VISUAL-OPSD: CROSS-MODAL ON-POLICY SELF-DISTILLATION FOR EFFICIENT UNIFIED MULTIMODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified multimodal models (UMMs) interleave generated "visual thoughts" (VTs) with text reasoning to improve spatial tasks, at roughly an order-of-magnitude inference cost from multi-step diffusion. We find that this cost buys little at inference: on ThinkMorph, removing or noising the intermediate VTs leaves accuracy essentially unchanged across nine benchmarks, and once a VT is rendered the subsequent reasoning attends to it regardless of what it contains. A KL diagnostic nonetheless shows that conditioning on a privileged VT trace shifts the model's completion distribution, which places the useful reasoning in the generation pathway rather than in the rendered pixels. We exploit that gap with Visual On-Policy Self-Distillation (Visual-OPSD). Teacher and student share identical weights and differ only in context: the teacher attends to privileged VTs, the student only to the problem image and question. Token-level JSD distillation along on-policy student trajectories moves the teacher's reasoning into a text-only student. Across nine benchmarks, Visual-OPSD improves on its generative teacher by $+3.40$pp while running $14.3\times$ faster (10.0s vs. 142.8s per sample), and exceeds same-scale VLMs by $+63.83$pp on VSP. Replacing real VTs with Gaussian noise recovers only $+0.40$pp of the $+10.28$pp gain, and distillation closes 58.4% of the measured KL gap on held-out samples.

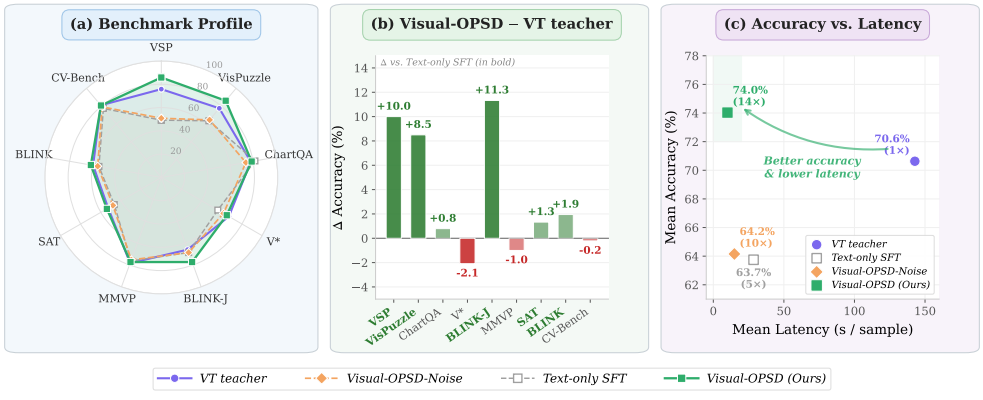


Figure 1: **Visual-OPSD matches its VT-generating teacher at $14\times$ lower latency.** **(a)** Radar over 9 tasks: Visual-OPSD (green) $\geq$ teacher (purple) on 6/9. **(b)** Largest per-task gains: VSP $+10.0$, VisPuzzle $+8.5$, BLINK-J $+11.3$. **(c)** Accuracy–latency Pareto: Visual-OPSD 74.0%/10.0s vs. teacher 142.8s.

1 INTRODUCTION

Unified multimodal models (UMMs) (Deng et al., 2025; Gu et al., 2026; Chameleon Team, 2024; Wang et al., 2024) handle visual understanding and generation within a single set of weights, and they turn out to support interleaved visual chain-of-thought reasoning: the model alternates text

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.

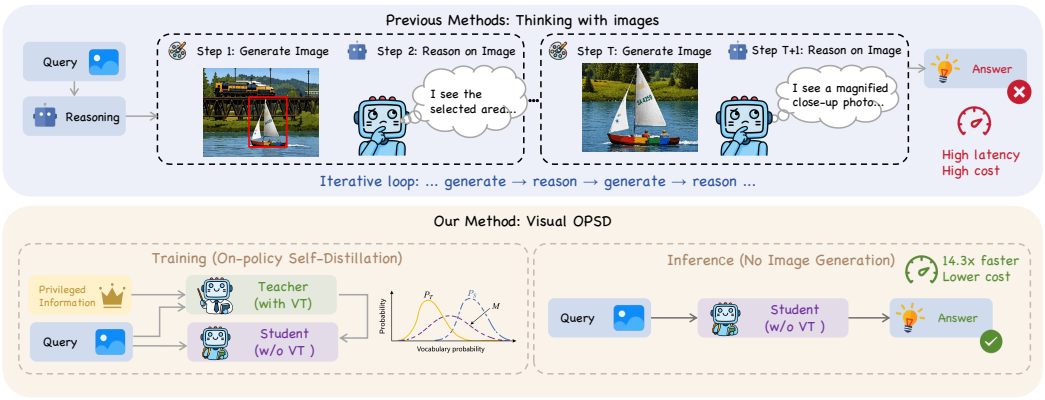


Figure 2: **Prior interleaved visual CoT vs. Visual-OPSD. (Top)** Previous methods iterate *generate-then-reason*, rendering each visual thought (VT) via 50-step diffusion at high latency and cost. **(Bottom)** Visual-OPSD distills the generation pathway into a text-only student via cross-modal on-policy self-distillation, yielding $+3.40$pp accuracy at $14.3\times$ speedup with no image generation at inference.

segments with intermediate "visual thoughts" (VTs) generated by multi-step diffusion before committing to an answer. ThinkMorph (Gu et al., 2026) reports consistent spatial-reasoning gains from this protocol over text-only reasoning, which it attributes to richer intermediate representations.

