UNCOVERING UNDERSTANDING–GENERATION SYNERGY IN NATIVE UNIFIED MULTIMODAL MODELS: FROM REPRESENTATION, TASK TO SYSTEM

Anonymous authors
Paper under double-blind review

ABSTRACT

While unified multimodal models (UMMs) jointly perform visual understanding and generation within a single model, functional unification does not guarantee learning synergy: the two objectives may reinforce each other, compete for capacity, or merely coexist. We investigate their relationship at the representation, task, and system levels in a controlled, structurally native setting without pretrained vision priors. At the representation level, we find that each objective provides useful signal to the other: generation enriches the visual features learned for understanding, while understanding strengthens vision–language alignment for generation. However, when both objectives are forced through the same computation path, one tends to dominate. A task-decoupled architecture that specializes conflicting visual computation while preserving semantic interaction avoids this asymmetric degradation. At the task level, through three case studies, we find positive bidirectional transfer when understanding and generation tasks rely on shared knowledge. At the system level, we show that an end-to-end UMM outperforms a matched planner–executor pipeline on complex tasks that explicitly require both image understanding and generation. Together, these results show that the value of UMMs extends beyond a unified interface: appropriate specialization, shared task knowledge, and end-to-end optimization can turn coexistence into synergy.

1 INTRODUCTION

Unified multimodal models (UMMs) bring visual understanding and generation into a single model, enabling interleaved text and image input and output (Chameleon Team, 2024; Zhou et al., 2025; Xie et al., 2025; Wu et al., 2025a; Deng et al., 2025; Liu et al., 2025; Qiu et al., 2026; Diao et al., 2026c). This supports interactions difficult for single-purpose models, such as reasoning through mixed visual–textual generation and iteratively interpreting and modifying images within one sequence.

However, a unified interface does not establish that the two capabilities intrinsically benefit each other. Their objectives may reinforce shared visual representations, compete for capacity, or simply coexist. Existing UMMs show that both can be supported at scale, but provide limited and conflicting evidence about their interaction (Kang et al., 2026; Han et al., 2026; Liu et al., 2025; Chen et al., 2025a; Pan et al., 2025). We therefore ask: *when and why do visual understanding and generation help each other, and when does unification create interference?*

To study this question under controlled conditions, we adopt a *native* pixel-in, pixel-out setting. Raw image patches enter without a pretrained vision encoder, while images are generated without a tokenizer or VAE (Kingma & Welling, 2014), minimizing external visual priors and allowing us to isolate the interaction between understanding and generation. We combine autoregressive text modeling with flow matching for images and, from the same pretrained language model, compare understanding-only, generation-only, and joint variants under controlled routing, parameter sharing, data, and training.

We organize the investigation at three levels. At the **representation level**, we study how parameter sharing and specialization determine transfer or interference. At the **task level**, we test whether paired tasks benefit from shared knowledge or capabilities. At the **system level**, we compare an end-to-end UMM with a pipeline of separate understanding and generation models.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

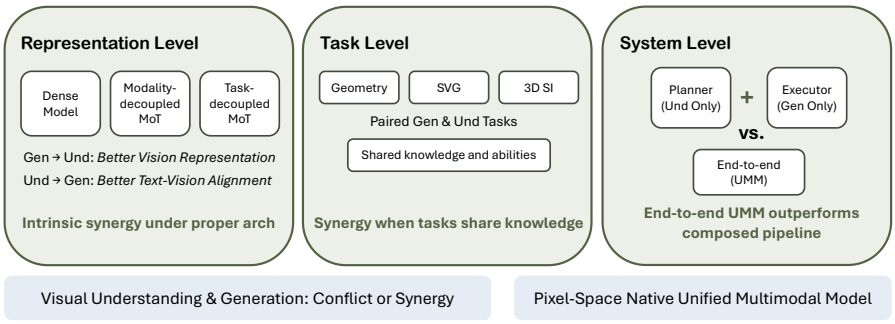


Figure 1: Overview of our three-level study of understanding–generation synergy in native UMMs.

This progression separates three questions: whether the objectives learn compatible representations, whether supervision transfers across related tasks, and whether unification remains advantageous when both capabilities are required within a single process.

At the representation level, we study two routing settings. A *dense shared model* processes text, clean understanding tokens, and noised generation tokens in the same LLM. A *modality-decoupled* MoT (Liang et al., 2024) instead routes all visual tokens through a scratch-trained branch. They exhibit opposite asymmetries: joint training improves understanding but degrades generation in the dense model, and does the reverse in the modality-decoupled model. Probing nevertheless reveals useful transfer in both directions: generation enriches visual representations for understanding, while understanding improves vision–language alignment for generation. We attribute the trade-off to forcing both objectives through the same computation path, which allows one to dominate. These opposite failure modes show that synergy depends not simply on sharing parameters, but on where sharing occurs and how each objective accesses pretrained semantics. Our *task-decoupled* MoT instead keeps understanding tokens in the language branch and assigns generation tokens to a specialized branch. Shared attention maintains soft semantic alignment while specialized computation reduces conflict without sacrificing the pretrained language semantics needed for understanding.

At the task level, we study geometry reasoning, vector graphics, and 3D spatial intelligence. Pairing geometry problem solving with diagram generation and editing improves both geometry and general mathematical reasoning. Joint training also yields bidirectional gains in vector graphics and 3D spatial intelligence, showing that synergy emerges when tasks share knowledge or capabilities.

At the system level, we study reasoning-intensive image editing, which requires understanding a source image and a textual request, deriving an explicit edit, and generating the result. We compare an end-to-end UMM with a matched planner–executor pipeline in which an understanding model first produces the explicit edit and a generation model executes it. Under the same base model, task data, and supervision, the end-to-end UMM performs better, showing an advantage when the two capabilities must interact tightly throughout a task.

Together, these results show that unification is beneficial but conditional: architecture determines whether transfer becomes mutual improvement, shared task knowledge enables bidirectional gains, and end-to-end modeling helps when the capabilities must interact closely. They recommend specializing computation under objective conflict and unifying learning around shared knowledge and decisions.

Our contributions are fourfold:

• We present a controlled and comprehensive study of understanding–generation interaction at the representation, task, and system levels in a native multimodal setting.

• We reveal that the two capabilities provide useful learning signals to each other, while their translation into mutual gains depends on architecture.

• We show through case studies from geometric reasoning to spatial intelligence that task-level synergy emerges when understanding and generation rely on shared knowledge or capabilities.

• We establish a system-level advantage of unification by showing that an end-to-end UMM outperforms a matched planner–executor pipeline on reasoning-intensive image editing.

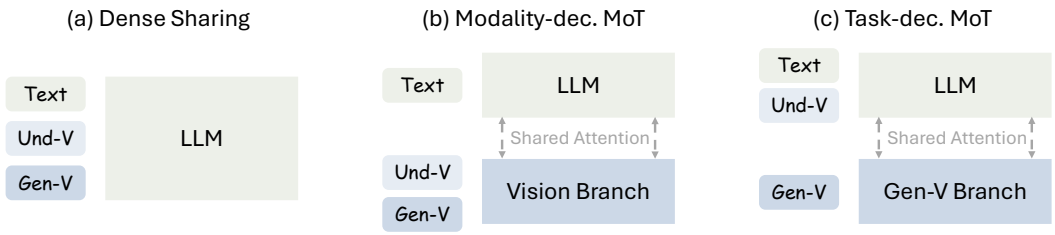


Figure 2: **Illustration of different architectures in our study.** (a) Dense sharing sends all tokens through the same LLM decoder. (b) Modality-decoupled MoT keeps text in the LLM and sends all visual tokens through a scratch-trained branch. (c) Task-decoupled MoT keeps UND-V language-anchored while specializing GEN-V. For simplicity, we omit the pre-buffer layers here.

2 NATIVE UNIFIED MODELING

2.1 SCOPE AND FORMULATION

Modeling choice. Unified multimodal models span several modeling paradigms. Images and text can both be predicted autoregressively with images being discretized (Chameleon Team, 2024; Chen et al., 2025b; Cui et al., 2025) or predicted using a flow-matching head (Huang et al., 2025; Wu et al., 2025b); both modalities can also be trained with a discrete denoising objective (Xin et al., 2025; Bie et al., 2026); or autoregressive text modeling can be combined with continuous diffusion or flow matching for images (Zhou et al., 2025; Deng et al., 2025). We adopt the last formulation, which is currently a widely used and empirically strong choice for jointly supporting language understanding and high-quality image generation. Specifically, the multimodal context is modeled causally, while the visual tokens within each generated image are modeled bidirectionally with flow matching.

