion models [6, 40, 52], originating from generative modeling, offer a powerful paradigm for iterative refinement under uncertainty. Their core process of transforming noise into structure provides a principled mechanism for evolving a system from a high-entropy (uncertain) to a low-entropy (coherent) state. This has been extended to non-generative domains like trajectory planning [33], policy optimization [39] and knowledge reasoning [45]. Building on this insight, we are the first to frame iterative, evidence-driven belief updating in immersive VQA as a denoising process, where the latent state of a diffusion model represents a noisy, high-entropy reasoning hypothesis that is progressively denoised into a coherent, low-entropy state by conditioning on actively-acquired visual evidence. This provides a probabilistically grounded alternative to rigid, deterministic pipelines and allows for robust reasoning in complex 360° scenes.

### 3 Methodology

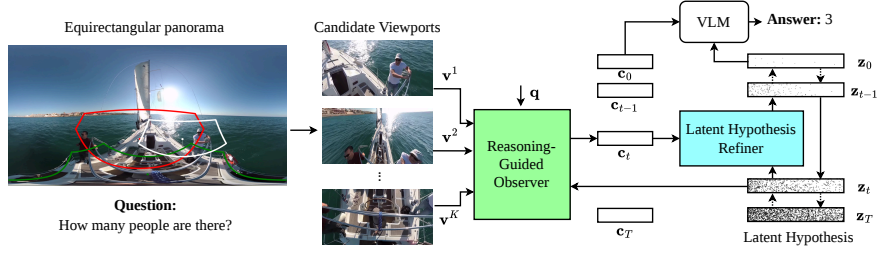
Reasoning in immersive environments is fundamentally a process of active exploration and belief updating. An agent in a 360° scene must decide where to look, what to remember, and how to incrementally build a coherent understanding from a series of partial views.

At its core, DIVR is founded on the principle of *reasoning as denoising*. The model’s evolving understanding of how to solve the task is encapsulated by a *latent reasoning hypothesis*, a vector  $\mathbf{z}_t \in \mathbb{R}^D$  in a continuous semantic space. This hypothesis vector represents the embedding of a textual reasoning process, which outlines the steps required to solve the given task.

The reasoning process begins at timestep  $t = T$  with a hypothesis  $\mathbf{z}_T$  sampled from a standard Gaussian distribution,  $\mathcal{N}(\mathbf{0}, \mathbf{I})$ , representing a state of maximal uncertainty. The framework’s objective is to iteratively denoise this vector over  $T$  steps, transforming it into a coherent, evidence-grounded reasoning hypothesis  $\mathbf{z}_0$  that guides the VLM to the final answer. This refinement occurs within a fundamental reasoning-perception cycle, illustrated in Fig. 2. Each cycle, corresponding to a single reverse diffusion step from  $t$  to  $t - 1$ , is composed of two processes:

1. The **Latent Hypothesis Refiner (LHR)** performs a single step of conditional denoising. It takes the newly acquired visual evidence from the RGO as a condition and updates the latent hypothesis from  $\mathbf{z}_t$  to a more refined





**Fig. 2:** Our DIVR framework iteratively refines a latent reasoning hypothesis  $\mathbf{z}_t$  over  $T$  denoising steps. At each step  $t$ , the RGO uses the current, uncertain hypothesis  $\mathbf{z}_t$  to actively select the most informative viewport, yielding a conditional vector  $\mathbf{c}_t$  encoding the question and the visual evidence. The LHR then conditions on  $\mathbf{c}_t$  to perform a single denoising step, producing a more refined hypothesis  $\mathbf{z}_{t-1}$ . This loop transforms an initially random guess  $\mathbf{z}_T$  into a final, evidence-grounded reasoning process  $\mathbf{z}_0$ .

state  $\mathbf{z}_{t-1}$ . This step directly operationalizes the reverse diffusion process, steering the reasoning trajectory toward a state that is more consistent with the observed world.

2. The **Reasoning-Guided Observer (RGO)** executes a learned policy network to actively select a viewport. Conditioned on the current hypothesis  $\mathbf{z}_t$  and the question  $\mathbf{q}$ , the policy chooses the viewport that is predicted to yield the most informative visual evidence for reducing hypothesis uncertainty. This transforms perception from a passive process into an active, goal-directed search for information.

This loop repeats iteratively. The refined hypothesis  $\mathbf{z}_{t-1}$  becomes the new reasoning state for the RGO in the subsequent cycle, allowing the gaze policy to become increasingly focused as the reasoning becomes clearer. This virtuous cycle enables the framework to dynamically explore the scene, gather visual information in a targeted manner, and converge on a robust and accurate solution. The final, fully-denoised hypothesis  $\mathbf{z}_0$  is then passed to a VLM decoder as prefix tokens to generate the answer.

### 3.1 Reasoning as Denoising

We formalize the reasoning process at any denoising step  $t \in \{0, 1, \dots, T\}$  as a conditional generative process over the latent variable  $\mathbf{z}_t$ . We define a conceptual *forward process* that progressively injects Gaussian noise into a clean reasoning hypothesis  $\mathbf{z}_0$ . This ground-truth  $\mathbf{z}_0$ , which represents the ideal reasoning process, is derived via knowledge distillation from a large VLM (detailed in Sec. 3.4).