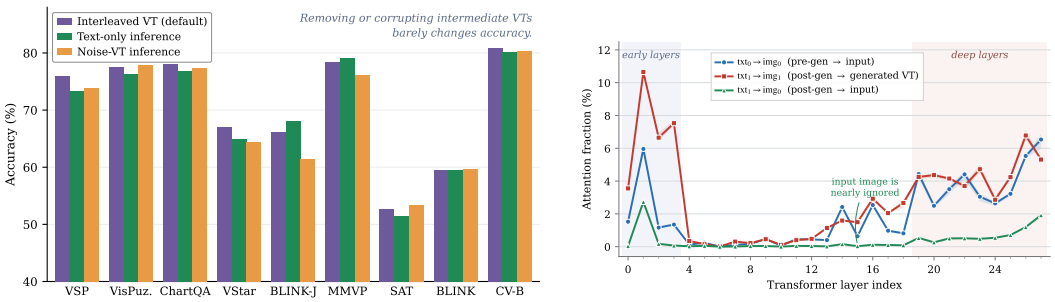
The protocol is expensive. Each VT costs 50 diffusion denoising steps, putting per-sample latency roughly an order of magnitude above text-only reasoning. The cost raises a question prior to efficiency: do the rendered VT pixels carry load-bearing information at all, or do the trajectory-level gains come from somewhere else?

We investigate this with a controlled pilot study on ThinkMorph (Figure 3a) that intervenes on intermediate VTs at inference time without retraining. Removing the VTs entirely and letting the model reason in text alone preserves most of the accuracy across nine benchmarks (69.90 vs. 70.63 on average), and on BLINK-J and MMVP text-only reasoning is slightly ahead of the full interleaved setting; replacing real VTs with Gaussian noise behaves much the same. A per-layer attention analysis on V* (Figure 3b) adds that the text reasoning following a VT attends almost exclusively to that VT and largely ignores the original input, whatever the VT depicts. The rendered pixels therefore contribute little beyond what text-only reasoning already recovers, and the diffusion cost is not commensurate with their direct benefit. What generation training does buy shows up elsewhere: even in text-only mode ThinkMorph averages 69.90, well above same-scale dedicated VLMs that were never trained with a generation pathway (InternVL3.5-8B 58.04, Qwen3-VL-8B 64.58). The value of that pathway thus sits somewhere other than in the pixels it renders.

Our hypothesis is that it sits in the internal representations that generation training shaped. To test it we measure the KL divergence between the model's completion distributions with and without a privileged VT trace. The divergence is large in every task category (Section 2.2), so a distributional gap is available for distillation to exploit even though the VTs themselves are not load-bearing at inference.

Building on these observations, we propose Visual On-Policy Self-Distillation (Visual-OPSD), a cross-modal on-policy self-distillation framework that exploits the above distributional gap within a single set of weights. Teacher and student share identical parameters and differ only in conditioning context: the teacher attends to privileged VT images, the student only to the problem image and question. Token-level JSD distillation along on-policy student trajectories transfers the generation pathway's distributional knowledge into the student. At inference the student runs in text-only mode, with no diffusion steps, architectural changes, or additional parameters.

Switching ThinkMorph to text-only inference already removes the diffusion latency at no cost, so Visual-OPSD has to justify itself against that option and not only against the interleaved teacher.



(a) Inference-time VT intervention on ThinkMorph.

(b) Per-layer cross-modal attention on V*.

Figure 3: **Two diagnostics on ThinkMorph that motivate Visual-OPSD. (a)** Removing or corrupting intermediate VTs at inference leaves accuracy largely unchanged across all nine benchmarks. **(b)** Once generated, a VT dominates the subsequent reasoning attention regardless of its content.

It does: the distilled student averages 74.03 against 69.90 for zero-shot text-only inference on the same weights, a gain of +4.13pp.

Our contributions are as follows:

- **Finding.** Controlled interventions on ThinkMorph show that rendered VT pixels are not load-bearing at inference, while the generation pathway still carries a large distributional signal that a KL diagnostic can measure. Visual generation training and inference utility come apart in UMMs, which to our knowledge has not been examined.
- **Method.** We propose Visual-OPSD, a cross-modal on-policy self-distillation framework that transfers this distributional knowledge from a VT-conditioned teacher to a text-only student within a single model. To our knowledge, this is the first OPSD instance bridging the asymmetry between generation and understanding in a unified multimodal model.
- **Results.** Across nine benchmarks, Visual-OPSD preserves or improves accuracy on 6 of 9 tasks (+3.40pp on average) while cutting per-sample inference time by 14.3×. A noise control gains only +0.40pp over text-only fine-tuning, against +10.28pp for Visual-OPSD, which places the source of the gain in the semantic content of the generation pathway rather than in regularization.

2 METHOD

2.1 PRELIMINARIES

Unified multimodal model. Visual-OPSD applies to any UMM that supports both visual understanding and generation. We instantiate it on ThinkMorph (Gu et al., 2026), built on the BAGEL architecture (Deng et al., 2025), which fuses a Qwen2.5 LLM backbone with MoT decoder layers, a SigLIP-so400m NaViT vision encoder, and a FLUX VAE for latent image encoding and generation, so that one set of weights covers both image→text understanding and text→image generation.

Interleaved chain-of-thought protocol. The model performs reasoning via an interleaved protocol in which text and generated images alternate. Writing x for the problem input (image and question), the thought sequence is $\mathcal{T} = (\hat{m}_1, \ldots, \hat{m}_n)$ with $\hat{m}_i \sim \mathcal{P}_\theta(\cdot \mid x, \hat{m}_{<i})$, where each $\hat{m}_i \in \{\hat{t}_i, \hat{v}_i\}$ is either a text thought or a visual thought. Special tokens are omitted from this notation, but modality transitions are controlled in practice by delimiters: image thoughts are bracketed by `<image_start>` and `<image_end>`, which lets the model switch between textual and visual reasoning within a single sequence. Section 2.3 uses the same pair to suppress visual thoughts during sampling.

Generation cost. Each VT costs 50 denoising steps through the diffusion pathway, roughly a 14× latency overhead: 142.8s per sample with VT generation against 10.0s without. The goal is to keep the knowledge that generation training produced while dropping the generation step at inference.