Native visual interface. Our use of *native* in this study refers specifically to how visual information enters and leaves the model, rather than to multimodal pretraining from scratch. The pixel-in, pixel-out design avoids importing visual representations learned by external models, which could otherwise mask whether changes in visual features arise from understanding–generation joint training or from the pretrained visual prior itself.

2.2 TWO CONTROLLED ROUTING SETTINGS

We start from the pretrained Qwen3-1.7B language model (Yang et al., 2025a) and augment it with a stack of randomly initialized pre-buffer layers, following the native multimodal design of Diao et al. (2026a). Image inputs are first converted into patch tokens by two convolutional layers with a patch size of 32 and then passed to the LLM decoder layers. Attention is causal across the overall multimodal sequence, while tokens within the same image attend bidirectionally to one another. MRoPE-I (Huang et al., 2026) is used to encode multimodal positions while preserving the 2D spatial structure of images and the positional priors of the pretrained LLM. To isolate how access to pretrained language parameters affects the interaction between understanding and generation, we compare the two routing settings illustrated in Figure 2(a–b).

Dense sharing. The most direct design sends every token type—text, clean visual tokens for understanding, and noised visual tokens for generation—through the same LLM layers. This setting maximizes parameter sharing and gives both types of visual tokens direct access to pretrained language semantics. We train three matched variants: UND uses only understanding data, GEN uses only generation data, and UND+GEN uses both.

Modality-decoupled MoT. To isolate visual representation learning from the pretrained language backbone, we construct a modality-decoupled variant following the token-routed principle of MoT (Liang et al., 2024). The pretrained branch processes text only, while a parallel branch initialized from scratch processes all visual tokens, including both understanding tokens (UND-V) and noised generation tokens (GEN-V). The two branches retain global attention over the interleaved sequence.

Table 1: Understanding and generation results across routing architectures. Understanding columns report category averages. Atom-level score is reported for GenEval2. Higher is better for all metrics.

ID	Model		Understanding			Generation			
	Architecture	Training	General	OCR	V-Centric & SI	GenEval2	DPG	HPSv3	Aesthetic
1	Dense	UND/GEN	67.82	71.96	60.72	51.16	79.11	7.05	5.12
2		UND+GEN	68.69	72.12	61.77	51.03	78.12	5.92	4.94
3	Modality-dec. MOT	UND/GEN	64.20	66.84	59.54	57.55	82.06	7.72	5.37
4		UND+GEN	61.58	62.11	57.49	63.47	83.09	7.97	5.48
5	Task-dec. MOT	UND+GEN	69.02	72.09	62.38	63.96	82.31	7.90	5.50

3 REPRESENTATION-LEVEL SYNERGY

We first study how joint training reshapes the foundational visual representations. Specifically, we investigate whether parameter sharing intrinsically fosters representational synergy or induces feature interference, and how routing choices affect this dynamic.

3.1 TRAINING SETTINGS AND EVALUATION

Training protocol. We mainly select understanding and generation data from SenseNova-U1 (Diao et al., 2026c) for this study. The understanding data covers diverse types of multimodal instruction tuning data. The generation stream contains text-to-image generation and image editing data. Detailed training data statistics are reported in Section B.1. All variants use single-stage training with 210k steps and a constant learning rate of 1×10^{-4}. More training details are provided in Section B.3.

Evaluation. We group understanding benchmarks into three categories. **General** contains MME (Fu et al., 2023), MMBench (Liu et al., 2024a), MMStar (Chen et al., 2024b), SEED Bench (Li et al., 2023), and MMMU (Yue et al., 2024). **OCR** contains DocVQA (Mathew et al., 2021), ChartQA (Masry et al., 2022), InfoVQA (Mathew et al., 2022), OCRBench (Liu et al., 2024b), and AI2D (Kembhavi et al., 2016). **Vision-centric & SI (Spatial Intelligence)** contains PerceptionBench (Lin et al., 2026b), P2GB (Chen et al., 2024a), BLINK (Fu et al., 2024), MME-Realworld (Zhang et al., 2025c), DA2K (Yang et al., 2024), CV-Bench (Tong et al., 2024), MindCube (Yin et al., 2025), 3DSR (Ma et al., 2025a), and ViewSpatial (Li et al., 2026). Generation is evaluated on GenEval2 (Kamath et al., 2025), DPGBench (Hu et al., 2024), HPSv3 (Ma et al., 2025b), and Aesthetic Score (discus0434, 2024). Full per-benchmark results and evaluation details are reported in Section B.4.

3.2 RESULTS OF DENSE SHARING

As shown in Table 1 (Rows 1–2), for visual understanding, joint training achieves better performance compared with the UND baseline, with the largest gain on Vision-centric & SI. The effect is nevertheless strongly asymmetric: compared with the GEN model, the UND+GEN model performs worse across the generation metrics.

To understand how joint training changes the visual representations for visual understanding, we conduct a series of layer-wise probes on the frozen backbones. We measure global semantics with ImageNet (Deng et al., 2009) classification and dense features with semantic segmentation on ADE20K (Zhou et al., 2017) and monocular depth estimation on NYU-Depth V2 (Silberman et al., 2012). We additionally visualize PCA projections of the visual features for qualitative comparison. From the probing results (Figure 3) and feature visualization (Figure 4), we observe that the generation supervision improves the visual representations learned for visual understanding.

Recent work shows that image and video generation training can learn strong, general-purpose visual representations, supporting zero-shot or data-efficient transfer to a broad range of visual understanding tasks (Gabeur et al., 2026; Wang et al., 2026b). Our probing results are consistent with this broader observation: generation supervision enriches the visual features for visual understanding. The dense setting, however, reveals that better representations for understanding do not automatically yield balanced understanding and generation within a unified model. We hypothesize that the pretrained LLM acts as a strong semantic anchor, allowing understanding to dominate the shared feature space.

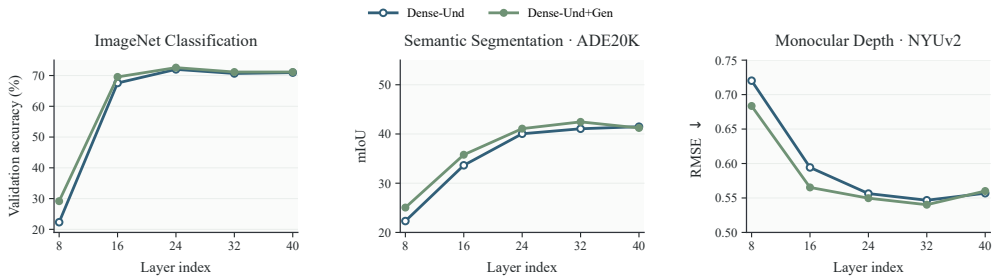


Figure 3: Layer-wise probing experiments of visual representations. Joint training strengthens visual representations, with the clearest gains in early and middle layers.

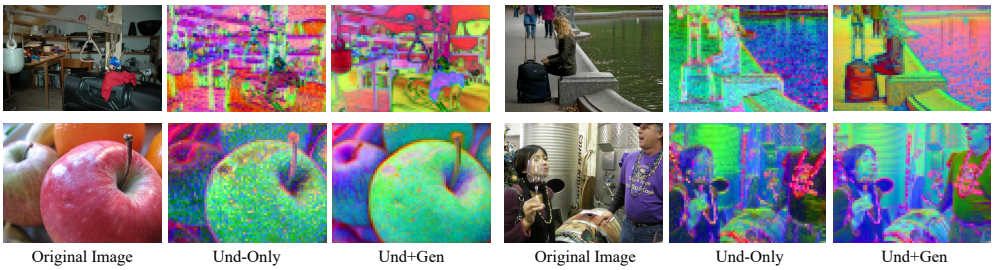


Figure 4: **PCA visualization of dense visual features.** Compared with understanding-only training, joint training produces more coherent object regions and clearer spatial structure.

Generation consequently serves as a useful auxiliary constraint for representation learning, but its own capability is degraded by competition within the shared parameters.

3.3 Results of Modality-decoupled MoT

As shown in Table 1 (Rows 3–4), when all visual tokens move to the scratch-trained visual-only branch, Und+Gen is worse than Und on visual understanding benchmarks, while generation performance becomes significantly better than Gen. In contrast to the dense architecture, generation dominates the shared visual computation in this setting, while understanding primarily serves as auxiliary supervision. Moreover, the modality-decoupled Und model is below its dense counterpart, particularly on General and OCR, indicating that, for tasks that rely on strong vision–text alignment, processing visual and textual tokens with shared parameters is more effective than separating them into modality-specific branches.