The forward process is a fixed Markovian process defined by:

$$q(\mathbf{z}_t | \mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{1 - \beta_t} \mathbf{z}_{t-1}, \beta_t \mathbf{I}), \quad (1)$$

where  $\{\beta_t\}_{t=1}^T$  is a predefined variance schedule controlling the noise level at each step. A notable property of this process is that we can sample  $\mathbf{z}_t$  directly from  $\mathbf{z}_0$  at any arbitrary timestep  $t$  in a closed form:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t} \mathbf{z}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}, \quad (2)$$

where  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ,  $\alpha_t = 1 - \beta_t$  and  $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ .

The core of our framework is to learn the *reverse process*  $p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{c}_t)$ , which represents a single step of reasoning. This process denoises the latent hypothesis conditioned on a set of time-varying information  $\mathbf{c}_t$ , which contains both the input question and the visual evidence gathered by the RGO. Following the standard derivation for diffusion models [17], the learning objective can be simplified to training a network  $s_\theta$  to predict the noise component  $\boldsymbol{\epsilon}$  from the noisy hypothesis  $\mathbf{z}_t$ . The objective is to minimize the expected L2 error:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t, \mathbf{z}_0, \boldsymbol{\epsilon}} \left[ \|\boldsymbol{\epsilon} - s_\theta(\mathbf{z}_t, t, \mathbf{c}_t)\|^2 \right], \quad (3)$$

where  $t$  is sampled uniformly from  $\{1, \dots, T\}$ ,  $\mathbf{z}_0$  is the target hypothesis, and  $\mathbf{z}_t$  is generated using Eq. (2).

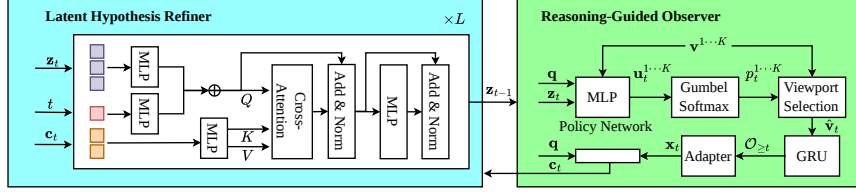
The key novelty of our DIVR framework lies in the dynamic construction and composition of the conditioning information  $\mathbf{c}_t$  at each refinement step. To maintain valid diffusion training, we define it as a deterministic mapping over states and other constants, concatenating both static and dynamic elements:

$$\mathbf{c}_t = [\mathbf{q}; \mathbf{x}_t], \quad (4)$$

where  $\mathbf{q} \in \mathbb{R}^{D_q}$  is the language embedding of the input question, and  $\mathbf{x}_t$  is the dynamic embedding of visual tokens accumulated at the denoising step  $t$ . This ensures that the refinement of the reasoning hypothesis is actively and continuously guided by both the overarching linguistic query and the incrementally acquired visual evidence from the scene. The specific architecture for incorporating  $\mathbf{c}_t$  into the denoiser  $s_\theta$  is detailed in the following section.

### 3.2 Latent Hypothesis Refiner

The LHR is the core reasoning module of our framework, designed with a Transformer-based architecture for conditional denoising. We implement its denoiser network  $s_\theta$  using an adapted Transformer architecture. The input to the LHR is a sequence of embedding vectors, including the latent hypothesis  $\mathbf{z}_t$  and the condition vector  $\mathbf{c}_t$ . The current timestep  $t$  is also encoded into a vector using sinusoidal positional embeddings [44]. As illustrated in Fig. 3 (Left), its architecture is composed of  $L = 6$  stacked blocks. Within each block, the refinement is driven by a cross-attention mechanism. Specifically, the latent hypothesis  $\mathbf{z}_t$  and the timestep embedding for  $t$  are projected by separate MLPs and then combined to form the *Query* ( $Q$ ). Concurrently, the conditioning vector  $\mathbf{c}_t$  is projected by an MLP to generate the *Key* ( $K$ ) and *Value* ( $V$ ).



**Fig. 3:** Architectures of the LHR and RGO. **Left:** The LHR is a Transformer-based denoiser. It refines the latent hypothesis  $\mathbf{z}_t$  by using it and the timestep  $t$  to form a query that cross-attends to the conditioning vector  $\mathbf{c}_t$ . This process is repeated for  $L$  blocks to produce the refined hypothesis  $\mathbf{z}_{t-1}$ . **Right:** The RGO acts as an active perception policy. It uses the current hypothesis  $\mathbf{z}_t$  and the question  $\mathbf{q}$  to score candidate viewports, selects one via Gumbel-Softmax, and aggregates the new visual evidence using a GRU, followed by an Adapter network converting the output into VLM tokens.

The core operation is a standard cross-attention layer, followed by a feed-forward network (MLP), with residual connections and layer normalization applied after each step. This mechanism enables the latent hypothesis to directly attend to the most relevant elements of the visual evidence, dynamically guiding the denoising trajectory. After processing through all  $L$  blocks, the final output embedding corresponding to the hypothesis token is passed through a linear layer to predict the noise vector  $\epsilon$ , which is then used to reconstruct the denoised reasoning hypothesis  $\mathbf{z}_{t-1}$ . This network design provides a powerful and flexible mechanism for fusing multi-modal, dynamically-acquired evidence to guide the iterative reasoning process.

