# Ψ-0.6: A UNIFIED VISION-LANGUAGE WORLD MODEL WITH ADAPTIVE STREAMS-OF-THOUGHT

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Intelligence in the physical world requires self-directed, task-adaptive reasoning through different representations of the world as a problem unfolds. We introduce Ψ-0.6, a unified vision-language world model that adaptively selects the representations through which it reasons. Ψ-0.6 reasons through *streams-of-thought*: sequences interleaving images and text with geometrically-grounded, interpretable visual abstractions such as camera pose, depth, optical flow, and point tracks. A shared token interface supports *any-to-any* prompting, allowing the same model to generate, interpret, and compose these representations as intermediate reasoning steps. Ψ-0.6 matches state-of-the-art zero-shot performance on MindCube, a challenging benchmark of multi-view spatial reasoning. Composing these capabilities – estimating camera motion, selecting the relevant observed view, and answering from it – further improves performance. World modeling enables *counterfactual imagination*: the model translates a specified object motion into a flow prediction, generates the resulting scene, and answers from it. On FlowBench-Real, which tests this ability, Ψ-0.6 shows better generalization than Qwen. Through reinforcement learning from task feedback, Ψ-0.6 learns *adaptive streams-of-thought* that select visual representations for each input and task. This improves visual question-answering accuracy, with gains transferring to unseen benchmarks. By *unifying* perception and imagination, geometry and language, as operations of the same model, Ψ-0.6 uses them for *representation-adaptive reasoning*.

## 1 INTRODUCTION

Intelligence in the physical world requires self-directed reasoning through different representations of the world that adapt as a problem unfolds. When assembling furniture, we read instructions, change viewpoint, estimate geometry, and mentally rotate a panel to imagine whether it will fit. Words express goals and relations; geometric representations expose spatial structure; motion describes change; and imagined scenes reveal action consequences (Shepard & Metzler, 1971; Battaglia et al., 2013). Each provides a different projection of the same underlying world, making some relationships explicit while abstracting others. General physical reasoning therefore requires not just a repertoire of representations, but the ability to *compose* and decide *which representation to use at each step*.

This is challenging because today's AI is fragmented across specialized systems: vision models estimate structure, vision-language models (VLMs) describe, and world models (WMs) predict. This separation makes it difficult for a reasoner to treat these capabilities as composable operations, interpret its own intermediate predictions, and optimize from downstream task feedback how those predictions should be generated and used. Our central question is: *how can visual, geometric, linguistic, and world-model predictions become native reasoning operations of a single model?*

A powerful blueprint for flexible, self-adaptive reasoning already exists – in large language models (LLMs): (i) a unified model provides a broad repertoire of capabilities, (ii) a generic prompting interface makes those capabilities accessible and composable across tasks (Brown et al., 2020; Kojima et al., 2022), and (iii) reinforcement learning on task outcomes shapes the intermediate reasoning strategies used to solve a problem (DeepSeek-AI et al., 2025). Chain-of-thought has been central to the remarkable reasoning capabilities of modern LLMs, motivating the question of whether similarly powerful self-directed, task-adaptive reasoning can extend to the physical world. Extending this



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: **Ψ-0.6: representation-adaptive reasoning across vision, language, and world modeling.** **(A)** *Capability unification.* Diverse representations share a common token interface. Different choices of supplied and predicted tokens express perception, geometry, language, and imagination within a single transformer. **(B)** *Compositional unification.* Streams-of-thought combine model capabilities. **(C)** *Optimization unification.* Because intermediate representations are predictions of the same model, choosing which representations to generate and use becomes a learnable action.

recipe beyond language requires making visual and geometric predictions available not only as task outputs, but as intermediate operations that a model can generate, compose, and learn to use.

We extend this recipe with Ψ-0.6, a single transformer that reasons across vision, language, and world modeling and adaptively selects the representations through which it reasons. Its core mechanism is reasoning through *streams-of-thought* (Fig. 1), sequences that interleave images and text with explicit visual abstractions including camera pose, depth, optical flow, and point tracks. All of these representations are expressed as tokens predicted by the same transformer. A depth estimate, camera-motion prediction, generated view, or textual answer is therefore not merely the endpoint of a specialized task; each can become an intermediate step in a larger reasoning process. In this sense, *unification turns representations into reasoning operations*: the same model can generate them, condition on them, compose them, and learn when they are useful.

We build a broad repertoire of capabilities through *capability unification*. Ψ-0.6 is first trained to learn appearance, geometry, and motion from video – building on Probabilistic Structure Integration (Kotar et al., 2025) – then further trained with language and instruction following (§2.4). Images, text, camera pose, depth, optical flow, and point tracks share a common token interface, so perception, description, geometric estimation, and imagination become different conditional predictions of one model. A generic *any-to-any* prompting interface flexibly accesses Ψ-0.6's unified capabilities by varying which representations are supplied and which are predicted: an image pair can prompt a camera-pose estimate, while an image and camera pose can prompt a new view (§2.2).

Our unified model exhibits strong zero-shot spatial intelligence. On MindCube (Wang et al., 2026), a challenging benchmark for reasoning across multiple views of a scene, Ψ-0.6 achieves $49.7\%$ zero-shot accuracy through direct question answering, matching state-of-the-art zero-shot performance (§3; Fig. 3). Providing explicit visual intermediates such as camera pose or depth as additional context improves accuracy to $51.3\%$, showing that these representations are useful not only as task outputs but as computational intermediates for downstream reasoning.

Unification becomes more powerful when these capabilities are chained together through *compositional unification*. Because vision-language and world-model predictions share the same model and interface, the output of one operation can become context for the next within a single stream-of-thought. On MindCube, Ψ-0.6 predicts camera motion, uses that prediction to select the relevant observed view, and then answers from that view. This structured inference raises zero-shot accuracy from $51.3\%$ to $57.7\%$ (§3), demonstrating that capabilities of a unified model can become stronger when used as mutually composable reasoning operations rather than independent endpoints.

World modeling extends this compositional reasoning beyond observed scenes through *counterfactual imagination* (§4). Asked whether moving an object would obstruct a pathway, Ψ-0.6 can translate the specified displacement into a motion prediction, generate the resulting scene, and answer using its own prediction. Crucially, the *reasoner and imaginer are the same model*: predicted scenes are generated and interpreted within the same token interface rather than supplied by a separate visual generator. Streams-of-thought therefore allow language, geometry, motion, and predicted appearance to participate in a single reasoning trajectory. On FlowBench, which tests counterfactual imagination, Ψ-0.6 shows better generalization than Qwen to unseen domains.

Finally, Ψ-0.6 moves beyond prescribed reasoning pathways to learn *representation-adaptive* streams-of-thought through reinforcement learning on task feedback. This is an elegant consequence of unification: because camera motion, depth, optical flow, imagined views, and text are all model actions, the reasoning process can be optimized end to end (*optimization unification*), shaping both which intermediate representations are generated and their contents. Optimizing this policy on downstream outcomes improves visual question-answering performance, with gains that transfer to unseen benchmarks (§5). The choice and generation of reasoning representations thus become actions of the same learned policy rather than components of a hand-designed pipeline.

Taken together, our core thesis is that by *unifying* perception, geometry, language, and imagination as operations of one model, Ψ-0.6 can use them as intermediate *reasoning* actions – and task feedback can teach it which to invoke when. At the level of capabilities, one model predicts language, geometry, motion, and appearance; at the level of composition, these predictions become intermediate steps in streams-of-thought; at the level of optimization, the model learns which representations to generate and use. The same model therefore serves as perceiver, VLM, world model, reasoner, and policy. Self-adaptive reasoning has been central to generalizable, reasoning intelligence in language (Wei et al., 2022; Kojima et al., 2022). Streams-of-thought extend this beyond language, allowing a unified model to reason through whichever projection of the physical world best serves each step.

More broadly, reinforcement learning over streams-of-thought provides a remarkably general recipe for representation-adaptive reasoning across diverse domains and modalities. Within a shared predictive interface, observations in any modality can provide feedback for learning how to reason through others. For example, given a video, a reasoning trajectory through depth, optical flow, or camera motion can be rewarded by how well it supports predictions of future visual input. This provides a label-free, task-independent source of reward without human annotations or separate RL environments. More generally, once perception and prediction are part of the reasoner, observations of the world are *sufficient* for supervising not only *what* a model predicts, but potentially *how it learns to reason* – this opens a path toward agents that continually improve their reasoning from experience.

## 2 Ψ-0.6: A Unified Model with Adaptive Streams-of-Thought

Ψ-0.6 combines three conceptual components, extending a familiar LLM recipe beyond language: (i) a unified model provides a broad repertoire of visual, geometric, and linguistic capabilities; (ii) a generic prompting interface flexibly accesses and composes these capabilities; and (iii) reinforcement learning teaches the model to direct its own reasoning from task feedback. These components support *capability unification*, *compositional unification*, and *optimization unification*, respectively.

### 2.1 Unified Multimodal Model

**Architecture: one vocabulary, one transformer.** Images, camera pose, depth, optical flow, point tracks, and text are serialized as tokens in a sequence processed by our 8B-parameter autoregressive transformer. Perception, description, and imagination thus become different conditional predictions.