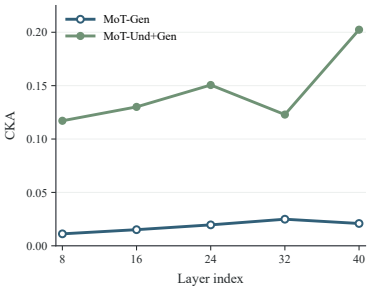


Figure 5: Joint learning improves vision–language alignment for Gen.

We further examine whether understanding supervision changes the generative visual representations. We measure layer-wise linear CKA (Kornblith et al., 2019) between pooled text features and generative image features from the visual branch at $t = 0.7$, where $t = 0$ denotes pure noise, using 5,000 text–image pairs from the COCO validation set (Lin et al., 2014). As shown in Figure 5, joint training produces consistently higher text-image CKA than generation-only training. This result indicates that understanding supervision improves generation by better aligning the generation representations with language semantics.

3.4 A Third Design: Task-decoupled MoT

The opposite failure modes motivate a third design: the task-decoupled MoT illustrated in Figure 2(c). It routes text and clean image tokens (for understanding or serving as context) through the pretrained LLM branch, while the noised image tokens (for generation) use a generation-specialized

visual branch. Thus, understanding retains direct access to pretrained language semantics, whereas generation receives parameters specialized for the generation objective.

The visual representations in the two branches remain softly aligned in two ways. First, they attend to the same textual features, providing a shared semantic context. Second, with image editing and interleaved samples, the context image in the understanding branch acts as a condition for the generation branch via shared attention at each layer. To further strengthen the connection between them, we allocate an additional 20% of the training mixture to an image reconstruction task formulated as image editing. The target images are drawn from the text-to-image data, while 50% of the visual tokens in each corresponding context image are randomly masked. The ablation in Section B.5 shows that this objective has little effect on Dense-Joint, while in task-decoupled MoT it mainly improves the General and V-Centric & SI understanding categories.

As shown in Table 1 (Row 5), the task-decoupled MoT avoids the opposite trade-offs observed in the dense and modality-decoupled models. This result suggests that the interference is not intrinsic to unification. Instead, it can be mitigated by specializing the visual computation required by the two objectives while retaining pathways for semantic interaction. We do not claim that task-decoupled routing is the unique optimal solution for unified modeling; other conditional architectures, such as mixtures-of-experts (MoE) (Shazeer et al., 2017) with task- or token-dependent routing, may also achieve a balance between specialization and sharing, which we leave for future work.

Finding 1 (Representation Level): Visual understanding and generation can benefit each other, but mutual gains depend on the architecture. Selectively specializing conflicting visual pathways while preserving semantic interaction enables both capabilities to improve without either dominating the other.

We therefore use the task-decoupled MoT as the default base model for the following studies.

4 Task-Level Synergy

Having obtained a model that avoids conflict between visual understanding and generation, we next move from representations to tasks and ask: when these two learning objectives rely on the same domain knowledge or underlying capability, can supervision in one direction improve the other? We study this question through three cases.

Across all three cases, each single-task model samples general-purpose and task-specific data at a 1:1 ratio, whereas the joint model samples general, task-specific understanding, and task-specific generation data at a 1:1:1 ratio. We keep the batch size of each corresponding data stream identical between the joint and single-task models. Domain-related examples are removed from the general-purpose data to prevent overlap. Further details are provided in Section C.1.

4.1 Case I: Geometry Problem Solving

Our first case asks whether the generation objective can improve geometry reasoning. Geometry problem solving requires the model to identify entities, relations, and constraints in geometric diagrams, while generating or editing a geometric figure requires it to construct visual content that satisfies the same structure. The two objectives therefore partially share the same domain knowledge.

We construct a geometry-related understanding data mixture which contains 965K geometry problem-solving data and general mathematical-reasoning data selected from public datasets (Lin et al., 2026a; Lu et al., 2021; Zhang et al., 2025b; Wiedmann et al., 2025; Li et al., 2024; Wang et al., 2025; Zhang et al., 2026). The generation part consists of two generation datasets: text-to-image examples from MathCanvas-Imagen (Shi et al., 2026) and geometry-diagram editing examples from MathCanvas-Edit (Shi et al., 2026). Our primary comparison is between understanding-only training (Und) and joint training (Und+Gen) to see whether the generation objective can effectively improve the geometry understanding and reasoning abilities.

We then ask a more specific question: with the same data, is the generation objective a more effective way than the corresponding understanding objective in this case? To test this, we convert exactly the

same generation data into an understanding format. Text-to-image pairs become image-description tasks, and editing pairs become tasks that describe the transformation between the source and target images. This Textual-Desc baseline controls the source data while changing only how the model learns from it: through text prediction instead of image generation.

Table 2: **Geometry problem solving and general mathematical reasoning.** UND+GEN adds the geometric image-generation objectives. UND + Textual-Desc uses the same examples after converting them into image-description and edit-description tasks. Higher is better for all metrics.

Training	Geometry		General Math	
	Geometry3K	PGPS9K	$\text{MathVerse}_{\text{testmini}}$	$\text{MathVista}_{\text{testmini}}$
UND	59.90	72.40	60.13	71.80
UND + Textual-Desc	63.56	73.90	59.11	73.60
UND+GEN	65.39	73.40	63.15	74.30

We evaluate the models on geometry problem solving benchmarks including Geometry3K (Lu et al., 2021) and PGPS9K (Zhang et al., 2023), and general math benchmarks including MathVerse (Zhang et al., 2024) and MathVista (Lu et al., 2024). As shown in Table 2, adding the generation objective improves the understanding model on all four benchmarks. The gains also transfer beyond the directly paired geometry tasks into general math reasoning. Converting the same additional examples into understanding tasks produces a smaller and less consistent benefit. This indicates that learning to generate and edit the diagrams is a more effective form of auxiliary supervision for geometry and general mathematical reasoning than learning text-output descriptions.

4.2 CASE II: SVG UNDERSTANDING AND GENERATION

Scalable Vector Graphics (SVG) provides a naturally bidirectional domain for studying task-level transfer: the same icon can be represented either as a rendered image or as an executable vector program. We ask whether learning both mappings jointly improves the model's correspondence between symbolic SVG structure and visual appearance.

Task setup. In the understanding direction, the model receives a rendered icon image and outputs its SVG code. In the generation direction, the model receives SVG code and directly renders the corresponding image. We use the 359K SVG code-image pairs from UniSVG (Li et al., 2025) to construct the corresponding understanding and generation training data. For evaluation, we use the test data of UniSVG and adopt its metric to evaluate the directly generated or code-rendered icon, which is a weighted sum of SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), and CLIP similarity (Radford et al., 2021).

Bidirectional transfer. We compare understanding-only, generation-only, and joint training in Table 3. Joint training improves both SVG image understanding and SVG-conditioned image generation over their corresponding single-direction models. This mutual improvement suggests that the two tasks share more than generic image–text semantics: supervision in either direction helps the model learn the executable relationship between vector operations and their visual consequences.

Table 3: SVG understanding and generation results. Higher is better.

Training	Image→SVG	SVG→Image
UND/GEN	80.64	80.21
UND+GEN	81.70	86.52

Diagnostic: mental visualization from SVG code. We test whether joint training helps the model infer visual appearance directly from SVG code using two diagnostics: code-conditioned blind VQA and the predicted clean image after the first generation step. The UND+GEN model performs better on the former and forms recognizable contours and layouts earlier on the latter, suggesting a stronger internal simulation of SVG execution. Experimental details and visualizations are provided in Section C.2.

Table 4: Summary of 3D SI understanding and generation results. The understanding score is averaged over nine benchmarks. Full results for understanding tasks are reported in Table 11.

Training	Understanding	Spatial-Edit		View Sequence Completion		
	Avg. ↑	VE ↓	FE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Single-task	57.15	0.4062	0.6564	16.76	0.6365	0.3828
UND+GEN	59.01	0.3766	0.6297	19.15	0.6674	0.3078

4.3 CASE III: 3D SPATIAL INTELLIGENCE

For our third case study, we move beyond the specialized domains of geometry problems and SVGs to 3D spatial intelligence, a broader and more widely required capability that is fundamental to many applications like embodied agents and interactive world modeling. We examine whether understanding and generation can improve each other on general 3D spatial intelligence tasks.

Task setup. The understanding training mixture for 3D SI contains single-image, multi-image, and video-based 3D VQA, covering spatial relations, viewpoint reasoning, depth, and camera motion. For generation, we train on four complementary tasks—ego-motion transition, multi-view reconstruction, view sequence completion, and layout-to-image generation—covering camera transformation, novel-view synthesis, trajectory completion, and layout-conditioned rendering. Full definitions, data sources, sample counts, and examples are provided in Section C.3.