### 3.3 Reasoning-Guided Observer

The LHR’s conditional denoising process is contingent on a stream of relevant visual evidence. The RGO’s role is to provide this stream by actively deciding where to look next. As shown in Fig. 3 (Right), it operationalizes a learned policy network,  $\pi_\phi$ , that transforms the passive act of observation into an active, goal-directed search.

The policy network is implemented as a 3-layer MLP with GELU activation functions, Layer Normalization, and a final Dropout layer for regularization. It computes a scalar utility score  $u_k$  for the  $k$ -th candidate viewport by processing the concatenated representation of the current reasoning state ( $\mathbf{z}_t$ ), textual ( $\mathbf{q}$ ) and the visual embeddings ( $\mathbf{v}_k$ ):

$$u_k = \text{MLP}_\phi([\mathbf{z}_t; \mathbf{q}; \mathbf{v}_k]), \quad (5)$$

where  $[\cdot; \cdot]$  denotes vector concatenation.

To enable end-to-end training via backpropagation, we utilize the Gumbel-Softmax reparameterization trick [21], a technique that provides a continuous and differentiable approximation to sampling from the categorical distribution

defined by the policy logits. The probability  $p_k$  for selecting viewport  $k$  is computed as:

$$p_k = \frac{\exp((u_k + g_k)/\tau)}{\sum_{j=1}^K \exp((u_j + g_j)/\tau)}, \quad (6)$$

where  $g_k \sim \text{Gumbel}(0, 1)$  are i.i.d. random variables and  $\tau$  is a temperature hyperparameter. During training, gradients from the final loss can flow through this sampling process to update the policy network parameters  $\phi$ .

The generated distribution  $\mathbf{p}_t$  is used to perform a weighted selection over the set of pre-computed visual features for all candidate viewports, producing the feature vector of the chosen viewport,  $\hat{\mathbf{v}}_t$ . To maintain a memory of observations,  $\hat{\mathbf{v}}_t$  is passed as the current input to a Gated Recurrent Unit (GRU), which updates its hidden state to produce an aggregated observation history  $\mathbf{O}_{\geq t}$ . This history is then projected by an Adapter network into an embedding  $\mathbf{x}_t$ , which is concatenated with the static question embedding  $\mathbf{q}$  to form the complete conditioning vector  $\mathbf{c}_t = [\mathbf{q}; \mathbf{x}_t]$  for the LHR’s next denoising step.

### 3.4 Training and Optimization

The LHR and RGO modules are trained end-to-end to ensure that the LHR learns to perform coherent reasoning, while the RGO learns a policy that is effective for the reasoning task. The composite loss function is a weighted sum of a diffusion objective and a task-specific objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diff}} + \lambda \mathcal{L}_{\text{cls}}, \quad (7)$$

where  $\lambda$  is a hyperparameter that balances the two terms.

The primary objective,  $\mathcal{L}_{\text{diff}}$ , defined in Eq. 3, trains the LHR’s denoiser network  $s_\theta$  to faithfully reconstruct the target reasoning hypothesis  $\mathbf{z}_0$ . The secondary objective,  $\mathcal{L}_{\text{cls}}$ , provides the end-to-end signal to train the model for the reasoning task. For example, in VQA, this loss is the standard cross-entropy between the predicted answer distribution and the ground-truth answer:

$$\mathcal{L}_{\text{cls}} = -\log p(\mathbf{y}|\mathbf{z}_0, \mathbf{c}_0), \quad (8)$$

where the final answer distribution is produced by the VLM’s classifier head, based on the fully-denoised hypothesis  $\mathbf{z}_0$  and condition  $\mathbf{c}_0$ .

Training the LHR and RGO modules requires a dataset of high-quality ground-truth signals for  $\mathbf{z}_0$ . We therefore employ a knowledge distillation strategy, using a powerful teacher VLM (e.g., GPT-5 [35]) to generate this supervisory signal. We emphasize that GPT-5 serves as a passive annotation tool for generating training targets, not as an active test-time reasoner. To do this, we construct a multi-modal prompt containing the equirectangular 360° image, the question, the ground-truth answer, and a set of the most salient viewports. These viewports are selected based on a saliency map derived from eye-tracking data [22] or a state-of-the-art saliency model pretrained on the Wild-360 dataset [11]. Given a saliency map of the 360° image, viewports with the highest area-normalized

saliency values are selected iteratively, with an inhibition-of-return mask applied to the selected areas to avoid overlapping [49]. Prompted with this context, the teacher generates a brief, natural-language rationale outlining the reasoning steps to the answer. The language embedding of this generated rationale serves as the ground-truth  $\mathbf{z}_0$  for our diffusion model. This distillation strategy allows us to train DIVR not just on the final answer, but on the entire reasoning process, enabling it to learn the multi-step, goal-directed behavior essential for complex immersive tasks. The full prompting methodology and our data validation protocol are detailed in the supplementary material.

## 4 Experiments

We conduct a comprehensive evaluation to validate the efficacy of our DIVR framework. We demonstrate its superiority over state-of-the-art methods on diverse 360° VQA tasks, perform rigorous ablation studies to isolate the contributions of its core components, and provide in-depth qualitative and quantitative analyses of its reasoning process. More implementation details and analyses are provided in the supplementary material.