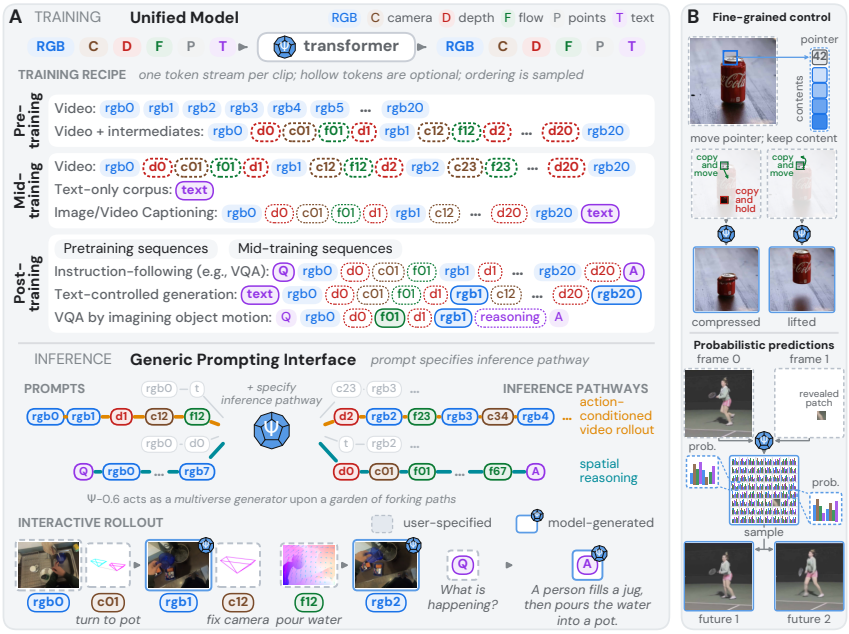


Figure 2: **Architecture, training recipe, and prompting interface. (A)** Ψ-0.6 is a unified model across diverse modalities. Pretraining, mid-training, and post-training sample different sequence families; hollow tokens denote stochastically included sequence elements. Generic prompting interface: prompts specify the available information and requested inference pathway. Ψ-0.6 is interactive: it can generate a view based on a user-specified camera pose, then use that generated view as context for subsequent reasoning. **(B)** Fine-grained control: each patch is represented by a pointer token and content tokens. Pointers specify modality and spatial and temporal locations, enabling targeted conditioning and scene edits. Probabilistic prediction: Ψ-0.6 represents a distribution over compatible continuations, allowing the same prompt to yield alternative predictions through sampling.

### 2.2 Generic Prompting Interface: Streams-of-Thought

A generic prompting interface makes the model's capabilities available as flexible reasoning operations. Ψ-0.6 reasons through *streams-of-thought*: sequences interleaving images and text with explicit visual abstractions, including camera pose, depth, optical flow, and point tracks. These representations offer complementary *projections of the world*, allowing each reasoning step to use a representation suited to the problem – text, geometry, or a predicted scene.

**Any-to-any prompting: all tasks are token orderings.** A prompt specifies what is given and what is requested. Different token choices and orderings express different inference pathways through the same model. Any modality combination can condition prediction, enabling richly controllable world modeling over spatial, temporal, and semantic aspects of a scene. Let $\mathrm{RGB}_0$ denote an image, while $C_{0\to1}$ and $F_{0\to1}$ denote camera motion and optical flow. Then $\mathrm{RGB}_0, C_{0\to1} \to \mathrm{RGB}_1$ and $\mathrm{RGB}_0, \mathrm{RGB}_1 \to C_{0\to1}$ express novel view synthesis and pose estimation; replacing $C$ with $F$ gives motion-conditioned generation and flow estimation. These are conditional predictions of the same model. A camera pose can therefore serve as an *estimate* when inferred from images, a *command* when supplied to generate a view, or an *intermediate* when used to answer a spatial question.

**Compositional reasoning: predictions become context.** Streams-of-thought compose these conditional predictions into multistep reasoning processes, with the output of one prediction providing context for the next. For example, the model can generate a view from an image and a user-specified camera pose, then use the generated view as context for answering a spatial question (Fig. 2A2). Visual and geometric predictions are therefore not only task outputs, but intermediate operations that connect perception, language, and imagination within one reasoning process. A prompt can prescribe these operations; their selection can also become part of what the model learns.

Ψ-0.6 builds on Probabilistic Structure Integration (PSI) (Kotar et al., 2025), inheriting two further properties that make its prompting interface unusually flexible: *fine-grained control* over individual

patches and *probabilistic prediction*, which allows unspecified content to remain open to multiple plausible continuations. We kindly refer readers to Appendix Sec B.1 for details.

### 2.3 Reinforcement Learning: Adaptive Streams-of-Thought

Beyond executing prescribed reasoning pathways, Ψ-0.6 learns *adaptive streams-of-thought* – choosing which representation to use and when – from task feedback. Within the shared token interface, these choices become policy actions, making the reasoning strategy itself a target of optimization. We instantiate this approach through reinforcement learning and apply it to visual question answering. The policy selects visual representations, including depth, camera motion, and optical flow, according to the input and question, and receives reward based on downstream answer quality. Section §5 specifies the learning formulation, training procedure, and evaluation.

### 2.4 Training Recipe: Pretraining, Mid-training, and Post-training

Pretraining builds a broad base of predictive knowledge, mid-training integrates additional representations and data, and post-training teaches task-directed use of the resulting capabilities. In each supervised stage, we form different dataset groups (e.g., image captioning, video question answering), each with its own datasets and sequence notations, and allocate a training ratio to each. Training on these dataset groups is unified under a simple *next-token prediction* objective. Thus, sequence construction determines both what the model observes and what it learns to predict.

**Pretraining: visual world modeling and geometric integration.** The model is pretrained on video spanning natural scenes, human activity, and robot interaction. Training begins with RGB and camera conditioning and continues with depth and flow integrated into the sequence. Pure RGB sequences train appearance and motion prediction; interleaved sequences additionally train relationships between images and geometric structure. Keeping both in the mixture exposes the model to prediction with different amounts and types of conditioning information. This stage extends Probabilistic Structure Integration (PSI) (Kotar et al., 2025), part of a lineage of models demonstrating state-of-the-art visual understanding and world modeling across optical flow and point tracking (Kim et al., 2025), depth estimation and novel-view synthesis (Lee et al., 2025b), object segmentation (Venkatesh et al., 2025), physical object discovery and 3D manipulation (Venkatesh et al., 2026), and perceptual simulation (Lee et al., 2026). Ψ-0.6 extends these visual prediction and control capabilities to language-grounded reasoning in subsequent stages.

**Mid-training: vision-language alignment.** The next step is learning language through text-only data from FineWeb-Edu and DataComp-LM (DCLM), followed by vision-language alignment by adding images and videos paired with captions, both global and temporally indexed.

**Post-training: useful, controllable, user-directed reasoning.** Instruction tuning introduces image and video question answering and text-to-video control of camera and object motion. Post-training also teaches spatial reasoning (§3) and answering questions by counterfactually imagining object motion (§4). Pretraining and mid-training sequences are retained to preserve the model's broad predictive capabilities. Beyond these next-token prediction stages, reinforcement learning trains adaptive streams-of-thought from task feedback, improving how the model selects intermediate representations for downstream reasoning (§5).

## 3 Capability and Compositional Unification: Spatial Intelligence

Ψ-0.6 represents our hypothesis for building unified VLMs with capable spatial intelligence. First, we start from a visual world-model backbone trained to jointly model appearance, geometry, and motion (Kotar et al., 2025), then extend it with language. Standard VLMs typically attach a visual encoder to a pretrained LLM, whose initial predictive knowledge comes from text (Liu et al., 2024a; Alayrac et al., 2022). Language pretraining does not directly supervise how appearance changes with viewpoint or object motion. We hypothesize that learning these relationships through visual and geometric prediction provides a stronger foundation for spatial reasoning, with language supplying an interface to these capabilities. Second, we ground reasoning in explicit geometric abstractions, whereas standard VLMs reason and answer primarily in text. A single transformer interleaves RGB frames, camera pose, depth, optical flow, point tracks, and text as a unified token sequence. These




Figure 3: **Spatial intelligence reasoning on MINDCUBE. (A)** Benchmark categories: Rotation, Around, and Among, adapted from Wang et al. (2026). **(B)** Accuracy for zero-shot and benchmark-trained systems; some trained comparisons also use RL. **(C)** Two structured-prompting routes: infer relative poses and select a real input view, or synthesize RGB at a target pose, then read an answer.

visual *streams-of-thought* express spatial structure and motion in intermediate predictions that can be inspected to diagnose errors. Third, unifying world-model and vision-language capabilities yields compositional, end-to-end reasoning. Flexible token orderings express pose estimation, novel view synthesis, and question answering, allowing one task's output to condition the next. A spatial question can thus combine capabilities of the same model architecture—estimate the camera motion the question describes, get to the view that reveals the answer, read it.

MINDCUBE tests spatial reasoning across multiple views of a scene (Wang et al., 2026). It is challenging in two ways: (i) its questions involve counterfactual camera motions (e.g., turn left) rather than answering from static images/video, and (ii) its images are separated by large angles (e.g., 90 degrees), which is unlike how most video models and VLMs are trained. MINDCUBE is organized into three categories (Fig. 3A), all requiring integrating observations across large viewpoint changes and reasoning in the reference frame specified by the question.

**Zero-shot.** Frozen Ψ-0.6 reaches 49.7–51.3% across RGB-only, RGB-depth, and RGB-camera configurations (Fig. 3B). Ψ-0.6 is state-of-the-art relative to several proprietary models (GPT-5, Gemini-2.5-Pro, Claude-4-Sonnet), open-source generalist VLMs (DeepSeek-VL2-Small, LLaVA-OneVision-7B, Qwen3-VL-8B), and spatial specialists (RoboBrain, Space-Qwen, Spatial-MLLM).