We evaluate 3D spatial understanding on nine benchmarks including VSI-Bench (Yang et al., 2025b), SPAR-Bench (Zhang et al., 2025a), CV-Bench (Tong et al., 2024), DA-2K (Yang et al., 2024), MindCube (Yin et al., 2025), ViewSpatial (Li et al., 2026), 3DSRBench (Ma et al., 2025a), MMSI (Yang et al., 2026), All-Angles (Yeh et al., 2026). For generation, we evaluate ego-motion transition on the SpatialEdit benchmark using Viewpoint Error (VE) and Framing Error (FE). We also construct a test set of the VSC task and evaluate using PSNR, SSIM, and LPIPS as metrics.

Bidirectional transfer. As shown in Table 4, joint training improves the averaged understanding score from 57.15 to 59.01. It also reduces both Spatial-Edit errors and improves all three VSC metrics, showing that understanding and generation benefit each other when their underlying 3D knowledge overlaps.

Finding 2 (Task Level): Shared underlying knowledge or capabilities make understanding and generation tasks mutually reinforcing.

This finding motivates exploring unified learning in broader settings where multiple capabilities depend on shared latent knowledge or structure. One direct example is world modeling and policy/action learning, where understanding the current state, predicting future observations, and generating actions all rely on a common model of the physical world and its dynamics.

5 SYSTEM-LEVEL SYNERGY

We next consider tasks in which understanding and generation must interact sequentially. Using reasoning-intensive image editing, we compare an end-to-end UMM with a planner–executor pipeline under matched conditions. Given a source image and an implicit request, both systems must understand the source, infer the concrete edit, and generate the target image.

5.1 CONTROLLED COMPARISON SETUP

To compare end-to-end unification and modular composition under controlled conditions, we initialize all models from the same task-decoupled MOT checkpoint and transform the same set of reasoning-editing data, 263K samples collected from UniReason (Wang et al., 2026a) and SenseNova-U1, into three training formats: (1) **end-to-end**, which reasons, produces an explicit edit instruction, and generates the target; (2) **planning**, which outputs only the explicit instruction; and (3) **execution**, which generates the target conditioned on that instruction. At inference, the pipeline composes (2)

Table 5: Reasoning-intensive image editing results. Higher is better.

System	RISEBench					KRIS-Bench			
	Temporal	Causal	Spatial	Logical	**Overall**	Factual	Conceptual	Procedural	**Overall**
Planner → Executor	18.82	22.22	14.00	11.76	**16.66**	61.40	71.17	64.58	**66.48**
End-to-end UMM	22.35	27.77	16.00	9.41	**18.88**	64.73	72.32	66.16	**68.33**

and (3), while (1) completes the process directly. Each mixture additionally contains the same 50% general-purpose data used in representation-level training.

We evaluate both systems on RISEBench (Zhao et al., 2025) and KRIS-Bench (Wu et al., 2025c). As shown in Table 5, the end-to-end UMM outperforms the planner–executor pipeline on both benchmarks. Under the matched base model and task data, this result shows that end-to-end unification can provide a performance advantage when visual understanding, reasoning, and generation must interact closely within a task.

Finding 3 (System Level): End-to-end unification can outperform modular composition for complex tasks that explicitly require both visual understanding and generation.

A future direction is to internalize broader agentic processes, such as multimodal search and iterative visual refinement, within UMMs.

6 Related Work

Unified multimodal models. Unified multimodal models (UMMs) integrate visual understanding and generation within a single model (Chameleon Team, 2024; Xie et al., 2025; Wu et al., 2025a; Zhou et al., 2025; Deng et al., 2025; Liu et al., 2025; Qiu et al., 2026; Diao et al., 2026c), enabling interleaved multimodal generation and visual reasoning through generation. Such functional unification, however, does not establish whether the underlying capabilities reinforce each other or merely coexist. We study when joint learning produces synergy and when sharing causes interference.

Native multimodal models. The term *native* can refer to training on multimodal data from the start (Kimi Team et al., 2026; Tong et al., 2026; Thinking Machines Lab, 2026; Han et al., 2026), or structurally processing visual inputs and outputs directly in the backbone without external encoders, tokenizers, or VAEs (Bavishi et al., 2023; Thinking Machines Lab, 2026; Diao et al., 2026b;c). We use the latter definition: our pixel-in, pixel-out setting removes external visual priors and isolates understanding–generation interaction. We initialize from a pretrained LLM rather than training entirely from scratch, retaining the language and reasoning capabilities needed for controlled analysis.

Understanding and generation: conflict or synergy? Existing work provides no consensus on whether understanding and generation benefit each other. Some studies report interference and motivate decoupled representations or pathways (Wu et al., 2025a; Pan et al., 2025). Some works show conditional mutual benefits on carefully designed tasks (Kang et al., 2026; Han et al., 2026). Tuna (Liu et al., 2025) and Hydra (Qiu et al., 2026) study the synergy with two sequentially connected pre-trained vision encoders. These suggest that synergy is possible but conditional. Because prior effects are entangled with specialized modules, auxiliary losses, or changed visual components, we study the relationship in a controlled native setting across representation, task, and system levels.

7 Conclusion

We studied understanding–generation synergy in native UMMs across representations, tasks, and systems. Mutual benefit is architecture-dependent: task-aware specialization avoids asymmetric degradation, while joint training improves both directions when tasks share knowledge. For tightly coupled understanding, reasoning, and generation, an end-to-end UMM outperforms a matched modular pipeline. Thus, UMMs offer value beyond a unified interface through cross-capability knowledge transfer and end-to-end optimization.

8 AI USE STATEMENT

GPT-5.6-luna was used to annotate 3,000 multiple-choice examples during the construction of our SVG Blind VQA diagnostic; the subsequent filtering process is described in Section C.2. Generative AI tools were also used to assist with language polishing and LaTeX formatting. All scientific claims, experimental results, citations, and AI-assisted content were reviewed and verified by the authors, who take full responsibility for the final manuscript.

9 ETHICS STATEMENT

This work does not involve human-subject experiments. We use existing research datasets and benchmarks in accordance with their intended research use. As with other multimodal models, the studied systems may inherit biases from their training data or be misused to generate misleading content; responsible deployment should therefore include appropriate safeguards and human oversight.

10 REPRODUCIBILITY STATEMENT

We report the model architectures, data mixtures, training configurations, evaluation protocols, and metrics in the main text and appendix. The appendix also provides full benchmark results and additional task-construction details to facilitate reproduction and comparison.

REFERENCES

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *NeurIPS Datasets and Benchmarks Track*, 2021.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. URL `https://www.adept.ai/blog/fuyu-8b`.

Tiwei Bie, Haoxing Chen, Tieyuan Chen, Zhenglin Cheng, Long Cui, Kai Gan, Zhicheng Huang, Zhenzhong Lan, Haoquan Li, et al. Llada2. 0-uni: Unifying multimodal understanding and generation with diffusion large language model. *arXiv preprint arXiv:2604.20796*, 2026.

Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3d: A large benchmark and model for 3d object detection in the wild. In *CVPR*, 2023.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Jiaxing Chen, Yuxuan Liu, Dehu Li, Xiang An, Weimo Deng, Ziyong Feng, Yongle Zhao, and Yin Xie. Plug-and-play grounding of reasoning in multimodal large language models. *arXiv preprint arXiv:2403.19322*, 2024a.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? In *NeurIPS*, 2024b.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025b.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, et al. Emu3. 5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. *arXiv preprint arXiv:2212.08051*, 2022.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Haiwen Diao, Mingxuan Li, Silei Wu, Linjun Dai, Xiaohua Wang, Hanming Deng, Lewei Lu, Dahua Lin, and Ziwei Liu. From pixels to words–towards native vision-language primitives at scale. In *ICLR*, 2026a.

Haiwen Diao, Jiahao Wang, Penghao Wu, Yuhao Dong, Yuwei Niu, Yue Zhu, Zhongang Cai, Weichen Fan, Linjun Dai, Silei Wu, et al. From pixels to words–towards native one-vision models at scale. *arXiv preprint arXiv:2605.28820*, 2026b.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, et al. Sensenova-u1: Unifying multimodal understanding and generation with neo-unify architecture. *arXiv preprint arXiv:2605.12500*, 2026c.

discus0434. Aesthetic predictor v2.5. `https://github.com/discus0434/aesthetic-predictor-v2-5`, 2024. SigLIP-based Aesthetic Score Predictor.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *ECCV*, 2024.

Valentin Gabeur, Shangbang Long, Songyou Peng, Paul Voigtlaender, Shuyang Sun, Yanan Bao, Karen Truong, Zhicheng Wang, Wenlei Zhou, Jonathan T Barron, et al. Image generators are generalist vision learners. *arXiv preprint arXiv:2604.20329*, 2026.