### 4.1 Experimental Setup

**Datasets.** We evaluate on three challenging benchmarks to assess various facets of immersive reasoning. IQVA [22]: An immersive video dataset with human attention annotations, which we use to evaluate reasoning on real-world dynamic scenes by extracting key frames. 360+X [9]: An activity-centric dataset with detailed annotations. We convert its recognition task into a structured VQA format to benchmark perceptual alignment and reasoning. Zillow Indoor (ZInD) [13]: A 360° indoor scene dataset with room layout annotations. We formulate VQA tasks centered on spatial relationships, categorized into ‘Easy’ (single-viewport) and ‘Hard’ (multi-viewport) questions to explicitly test evidence aggregation.

**Compared Models.** We benchmark DIVR against a strong and diverse set of state-of-the-art models: SPHTR [27]: A representative model using spherical attention mechanisms designed specifically for 360° imagery. SLiME-8B [50]: A recent high-resolution model using adaptive patching, testing whether performance gains can be achieved via improved spatial coverage alone. GPT-5 (Zero-shot) [35]: A frontier general-purpose VLM, evaluated in a zero-shot setting to establish a powerful, generalist baseline.

**Input Settings.** The lack of a standard protocol for feeding 360° images to VLMs necessitates a controlled comparison. We evaluate baselines using three settings. Equirectangular: The full equirectangular image is fed to the model. This approach preserves global spatial continuity but introduces significant distortions and information overload, challenging the model’s ability to ground details. Cubemap: The surrounding scene is projected onto six distortion-reduced cubemap faces, which are concatenated as a multi-image input. This mitigates distortion, but increases the amount of visual tokens, posing a challenge in terms

**Table 1:** Quantitative comparison of model accuracy (%) on immersive reasoning datasets. Best results are in **bold**.

| Model                      | IQVA 360+X   |              | ZInD         |              |
|----------------------------|--------------|--------------|--------------|--------------|
|                            |              |              | Easy         | Hard         |
| SPHTR (Equirectangular)    | 45.96        | 48.26        | 57.73        | 23.55        |
| SLiME-8B (Equirectangular) | 41.78        | 48.62        | 61.48        | 28.46        |
| SLiME-8B (Cubemap)         | 46.50        | 54.18        | 63.90        | 33.35        |
| SLiME-8B (Saliency)        | 53.43        | 54.41        | 65.85        | 34.14        |
| GPT-5 (Equirectangular)    | 47.78        | 50.36        | 61.04        | 33.30        |
| GPT-5 (Cubemap)            | 53.42        | 56.47        | 65.82        | 35.49        |
| GPT-5 (Saliency)           | 55.84        | 58.62        | 68.01        | 35.64        |
| DIVR + GLM-4.1V-9B         | 56.61        | 63.77        | <b>71.33</b> | 41.36        |
| DIVR + Qwen2.5-VL-7B       | <b>56.75</b> | <b>64.72</b> | 70.45        | <b>41.89</b> |

**Table 2:** Comparison of VLM model accuracy (%) across different input strategies. Best results are in **bold**.

| Input     | IQVA 360+X   |              | ZInD         |              | IQVA 360+X    |              | ZInD         |              |
|-----------|--------------|--------------|--------------|--------------|---------------|--------------|--------------|--------------|
|           |              |              | Easy         | Hard         |               |              | Easy         | Hard         |
|           | GLM-4.1V-9B  |              |              |              | Qwen2.5-VL-7B |              |              |              |
| Equirect. | 43.89        | 50.06        | 59.84        | 31.73        | 43.54         | 49.75        | 59.73        | 31.49        |
| Cubemap   | 49.07        | 54.69        | 63.74        | 33.56        | 49.20         | 54.36        | 63.45        | 33.26        |
| Saliency  | 52.92        | 56.22        | 64.40        | 34.09        | 53.22         | 55.92        | 64.88        | 33.97        |
| DIVR      | <b>56.61</b> | <b>63.77</b> | <b>71.33</b> | <b>41.36</b> | <b>56.75</b>  | <b>64.72</b> | <b>70.45</b> | <b>41.89</b> |

of computational cost and the model’s ability to integrate information across disconnected views. **Saliency:** This setting simulates an attention-guided, decoupled pipeline. We use ground-truth eye-fixation maps (on IQVA [22]) or a pre-trained saliency model [11] (on 360+X [9] and ZinD [13]) to identify and crop the most salient regions. This provides a strong baseline for attention-aware systems. Unlike these settings, our DIVR dynamically selects viewports from a grid of  $K = 36$  overlapping candidate viewports [34], covering several different FoV sizes uniformly distributed across the  $360^\circ$  sphere.

**Implementation Details.** We instantiate DIVR with two powerful open-source VLMs: Qwen2.5-VL-7B [5] and GLM-4.1V-9B [18]. The reasoning hypothesis evolves over  $T = 20$  diffusion steps with a cosine variance schedule. We train end-to-end using the AdamW optimizer with a learning rate of  $10^{-4}$  and a batch size of 8 across 8 NVIDIA H100 GPUs. To ensure a fair comparison, all models (except zero-shot GPT-5) undergo a two-stage fine-tuning protocol: first jointly supervised with the answer and the distilled reasoning ground-truth on IQVA, then on downstream datasets with answer supervision. Performance is measured by standard VQA accuracy.