**Structured Prompting.** Beyond directly answering the question, Ψ-0.6 can combine its VLM capabilities with its vision and world-modeling capabilities (Figure 3C). Ψ-0.6 uses *inverse dynamics*: predict relative camera poses, select the real input view matching the question's requested turn, and read the answer from it. Alternatively, Ψ-0.6 uses *novel view synthesis* (NVS): generate an RGB frame at the target camera pose and read the answer from this imagined view. Inverse dynamics requires the viewpoint to be one of the inputs, but NVS additionally allows a viewpoint absent from the inputs. Using inverse dynamics, Ψ-0.6 reaches 57.7%, compared to 51.3% zero-shot.

**SFT.** MINDCUBE fine-tuning yields 68.7% with RGB-only inputs and 76.3% with RGB+camera inputs on the held-out Tiny test split (Fig. 3B). The RGB-only result exceeds the displayed SFT-only Qwen2.5-VL and 3DThinker baselines. The RGB+camera result exceeds InternVL3's map-then-reason configuration and Qwen2.5-VL with SFT+RL, is comparable to 3DThinker with SFT+RL (76.0%), and trails GeoThinker (83.0%) (Wang et al., 2026; Chen et al., 2026; Li et al., 2026).

## 4 FROM COMPOSITIONS TO COUNTERFACTUALS

Beyond answering questions about static scenes, Ψ-0.6's WM capability lets it answer questions requiring *counterfactual imagination*, reasoning about worlds that are not directly observed. The model intervenes on a scene, predicts the consequence, and reads the answer off its own prediction.

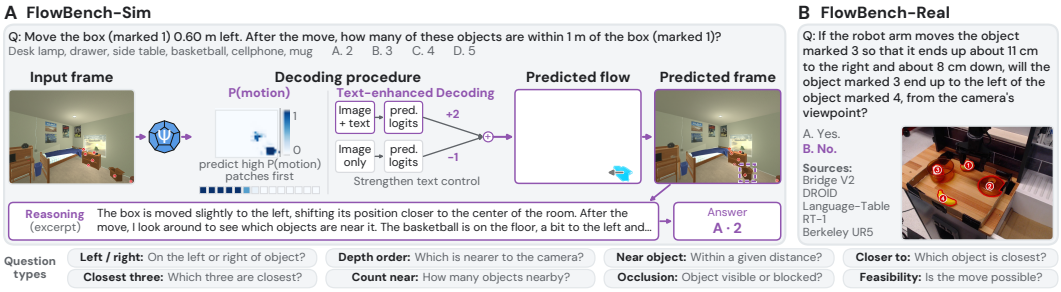


Figure 4: **Counterfactual imagination of object motion.** (A) Stream of Thought on FlowBench-Sim with pMotion and CFG. (B) Ψ-0.6 shows stronger generalization than Qwen on FlowBench-Real.

Table 1: Optical flow prediction: expert models vs. PSI-0.6 variants. "$\text{rEPE}_{\text{move}}$" is the relative endpoint error on moving pixels; static pixels are excluded. "Acc" denotes what percentage of predictions have rEPE lower than a threshold. "Blank" stands for the percentage of the model predicting zero flow. "Dir" is the 2-D angle between predicted and ground-truth flow on pixels where both move. "ΔPSNR" shows the improvement of the generated imagination image after the move on moving pixels compared with zero flow. $n = 622$.

| Method | Acc↑ | $\text{rEPE}_{\text{move}}$ ↓ | $\text{rEPE}_{\text{move}}$ med ↓ | Blank ↓ | Dir (°) ↓ | ΔPSNR ↑ | Oracle Input |
|---|---|---|---|---|---|---|---|
| Zero flow | 0.0% | 1.000 | 1.000 | 100.0% | – | 0.00 | – |
| Modular (Grounding DINO + SAM2 + Depth Pro) | 50.8% | 0.552 | 0.358 | 15.0% | **1.4** | – | Camera intrinsics |
| FOFPred (Ranasinghe et al., 2026) | 2.4% | 0.915 | 0.958 | **0.0%** | 25.9 | – | – |
| Qwen-Image-Edit (Wu et al., 2025a) | 7.9% | 2.154 | 1.014 | 2.9% | 64.8 | -1.63 | GT frame-0 mask |
| IP2P (Brooks et al., 2023) (plain MSE) | 10.5% | 2.659 | 1.221 | 2.9% | 71.0 | -3.35 | GT frame-0 mask |
| IP2P (Brooks et al., 2023) (SoM target) | 13.7% | 2.938 | 1.122 | 3.1% | 64.4 | -0.96 | GT frame-0 mask |
| IP2P (Brooks et al., 2023) (mask-weighted) | 16.4% | 2.655 | 1.000 | 2.7% | 57.3 | -4.00 | GT frame-0 mask |
| Ψ-0.6 Default seq 32 | 31.8% | 0.635 | 0.639 | 27.5% | 14.1 | **+1.60** | – |
| Ψ-0.6 + pMotion + CFG2.0 | **56.4%** | **0.454** | **0.356** | 0.3% | 8.1 | +0.94 | – |

While §3 moved the camera, here the intervention moves an object: the question states a displacement of a marked object and asks where things end up – occluded or not, closer to which landmark, touching a wall, still feasible to push further, etc. This is the rung above association on the ladder of causation (Pearl, 2009), and it is exactly where existing VLMs are near chance before fine-tuning.

**Benchmark.** FLOWBENCH (Liu et al., 2026b) evaluates visual world models on spatial questions involving counterfactual object motions. It comprises two splits: FLOWBENCH-SIM, featuring scalable simulated indoor scenes (Kolve et al., 2017; Deitke et al., 2022), and FLOWBENCH-REAL, augmenting five robot datasets (DROID, Bridge V2, Language-Table, RT-1, and Berkeley UR5) with counterfactual motion questions across diverse embodiments and environments (Appendix A.2).

**World modeling: Ψ-0.6 outperforms state-of-the-art expert models for optical flow prediction.** Text-conditioned future prediction is one of the pressing abilities needed in world modeling, with promising applications like planning in robotics. To improve text-to-flow predictions, Ψ-0.6 further employs two powerful decoding techniques: pMotion and CFG. Details are introduced in Appendix A.5. Critically, these are enabled by Ψ-0.6's fine-grained control and probabilistic predictions. Ψ-0.6 achieves robust, coherent flow prediction as shown in Fig. 4 and Tab. 1.

We compare our models with various baseline models: FOFPred (Ranasinghe et al., 2026), Instruct-Pix2Pix (Brooks et al., 2023), Qwen-Image-Edit (Wu et al., 2025a), and a Modular design with SoTA experts (Liu et al., 2024b; Ravi et al., 2025; Bochkovskii et al., 2025) (Appendix A.3). Against baselines, Ψ-0.6 establishes the strongest overall flow prediction without privileged inputs. Its best configuration reaches $56.4\%$ accuracy, compared with $16.4\%$ for the strongest image-editing baseline. It also reduces $\text{rEPE}_{\text{move}}$ to $0.454$, versus $0.915$ for FOFPred. Even the modular geometric pipeline reaches only $50.8\%$ accuracy and $0.552$ $\text{rEPE}_{\text{move}}$, despite its expertise in each task and the oracle input. Finally, every evaluated Ψ-0.6 configuration improves moving-region image fidelity over the zero-flow reference ($+0.25$ to $+1.60$ dB), whereas the image editors decrease it ($-0.96$ to $-4.00$ dB). Together, these results show that Ψ-0.6 predicts both where an object moves and how its appearance should change more reliably than task-specific generative experts, while requiring no oracle inputs.

**Strong generalization abilities of Ψ-0.6.** Fine-tuning on FLOWBENCH-SIM substantially improves the VQA performance of Ψ-0.6 across all eight categories (Table 3), trailing Qwen3-VL-8B fine-tuned

Table 2: Accuracy on FlowBench-Real (%). Note: No FlowBench-Real (%) examples are used in training, so this table evaluates the generalization ability of models. Closest-3 requires the exact set of three. Cat. avg is the unweighted mean over the 7 categories.

| Model | L/R | Near | Closer | Depth | Closest-3 | Count | Occl. | Avg | Cat. Avg |
|---|---|---|---|---|---|---|---|---|---|
| Chance | 50.0 | 50.0 | 26.5 | **50.0** | 7.1 | 25.1 | 50.0 | 37.6 | 36.9 |
| Qwen3-VL-8B (Bai et al., 2025) | 53.9 | 54.9 | 56.6 | 28.1 | 14.8 | 44.3 | 40.0 | 43.8 | 41.8 |
| Qwen3.5-9B (Qwen Team, 2026) | 59.6 | 61.1 | 60.3 | 40.4 | 0.0 | 25.2 | 46.0 | 44.6 | 41.8 |
| Qwen3-VL-32B (Bai et al., 2025) | 59.1 | 59.1 | 57.1 | 26.7 | 0.7 | 38.9 | 34.0 | 42.4 | 39.4 |
| Qwen-8B FlowBench-FT | 73.6 | 68.4 | **76.7** | 29.5 | **40.0** | 38.2 | 49.0 | 56.6 | 53.6 |
| Ψ-0.6 Base | 50.3 | 51.3 | 23.3 | **50.0** | 7.4 | 26.7 | 50.0 | 37.5 | 37.0 |
| Ψ-0.6 FlowBench-FT | **80.3** | **79.3** | 36.5 | 33.6 | 35.6 | **71.8** | **57.0** | **57.5** | **56.3** |