2.2 MEASURING GENERATION KNOWLEDGE: KL DIAGNOSTIC

We first formalize and measure the "generation knowledge" hypothesis. Consider the frozen unified model M with parameters θ. We construct two forward passes that share *identical completion tokens* $\mathbf{c} = (c_1, \ldots, c_T)$ but differ in their conditioning context:

$$\mathcal{C}_T = [\text{sys},\ \text{ViT(img)},\ \text{question},\ \text{ref_intro},\ (\text{ViT}(\text{VT}_i))^+,\ \text{transition}] \tag{1}$$

$$\mathcal{C}_S = [\text{sys},\ \text{ViT(img)},\ \text{question}] \tag{2}$$

The teacher context $\mathcal{C}_T$ prepends a *strictly visual-only* privileged reasoning trace before the completion. Here, ref_intro is a preamble that frames the subsequent images as privileged visual references, and transition instructs the model to now reason independently from this privileged context (full prompts in Appendix E). The student context $\mathcal{C}_S$ contains only the problem image and question. Both paths autoregressively process the *same* completion tokens $\mathbf{c}$ yet produce different next-token distributions at each position, $p_\theta(c_t \mid c_{<t}, \mathcal{C}_T) \neq p_\theta(c_t \mid c_{<t}, \mathcal{C}_S)$. We define the generation knowledge for a sample as the average per-token KL divergence over the shared completion span:

$$\mathcal{K}_{\text{gen}} \triangleq \frac{1}{T} \sum_{t=1}^{T} D_{\text{KL}}\Big(p_\theta(\cdot \mid c_{<t}, \mathcal{C}_T) \ \Big\|\ p_\theta(\cdot \mid c_{<t}, \mathcal{C}_S) \Big) \tag{3}$$

This quantity measures how much the VT reasoning trace shifts the model's next-token predictions on a fixed completion, and serves as a proxy for the size of the distributional gap that Visual-OPSD training seeks to close. The noise-control variant (Section 3.3) and the post-distillation gap-closing analysis (Appendix I) further attribute this gap to the semantic content of the VT reasoning trace.

We evaluate the diagnostic on a held-out set of 1,000 examples (250 per category) spanning the four task categories, disjoint from the data used for training:

Table 1: KL diagnostic on shared completion tokens, measured on 1,000 held-out examples. Large $\mathcal{K}_{\text{gen}}$ indicates substantial VT-encoded knowledge in the completion distribution, which is the learning signal available to Visual-OPSD.

Task Category	**Samples**	$\mathcal{K}_{\text{gen}}$ **(nats/token)**
Visual Search	250	4.23
Spatial Navigation	250	3.96
Chart Refocus	250	3.51
Jigsaw Assembly	250	6.84
Overall	**1,000**	**4.64**

All categories exhibit $\mathcal{K}_{\text{gen}} \gg 0$ (overall 4.64 nats/token), indicating that the VT reasoning trace systematically shifts the model's completion predictions. Jigsaw Assembly shows the largest gap (6.84 nats/token), consistent with spatial-manipulation tasks deriving the most benefit from intermediate visual reasoning.

A per-token analysis (Figure 6 in Appendix F) shows the divergence to be *non-uniform*: it concentrates on tokens encoding spatial relations, quantities, and visually grounded answers, while function words carry near-zero divergence. VT reasoning therefore informs the predictions that decide task success rather than shifting overall style. Appendix I shows that Visual-OPSD training closes 58.4% of this gap on the same held-out set.

2.3 VISUAL-OPSD: CROSS-MODAL ON-POLICY SELF-DISTILLATION

Cross-modal information gap. Visual-OPSD distills generation knowledge by exploiting the cross-modal information gap between teacher and student contexts within the same model. The two sequences are $\mathcal{C}_T$ and $\mathcal{C}_S$ from Eqs. 1–2, each followed by the same completion span, and loss is active on that span alone. The privileged channel is *strictly visual-only*: the teacher's extra context holds the intermediate VT images (encoded via ViT) and nothing else. Since the completion tokens are identical on both sides, the teacher's logits differ only because its KV cache encodes those images, and that difference is the generation pathway's distillable knowledge. Figure 4 shows the training loop; Algorithm 1 in Appendix C gives the pseudocode.

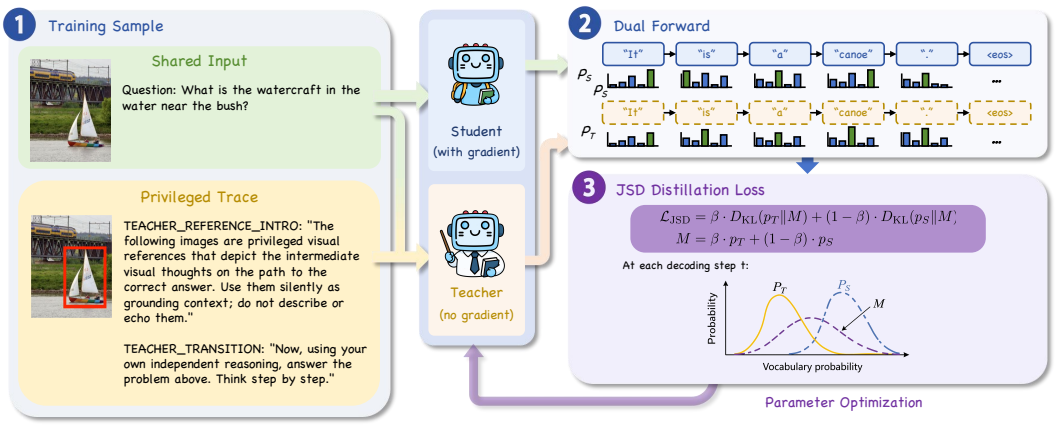


Figure 4: **Overview of Visual-OPSD.** From the same UMM, a student $\pi_\theta(\cdot|\mathcal{C}_S)$ (gradients on) sees only $[\text{sys}, \text{ViT}(x), q]$, while an EMA teacher $\pi_{\tilde{\theta}}(\cdot|\mathcal{C}_T)$ (no gradient) additionally receives privileged visual thoughts $(\text{ViT}(\hat{v}_i))^+$. The student samples $\hat{c} \sim \pi_\theta$ on-policy; both policies rescore the shared completion to yield $p_S^{(t)}, p_T^{(t)}$, optimized by per-token JSD. At inference, the student runs text-only with no VT generation, $14.3\times$ faster, and $+3.40$pp over the generative teacher.