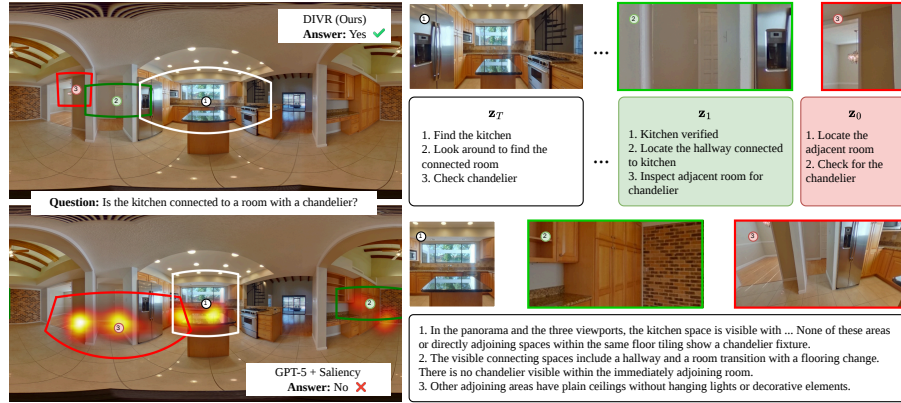
## 4.2 Quantitative Results

Our proposed DIVR framework consistently and significantly outperforms all baselines across every benchmark (see Tab. 1). Instantiated with either VLM backbone, DIVR establishes a new state-of-the-art, providing clear evidence that an iterative, coupled perception-reasoning process is fundamentally superior for immersive reasoning. DIVR surpasses specialized models like SPHTR, indicating that addressing geometric distortion alone is insufficient. More critically, DIVR outperforms strong baselines like SLIME and GPT-5 even when they are provided with saliency-guided viewports. This result shows that the ability to dynamically decide where and when to look, guided by an evolving internal reasoning hypothesis, is more effective than reasoning over a pre-determined, static set of salient views.

The performance gains are most pronounced on the ZInD ‘Hard’ split, where DIVR achieves an absolute gain of over 6% compared to the strongest baseline (41.89% vs. 35.64%). This highlights DIVR’s core strength: its active, multi-step evidence gathering is essential for complex questions requiring information integration across multiple views. As shown in Tab. 2, while shifting from equirectangular to cubemap or saliency inputs provides marginal gains, these improvements are dwarfed by the gains achieved with DIVR’s active perception loop. This controlled comparison allows us to conclude that DIVR’s performance gains are not attributable to a superior input data representation but to a fundamentally more powerful reasoning paradigm. Results are averaged over three seeds with standard deviations below 0.4%. Unless otherwise specified, all subsequent analyses are conducted using Qwen2.5-VL-7B [5] as our foundational VLM.

## 4.3 Qualitative Analysis

The top of Fig. 4 visualizes DIVR’s closed-loop reasoning on a challenging ZInD example: “Is the kitchen connected to a room with a chandelier?”. The process begins with a generic high-level reasoning hypothesis  $\mathbf{z}_T$ . The RGO first grounds the primary subject by selecting Viewport 1 (the kitchen). This observation allows the LHR to denoise the hypothesis, which in turn re-directs the RGO’s policy to continue its search. At step  $t = 1$ , the RGO now correctly ignores salient objects within the kitchen and selects Viewport 2, a non-salient but logically critical doorway. At the final step, the hypothesis converges to a fully-denoised, evidence-grounded state  $\mathbf{z}_0$ , focusing on the search for the chandelier in the adjacent room. The RGO, guided by this precise instruction, successfully finds Viewport 3, which provides a clear view of the chandelier. In contrast, the bottom of Fig. 4 illustrates why a powerful GPT-5 model with a static, saliency-based approach fails. The pre-determined views might capture the most salient regions but they miss the critical, non-salient chandelier. The model is forced to draw a conclusion from the incomplete observation. Its generated reasoning process reveals its failure: unable to find the chandelier, it incorrectly concludes one does not exist. This detailed comparison illustrates that immersive reasoning is not a static pattern recognition task but a dynamic, sequential process of hypothesis refinement and belief updating.



**Fig. 4:** Qualitative comparison on a complex multi-step reasoning task from ZInD. **Top:** Our DIVR framework executes a dynamic, active reasoning that iteratively reduces the uncertainty of reasoning hypothesis and the precision of viewports selection, enabling accurate and reliable answers. **Bottom:** A strong VLM baseline using static, saliency-based viewports fails, as its pre-determined views are disconnected and insufficient for the reasoning required. The heatmap visualizes human gaze patterns predicted by the saliency model.

#### 4.4 Ablation Analysis

Tab. 3 reports the ablation study to dissect our DIVR framework and quantify the contribution of its core components.

First, we isolate the RGO by comparing its learned policy against two strong alternatives for viewport selection: (1) a static *Saliency*-based approach that pre-selects a fixed set of the most salient viewports, and (2) a dynamic but heuristic *CLIP*-similarity approach, where the next viewport is chosen by finding the one with the highest *CLIP* [37] similarity with the current latent hypothesis  $z_t$ . The results show that a dynamic policy is superior to a static one, and our learned, task-specific RGO policy significantly outperforms the generic *CLIP*-based heuristic. This confirms that the RGO learns what visual information is most useful for reducing uncertainty, not just what is semantically relevant.