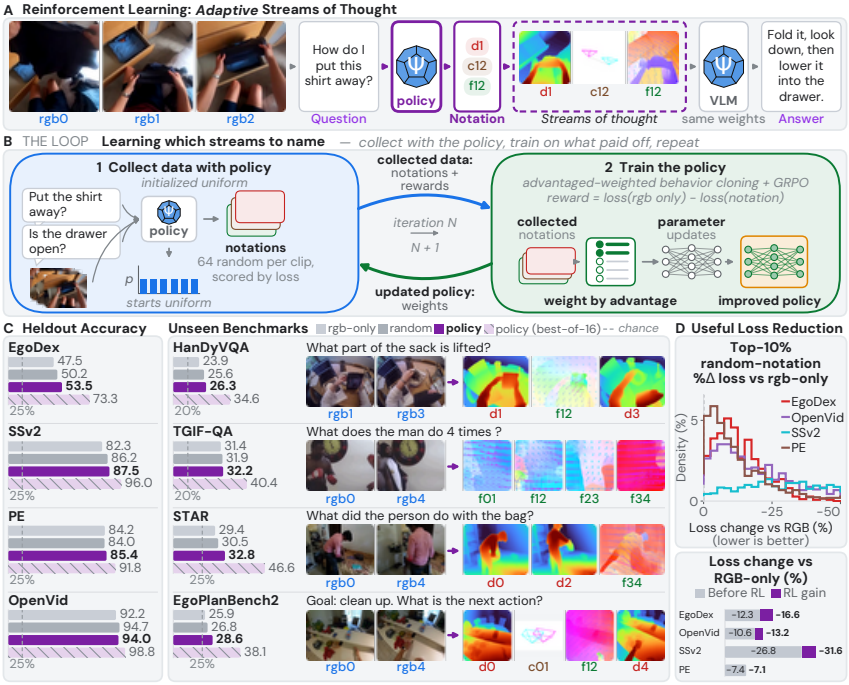


Figure 5: **Learning adaptive streams of thought.** **(A)** Given a video and question, Ψ-0.6 functions as both a policy – emitting a notation naming the stream-of-thought to use – and as a VLM, answering with those streams in context. **(B)** The RL loop: collect notations with the current policy, score each by probability of groundtruth answer relative to RGB-only, and train on the beneficial notations. **(C)** The learned Ψ-0.6 policy improves video question-answering accuracy, with gains generalizing to unseen benchmarks. **(D)** These accuracy gains are a result of loss reductions on the ground truth.

on the same dataset by 4.8 points. To evaluate transfer, we test the models on FLOWBENCH-REAL without using any of its examples during training (Table 2). Ψ-0.6 fine-tuned on FLOWBENCH-SIM achieves the highest overall and category-averaged accuracies among the evaluated models, reaching 57.5% and 56.3%, respectively. These results improve upon its base model by 20.0 and 19.3 percentage points and exceed the fine-tuned Qwen3-VL-8B by 0.9 and 2.7 points. Ψ-0.6 leads in four of the seven categories – L/R, Near, Count, and Occlusion. This shows that supervised fine-tuning of Ψ-0.6 on FLOWBENCH enables *sim-to-real* transfer for its spatial question-answering capabilities to unseen data. Viewed alongside its world-modeling capabilities, these results support the use of Ψ-0.6 as a unified framework for world modeling and visual question answering.

## 5 OPTIMIZATION UNIFICATION: RL FOR ADAPTIVE STREAMS-OF-THOUGHT

Whereas Ψ-0.6 used prescribed reasoning pathways in the previous sections, here we show that it can learn self-directed streams-of-thought that are *self-generated*, *representation-adaptive*, and unfold over time. As reasoning proceeds, the model learns not only what intermediate content to generate, but which representation to invoke and when. Unification makes both the choice and generation of these intermediates actions of the same policy, enabling reinforcement learning on task feedback to optimize the reasoning process end to end (*optimization unification*).

- **Self-generated intermediates.** Unlike tool-augmented LLMs and VLMs that invoke external models or programs (Schick et al., 2023; Surís et al., 2023; Hu et al., 2024), Ψ-0.6 generates its intermediates within the reasoner itself. This enables: (i) *co-adaptation* of the reasoner and its intermediate predictions through gradient-based credit assignment, (ii) generation of states unavailable through existing tool calls, including future or counterfactual states, and (iii) task-dependent intermediates whose appropriate form may itself be ambiguous. For example, a bottle and its cap are one segment for object recognition and separate segments for robotic manipulation.
- **Representation-adaptive computation.** Ψ-0.6 uses *representation*-adaptive *test-time compute* through explicit geometric projections of the world. In text chain-of-thought, the representational medium is fixed – language; in Ψ-0.6, the medium itself becomes part of the computation.
- **Unfolds over time.** Different representations can be called into computation as the situation evolves: depth for spatial structure, optical flow for motion, camera pose across viewpoints, or a predicted scene to imagine what may happen next. This is particularly relevant to embodied agents and robotics, where perception and action unfold continuously in the world.

**Learning to Select Streams-of-Thought** Although our framework is generic, here we show a proof-of-concept specifically for visual question-answering. Given a video $v$ and question $q$, the policy emits a notation $z$ selecting from depth, camera-motion, and optical-flow streams (Fig. 5A). For example, $d_0, c_{2\to3}, f_{2\to3}$ requests depth at the first frame and camera motion and flow between two later frames. This turns Ψ-0.6's generic prompting interface into a policy over reasoning representations. Reinforcement learning alternates between two phases (Fig. 5B):

- **Collect data with the policy.** For each question, sample notations $z$ from the current policy (initialized uniform), and score each by how much it lowers the ground truth's loss.
- **Train the policy on the collected data.** Initially, we use filtered behavior cloning (rejection-sampling fine-tuning) on the sampled notations that beat RGB-only to teach the model to emit notations; subsequently, we use on-policy GRPO to update the policy's LoRA parameters with rewards relative to the other samples in the group, refining which streams it selects.

**Learned policy generalizes accuracy improvements to unseen benchmarks.** The RL-trained policy improves accuracy on held-out examples from the training distributions, with the largest gains on EgoDex, the most challenging dataset. Encouragingly, this suggests that adaptive reasoning may provide greater benefits on harder problems, where additional intermediate computation is most useful. Critically, these gains transfer to four completely unseen benchmarks, providing evidence that the learned policy generalizes beyond its training data. Another encouraging signal is the best-of-$k$ performance is substantially higher than single-rollout (Fig. 5C), indicating further headroom for policy optimization. The selected streams are diverse and often interpretable: repeated-action questions elicit sequences of optical-flow predictions, object interactions combine depth and flow, and observer-motion questions combine depth with camera motion.

## 6 Conclusion

We introduce Ψ-0.6, a unified vision-language world model that turns visual, geometric, linguistic, and world-model predictions into composable reasoning operations. Through capability, compositional, and optimization unification, a single model can perceive, imagine, reason, and learn which representations to invoke as a problem unfolds. Streams-of-thought support stronger spatial reasoning, counterfactual imagination, and representation-adaptive reasoning learned from task feedback. More broadly, the medium of reasoning need not be fixed: models can learn not only *what* to predict, but *which representation* to reason through and *when*. This opens a path toward agents whose reasoning continually reorganizes itself across representations as they learn from embodied experience.

**Limitation and future work: scale remains an important open dimension.** Our results establish the conceptual and functional consequences of unifying vision, language, and world modeling, rather than exhaust the performance achievable by this framework. Ψ-0.6 was developed at substantially smaller scale in training data, compute, research, and engineering effort than contemporary frontier multi-modal systems. We expect scaling these resources will strengthen all of Ψ-0.6's capabilities; our preliminary experiments already suggest continued gains with scale, motivating further study.

### AI USE STATEMENT

In this work, we used generative AI tools for the following: (i) aid or polish writing – over multiple rounds of back-and-forth iteration, (ii) retrieval and discovery (e.g., finding related work), (iii) research ideation or execution, and (iv) draft sections of the paper. We have reviewed AI-assisted work. For example, all the text in this paper was checked by us manually as authors.

## REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a Visual Language Model for Few-Shot Learning, 2022. URL http://arxiv.org/abs/2204.14198. arXiv:2204.14198 [cs].

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Shikun Ban, Juling Fan, Xiaoxuan Ma, Wentao Zhu, Yu Qiao, and Yizhou Wang. Real-time holistic robot pose estimation with unknown states. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2024.

Peter W. Battaglia, Jessica B. Hamrick, and Joshua B. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013. https://doi.org/10.1073/pnas.1306572110. URL https://doi.org/10.1073/pnas.1306572110.

Yutong Bian, Dongjie Cheng, Heming Xia, Yongqi Li, and Wenjie Li. Optical Reasoning: Rethinking Images as an Expressive Reasoning Medium Beyond Text, June 2026. URL http://arxiv.org/abs/2606.09585. arXiv:2606.09585 [cs.AI].

Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.02073.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402. IEEE, 2023.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2005.14165.

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*, 2025.

Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Xiang An, Yan Feng, Peng Pei, Xunliang Cai, and Ruqi Huang. Think with 3D: Geometric Imagination Grounded Spatial Reasoning from Limited Views, March 2026. URL http://arxiv.org/abs/2510.18632. arXiv:2510.18632 [cs.CV].

Ethan Chern, Zhulin Hu, Steffi Chern, Siqi Kou, Jiadi Su, Yan Ma, Zhijie Deng, and Pengfei Liu. Thinking with Generated Images, May 2025. URL http://arxiv.org/abs/2505.22525. arXiv:2505.22525 [cs.CV].

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale, April 2020. URL http://arxiv.org/abs/1911.02116. arXiv:1911.02116 [cs.CL].