Junlin Han, Shengbang Tong, David Fan, Minghao Chen, Philip Torr, Filippos Kokkinos, and Mike Lewis. Towards physics of multimodal pretraining: Knowledge flow, modality synergy, early unification, and recipes. *arXiv preprint arXiv:2608.05000*, 2026.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Jie Huang, Xuejing Liu, Sibo Song, Ruibing Hou, Hong Chang, Junyang Lin, and Shuai Bai. Revisiting multimodal positional encoding in vision–language models. In *ICLR*, 2026.

Ziyuan Huang, DanDan Zheng, Cheng Zou, Rui Liu, Xiaolong Wang, Kaixiang Ji, Weilong Chai, Jianxin Sun, Libin Wang, Yongjie Lv, et al. Ming-univision: Joint image understanding and generation with a unified continuous tokenizer. *arXiv preprint arXiv:2510.06590*, 2025.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. Geneval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025.

Jiwon Kang, Heeji Yoon, Jaewoo Jung, Jaewon Min, Minkyeong Jeon, Biyeon Hwang, Sangwon Jung, and Seungryong Kim. Transferability between understanding and generation in unified multimodal models. *arXiv preprint arXiv:2607.04423*, 2026.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *ICML*, 2019.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, et al. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. In *ECCV*, 2026.

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath. `https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf`, 2024.

Jinke Li, Jiarui Yu, Chenxing Wei, Hande Dong, Qiang Lin, Liangjing Yang, Zhicai Wang, and Yanbin Hao. Unisvg: A unified dataset for vector graphic understanding and generation with multimodal large language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025.

Weixin Liang, Lili Yu, Liang Luo, Srini Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, et al. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *Transactions on Machine Learning Research*, 2024.

Honglin Lin, Zheng Liu, Yun Zhu, Chonghan Qin, Juekai Lin, Xiaoran Shang, Conghui He, Wentao Zhang, and Lijun Wu. Mmfinereason: Closing the multimodal reasoning gap via open data-centric methods. *arXiv preprint arXiv:2601.21821*, 2026a.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Zichao Lin, Yifeng Xie, Bowen Qu, Haiming Wang, Jia Li, Haoning Wu, Yuhao Dong, Zuhao Yang, Jinguo Zhu, Haoyu Lu, et al. Perceptionbench: Evaluating atomic visual perception in multimodal large language models. *arXiv preprint arXiv:2607.24957*, 2026b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, 2024a.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 2024b.

Zhiheng Liu, Weiming Ren, Haozhe Liu, Zijian Zhou, Shoufa Chen, Haonan Qiu, Xiaoke Huang, Zhaochong An, Fanny Yang, Aditya Patel, et al. Tuna: Taming unified visual representations for native unified multimodal models. *arXiv preprint arXiv:2512.02014*, 2025.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *ACL*, 2021.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *ICCV*, 2025a.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *ICCV*, 2025b.

Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *ACL Findings*, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, 2022.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition, 2026. URL `https://openai.com/index/gpt-5-6/`. Large language model.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

Xuerui Qiu, Yutao Cui, Guozhen Zhang, Junzhe Li, JiaKui Hu, Xiao Zhang, Yang Li, Songtao Liu, Miles Yang, Yu Shi, et al. Hydra: Unifying multi-modal generation and understanding via representation-harmonized tokenization. *arXiv preprint arXiv:2603.15228*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Weikang Shi, Aldrich Yu, Rongyao Fang, Houxing Ren, Ke Wang, Aojun Zhou, Changyao Tian, Xinyu Fu, Yuxuan Hu, Zimu Lu, et al. Mathcanvas: Intrinsic visual chain-of-thought for multimodal mathematical reasoning. In *ACL*, 2026.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

Thinking Machines Lab. Interaction models: A scalable approach to human-ai collaboration. *Thinking Machines Lab: Connectionism*, May 2026. doi: 10.64434/tml.20260511. https://thinkingmachines.ai/blog/interaction-models/.

Shengbang Tong, Ellis L Brown II, Penghao Wu, Sanghyun Woo, Adithya Jairam Iyer, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*, 2024.

Shengbang Tong, David Fan, John Nguyen, Ellis Brown, Gaoyue Zhou, Shengyi Qian, Boyang Zheng, Théophane Vallaeys, Junlin Han, Rob Fergus, et al. Beyond language modeling: An exploration of multimodal pretraining. *arXiv preprint arXiv:2603.03276*, 2026.

Dianyi Wang, Chaofan Ma, Feng Han, Size Wu, Wei Song, Yibin Wang, Zhixiong Zhang, Tianhang Wang, Siyuan Wang, Zhongyu Wei, et al. Unireason 1.0: A unified reasoning framework for world knowledge aligned image generation and editing. *arXiv preprint arXiv:2602.02437*, 2026a.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025.

Letian Wang, Chuhan Zhang, Rishabh Kabra, Jasper Uijlings, Steven Waslander, Andrew Zisserman, Joao Carreira, Kaiming He, Misha Andriluka, Eduard Gabriel Bazavan, et al. Video generation models are general-purpose vision learners. *arXiv preprint arXiv:2607.09024*, 2026b.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004.

Luis Wiedmann, Orr Zohar, Amir Mahla, Xiaohan Wang, Rui Li, Thibaud Frere, Leandro von Werra, Aritra Roy Gosthipaty, and Andrés Marafioti. Finevision: Open data is all you need. *arXiv preprint arXiv:2510.17269*, 2025.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *CVPR*, 2025a.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation. In *ICCV*, 2025b.

Yongliang Wu, Zonghui Li, Xinting Hu, Xinyu Ye, Xianfang Zeng, Gang Yu, Wenbo Zhu, Bernt Schiele, Ming-Hsuan Yang, and Xu Yang. Kris-bench: Benchmarking next-level intelligent image editing models. *NeurIPS*, 2025c.

Yicheng Xiao, Wenhu Zhang, Lin Song, Yukang Chen, Wenbo Li, Nan Jiang, Tianhe Ren, Haokun Lin, Wei Huang, Haoyang Huang, et al. Spatialedit: Benchmarking fine-grained image spatial editing. *arXiv preprint arXiv:2604.04911*, 2026.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *ICLR*, 2025.

Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, et al. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*, 2025b.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. Mmsi-bench: A benchmark for multi-image spatial intelligence. In *ICLR*, 2026.

Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Ruoyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in mllms. In *AAAI*, 2026.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, 2023.

Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at ICCV'25*, 2025.

Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024.

Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Jilin Mei, Chunhui Chen, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3d. In *NeurIPS*, 2025a.

Ming-Liang Zhang, Fei Yin, and Cheng-Lin Liu. A multi-modal neural geometric solver with textual clauses parsed from diagram. In *IJCAI*, 2023.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Shanghang Zhang, Gao Peng, et al. Mavis: Mathematical visual instruction tuning with an automatic data engine. In *ICLR*, 2025b.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Yi Zhang, Bolin Ni, Xin-Sheng Chen, Hengrui Zhang, Yongming Rao, Houwen Peng, Qinglin Lu, Winston Hu, Meng-Hao Guo, and Shi-Min Hu. Bee: A high-quality corpus and full-stack suite to unlock advanced fully open mllms. In *ICLR*, 2026.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? In *ICLR*, 2025c.

Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Xiaorong Zhu, Hao Li, Wenhao Chai, Zicheng Zhang, Renqiu Xia, Guangtao Zhai, Junchi Yan, et al. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. *NeurIPS*, 2025.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *ICLR*, 2025.

APPENDIX

A LIMITATIONS

Our study focuses on the prevalent formulation combining discrete autoregressive text modeling with continuous diffusion-based image modeling. Although our conclusions may transfer to fully discrete or fully autoregressive multimodal formulations, this remains to be empirically validated. In addition, our experiments primarily use a task-decoupled architecture to demonstrate mutual benefits between understanding and generation. We do not extensively explore the broader design space; identifying the optimal balance between sharing and specialization remains an important direction for future work.

B REPRESENTATION-LEVEL DETAILS

B.1 DATA

We draw the training data from SenseNova-U1 using a tailored sampling mixture. The category distribution of the visual-understanding data is summarized in Figure 6. For visual generation, we exclude infographic data and downweight samples that emphasize text rendering. We only include generation samples with English prompts. For joint training, visual-understanding, text-to-image, and image-editing data are sampled with a relative ratio of $3 : 6 : 1$. The task-decoupled MoT additionally includes two parts of reconstruction data. For fair comparison, we keep the batch sizes for understanding and generation data identical between the joint models and their corresponding UND-only and GEN-only counterparts. Training for 210K steps exposes each model to approximately 78M understanding samples, 159M text-to-image samples, and 26.5M image-editing samples. For the 3D spatial-intelligence case study, we retrain the base model after removing all SI-related examples, preventing data overlap and isolating transfer from the task-specific data introduced in that study.