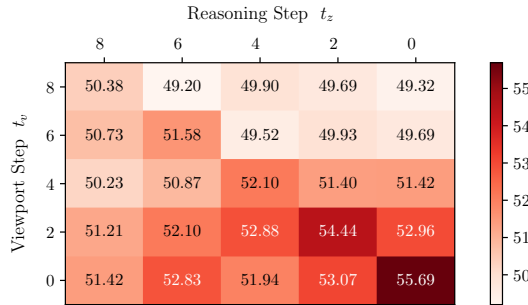
Second, we validate the LHR by comparing our full model to a variant without it (w/o LHR), which reasons directly over the accumulated visual evidence. Even with identical viewport selection policies, explicitly modeling the reasoning process as the iterative denoising of a latent hypothesis  $z_t$  provides a significant performance gain. It demonstrates that  $z_t$  serves as a more powerful and structured representation for the model’s evolving beliefs than raw observation.

Finally, the fact that our full DIVR model significantly outperforms any of the ablated variants confirms that the LHR and RGO are highly synergistic. Principled reasoning is necessary to guide intelligent perception, and in turn, targeted perception is required to fuel effective reasoning.



**Table 3:** Ablation study on the contributions of the LHR and RGO modules to the model accuracy (%). Best results are in **bold**.

| Reasoning  | Perception | IQVA         | 360+X        | ZInD         |              |
|------------|------------|--------------|--------------|--------------|--------------|
|            |            |              |              | Easy         | Hard         |
| w/o LHR    | Saliency   | 52.69        | 56.72        | 67.30        | 36.98        |
| w/o LHR    | CLIP       | 53.22        | 57.90        | 67.57        | 39.23        |
| w/o LHR    | RGO (Ours) | 54.24        | 58.51        | 68.34        | 39.44        |
| LHR (Ours) | Saliency   | 53.46        | 57.71        | 66.39        | 37.58        |
| LHR (Ours) | CLIP       | 55.14        | 58.59        | 67.40        | 38.68        |
| LHR (Ours) | RGO (Ours) | <b>56.75</b> | <b>64.72</b> | <b>70.45</b> | <b>41.89</b> |

**Fig. 5:** Analysis of the interplay between perception and reasoning throughout the denoising process on IQVA. Each cell shows the final VQA accuracy (%) when a reasoning hypothesis at step  $t_z$  is combined with evidence accumulated up to step  $t_v$ .

#### 4.5 Perception-Reasoning Interplay

To probe the temporal coupling between perception and reasoning, we evaluate the model’s final answer accuracy when conditioning its answer decoder on mismatched intermediate states: visual evidence accumulated up to step  $t_v$  and a reasoning hypothesis from step  $t_z$ . We set the number of reasoning steps to  $T=10$ , a value sufficient to capture the iterative refinement behavior while maintaining efficiency. The performance grid in Fig. 5 reveals a strong coupling: accuracy is highest along the diagonal ( $t_v \approx t_z$ ), confirming that perceptual and reasoning updates must progress in parallel for optimal performance. Performance at later stages remains high even with mismatched inputs, implying that information is robustly accumulated and retained throughout the diffusion process.

#### 4.6 Comparison with Human Attention

To further analyze the RGO’s learned policy, we compare the similarity of its selected viewports to ground-truth human gaze patterns on IQVA. For the selected

**Table 4:** Comparison between selected viewports and human fixation maps on the IQVA dataset.

| Method      | NSS  | KL   |
|-------------|------|------|
| Saliency    | 0.36 | 2.68 |
| DIVR (Ours) | 0.49 | 2.34 |

**Table 5:** Efficiency and performance trade-off analysis on the ZInD ‘Hard’ subset. For reference, baselines achieve 2.83 QPS (SLiME-8B Equirectangular) and 3.82 QPS (SPHTR).

| Depth    | Accuracy (%) | Speed (QPS) |
|----------|--------------|-------------|
| $T = 5$  | 38.46        | 3.57        |
| $T = 10$ | 40.63        | 1.91        |
| $T = 20$ | 41.89        | 1.08        |

viewports, a saliency map is computed by considering the center of each viewport as a fixation point [23]. Both maps are smoothed with the same Gaussian kernel ( $1^\circ$  FoV). As shown in Tab. 4, DIVR’s selections exhibit a significantly higher Normalized Scanpath Saliency (NSS) and lower Kullback-Leibler (KL) divergence [7] than a static saliency baseline. This suggests that our model learns a dynamic, task-oriented exploration that more closely mimics human question-answering strategies in immersive environments.

#### 4.7 Computational Cost

The iterative nature of DIVR, while effective, is computationally more expensive than single-pass models. However, a key advantage of DIVR is its flexible control over the number of diffusion steps  $T$ . As shown in Tab. 5 for the ZInD ‘Hard’ subset, with only 5 diffusion steps, the model achieves 38.46% accuracy, already outperforming the state-of-the-art models. Increasing  $T$  from 5 to 20 improves accuracy by a substantial 3.43%, at a linear cost to efficiency, measured with queries per second (QPS). This highlights a practical trade-off, demonstrating that deeper, iterative refinement is critical for complex questions.