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *Nature*, 645(8081):633–638, September 2025. ISSN 0028-0836, 1476-4687. https://doi.org/10.1038/s41586-025-09422-z. URL http://arxiv.org/abs/2501.12948. arXiv:2501.12948 [cs.CL].

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-Scale Embodied AI Using Procedural Generation. In *NeurIPS*, 2022. Outstanding Paper Award.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19790–19800. IEEE, 2024. https://doi.org/10.1109/cvpr52733.2024.01871. URL http://dx.doi.org/10.1109/CVPR52733.2024.01871.

Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. AdaWorld: Learning Adaptable World Models with Latent Actions, June 2025. URL http://arxiv.org/abs/2503.18938. arXiv:2503.18938 [cs.AI].

Raktim Gautam Goswami, Prashanth Krishnamurthy, Yann LeCun, and Farshad Khorrami. Robopepp: Vision-based robot pose and joint angle estimation through embedding predictive pre-training. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6930–6939. IEEE, 2025.

Ziyu Guo, Xinyan Chen, Renrui Zhang, Ruichuan An, Yu Qi, Dongzhi Jiang, Xiangtai Li, Manyuan Zhang, Hongsheng Li, and Pheng-Ann Heng. Are Video Models Ready as Zero-Shot Reasoners? An Empirical Study with the MME-CoF Benchmark, October 2025. URL http://arxiv.org/abs/2510.26802. arXiv:2510.26802 [cs.CV].

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. Visual Sketchpad: Sketching as a Visual Chain of Thought for Multimodal Language Models, November 2024. URL http://arxiv.org/abs/2406.09403. arXiv:2406.09403 [cs.CV].

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labeling real videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–10. IEEE, 2025.

Seungwoo Kim, Khai Loong Aw, Klemen Kotar, Cristobal Eyzaguirre, Wanhee Lee, Yunong Liu, Jared Watrous, Stefan Stojanov, Juan Carlos Niebles, Jiajun Wu, and Daniel L. K. Yamins. Taming generative video models for zero-shot optical flow extraction, November 2025. URL http://arxiv.org/abs/2507.09082. arXiv:2507.09082 [cs.CV].

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2205.11916.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv*, 2017.

Klemen Kotar, Wanhee Lee, Rahul Venkatesh, Honglin Chen, Daniel Bear, Jared Watrous, Simon Kim, Khai Loong Aw, Lilian Naing Chen, Stefan Stojanov, Kevin Feigelis, Imran Thobani, Alex Durango, Khaled Jedoui, Atlas Kazemian, and Dan Yamins. World Modeling with Probabilistic Structure Integration, September 2025. URL http://arxiv.org/abs/2509.09737. arXiv:2509.09737 [cs.CV].

Phillip Y. Lee, Jihyeon Je, Chanho Park, Mikaela Angelina Uy, Leonidas Guibas, and Minhyuk Sung. Perspective-aware reasoning in vision-language models via mental imagery simulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025a. arXiv:2504.17207.

Wanhee Lee, Klemen Kotar, Rahul Mysore Venkatesh, Jared Watrous, Honglin Chen, Khai Loong Aw, and Daniel L. K. Yamins. 3D Scene Understanding Through Local Random Access Sequence Modeling, April 2025b. URL http://arxiv.org/abs/2504.03875.

Wanhee Lee, Klemen Kotar, Rahul Mysore Venkatesh, Jared Watrous, and Daniel L. K. Yamins. Perceptual 3D Simulation With Physical World Modeling, June 2026. URL http://arxiv.org/abs/2606.27575. arXiv:2606.27575 [cs.CV].

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent Visual Reasoning, October 2025a. URL http://arxiv.org/abs/2509.24251. arXiv:2509.24251 [cs.CV].

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while Reasoning in Space: Multimodal Visualization-of-Thought, January 2025b. URL http://arxiv.org/abs/2501.07542. arXiv:2501.07542 [cs.CL].

Haoyuan Li, Qihang Cao, Tao Tang, Kun Xiang, Zihan Guo, Jianhua Han, Hang Xu, and Xiaodan Liang. Thinking with geometry: Active geometry integration for spatial reasoning, 2026. URL https://arxiv.org/abs/2602.06037v1.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning, 2024a. URL http://arxiv.org/abs/2310.03744. arXiv:2310.03744 [cs].

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024b.

Yifan Liu, Fangneng Zhan, Wanhua Li, Haowen Sun, Katerina Fragkiadaki, and Hanspeter Pfister. Robotag: End-to-end robot configuration estimation via topological alignment graph. *arXiv preprint arXiv:2511.07717*, 2025.

Yifan Liu, Fangneng Zhan, Kaichen Zhou, Yilun Du, Paul Pu Liang, and Hanspeter Pfister. Abstract 3d perception for spatial intelligence in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38647–38656, 2026a.

Yifan Liu, Baihan Zhang, Khai Loong Aw, Klemen Kotar, and Daniel L. K. Yamins. Flowbench. https://github.com/liuyifan22/FlowBench, 2026b. Yifan Liu, Baihan Zhang, and Khai Loong Aw contributed equally.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. The Flan Collection: Designing Data and Methods for Effective Instruction Tuning, February 2023. URL http://arxiv.org/abs/2301.13688. arXiv:2301.13688 [cs.AI].

Moire. Language models are multiverse generators, January 2021. URL https://generative.ink/posts/language-models-are-multiverse-generators/. Section: GPT-3.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Kanchana Ranasinghe, Honglu Zhou, Yu Fang, Luyu Yang, Le Xue, Ran Xu, Caiming Xiong, Silvio Savarese, Michael S. Ryoo, and Juan Carlos Niebles. Future optical flow prediction improves robot control and video generation. In *Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, 2026.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Zhongwei Ren, Yunchao Wei, Xun Guo, Yao Zhao, Bingyi Kang, Jiashi Feng, and Xiaojie Jin. VideoWorld: Exploring Knowledge Learning from Unlabeled Videos, March 2025. URL http://arxiv.org/abs/2501.09781. arXiv:2501.09781 [cs.CV].

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language Models Can Teach Themselves to Use Tools, February 2023. URL http://arxiv.org/abs/2302.04761. arXiv:2302.04761 [cs.CL].

Roger N. Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171 (3972):701–703, 1971. https://doi.org/10.1126/science.171.3972.701. URL https://doi.org/10.1126/science.171.3972.701.

Yang Shi, Yuhao Dong, Yue Ding, Yuran Wang, Xuanyu Zhu, Sheng Zhou, Wenting Liu, Haochen Tian, Rundong Wang, Huanqian Wang, Zuyan Liu, Bohan Zeng, Ruizhe Chen, Qixun Wang, Zhuoran Zhang, Xinlong Chen, Chengzhuo Tong, Bozhou Li, Qiang Liu, Haotian Wang, Wenjing Yang, Yuanxing Zhang, Pengfei Wan, Yi-Fan Zhang, and Ziwei Liu. RealUnify: Do Unified Models Truly Benefit from Unification? A Comprehensive Benchmark, March 2026. URL http://arxiv.org/abs/2509.24897. arXiv:2509.24897 [cs.AI].

Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual Inference via Python Execution for Reasoning, March 2023. URL http://arxiv.org/abs/2303.08128. arXiv:2303.08128 [cs.CV].

Jingqi Tong, Yurong Mou, Hangcheng Li, Mingzhe Li, Yongzhuo Yang, Ming Zhang, Qiguang Chen, Tianyi Liang, Xiaomeng Hu, Yining Zheng, Xinchi Chen, Jun Zhao, Xuanjing Huang, and Xipeng Qiu. Thinking with Video: Video Generation as a Promising Multimodal Reasoning Paradigm, April 2026. URL http://arxiv.org/abs/2511.04570. arXiv:2511.04570 [cs.CV].

Rahul Venkatesh, Klemen Kotar, Lilian Naing Chen, Seungwoo Kim, Luca Thomas Wheeler, Jared Watrous, Ashley Xu, Gia Ancone, Wanhee Lee, Honglin Chen, Daniel Bear, Stefan Stojanov, and Daniel Yamins. Discovering and using Spelke segments, July 2025. URL http://arxiv.org/abs/2507.16038. arXiv:2507.16038 [cs].

Rahul Venkatesh, Klemen Kotar, Lilian Naing Chen, Wanhee Lee, Gia Ancone, Seungwoo Kim, Luca Thomas Wheeler, Jared Watrous, Honglin Chen, Daniel Bear, Stefan Stojanov, and Daniel LK Yamins. Physical Object Understanding with a Physically Controllable World Model, May 2026. URL http://arxiv.org/abs/2606.00439. arXiv:2606.00439 [cs.CV].

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Jiajun Wu, Li Fei-Fei, and Manling Li. MindCube: Spatial Mental Modeling from Limited Views, March 2026. URL http://arxiv.org/abs/2506.21458. arXiv:2506.21458 [cs.AI].

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in Latent Visual Space Beyond Images and Language, November 2025. URL http://arxiv.org/abs/2511.21395. arXiv:2511.21395 [cs.CV].

Yihan Wang, Lahav Lipson, and Jia Deng. Sea-raft: Simple, efficient, accurate raft for optical flow. In *European Conference on Computer Vision*, pp. 36–54. Springer, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2201.11903.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners, September 2025. URL http://arxiv.org/abs/2509.20328. arXiv:2509.20328 [cs.LG].

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025a. URL https://arxiv.org/abs/2508.02324.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12966–12977. IEEE, 2025b.