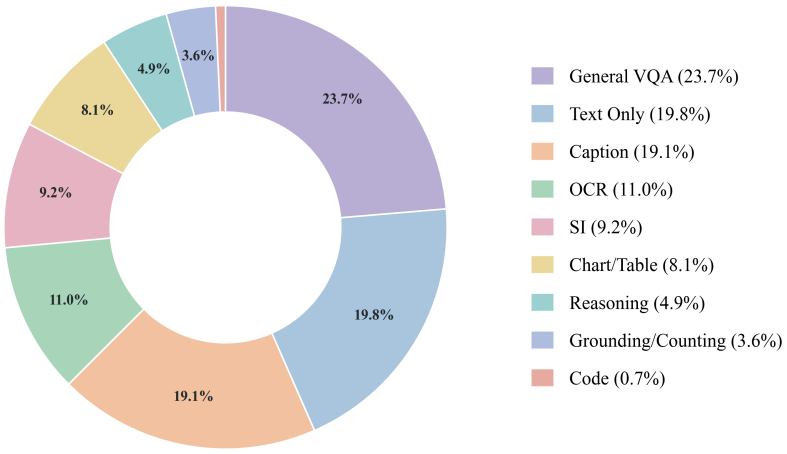


Figure 6: Category distribution of the understanding training data.

B.2 MODEL

All models are initialized from Qwen3-1.7B with a 16-layer pre-buffer. We use an MLP-based timestep embedder and noise scale embedder to extract embeddings and directly add them to the visual inputs, similar to SenseNova-U1. A two layer MLP is used as the image generation head. Different from SenseNova-U1, we do not have additional norm layers for visual hidden states as we choose a different multimodal positional embedding scheme.

B.3 TRAINING

Table 6: Representation-level training configuration.

Item	Value
Peak learning rate	1×10^{-4}
Learning-rate scheduler	Constant
Optimizer	AdamW (β_1 = 0.9, β_2 = 0.95, $\epsilon = 10^{-8}$)
Weight decay	0.0
Gradient-norm clipping	1.0
EMA ratio	0.9999
Training steps	210K
Warmup steps	2000
Loss weight (CE : MSE)	0.1:1
Understanding resolution	$[256^2, 2048^2]$
Generation resolution	$[512^2, 1024^2]$
Sequence length	16K
Timestep shift	$\mu = -0.8, \sigma = 0.8$ in the logit-normal t-sampler
Noise scale	$\sqrt{N/64}$, where N is # of tokens

Training details in our representation-level studies are provided in Table 6.

B.4 EVALUATION

Table 7: Full results for visual understanding benchmarks. Higher is better for all metrics.

Group	Benchmark	Dense-U	Dense-U+G	Mod.-dec. MoT U	Mod.-dec. MoT U+G	Task-dec. MoT U+G
General	MME	1930.46	1967.90	1773.31	1666.54	1992.06
	MMBench	71.39	69.58	64.43	66.66	69.84
	MMStar	52.20	55.82	50.72	47.69	56.51
	SEED Bench-I	76.56	76.91	74.33	71.91	76.42
	MMMU	42.44	42.77	42.89	38.33	42.77
OCR	DocVQA	94.37	93.13	90.05	85.29	94.63
	ChartQA	45.40	46.84	44.68	40.20	45.24
	InfoVQA	68.75	67.52	52.65	44.59	68.30
	OCRBench	67.50	68.90	63.90	59.90	67.80
	AI2D	83.80	84.20	82.90	80.60	84.50
V-Centric & SI	$\text{PerceptionBench}_{\text{mcq}}$	31.30	32.59	29.66	29.66	34.02
	P2GB	72.63	73.94	66.40	63.40	72.07
	MME-RealWorld	48.93	49.81	46.63	41.01	50.39
	BLINK	63.92	64.09	60.07	59.02	63.48
	DA-2K	74.85	74.85	77.22	72.24	77.46
	CV-Bench	81.88	81.27	79.15	79.90	82.44
	MindCube	58.57	61.81	62.67	58.19	64.85
	3DSR	59.01	61.78	59.78	59.14	61.02
	ViewSpatial	55.41	55.86	54.32	54.90	55.75

The full per-benchmark results for visual understanding benchmarks are provided in Table 7. We use the image subset for SEED-Bench, the lite subset for MME-RealWorld, and the mini subset for MindCube. We only use the MCQ format questions in PerceptionBench for more stable evaluation. All benchmarks are evaluated with temperature set to 0. Detailed results by skill and level-1 category for GenEval2 and DPG-Bench are reported in Table 8 and Table 9, respectively.

For generation-related evaluation, we generate images at a resolution of 1024×1024 using a classifier-free guidance scale of 4.0, 50 sampling steps, and a timestep shift of 3.0.

B.5 MASKED-RECONSTRUCTION ABLATION

To separate the effect of task-decoupled routing from that of the auxiliary masked-image reconstruction objective, we add the same reconstruction data to the dense joint model and remove it from the task-decoupled MoT model. Table 10 reports the resulting visual understanding category averages.

Table 8: GenEval2 results by skill. We report atom-level scores; higher is better.

Model	Object	Attribute	Count	Position	Verb	**Overall**
Dense-G	67.91	56.08	34.34	33.35	9.02	51.16
Dense-U+G	67.08	53.97	34.61	33.67	7.45	51.03
Mod.-dec. MoT-G	77.65	58.69	39.76	42.81	13.75	57.55
Mod.-dec. MoT-U+G	84.60	63.26	43.95	53.05	9.28	63.47
Task-dec. MoT-U+G	82.97	64.45	45.60	56.11	15.77	63.96

Table 9: DPG-Bench results by level-1 category. Higher is better.

Model	Global	Entity	Attribute	Relation	Other	**Overall**
Dense-G	91.95	85.55	88.83	87.52	79.49	79.11
Dense-U+G	83.58	85.77	87.25	89.45	84.63	78.12
Mod.-dec. MoT-G	88.02	89.82	88.58	90.42	89.20	82.06
Mod.-dec. MoT-U+G	92.14	88.12	90.39	92.03	90.77	83.09
Task-dec. MoT-U+G	87.30	90.50	89.09	90.65	91.17	82.31

Table 10: Ablation of masked-image reconstruction for visual understanding.

Model	General	OCR	V-Centric & SI
Dense-Joint	68.69	72.12	61.77
Dense-Joint w/ Masked Rec.	68.46	72.19	62.15
Task-dec. MoT w/o Masked Rec.	68.21	72.41	61.76
Task-dec. MoT w/ Masked Rec.	69.02	72.09	62.38

Adding masked reconstruction to Dense-Joint yields performance similar to the original Dense-Joint model, indicating that this auxiliary objective alone does not account for the gains of task-decoupled routing. Within the task-decoupled MoT, masked reconstruction mainly improves the General and V-Centric & SI categories, with a smaller effect on OCR. Thus, masked reconstruction provides complementary benefits, while the architectural separation of conflicting visual computation remains the primary factor behind the balanced understanding–generation performance.

B.6 Probing Experiments

For ImageNet classification, we follow the standard linear-probing protocol. We freeze the multimodal backbone, globally pool the visual features, and train a linear classification head using the LARS (You et al., 2017) optimizer for 100 epochs.

For semantic segmentation and monocular depth estimation, we conduct dense probing with the backbone similarly frozen. Because the model produces visual features at a relatively large spatial downsampling factor of 32, the probing head contains three consecutive convolution-and-upsampling blocks to recover spatial resolution. We train the segmentation and depth heads for 30 epochs, optimizing only the parameters of the corresponding probing head.

C Task-level Details

C.1 Fine-tuning Configuration

Optimization and data mixing. All task-level models are fine-tuned for 20K steps using AdamW with a constant learning rate of 5×10^{-5}. Other training configurations follow the representation-level stage. For each single-task model, general-purpose and task-specific data are sampled at a 1:1 ratio. For each Und+Gen model, general-purpose, task-specific understanding, and task-specific generation data are sampled at a 1:1:1 ratio. The batch size of each data stream is kept identical to that in its corresponding Und-only or Gen-only model. Thus, all comparisons are step-matched and per-stream batch-matched.

C.2 SVG Diagnostics

We construct an SVG Blind VQA diagnostic to evaluate whether the model can infer visual appearance directly from SVG code. Starting from the UniSVG test set, we use GPT-5.6-luna (OpenAI, 2026) to annotate 3,000 multiple-choice VQA examples. To retain nontrivial questions, we evaluate them with our base model, which has not been trained on SVG-specific data, and keep only the examples it answers incorrectly. This filtering yields 1,370 examples. Examples are shown in Figure 8.