## 5 Conclusion

We have introduced a fundamental paradigm shift for immersive visual reasoning, moving beyond static approaches by reframing the task as an iterative process of probabilistic belief updating. Our proposed DIVR framework uses a tightly-coupled, closed-loop architecture where a Latent Hypothesis Refiner (LHR) guides a Reasoning-Guided Observer (RGO) to seek the most informative evidence, enabling a virtuous cycle of belief-guided perception. Our extensive experiments provide conclusive evidence that this active, goal-directed paradigm is not merely beneficial but essential, establishing a new state-of-the-art and decisively outperforming passive methods on complex reasoning tasks that require dynamic evidence gathering. While the iterative nature of our approach presents a natural trade-off between reasoning depth and computational cost, we believe DIVR offers a powerful and principled framework for a broad class of AI challenges that require sequential decision-making under partial observability, from embodied AI to robotics.

## Acknowledgements

This work was supported by the Fasikl.

## Declaration

We explicitly note that major portions of the text, methodologies, and findings presented in this manuscript are derived from the previously published academic thesis [51] of the co-author, Yifeng Zhang, hosted on the ProQuest repository (ID: 3297291829). All permissions under institutional guidelines have been observed.

## References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6077–6086 (2018)
2. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Neural module networks. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 39–48 (2016)
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual question answering. In: Int. Conf. Comput. Vis. pp. 2425–2433 (2015)
4. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond (2023), arXiv:2308.12966
5. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-VL technical report (2025), arXiv:2502.13923
6. Beal, G.M., Bohlen, J.M.: The diffusion process. Tech. rep., Iowa State College (1956)
7. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? IEEE Trans. Pattern Anal. Mach. Intell. **41**(3), 740–757 (2018)
8. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 11621–11631 (2020)
9. Chen, H., Hou, Y., Qu, C., Testini, I., Hong, X., Jiao, J.: 360+X: A panoptic multimodal scene understanding dataset. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 19373–19382 (2024)
10. Chen, J., Luo, X., Hu, M., Wu, D., Zhou, Y.: Sparkle: User-aware viewport prediction in 360-degree video streaming. IEEE Trans. Multimedia **23**, 3853–3866 (2020)
11. Cheng, H.T., Chao, C.H., Dong, J.D., Wen, H.K., Liu, T.L., Sun, M.: Cube padding for weakly-supervised saliency prediction in 360° videos. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1420–1429 (2018)
12. Chou, S.H., Chao, W.L., Lai, W.S., Sun, M., Yang, M.H.: Visual question answering on 360° images. In: IEEE/CVF Winter Conf. Appl. Comput. Vis. pp. 1607–1616 (2020)

13. Cruz, S., Hutchcroft, W., Li, Y., Khosravan, N., Boyadzhiev, I., Kang, S.B.: Zillow indoor dataset: Annotated floor plans with 360° panoramas and 3D room layouts. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2133–2143 (2021)
14. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the KITTI vision benchmark suite. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3354–3361. IEEE (2012)
15. Guo, Z., Xu, R., Yao, Y., Cui, J., Ni, Z., Ge, C., Chua, T.S., Liu, Z., Huang, G.: LLaVA-UHD: an LMM perceiving any aspect ratio and high-resolution images. In: Eur. Conf. Comput. Vis. pp. 390–406. Springer (2024)
16. Gupta, N., Lin, K., Roth, D., Singh, S., Gardner, M.: Neural module networks for reasoning over text. In: Int. Conf. Learn. Represent. (2020)
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural Inform. Process. Syst. **33**, 6840–6851 (2020)
18. Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., et al.: GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning (2025), arXiv:2507.01006
19. Hu, R., Andreas, J., Darrell, T., Saenko, K.: Explainable neural computation via stack neural module networks. In: Eur. Conf. Comput. Vis. pp. 53–69 (2018)
20. Hudson, D.A., Manning, C.D.: GQA: A new dataset for real-world visual reasoning and compositional question answering. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6700–6709 (2019)
21. Jang, E., Gu, S., Poole, B.: Categorical Reparameterization with Gumbel-Softmax. In: Int. Conf. Learn. Represent. (2017)
22. Jiang, M., Chen, S., Yang, J., Zhao, Q.: Fantastic answers and where to find them: Immersive question-directed visual attention. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2980–2989 (2020)
23. John, B., Raiturkar, P., Le Meur, O., Jain, E.: A benchmark of four methods for generating 360° saliency maps from eye tracking data. Int. J. Semantic Computing **13**(03), 329–341 (2019)
24. Lai, X., Chen, Y., Lu, F., Liu, J., Jia, J.: Spherical transformer for lidar-based 3D recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 17545–17555 (2023)
25. Lai, X., Liu, J., Jiang, L., Wang, L., Zhao, H., Liu, S., Qi, X., Jia, J.: Stratified transformer for 3D point cloud segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8500–8509 (2022)
26. Li, C., Xu, M., Jiang, L., Zhang, S., Tao, X.: Viewport proposal cnn for 360° video quality assessment. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10177–10186 (2019)
27. Li, J., Han, L., Zhang, C., Li, Q., Liu, Z.: Spherical convolution empowered viewport prediction in 360 video multicast with limited FoV feedback. ACM Trans. Multimedia Comput. Commun. Appl. **19**(1), 1–23 (2023)
28. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Int. Conf. Mach. Learn. pp. 19730–19742. PMLR (2023)
29. Li, K., Bacher, R., Schmidt, S., Leemans, W., Steinicke, F.: Reality fusion: Robust real-time immersive mobile robot teleoperation with volumetric visual data fusion (2024), arXiv:2408.01225
30. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Adv. Neural Inform. Process. Syst. pp. 34892–34916 (2023)
31. Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Adv. Neural Inform. Process. Syst. **32**, 13–23 (2019)