Linquan Wu, Tianxiang Jiang, Yifei Dong, Haoyu Yang, Fengji Zhang, Shichaang Meng, Ai Xuan, Linqi Song, and Jacky Keung. LaViT: Aligning Latent Visual Thoughts for Multi-modal Reasoning, August 2026a. URL http://arxiv.org/abs/2601.10129. arXiv:2601.10129 [cs.CV].

Penghao Wu, Haiwen Diao, Weichen Fan, Lewei Lu, Dahua Lin, and Ziwei Liu. Uncovering Understanding-Generation Synergy in Native Unified Multimodal Models: From Representation, Task to System, September 2026b. URL http://arxiv.org/abs/2609.01607. arXiv:2609.01607 [cs.CV].

Eric Xing, Mingkai Deng, and Jinyu Hou. Critique of World Model, June 2026. URL http://arxiv.org/abs/2507.05169. arXiv:2507.05169 [cs.LG].

Yuncong Yang, Jiageng Liu, Zheyuan Zhang, Siyuan Zhou, Reuben Tan, Jianwei Yang, Yilun Du, and Chuang Gan. Mindjourney: Test-time scaling with world models for spatial reasoning. *Advances in Neural Information Processing Systems*, 38:109855–109885, 2026.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine Mental Imagery: Empower Multimodal Reasoning with Latent Visual Tokens, June 2025. URL http://arxiv.org/abs/2506.17218. arXiv:2506.17218 [cs.CV].

Shoubin Yu, Yue Zhang, Zun Wang, Jaehong Yoon, Huaxiu Yao, Mingyu Ding, and Mohit Bansal. When and How Much to Imagine: Adaptive Test-Time Scaling with World Models for Visual Spatial Reasoning, May 2026. URL http://arxiv.org/abs/2602.08236. arXiv:2602.08236 [cs.CV].

Huanyu Zhang, Wenshan Wu, Chengzu Li, Ning Shang, Yan Xia, Yangyu Huang, Yifan Zhang, Li Dong, Zhang Zhang, Liang Wang, Tieniu Tan, and Furu Wei. Latent Sketchpad: Sketching Visual Thoughts to Elicit Multimodal Reasoning in MLLMs, October 2025. URL http://arxiv.org/abs/2510.24514. arXiv:2510.24514 [cs.CV].

Wanyue Zhang, Wenxiang Wu, Wang Xu, Jiaxin Luo, Helu Zhi, Yibin Huang, Shuo Ren, Zitao Liu, and Jiajun Zhang. World2VLM: Distilling World Model Imagination into VLMs for Dynamic Spatial Reasoning, April 2026a. URL http://arxiv.org/abs/2604.26934. arXiv:2604.26934 [cs.CV].

Yanbing Zhang, Bo Wang, Jianhui Liu, Nan Jiang, Jiaxiu Jiang, Haoze Sun, Yijun Yang, Shenghe Zheng, Lin Song, Haoyang Huang, Nan Duan, and Wenbo Li. Thinking with Novel Views: A Systematic Analysis of Generative-Augmented Spatial Intelligence, May 2026b. URL http://arxiv.org/abs/2605.10588. arXiv:2605.10588 [cs.CV].

Zirui Zhang, Haoyu Dong, Kexin Pei, and Chengzhi Mao. R-c2: Cycle-consistent reinforcement learning improves multimodal reasoning. *arXiv preprint arXiv:2603.25720*, 2026c.

Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12405–12414. IEEE, 2025.

# A SUPPLEMENTARY RESULTS

## A.1 SPATIAL INTELLIGENCE REASONING: MINDCUBE

An interesting observation is that different spatial tasks expose different bottlenecks. MINDCUBE separates in-place camera turns (Rotation), movement around a scene (Around), and relative object positions across views (Among; Fig. 3A). Direct RGB+depth prompting is strongest on Around and near chance on Rotation. Structured prompting improves the displayed Rotation and Among results most. The SFT configuration is strong on Around and Among but remains near chance on Rotation – our analysis shows that this is because the benchmark contains very limited scene diversity in its training data, making it challenging for learning to generalize to the test set.

## A.2 DETAILS OF FLOWBENCH

FLOWBENCH (Liu et al., 2026b) is a robotics and computer vision benchmark that evaluates the capability of visual world models to answer questions based on specified counterfactual motions. FLOWBENCH contains diverse categories of questions, such as "occlusion" ("if we move the table 0.50 meters to the right, will it occlude the view of the chair?") and "closer to" ("if we move the eraser 5cm to the left, which three objects would it be closest to?"). FLOWBENCH contains two main splits. First, FLOWBENCH-SIM is built in a simulator (Kolve et al., 2017; Deitke et al., 2022) and contains diverse indoor scenes with full rooms depicted, and involves moving objects with various scales from chairs to potatoes. Second, FLOWBENCH-REAL was built by adding counterfactual motion questions to five robot datasets: DROID, Bridge V2, Language-Table, RT-1, and Berkeley UR5. It spans five distinct robot embodiments, with scenes ranging from laboratory tabletops to toy kitchens and kitchen counters, and moved objects ranging from colored blocks and toy food to cups, bowls, bottles, utensils, and cloths.

## A.3 DETAILS FOR BASELINES IN FLOW PREDICTION

**Baselines.** We compare our model against expert models that can map one RGB frame plus a language instruction to a dense displacement field. Each is fine-tuned on the same FlowBenchtraining set used for training Ψ-0.6, and evaluated on the identical benchmark. We first report a zero-flow predictor: under a static camera only 0.6% of pixels move, so predicting nothing is already a fine approximation, and any method whose wrong vectors exceed the true magnitude scores worse. FOFPred (Ranasinghe et al., 2026) is the closest published match, image plus action phrase to HSV-encoded flow, fine-tuned with LoRA (Hu et al., 2021) after raising its magnitude normaliser so our displacements are not clipped. As instruction editors we fine-tune InstructPix2Pix (Brooks et al., 2023) in three variants differing only in supervision target (badge-free render, Set-of-Mark render, and a mask-weighted loss on the object's before/after footprints), and Qwen-Image-Edit (Wu et al., 2025a) with 4-bit QLoRA Dettmers et al. (2023). The mask weighting targets an identity shortcut: rgb0 and rgb1 differ in under 1% of pixels on average, so a plain reconstruction loss is minimised by copying the input. Finally, a non-generative pipeline — the only baseline using the instruction's metric distance — grounds the referring expression with Grounding DINO (Liu et al., 2024b) and SAM 2 (Ravi et al., 2025), predicts metric depth with Depth Pro (Ravi et al., 2025), and applies the displacement as a 3-D rigid translation.

**Read-out and oracles.** Editors emit images, not flow, and correspondence models cannot convert them here: with an identical camera the correct warp is the identity, so given the true rgb1, SEA-RAFT (Wang et al., 2024), RoMa (Edstedt et al., 2024) and CoTracker3 (Karaev et al., 2025) recover under 25% of the true magnitude. We instead localise the object by frame-difference proposals scored with DINOv2 descriptors (Oquab et al., 2023) and refined by SAM 2. This paints flow on the ground-truth frame-0 mask, so editors are given that mask while PSI and FOFPred must also decide which pixels move — the groups are not directly comparable and PSI's margin is understated. **Notably, Ψ-0.6 does not need any of these oracles to perform well on FlowBench**, as it is able to read the answer off its own prediction.

Table 3: Accuracy on FlowBench-Sim (%). Closest-3 requires the exact set of three. Cat. avg is the unweighted mean over the 8 categories.

| Model | Closer | Closest_3 | Contact | Count | Feasibility | L/R | Near | Occl. | Avg | Cat. Avg |
|---|---|---|---|---|---|---|---|---|---|---|
| Chance | 25.0 | 5.0 | 50.0 | 25.0 | 50.0 | 50.0 | 50.0 | 50.0 | 36.7 | 38.1 |
| Qwen3-VL-8B (Bai et al., 2025) | 47.9 | 0.9 | 40.0 | 26.7 | 55.7 | 80.2 | 62.4 | 68.5 | 47.1 | 47.8 |
| Qwen3.5-9B (Qwen Team, 2026) | 20.2 | 0.0 | 46.0 | 27.6 | 51.9 | 50.9 | 48.7 | 34.2 | 34.2 | 34.9 |
| Qwen3-VL-32B (Bai et al., 2025) | 27.7 | 0.0 | 28.0 | 24.1 | 59.4 | 81.9 | 55.6 | 45.2 | 41.1 | 40.2 |
| Qwen3-VL-8B FlowBench-FT | **80.9** | **41.9** | **88.0** | **46.6** | **83.0** | 94.8 | **85.5** | **82.2** | **73.6** | **75.4** |
| Ψ-0.6 Base | 23.4 | 0.0 | 24.0 | 12.1 | 9.4 | 45.7 | 50.4 | 31.5 | 24.5 | 24.6 |
| Ψ-0.6 FT | 64.9 | **41.9** | 84.0 | 43.1 | 82.1 | **97.4** | 74.4 | 74.0 | 68.8 | 70.2 |

### A.4 VQA WITH Ψ-0.6 ON FLOWBENCH.

Fine-tuning on FLOWBENCH substantially improves the VQA performance of Ψ-0.6 across all eight categories (Table 3). Overall accuracy increases from 24.5% to 68.8%. The fine-tuned model outperforms all evaluated off-the-shelf Qwen models on both aggregate metrics and achieves the second-highest overall accuracy, trailing Qwen3-VL-8B fine-tuned on the same dataset by 4.8 points. At the category level, Ψ-0.6 achieves the highest L/R accuracy (97.4%) and matches the best Closest-3 accuracy (41.9%), where a prediction is correct only if it recovers the exact set of three objects. Although fine-tuned Qwen3-VL-8B retains an advantage in the remaining six categories, these results demonstrate strong abilities in Ψ-0.6's spatial question answering, showcasing Ψ-0.6's capability serving as a VLM.