On-policy sampling. At each step the student generates a completion from its current policy rather than reading ground-truth text. When the student emits `<image_start>`, which would enter generation mode, we immediately append the closing `<image_end>` to the running context and resume text sampling, so the image span is emptied and the diffusion pathway is never invoked; this keeps sampling on-policy while preventing collapse into generation mode. The injected `<image_end>` steers sampling only and never enters the completion $\mathbf{c}$ that the two policies later rescore. No divergence term is therefore computed at that position, and the teacher is never asked to score a position where its context leads it to expect latent visual tokens.

Training objective. Given an on-policy completion $\mathbf{c} = (c_1, \ldots, c_T)$ sampled from the student, both teacher and student evaluate this shared sequence and produce next-token distributions $p_T^{(t)} \triangleq p_{\tilde{\theta}}(\cdot \mid c_{<t}, \mathcal{C}_T)$ and $p_S^{(t)} \triangleq p_\theta(\cdot \mid c_{<t}, \mathcal{C}_S)$ at each position t. We minimize the trajectory-averaged, token-wise generalized Jensen–Shannon divergence:

$$\mathcal{L}(\theta) = \frac{1}{T} \sum_{t=1}^{T} \text{JSD}_\beta \left(p_T^{(t)} \,\|\, p_S^{(t)} \right), \qquad \text{JSD}_\beta(p_T \| p_S) = \beta D_{\text{KL}}(p_T \| m) + (1-\beta) D_{\text{KL}}(p_S \| m), \quad m = \beta p_T + (1-\beta) p_S, \tag{4}$$

where $\beta = 0.5$ gives symmetric JSD (our default) and gradients flow only through p_S. Standard distillation interpolates soft targets with hard CE labels; here the teacher–student gap is one of *information* rather than capacity, so we use pure JSD and no CE term that would anchor the student to fixed ground-truth completions.

Top-K support. For memory we restrict the divergence at each position to the teacher's top-K token indices $\mathcal{V}_K^{(t)}$ ($K = 256$), cutting the vocabulary dimension from 152K to 256 with negligible effect on distillation quality (Appendix J). Both arguments are renormalized over that support, $\tilde{p}^{(t)}(v) = p^{(t)}(v) / \sum_{u \in \mathcal{V}_K^{(t)}} p^{(t)}(u)$, so each is a proper distribution on $\mathcal{V}_K^{(t)}$ and the divergence is well defined. A position whose sampled token c_t falls outside $\mathcal{V}_K^{(t)}$ is not skipped: the objective matches distributions rather than scoring the realized token, so such a position contributes like any other, with the off-support mass of $p_S^{(t)}$ removed by renormalization.

Pointwise clipping. Raw per-position divergence is heavily skewed, and a few extreme disagreements can dominate the gradient. We clip *pointwise in the vocabulary* rather than per token, giving

the implemented objective

$$\mathcal{L}(\theta) = \frac{1}{T}\sum_{t=1}^{T}\sum_{v\in\mathcal{V}_K^{(t)}} \min\big(d_t(v),\ \tau\big), \qquad \tau = 0.05,$$
$$d_t(v) = \beta\,\tilde{p}_T^{(t)}(v)\log\frac{\tilde{p}_T^{(t)}(v)}{m_t(v)} + (1-\beta)\,\tilde{p}_S^{(t)}(v)\log\frac{\tilde{p}_S^{(t)}(v)}{m_t(v)} \ \geq\ 0, \tag{5}$$

with $m_t = \beta\tilde{p}_T^{(t)} + (1-\beta)\tilde{p}_S^{(t)}$. At $\beta = 0.5$ a single entry contributes at most $\frac{1}{2}\log 2 \approx 0.347$ nats, so τ sits at 14% of the per-entry range and the clamp binds only where the two policies disagree sharply on one alternative. A position whose divergence is spread across many entries, which Figure 6 associates with content tokens, keeps its full token-level signal and its total may exceed τ by orders of magnitude, because the clamp applies entry by entry. Eq. 5 is a clipped surrogate rather than a divergence; it agrees with Eq. 4 exactly on positions where every entry falls below τ. Removing the clamp costs 3.2pp of average accuracy on the three spatial benchmarks of Appendix J.

2.4 A Distributional View of Teacher–Student Transfer

Visual-OPSD exceeds its generative teacher by $+3.40$pp while using less information at inference (Table 2). A VT-conditioned prediction carries both useful reasoning and whatever errors the generated image introduced, and Figure 3b shows that later reasoning concentrates on that image, which can amplify the errors. Ensemble distillation gives a distributional account of transferring teacher information without conditioning inference on any single generated VT.

Fix the input (X, Q), completion prefix $c_{<t}$, and teacher weights $\bar{\theta}$. For $V \sim p_{\text{gen}}(\cdot \mid X, Q)$, write $p_V(v) = p_{\bar{\theta}}(v \mid c_{<t}, \mathcal{C}_T(V))$ for the teacher's probability of vocabulary entry v, and define the arithmetic-mean teacher $\bar{p}(v) = \mathbb{E}_V[p_V(v)]$. Consider strictly positive teacher distributions on a finite vocabulary and a student distribution q that is freely optimized on the probability simplex.

Theorem 1 (Ensemble distillation targets)**.** *For every strictly positive student distribution q,*

$$\mathbb{E}_V[D_{\mathrm{KL}}(p_V\|q)] = \mathbb{E}_V[D_{\mathrm{KL}}(p_V\|\bar{p})] + D_{\mathrm{KL}}(\bar{p}\|q). \tag{6}$$

Thus the expected forward-KL objective has the unique minimizer $q^\star = \bar{p}$. For symmetric JSD,

$$\mathbb{E}_V[\mathrm{JSD}_{1/2}(p_V\|q)] \geq \mathrm{JSD}_{1/2}(\bar{p}\|q). \tag{7}$$

The JSD inequality gives a lower bound, not an identification of its minimizer. The proof is given in Appendix B.