32. Lv, H., Yang, Q., Li, C., Dai, W., Zou, J., Xiong, H.: SalGCN: Saliency prediction for 360-degree images based on spherical graph convolutional networks. In: ACM Int. Conf. Multimedia. pp. 682–690 (2020)
33. Mao, W., Xu, C., Zhu, Q., Chen, S., Wang, Y.: Leapfrog diffusion model for stochastic trajectory prediction. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5517–5526 (2023)
34. Nasrabadi, A.T., Samiei, A., Prakash, R.: Viewport prediction for 360° videos: a clustering approach. In: ACM Workshop Netw. Oper. Syst. Support Digit. Audio Video. pp. 34–39 (2020)
35. OpenAI: GPT-5 system card. Tech. rep., OpenAI (2025), <https://openai.com/index/gpt-5-system-card/>
36. Qiao, M., Xu, M., Wang, Z., Borji, A.: Viewport-dependent saliency prediction in 360° video. IEEE Trans. Multimedia **23**, 748–760 (2020)
37. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Int. Conf. Mach. Learn. pp. 8748–8763 (2021)
38. Ramakrishnan, S.K., Jayaraman, D., Grauman, K.: An exploration of embodied visual exploration. Int. J. Comput. Vis. **129**(5), 1616–1649 (2021)
39. Ren, A.Z., Lidard, J., Ankile, L.L., Simeonov, A., Agrawal, P., Majumdar, A., Burchfiel, B., Dai, H., Simchowitz, M.: Diffusion policy policy optimization. In: Int. Conf. Learn. Represent. (2025)
40. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10684–10695 (2022)
41. Shi, J., Zhang, H., Li, J.: Explainable and explicit visual reasoning over scene graphs. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8376–8384 (2019)
42. Stotko, P., Krumpen, S., Schwarz, M., Lenz, C., Behnke, S., Klein, R., Weinmann, M.: A VR system for immersive teleoperation and live exploration with a mobile robot. In: IEEE/RSJ Int. Conf. Intell. Robots Syst. pp. 3630–3637. IEEE (2019)
43. Tao, S., Xiang, F., Shukla, A., Qin, Y., Hinrichsen, X., Yuan, X., Bao, C., Lin, X., Liu, Y., Chan, T.k., Gao, Y., Li, X., Mu, T., Xiao, N., Gurha, A., Rajesh, V.N., Choi, Y.W., Chen, Y.R., Huang, Z., Calandra, R., Chen, R., Luo, S., Su, H.: ManiSkill3: GPU: parallelized robotics simulation and rendering for generalizable embodied AI. In: Robot.: Sci. Syst. (2025)
44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Adv. Neural Inform. Process. Syst. pp. 5998–6008 (2017)
45. Venkatraman, S., Khaitan, S., Akella, R.T., Dolan, J., Schneider, J., Berseth, G.: Reasoning with latent diffusion in offline reinforcement learning. In: Int. Conf. Learn. Represent. (2024)
46. Wu, Y., Zheng, W., Zuo, S., Huang, Y., Zhou, J., Lu, J.: EmbodiedOcc: Embodied 3D occupancy prediction for vision-based online scene understanding. In: Int. Conf. Comput. Vis. pp. 26360–26370 (2025)
47. Yew, A., Ong, S., Nee, A.: Immersive augmented reality environment for the teleoperation of maintenance robots. Procedia CIRP **61**, 305–310 (2017)
48. Yun, H., Yu, Y., Yang, W., Lee, K., Kim, G.: Pano-AVQA: Grounded audio-visual question answering on 360° videos. In: Int. Conf. Comput. Vis. pp. 2031–2041 (2021)
49. Zénou, A., Hamed, S.B., Duhamel, J.R., Olivier, E.: Attentional guidance relies on a winner-take-all mechanism. Vis. Res. **49**(12), 1522–1531 (2009)

50. Zhang, Y.F., Wen, Q., Fu, C., Wang, X., Zhang, Z., Wang, L., Jin, R.: Beyond LLaVA-HD: Diving into high-resolution large multimodal models. *IEEE Trans. Pattern Anal. Mach. Intell.* **48**(5), 5493–5504 (2026)
51. Zhang, Y.: Integrating Knowledge Into Visual Reasoning: Enhancing Generalizability, Explainability, and Bias Mitigation. Ph.D. thesis, University of Minnesota (2025)
52. Zhao, W., Rao, Y., Liu, Z., Liu, B., Zhou, J., Lu, J.: Unleashing text-to-image diffusion models for visual perception. In: *Int. Conf. Comput. Vis.* pp. 5729–5739 (2023)
53. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: Enhancing vision-language understanding with advanced large language models (2023), [arXiv:2304.10592](https://arxiv.org/abs/2304.10592)
54. Zhu, Y., Zhai, G., Yang, Y., Duan, H., Min, X., Yang, X.: Viewing behavior supported visual saliency predictor for 360 degree videos. *IEEE Trans. Circuits Syst. Video Technol.* **32**(7), 4188–4201 (2021)