### A.5 FLOW PREDICTION ABLATION

To improve text-to-flow predictions, Ψ-0.6 further employs two powerful decoding techniques. Critically, these are enabled by Ψ-0.6's fine-grained control and probabilistic predictions, attributes that are absent from many other world models. First, before generating flow, Ψ-0.6 first computes the probability of motion ("*pMotion*") at every patch with a single feedforward pass. Then, Ψ-0.6 generates flow predictions sequentially from patches most-likely to move to least-likely. The intuition behind this is In FLOWBENCH, motion typically involves a single, often small object, so predicting flow for the patches most likely to move first helps the model focus on the relevant region.. Second, our model uses classifier-free guidance (CFG) (Ho & Salimans, 2022): at every generated token, compare the flow prediction logits with and without conditioning on the text instruction, then amplify the difference – this amplifies the effect of the instruction.

Table 4: Ablation study of PSI-0.6 models: effect of pMotion, CFG scale, and sequence length on optical flow prediction. $n = 622$.

| Method | Acc↑ | $\text{rEPE}_{\text{move}}$ ↓ | $\text{rEPE}_{\text{move}}$ med ↓ | Blank ↓ | Dir (°) ↓ | ΔPSNR ↑ | Oracle Input |
|---|---|---|---|---|---|---|---|
| Ψ-0.6 Default seq 32 | 31.8% | 0.635 | 0.639 | 27.5% | 14.1 | **+1.60** | – |
| + pMotion | 44.2% | 0.577 | 0.474 | 34.6% | **6.6** | +1.59 | – |
| + pMotion + CFG1.5 | 54.0% | 0.463 | 0.363 | 5.6% | 7.7 | +1.49 | – |
| + pMotion + CFG2.0 | **56.4%** | **0.454** | 0.356 | 0.3% | 8.1 | +0.94 | – |
| + pMotion + CFG2.5 | **56.4%** | 0.463 | **0.347** | **0.0%** | 9.1 | +0.25 | – |
| + pMotion + CFG1.5 + seq 128 | 54.2% | 0.462 | 0.364 | 5.6% | 7.8 | +1.45 | – |
| + pMotion + CFG1.5 + seq full | 54.2% | 0.462 | 0.364 | 5.3% | 7.8 | +1.39 | – |

Table 4 shows that pMotion raises accuracy from $31.8\%$ to $44.2\%$ and reduces directional error from $14.1°$ to $6.6°$. CFG at scale 1.5 further improves accuracy to $54.0\%$, lowers $\text{rEPE}_{\text{move}}$ to $0.463$, and reduces the blank rate to $5.6\%$. Stronger guidance further improves flow metrics but compromises reconstruction: scale 2.5 eliminates blanks while reducing reconstruction gains from $+1.49$ to $+0.25$ dB. Decoding beyond 32 tokens yields no further improvement, likely due to FlowBench's sparse, single-object motions. These results favor pMotion with moderate CFG as the best overall configuration, without requiring longer decoding.

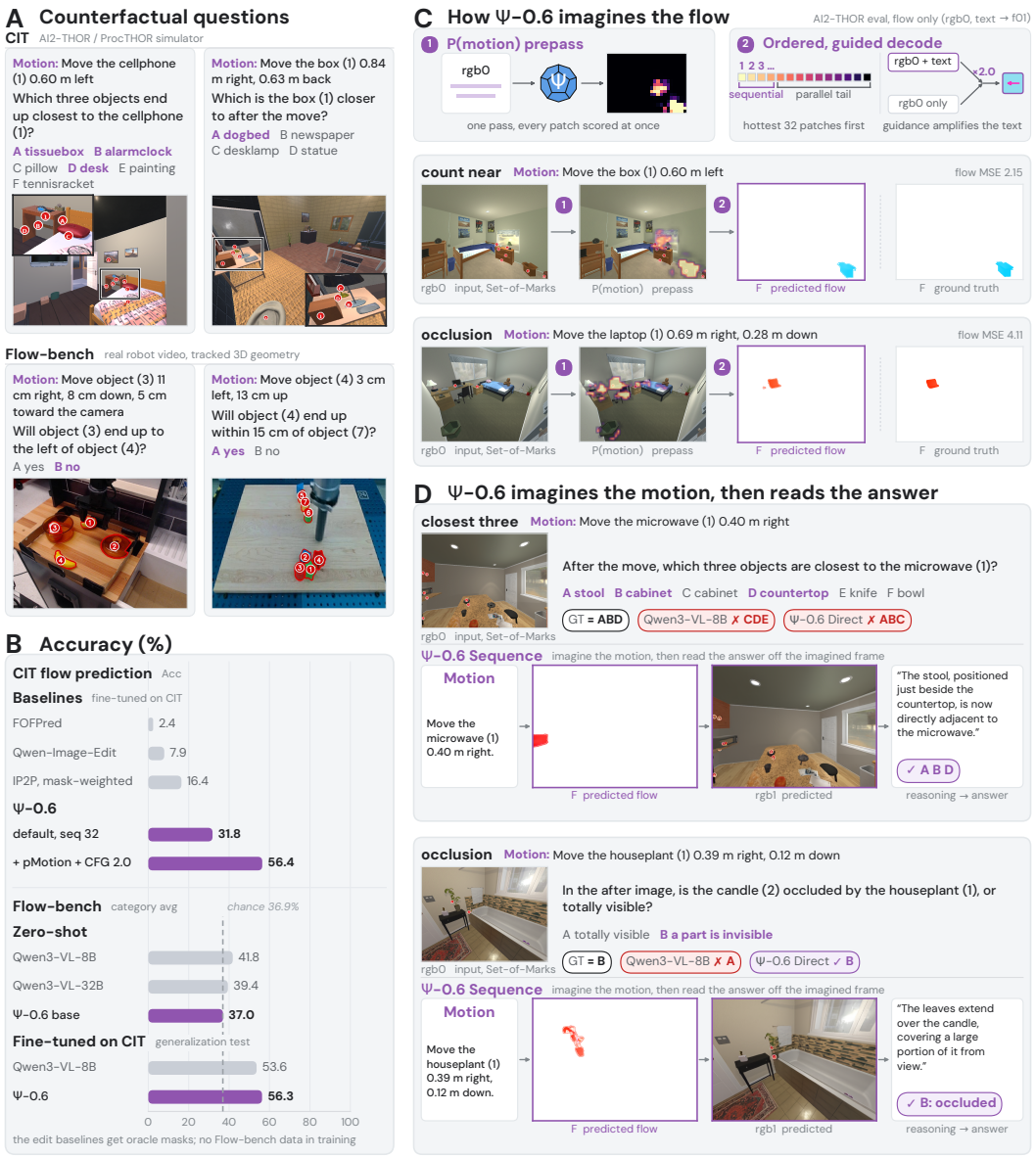


Figure 6: **Counterfactual imagination of object motion.** **(A)** The FlowBench benchmark (Liu et al., 2026b) has two parts: FlowBench, created in the AI2-THOR simulator, and Flow-bench, created using real-world robotics datasets. **(B)** Average accuracy. **(C)** Ψ-0.6 is capable of counterfactual imagination: it simulates the intervention, predicts the consequence, and reads the answer off its own prediction. The two examples show complete rollouts of Ψ-0.6's Sequence reasoning on questions of FlowBench.

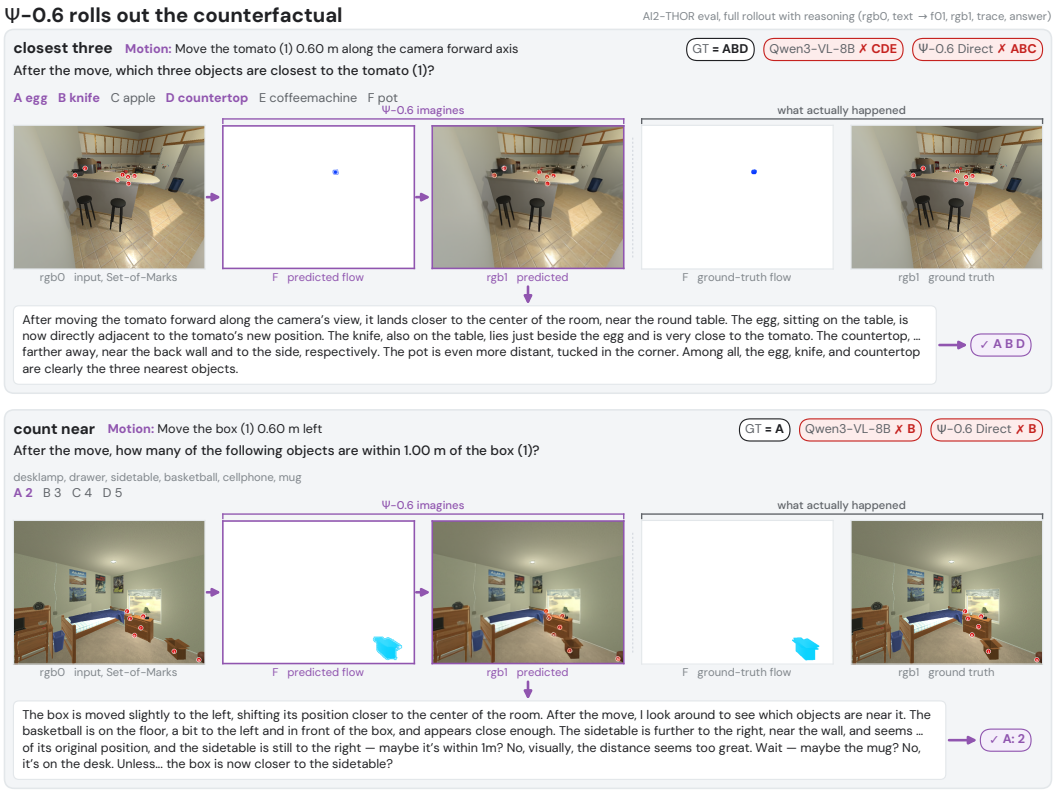


Figure 7: Ψ-0.6's rollout on FlowBenchquestions. The two examples show complete streams of thought of Ψ-0.6's Sequence reasoning on FlowBenchquestions. Ψ-0.6 first predicts the optical flow of the motion, then generates the imagination image, and finally answers the question with text reasoning – a process similar to human mindset.