The analysis concerns full distributions and a fixed teacher, whereas Visual-OPSD uses top-K truncation, clipping, and EMA updates. It supports reading distillation as the transfer of aggregate teacher information rather than the reproduction of any individual VT-conditioned prediction. The contrast between Visual-OPSD-Noise ($+0.40$pp) and Visual-OPSD ($+10.28$pp) points to semantic VT content as what makes that transfer work, and not depending on generated pixels at inference is one possible explanation for the student's advantage.

3 Experiments

3.1 Experimental Setup

Benchmarks. We evaluate on 9 benchmarks in two regimes. The **VT-useful** group covers spatial reasoning, where generation knowledge should matter most: VSP (Wu et al., 2024b) (n=120), VisPuzzle (Gu et al., 2026) (n=200), BLINK-Jigsaw (Fu et al., 2024) (n=150), SAT (Ray et al., 2024) (n=150), and BLINK (Fu et al., 2024) (n=671). The **VT-neutral** group covers general VLM tasks: ChartQA (Masry et al., 2022) (n=944), V* (Wu & Xie, 2024) (n=191), MMVP (Tong et al., 2024b) (n=150), and CV-Bench (Tong et al., 2024a) (n=700).

Compared. **BAGEL-7B** is the base UMM with no task-specific fine-tuning. **Text-only SFT** fine-tunes BAGEL-7B with CE loss on text reasoning traces without VT images for 2,000 steps; it starts

from BAGEL-7B rather than ThinkMorph-7B because its training data overlaps ThinkMorph's original training distribution, so re-fitting the same data with CE loss would measure memorization rather than the text-only training signal. Visual-OPSD-Noise and **Visual-OPSD (Ours)** both run for 1,000 steps and differ only in whether the teacher context holds Gaussian noise or real VT images. **ThinkMorph** is the original model with full VT generation at inference, which serves as teacher and cost reference; we additionally evaluate it with VT generation disabled, the zero-shot text-only mode of Figure 3a.

External Baselines. GPT-4o, GPT-5, InternVL3.5-8B/38B (Wang et al., 2025b), Qwen3-VL-8B/32B (Bai et al., 2025), Janus-Pro-7B (Chen et al., 2025), and Chameleon-7B (Chameleon Team, 2024).

Training Details. All Visual-OPSD variants are initialized from ThinkMorph-7B and trained on $8\times$H800 GPUs with FSDP (Hybrid Shard). Key hyperparameters: learning rate $1e{-}5$ with cosine decay, EMA decay 0.995, JSD $\beta{=}0.5$, top-$K{=}256$, token clip$=0.05$, max completion length 1024 tokens. The training data consists of $\sim$24K samples across 4 task categories. See Appendix C for full details.

3.2 Main Results

The student overtakes its teacher. Visual-OPSD exceeds the VT teacher on 6 of 9 benchmarks, $+3.40$pp on average, with the margin concentrated where generation knowledge should matter most: BLINK-J $+11.3$, VSP $+10.0$, VisPuzzle $+8.5$. Two of the remaining differences carry no information. ChartQA ($+0.79$pp) and CV-Bench (-0.22pp) both fall inside the run-to-run spread of ±1.27 and ±0.51 in Table 2, so we do not read a direction into them.

Transfer is task-specific. Against text-only SFT, Visual-OPSD gains $+10.28$pp on average, again concentrated on the VT-useful tasks: VSP $+36.7$, VisPuzzle $+22.5$, BLINK-J $+8.7$, SAT $+7.4$, BLINK $+7.1$. Among the VT-neutral tasks only V* moves comparably ($+8.9$); CV-Bench ($+3.3$) and MMVP ($+1.0$) move little and ChartQA drops by 2.87pp, which fits chart understanding having little to gain from spatial generation knowledge. The same spatial advantage separates Visual-OPSD from same-scale VLMs: on VSP it reaches 85.8 against 8.2 for InternVL3.5-8B and 22.0 for Qwen3-VL-8B, so generation training supplies spatial reasoning that dedicated VLMs at this scale do not have (per-task visualization in Appendix H).

Speed comes partly from brevity. Visual-OPSD runs at 10.0s per sample, $14.3\times$ faster than the VT teacher (142.8s) and $2.9\times$ faster than text-only SFT (28.5s), under identical greedy decoding (temperature$=0$, 1024-token cap) on the same hardware. Part of the second factor is output length: Visual-OPSD emits 201 tokens per sample on average against 411 for SFT and 452 for ThinkMorph's text stream (Appendix G).

That compression deserves a comment, since a student trained against a verbose teacher might be expected to inherit its verbosity. It does not, because the teacher never writes anything: it only rescores the student's own on-policy trajectory, so length is set by the student's sampling and the only pressure on it comes from the per-position targets. Two properties of the objective push those targets short. The loss is pure JSD with no CE term, so nothing anchors the student to the long reference traces that the SFT baseline is fitted on. And the teacher has already been shown the VTs, so it has little reason to spend mass on narrating what the image contains; where the student opens such a narration, matching the teacher pulls mass toward committing to an answer instead. We did not isolate this second effect on its own, but it matches the token-level picture in Figure 6, where divergence sits on content rather than function words, and it matches compression being largest on VisPuzzle and BLINK-J, the tasks whose baseline traces carry the most narration.

3.3 VT Information Quality Scaling

To separate generation knowledge from regularization, we vary the quality of the VT content in the teacher context and leave everything else fixed.

Table 2: Main results across 9 benchmarks. The three rows under *Ours* report mean ± standard deviation over 3 independent training and evaluation runs; every other row is a single evaluation. *ThinkMorph (text-only)* is the same checkpoint with VT generation disabled at inference. **Bold**/underline: best/second-best among open 7–8B models; larger and proprietary rows are for reference and excluded from ranking. VLM baselines on VStar use no external visual-search tools.