We complement blind VQA with a generation-side diagnostic that inspects the predicted clean image after only the first generation step, revealing how readily the model forms a global visual plan from

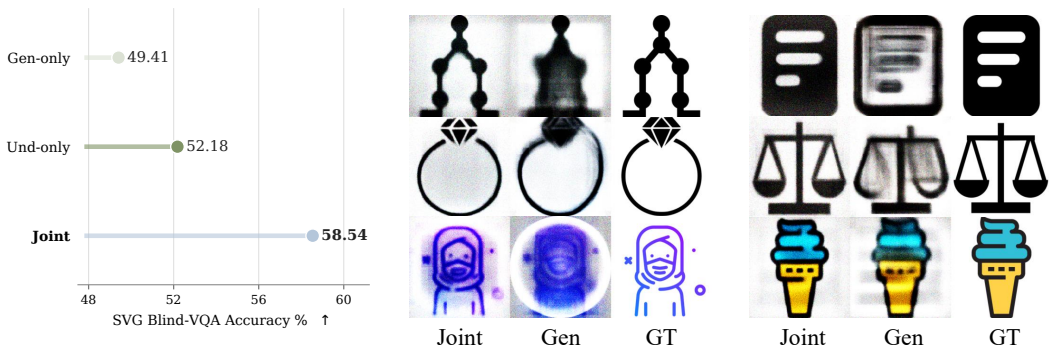


Figure 7: **Joint training improves direct visualization from SVG code.** Left: accuracy on code-conditioned blind VQA. Right: predicted clean images after the first generation step, showing the emergence of global shape and layout.

Table 11: Full understanding results on 3D SI benchmarks. Higher is better.

Model	VSI-Bench	SPAR-Bench	CV-Bench	DA-2K	MindCube-tiny	ViewSpatial	3DSRBench	MMSI	ALL-Angles	**Average**
UND only	53.98	38.80	77.67	70.69	75.71	58.75	59.84	36.40	42.50	**57.15**
UND+GEN	55.49	41.13	78.77	73.89	77.90	58.07	60.51	39.20	46.11	**59.01**

SVG code. As shown in Figure 7, the UND+GEN model achieves higher blind-VQA accuracy than the understanding-only model and recovers recognizable contours and spatial layouts earlier than the generation-only model. Together, the two diagnostics indicate that joint training strengthens an internal, visually grounded simulation of SVG execution.

C.3 GENERATION TASKS FOR 3D SI

We construct four generation tasks that require different aspects of 3D spatial intelligence:

- **Ego-motion Transition** predicts a target frame from a source frame and ego-motion descriptions. We use SpatialEdit-500K (Xiao et al., 2026) and construct an additional 1.6M samples from ScanNet (Dai et al., 2017) and ScanNet++ (Yeshwanth et al., 2023).
- **Multi-view Reconstruction** predicts a target object view from provided views. We create 248K samples from Objaverse (Deitke et al., 2022), considering the front, back, left, and right views and randomly selecting two context views and one target view.
- **View Sequence Completion (VSC)** predicts the middle frame from the first and last two frames of a five-frame sequence. We construct 880K samples from ScanNet (Dai et al., 2017), ScanNet++ (Yeshwanth et al., 2023), and ARKitScenes (Baruch et al., 2021).
- **Layout-to-image Generation** generates an image from either structured object sizes and camera-relative 3D positions or a natural-language description of the scene's 3D structure and spatial relations. We construct 332K samples in these two formats from Omni3D (Brazil et al., 2023).

Examples of the four task formats are shown in Figure 9 and Figure 10.

Mentally render the following SVG code and answer the visual question.

<svg_code>
<svg xmlns="http://www.w3.org/2000/svg" xml:space="preserve" viewBox="0 0 100 100"><path d="m88.553 43.441-10.276-1.013a29 29 0 0 0-2.936-7.059l6.561-7.993a2.16 2.16 0 0 0-.145-2.893l-6.24-6.243a2.153 2.153 0 0 0-2.894-.142L64.63 24.66a29 29 0 0 0-7.06-2.936l-1.011-10.278A2.16 2.16 0 0 0 54.412 9.5h-8.823a2.16 2.16 0 0 0-2.148 1.946L42.43 21.727a29 29 0 0 0-7.054 2.936l-7.998-6.564c-.4-.329-.884-.49-1.369-.49-.555 0-1.105.212-1.525.633l-6.24 6.241a2.16 2.16 0 0 0-.145 2.893l6.566 7.996a29 29 0 0 0-2.936 7.057l-10.281 1.013A2.154 2.154 0 0 0 9.5 45.587v8.825c0 1.11.843 2.04 1.947 2.146l10.281 1.013a28.9 28.9 0 0 0 2.936 7.054l-6.562 7.994a2.15 2.15 0 0 0 .14 2.895l6.24 6.242a2.155 2.155 0 0 0 2.894.142l7.993-6.562a29 29 0 0 0 7.06 2.938l1.011 10.279a2.16 2.16 0 0 0 2.148 1.946h8.823c1.11 0 2.039-.842 2.148-1.946l1.011-10.278a28.9 28.9 0 0 0 7.06-2.939l7.993 6.562a2.156 2.156 0 0 0 2.894-.142l6.24-6.242a2.15 2.15 0 0 0 .14-2.895l-6.557-7.991a29 29 0 0 0 2.936-7.057l10.276-1.013a2.157 2.157 0 0 0 1.947-2.146v-8.825a2.156 2.156 0 0 0-1.946-2.146M50.002 61.95c-6.603 0-11.953-5.351-11.953-11.95 0-6.602 5.351-11.951 11.953-11.951C56.6 38.049 61.95 43.398 61.95 50c.001 6.599-5.35 11.95-11.948 11.95"/></svg>
</svg_code>

Question: How many distinct projecting teeth surround the central opening?
Options:
A. Eight teeth
B. Six teeth
C. Twelve teeth
D. Ten teeth



Answer: A

Mentally render the following SVG code and answer the visual question.

<svg_code>
<svg xmlns="http://www.w3.org/2000/svg" viewBox="0 0 48 48"><defs><style>.a{fill:none;stroke:#000;stroke-linecap:round;stroke-linejoin:round}</style></defs><path d="M22.17 27.561c-2.116 1.063-3.63 3.916-3.63 7.268h10.92c0-3.352-1.513-6.204-3.628-7.267" class="a"/><ellipse cx="24" cy="25.666" class="a" rx="2.273" ry="3.21"/><path d="M11.333 30.624C9.856 31.38 9 32.277 9 33.24c0 2.697 6.716 4.883 15 4.883s15-2.186 15-4.883c0-.962-.856-1.86-2.333-2.616M21.745 25.268a5 5 0 0 0-.75-.47" class="a"/><path d="M17.332 24.797c-2.116 1.063-3.63 3.915-3.63 7.267H18.9" class="a"/><ellipse cx="19.163" cy="22.901" class="a" rx="2.273" ry="3.21"/><path d="M29.099 32.064h5.199c0-3.351-1.514-6.203-3.629-7.267M27.007 24.797a5 5 0 0 0-.752.471" class="a"/><ellipse cx="28.837" cy="22.901" class="a" rx="2.273" ry="3.21"/><path d="m26.87 6.041 2.865 2.868-2.867 2.867m-5.739 0L18.265 8.91l2.867-2.868M18.265 8.909h11.47M24 6.041v5.736" class="a"/><circle cx="24" cy="24" r="21.5" class="a"/><path d="M4.421 15.318h39.158" class="a"/></svg>
</svg_code>

Question: What does the symbol above the horizontal divider depict?
Options:
A. A circular arrow surrounding a dot
B. Two downward arrows joined at the center
C. A single upward-pointing arrow
D. A left-right double arrow crossed by a vertical bar



Answer: D

Mentally render the following SVG code and answer the visual question.