# B SUPPLEMENTARY TEXT

## B.1 PRELIMINARIES ON PROBABILISTIC STRUCTURE INTEGRATION (PSI)

Ψ-0.6 builds on Probabilistic Structure Integration (PSI) (Kotar et al., 2025), inheriting two further properties that make its prompting interface unusually flexible: *fine-grained control* over individual patches and *probabilistic prediction*, which allows unspecified content to remain open to multiple plausible continuations.

**Fine-grained control.** Ψ-0.6 represents visual fields as independently encoded spatial patches, using *local patch quantization* together with explicit *pointer tokens*. Each RGB, flow, or depth patch is represented by content tokens describing *what* the patch contains and a pointer token specifying its modality, spatial location, and temporal location—that is, *where* and *when* the content belongs. Because content and location are represented separately, a prompt can address only selected regions of a visual field, leave the remaining regions unspecified, or relocate existing content by changing its pointer while preserving its content tokens. This enables fine-grained conditioning and targeted scene edits. For example, in Fig. 2B, moving a patch on the upper part of a can downward while keeping the base fixed specifies a compression, whereas moving it upward without fixing the base specifies a lift. The same representation also allows patches to be generated in arbitrary spatial order, including sequentially (autoregressively), in parallel, or through a combination of both.

**Probabilistic prediction.** Prompts specify constraints rather than a single deterministic target. For any partially specified scene, Ψ-0.6 represents a distribution over compatible continuations, allowing the same prompt to yield alternative predictions through sampling. In this sense, the model acts as a *multiverse generator* over a *garden of forking paths* (Moire, 2021). Prompts control both *what is constrained* and *what is left for the model to infer*.

## B.2 RELATED WORK AND FURTHER DISCUSSION

**Streams-of-thought combine the twin advantages of interpretability and low-dimensionality.** If representations become reasoning operations, the form of those representations matters. Existing approaches to visual reasoning largely occupy two ends of a spectrum: "thinking with images/video" reasons by generating or editing *pixels*, making intermediate states directly interpretable (Wiedemer et al., 2025; Tong et al., 2026; Li et al., 2025b; Chern et al., 2025; Bian et al., 2026), whereas *latent* reasoning compresses intermediate computation into continuous representations that are difficult to inspect directly (Wang et al., 2025; Li et al., 2025a; Yang et al., 2025; Zhang et al., 2025; Wu et al., 2026a). Ψ-0.6 occupies a middle ground by *externalizing* compact intermediates as physically meaningful tokens. Camera pose, depth, optical flow, and point tracks are far lower-dimensional than pixels while remaining interpretable to both humans and the model itself. MME-CoF (Guo et al., 2025) finds that chain-of-frames video models can maintain short-horizon coherence while struggling with strict geometric constraints and long-horizon causal reasoning, motivating representations that expose geometry and motion more directly. PAN (Xing et al., 2026) similarly argues that discrete representations can preserve distinctions that latent-only objectives may collapse. Streams-of-thought therefore offer a complementary synthesis for visual reasoning: compact enough to abstract away unpredictable pixel-level detail, yet explicit enough to expose the physical structure on which reasoning depends.

**When the reasoner and the imaginer are the same model, imagination becomes a first-class reasoning operation.** The value of a unified predictive model becomes particularly clear when reasoning requires considering a world that has not yet been observed. A growing line of work augments language-based reasoners with external visual generators or world models. Mindjourney (Yang et al., 2026) invokes a frozen Stable Virtual Camera (Zhou et al., 2025); SandboxVLM further introduces expert models and 3D reconstruction (Liu et al., 2026a); Thinking with Novel Views (Zhang et al., 2026b) uses an external "Painter"; APC (Lee et al., 2025a) chains multiple foundation models; and AVIC (Yu et al., 2026) learns when to invoke a world model. These approaches demonstrate the value of visual imagination, but retain a separation between the system that reasons and the system that imagines. In Ψ-0.6, the model that interprets an observation is also the model that predicts its geometry, motion, and possible transformations. Simulation is therefore not a separately invoked tool: imagined states are generated and consumed within the same model and token interface as the rest of the reasoning process. World2VLM (Zhang et al., 2026a) finds benefits from distilling imagined

experience into a VLM rather than relying only on runtime coupling; Ψ-0.6 pushes this integration further, because the reader and the imaginer are already the same weights.

**Functional unification does not guarantee learning synergy.** The inference-time benefits of unification do not imply that every capability necessarily benefits from being trained together. Large-scale multitask language models provide broad evidence for positive transfer, with training across diverse tasks improving generalization beyond the individual training tasks (Longpre et al., 2023). At the same time, multilingual models exhibit a tradeoff between positive transfer and *capacity dilution* when many capabilities compete for fixed shared capacity (Conneau et al., 2020). The same tension is emerging in unified multimodal models. RealUnify finds that architectural unification alone does not reliably produce bidirectional synergy between understanding and generation (Shi et al., 2026), while controlled experiments find positive transfer when tasks rely on shared knowledge but interference when conflicting computations are forced through the same pathway (Wu et al., 2026b). Janus addresses such conflicts through partial specialization while retaining a shared transformer (Wu et al., 2025b), whereas Janus-Pro shows that increasing data and model scale can improve both understanding and generation (Chen et al., 2025). Ψ-0.6 demonstrates *functional unification* and inference-time gains from composing capabilities, but we do not isolate positive and negative transfer among those capabilities during training. In principle, competition need not be a fundamental cost of unification: with sufficient model capacity and training, a unified system can allocate computation to both shared and specialized functions, echoing the brain as an existence proof that an integrated system can support functionally specialized subsystems.

**Physical consistency could provide the verifier needed to discover new reasoning representations.** Beyond understanding these training-time interactions, a more speculative direction is to make the reasoning vocabulary itself learnable. Ψ-0.6 currently reasons through deliberately chosen abstractions—camera pose, depth, optical flow, and point tracks—whose physical meaning is known in advance. Adaptive streams-of-thought learn *which* of these representations to use and *when*; future systems could instead ask *which intermediate representations should exist at all*. Physical reasoning may provide an unusually rich source of supervision for this problem because different representations of the same world must satisfy shared constraints. Depth, camera motion, optical flow, point tracks, and appearance should agree across viewpoints and time, while forward and inverse predictions create additional cycles that can be checked. For example, a camera motion inferred from two views should generate a compatible second view when supplied back to the model. Recent work shows that cross-modal cycle consistency can itself provide a dense, label-free reward for multimodal reasoning (Zhang et al., 2026c). Such signals suggest a route toward generic *physical verifiers*: mechanisms that score intermediate reasoning through consistency with the structure of the world rather than requiring human labels for every trajectory. Verification could in turn support the *autodiscovery* of reasoning intermediates. A model could propose compact intermediate variables, evaluate whether reasoning through them improves predictive consistency or downstream task performance, and retain useful abstractions within its own reasoning interface. Work on latent-action world models already shows that useful transition representations can emerge from unlabeled video rather than being specified a priori (Gao et al., 2025; Ren et al., 2025). Extending this idea from action representations to reasoning representations suggests a closed loop: propose an abstraction, evaluate the consequences of using it, and reincorporate successful abstractions into future computation. In this view, not only the reasoning policy, but the vocabulary through which a model reasons, can itself become an object of learning.

**Adaptive streams-of-thought naturally extend from visual proxy actions to embodied control.** A complementary and more concrete future direction is to extend the framework from reasoning about actions to acting in the physical world. In Ψ-0.6, optical flow and point tracks already function as *proxy actions*: they specify how parts of a visual scene should move, but do not directly actuate the world. One immediate application to robotics can be achieved by training an inverse kinematics (IK) model on the flow prediction or point prediction, similar to the RGB-based IK models like Goswami et al. (2025); Liu et al. (2025); Ban et al. (2024). On the high level, robot joint commands, end-effector poses, gripper states, and other control signals could analogously become additional token modalities. Policy prediction, world simulation, and progress estimation could then become different queries to the same model, extending the unified interface from perception and imagination to physical control. Combining such action tokens with *adaptive streams-of-thought* would enable a richer form of embodied reasoning. Rather than mapping observations directly to commands, the model could

learn when to inspect geometry, predict motion, imagine the consequences of an action, evaluate progress, seek another observation, or act. Our current adaptive-stream experiments select among a predefined set of representations and optimize their use on visual question answering rather than physical control. They also optimize downstream task performance without explicitly accounting for the computational cost of reasoning. In an embodied agent, reward could jointly optimize *what action to take*, *what reasoning to perform before acting*, and *how much computation that reasoning should consume*. This would extend task-adaptive streams-of-thought from choosing how to reason about the world to choosing how to reason *and act* within it.