Method	VSP ↑	VisPuzzle ↑	ChartQA ↑	VStar ↑	BLINK-J ↑
Vision-Language Models (VLMs)					
InternVL3.5-8B	8.17	34.75	76.26	68.59	71.33
InternVL3.5-38B	20.16	36.50	80.44	76.96	80.67
Qwen3-VL-8B	22.00	37.00	**82.55**	**84.29**	68.66
Qwen3-VL-32B	43.50	50.75	83.52	84.80	70.66
GPT-4o	33.50	43.75	76.34	61.78	72.67
GPT-5	57.33	78.00	80.85	71.73	77.33
Unified Multimodal Models (UMMs)					
Janus-Pro-7B	0.00	33.50	43.08	38.22	50.67
Chameleon-7B	0.83	30.50	5.74	28.27	0.67
BAGEL-7B	0.83	35.00	61.82	55.49	67.33
ThinkMorph (text-only)	73.27	76.25	76.80	64.92	68.00
ThinkMorph	75.83	77.50	78.00	67.01	66.00
Ours					
Text-only SFT	49.17 ±1.06	63.50 ±1.12	81.66 ±1.16	56.02 ±0.46	68.67 ±0.72
Visual-OPSD-Noise	50.83 ±1.14	64.50 ±0.53	73.77 ±0.42	61.70 ±0.52	68.66 ±0.63
Visual-OPSD (Ours)	**85.83** ±0.89	**86.00** ±0.82	78.79 ±1.27	64.92 ±0.70	**77.33** ±0.63
Δ *vs.* ThinkMorph	+10.00	+8.50	+0.79	−2.09	+11.33
Δ *vs.* Text-only SFT	+36.66	+22.50	−2.87	+8.90	+8.66

Method	MMVP ↑	SAT ↑	BLINK ↑	CV-Bench ↑	Avg ↑	Lat. (s) ↓
Vision-Language Models (VLMs)						
InternVL3.5-8B	76.33	45.33	59.60	81.99	58.04	–
InternVL3.5-38B	80.33	49.33	62.65	85.96	63.67	–
Qwen3-VL-8B	77.66	**54.00**	**69.43**	**85.63**	64.58	–
Qwen3-VL-32B	80.00	76.00	66.22	87.51	71.44	–
GPT-4o	84.67	28.00	60.28	75.61	59.62	–
GPT-5	86.33	73.30	69.86	85.46	75.58	–
Unified Multimodal Models (UMMs)						
Janus-Pro-7B	63.33	22.00	38.51	67.83	39.68	–
Chameleon-7B	47.67	10.67	16.52	36.52	19.71	–
BAGEL-7B	70.33	44.67	47.66	76.03	51.02	–
ThinkMorph (text-only)	**79.00**	51.33	59.50	80.07	69.90	–
ThinkMorph	78.33	52.67	59.49	80.86	70.63	142.8
Ours						
Text-only SFT	76.33 ±0.37	46.63 ±0.79	54.39 ±0.67	77.37 ±0.97	63.75	28.5
Visual-OPSD-Noise	75.33 ±0.36	48.00 ±1.15	55.49 ±0.85	79.09 ±1.14	64.15	14.9
Visual-OPSD (Ours)	77.33 ±0.68	**54.00** ±1.21	61.44 ±0.69	80.64 ±0.51	**74.03**	**10.0**
Δ *vs.* ThinkMorph	−1.00	+1.33	+1.95	−0.22	+3.40	14.3×
Δ *vs.* Text-only SFT	+1.00	+7.37	+7.05	+3.27	+10.28	2.9×

Real VT yields +10.28pp over SFT, against +0.40pp for noise. Visual-OPSD-Noise holds every non-VT element of the teacher context fixed, including the system prompt, question, reference intro and transition prompt, and swaps only the VT pixels for Gaussian noise of the same shape. Neither JSD regularization nor the surrounding privileged structure can therefore account for Visual-OPSD's gains.

3.4 Analysis

Per-sample win/loss analysis. Aggregate accuracy hides how the two models disagree, so we compare their predictions sample by sample (Figure 5). On the complex spatial benchmarks (BLINK-J net +11.3pp, VSP +10.0pp, VisPuzzle +8.5pp) Visual-OPSD wins far more samples

Table 3: VT information quality scaling. **Avg (9)** is the mean over all 9 benchmarks of Table 2; the four columns to its left are representative tasks listed for readability.

Variant	Teacher VT	VSP	VisPuz.	BLK-J	VStar	Avg (9)	Δ SFT
ThinkMorph (text-only)	None (zero-shot)	73.27	76.25	68.00	64.92	69.90	+6.15
Text-only SFT	None	49.17	63.50	68.67	56.02	63.75	–
Visual-OPSD-Noise	Gaussian noise	50.83	64.50	68.66	61.70	64.15	+0.40
Visual-OPSD (Ours)	Real VT	**85.83**	**86.00**	**77.33**	**64.92**	**74.03**	**+10.28**

than it loses. These tasks need multi-step spatial reasoning, the quality of ThinkMorph's generated VTs is bounded by its diffusion model, and Figure 3b indicates that a rendered VT tends to dominate whatever reasoning follows it regardless of quality. Reasoning straight from the original input, Visual-OPSD avoids that dependence while keeping the generation pathway's knowledge in distribution (further examples in Appendix K).

ThinkMorph keeps a modest edge on three benchmarks (V* -2.1pp, MMVP -1.0pp, CV-Bench -0.2pp), all visually simpler, where its VTs come out cleaner. There the cost of losing explicit VT images outweighs the benefit of ignoring them: magnified object details and highlighted regions are cues that distributional distillation does not fully replace (failure cases in Appendix L).