<svg_code>
<svg xmlns="http://www.w3.org/2000/svg" xml:space="preserve" viewBox="0 0 297 297"><path d="M276.449 41.159h10.667c5.458 0 9.884-4.425 9.884-9.884s-4.425-9.884-9.884-9.884h-18.285a9.88 9.88 0 0 0-9.557 7.366l-10.947 41.557H9.884C4.425 70.315 0 74.741 0 80.199v100.235c0 19.9 14.568 36.453 33.598 39.575-9.771 4.745-16.523 14.767-16.523 26.338 0 16.135 13.127 29.262 29.262 29.262s29.261-13.127 29.261-29.262c0-11.153-6.274-20.865-15.476-25.801h108.073c-9.202 4.936-15.476 14.649-15.476 25.801 0 16.135 13.127 29.262 29.261 29.262 16.135 0 29.262-13.127 29.262-29.262 0-11.153-6.274-20.865-15.477-25.801h2.521a40.05 40.05 0 0 0 38.896-30.317zm-94.47 214.682c-5.235 0-9.494-4.259-9.494-9.495s4.258-9.494 9.494-9.494 9.495 4.258 9.495 9.494-4.26 9.495-9.495 9.495m-135.642 0c-5.235 0-9.495-4.259-9.495-9.495s4.259-9.494 9.495-9.494 9.494 4.258 9.494 9.494-4.259 9.495-9.494 9.495m171.684-70.472a20.31 20.31 0 0 1-19.736 15.409H40.11c-10.45 0-19.075-7.923-20.209-18.075h198.82l-.67 2.544q-.016.061-.03.122m5.908-22.433H19.767V90.082H243.12z"/></svg>
</svg_code>

Question: How is the raised handle positioned relative to the basket?
Options:
A. It rises vertically from the center to a rounded grip
B. It rises diagonally from the left side to a short horizontal grip
C. It rises diagonally from the right side to a short horizontal grip
D. It extends horizontally beneath the basket



Answer: C

Figure 8: **Examples from the SVG Blind VQA diagnostic.** The model is given only the SVG code, question, and answer options, and must infer the visual answer without observing the rendered icon. The renderings shown here are provided solely for readers.

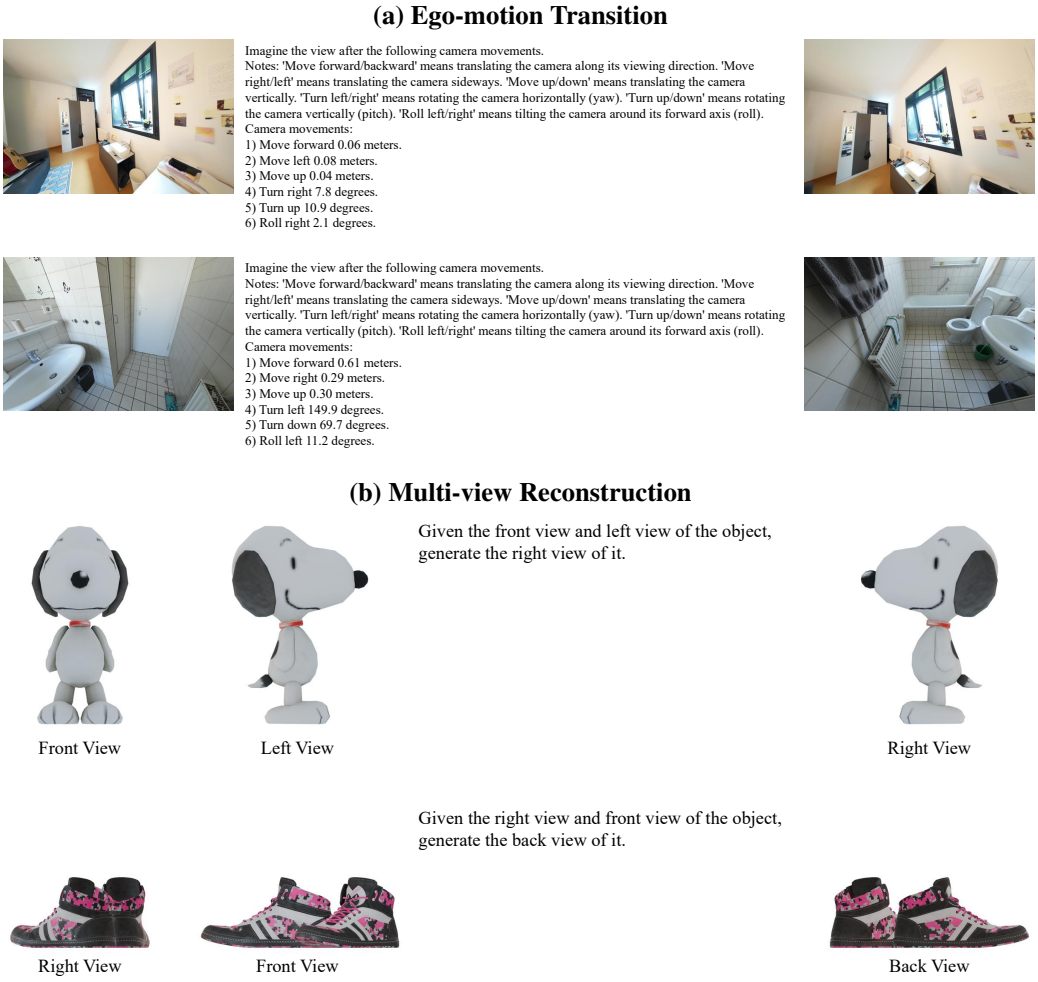
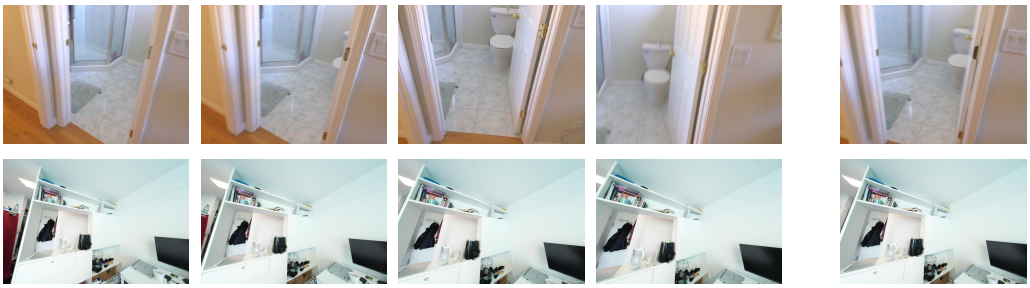


Figure 9: **Examples of 3D spatial-intelligence generation tasks (I).** Ego-motion transition requires applying explicit camera motion to a source view, while multi-view reconstruction infers an unobserved view from the provided object views.

(c) View Sequence Completion

Given frames 1, 2, 4, and 5 from a video clip, generate the missing frame 3. Ensure temporal consistency, realistic appearance, and coherent scene geometry.



Frame 1 Frame 2 Frame 4 Frame 5 Frame 3

(d) Layout-to-image Generation

An industrial-style loft bedroom featuring exposed brick walls, wooden ceiling beams, and a large white canopy bed.
3D object annotations:
Coordinate system: ego-camera coordinates, x=right, y=down, z=forward, unit=meter.
Objects:
- object_1: name=bed, center=(x=1.43, y=-0.18, z=5.09), size=(width=2.70, height=3.88, length=2.47)
- object_2: name=window, center=(x=3.86, y=0.23, z=6.28), size=(width=1.33, height=1.85, length=0.17)
- object_3: name=window, center=(x=-0.57, y=0.19, z=6.78), size=(width=1.33, height=1.85, length=0.17)
- object_4: name=bookcase, center=(x=-0.53, y=1.50, z=6.32), size=(width=0.85, height=0.85, length=0.30)
- object_5: name=chair, center=(x=-1.48, y=1.18, z=3.46), size=(width=0.62, height=0.49, length=0.53)
- object_6: name=chair, center=(x=-1.37, y=1.07, z=2.76), size=(width=0.62, height=0.49, length=0.53)
- object_7: name=pillow, center=(x=0.48, y=1.12, z=5.35), size=(width=0.60, height=0.41, length=0.36)
- object_8: name=pillow, center=(x=1.00, y=1.15, z=5.62), size=(width=0.60, height=0.40, length=0.35)
- object_9: name=pillow, center=(x=1.52, y=1.24, z=5.89), size=(width=0.62, height=0.43, length=0.30)
- object_10: name=pillow, center=(x=1.43, y=1.22, z=6.07), size=(width=0.62, height=0.50, length=0.21)



The image displays a kitchen counter setup from a direct frontal view. Above, wooden shelving units hold various supplies, including stacks of paper and cups, with a microwave visible in the upper left corner. The main surface is a red countertop. On the left side of the counter, a black toaster oven is positioned with a white cup sitting in front of it. Moving right, a dish drying rack with a black mug sits next to a stainless steel sink equipped with a faucet. Below the counter, wooden cabinets with silver handles are on the left. Underneath the sink area on the right, there is open storage containing blue water filtration equipment, specifically vertical blue cylinders and a large blue barrel, all resting on a tiled floor.



Figure 10: **Examples of 3D spatial-intelligence generation tasks (II).** View sequence completion predicts a temporally and geometrically consistent missing frame, while layout-to-image generation synthesizes an image from structured or textual descriptions of 3D scene layout.