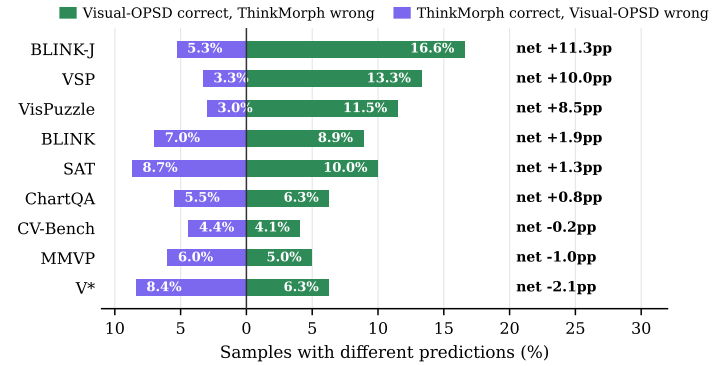


Figure 5: Per-sample win/loss between Visual-OPSD and ThinkMorph. Green: Visual-OPSD correct and ThinkMorph wrong; purple: the reverse. The deficits on ThinkMorph-leading benchmarks are small and near-symmetric.

4 Conclusion

We introduced Visual On-Policy Self-Distillation (Visual-OPSD), the first On-Policy Self-Distillation framework to operate across modalities within a single unified multimodal model. The visual generation pathway of a UMM turns out to hold reasoning knowledge that its rendered pixels do not deliver, and on-policy JSD moves that knowledge into the text understanding pathway with no architectural change. The Visual-OPSD student beats its generative teacher on 6 of 9 benchmarks ($+3.40$pp on average) at a $14.3\times$ speedup, and clears same-scale dedicated VLMs by a wide margin on spatial reasoning. The Visual-OPSD-Noise control ($+0.40$pp against $+10.28$pp) and the post-distillation KL analysis (58.4% against 3.5% gap closure) place the transferred signal in the semantic content of the generation pathway rather than in regularization.

Broader implications. The recipe should carry to any model with two pathways separated by an information asymmetry, suggesting cross-modal OPSD variants for audio-language and tool-using systems, and to other UMMs such as Chameleon, Emu3, and Janus-Pro.

AI USE STATEMENT

Generative AI tools were used only for language polishing and LaTeX formatting, including adaptation to the conference template. They were not used for translation, research methodology, experiment implementation, mathematical claims or proofs, data processing, or interpretation of results. The authors are responsible for the final manuscript and its scientific content.

REPRODUCIBILITY STATEMENT

The method and training objective are described in Section 2. Appendix C provides the architecture, training infrastructure, hyperparameters, data preparation, and training algorithm. The teacher prompts are given in Appendix E, and the evaluation protocol is specified in Appendix M. The theoretical setting and result are stated in Section 2.4, with the accompanying proof in Appendix B.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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

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.

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.

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. *arXiv preprint arXiv:2404.12390*, 2024.

Tommaso Furlanello, Zachary C. Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born-again neural networks. In *International Conference on Machine Learning*, pp. 1602–1611, 2018.

Jiawei Gu, Yunzhuo Hao, Huichen Will Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. ThinkMorph: Emergent properties in multimodal interleaved chain-of-thought reasoning. In *International Conference on Learning Representations*, 2026.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017, 2023.

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. *arXiv preprint arXiv:2406.09403*, 2024.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 1773–1781, 2023.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2023.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, et al. SAT: Dynamic spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024.

Daniel Rose, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chinmay Sonar, Diba Mirza, and William Yang Wang. Visual chain of thought: Bridging logical gaps with multimodal infillings. *arXiv preprint arXiv:2305.02317*, 2023.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. *arXiv preprint arXiv:2406.16860*, 2024a.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. *arXiv preprint arXiv:2401.06209*, 2024b.

Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: Similarity control and knowledge transfer. *Journal of Machine Learning Research*, 16(61):2023–2049, 2015.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. In *Advances in Neural Information Processing Systems*, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

Lai Wei, Liangbo He, Jun Lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang, Linghe Kong, Yue Wang, et al. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception. *arXiv preprint arXiv:2602.11858*, 2026.

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. *arXiv preprint arXiv:2410.13848*, 2024a.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024.

Qiucheng Wu, Handong Zhao, Michael Saxon, Trung Bui, William Yang Wang, Yang Zhang, and Shiyu Chang. VSP: Assessing the dual challenges of perception and reasoning in spatial planning tasks for VLMs. *arXiv preprint arXiv:2407.01863*, 2024b.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-OPD: Learning to see fine details for multimodal LLMs via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026.

Xinsen Zhang, Zhenkai Ding, Tianjun Pan, Run Yang, Chun Kang, Xue Xiong, and Jingnan Gu. OPSDL: On-policy self-distillation for long-context language models. *arXiv preprint arXiv:2604.17535*, 2026.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

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. *arXiv preprint arXiv:2408.11039*, 2024.

A Related Work

On-Policy Self-Distillation (OPSD) and its emerging family. Visual-OPSD belongs to the recently emerging family of on-policy self-distillation (OPSD) methods, in which a *single* model instantiates both teacher and student by conditioning on different contexts, and token-level distillation is performed along the student's own on-policy trajectories. The principle was introduced by Zhao et al. (2026) for text-only reasoning: the teacher conditions on the verified ground-truth solution while the student sees only the problem, and JSD on the student's rollouts transfers privileged reasoning knowledge without an external teacher model. Zhang et al. (2026) extends OPSD to long-context language modeling, using a short-context self-teacher to denoise long-context generation. Closest to us in modality is Vision-OPD (Yuan et al., 2026), which applies OPSD within the visual modality: a crop-conditioned teacher supervises a full-image student to transfer fine-grained regional perception. Visual-OPSD differs from all prior OPSD instances in a critical way: the teacher–student information gap is *cross-modal*, namely between the generation pathway (which has internalized visual reasoning via diffusion training) and the understanding pathway (which has not), rather than within text (OPSD), within long-vs-short context (OPSDL), or within visual crops (Vision-OPD). To our knowledge, Visual-OPSD is the first OPSD framework that bridges a generation–understanding gap within a unified multimodal architecture.

Knowledge Distillation and Privileged Information. Classical knowledge distillation (Hinton et al., 2015) transfers knowledge from a larger teacher to a smaller student. Born-Again Networks (Furlanello et al., 2018) showed that self-distillation between same-capacity networks can improve performance. CoT-style rationale distillation (Hsieh et al., 2023; Magister et al., 2023) extracts reasoning from larger models into smaller ones, while DPO-style methods (Rafailov et al., 2023) align via preference signals. Learning Using Privileged Information (LUPI) (Vapnik & Izmailov, 2015) formalizes the teacher–student gap in terms of information rather than capacity, and Zooming without Zooming (Wei et al., 2026) applies it to visual zooming with better data as the privileged signal. Visual-OPSD differs in kind from all of these: the teacher–student gap is neither in capacity nor in data quality, but in the *modality of conditioning* within a single shared parameter set. The teacher's privileged information is the activation pattern induced by the generation pathway on its own VT outputs, that is, an internal-state asymmetry rather than an external data asymmetry.

Unified Multimodal Models. Recent work converges on architectures that handle visual understanding and generation within a single model: BAGEL (Deng et al., 2025) fuses a Qwen2.5 LLM backbone with a SigLIP vision encoder and a FLUX VAE; ThinkMorph (Gu et al., 2026) adds interleaved visual chain-of-thought, generating intermediate images during reasoning; other notable UMMs include Chameleon (Chameleon Team, 2024), Emu3 (Wang et al., 2024), Janus (Wu et al., 2024a) and Janus-Pro (Chen et al., 2025), and Transfusion (Zhou et al., 2024). Any UMM whose generation pathway can act as a privileged knowledge source is a candidate teacher for cross-modal OPSD.

Visual Chain-of-Thought. Visual CoT (Rose et al., 2023) and Multimodal CoT (Zhang et al., 2023) explore generating intermediate visual representations during reasoning. Visual Sketchpad (Hu et al., 2024) uses code-generated sketches as reasoning aids. All of these methods retain visual generation at inference and therefore inherit its cost. Visual-OPSD instead distills the *knowledge* from visual CoT into the text understanding pathway, removing the inference-time generation cost without losing reasoning accuracy.

B Proof of Theorem 1

Proof. **(a) Forward-KL decomposition.** With the input, completion prefix, and teacher weights fixed, let $\bar{p}(v) = \mathbb{E}_V[p_V(v)]$. Splitting the log-ratio gives

$$\begin{aligned}\mathbb{E}_V[D_{\mathrm{KL}}(p_V\|q)] &= \mathbb{E}_V\left[\sum_v p_V(v)\log\frac{p_V(v)}{\bar{p}(v)}\right] + \sum_v \mathbb{E}_V[p_V(v)]\log\frac{\bar{p}(v)}{q(v)} \\ &= \mathbb{E}_V[D_{\mathrm{KL}}(p_V\|\bar{p})] + D_{\mathrm{KL}}(\bar{p}\|q).\end{aligned}$$

The first term is independent of q. Nonnegativity of KL, with equality only when $q = \bar{p}$, proves the unique minimizer. Boundary distributions with $q(v) = 0$ have infinite forward KL because $\bar{p}(v) > 0$.

(b) JSD lower bound. Write $m_V = (p_V + q)/2$ and $\bar{m} = (\bar{p} + q)/2 = \mathbb{E}_V[m_V]$. Joint convexity of KL yields

$$\begin{aligned}\mathbb{E}_V[\mathrm{JSD}_{1/2}(p_V\|q)] &= \tfrac{1}{2}\mathbb{E}_V[D_{\mathrm{KL}}(p_V\|m_V)] + \tfrac{1}{2}\mathbb{E}_V[D_{\mathrm{KL}}(q\|m_V)] \\ &\geq \tfrac{1}{2}D_{\mathrm{KL}}(\bar{p}\|\bar{m}) + \tfrac{1}{2}D_{\mathrm{KL}}(q\|\bar{m}) \\ &= \mathrm{JSD}_{1/2}(\bar{p}\|q).\end{aligned}$$

This proves Eq. 7. □

C Implementation Details

Model Architecture. ThinkMorph-7B, the checkpoint Visual-OPSD starts from, consists of: Qwen2.5 LLM backbone with Qwen2MoTDecoderLayer (∼7B parameters), SigLIP-so400m-14-980 NaViT vision encoder, and FLUX VAE for latent image encoding.

Training Infrastructure. 8×NVIDIA H800 80GB GPUs with FSDP (HYBRID_SHARD strategy). Activation checkpointing on all Qwen2MoTDecoderLayer modules. Optimizer state CPU offloading during sampling and forward/backward to manage memory constraints.

Hyperparameters. Table 4 lists the training configuration.

Table 4: Training hyperparameters for Visual-OPSD.

Parameter	Value	Parameter	Value
Learning rate	1e-5	JSD β	0.5
Min learning rate	1e-7	JSD temperature	1.0
LR scheduler	Cosine	JSD top-K	256
Warmup steps	200	Token clip	0.05
Total steps	1000	Loss	Pure JSD
AdamW (β_1, β_2)	(0.9, 0.95)	On-policy sampling	Yes
AdamW ϵ	1e-15	EMA decay	0.995
Max grad norm	1.0	Teacher mode	EMA
Gradient accum. steps	2	Max completion tokens	1024
Max forward tokens	10240	Image skips	1

Data. Training data consists of ∼24,990 samples across four task categories, identical to the ThinkMorph training distribution (Gu et al., 2026): Visual Search (6,990), Spatial Navigation (6,000), Jigsaw Assembly (6,000), Chart Refocus (6,000). Each sample contains a problem image, question, interleaved textual reasoning traces with VT images, and the answer. Training and evaluation data are disjoint: training samples are drawn from designated training splits, while all 9 evaluation benchmarks use their respective held-out test sets. No evaluation benchmark images or questions appear in the training data. The 1,000 samples used for the KL diagnostic of Section 2.2 and Appendix I are drawn from the same four categories but are held out from training. Images are processed with NaViT-style patching: stride 14, max size 980, min size 378, max pixels 2,007,040.

Visual token budget. The two caps interact as follows. A ViT grid is limited to 70 patches per side, so one image contributes at most $70 \times 70 = 4{,}900$ tokens; the max-pixel setting of 2,007,040 is an area cap that the side cap makes unreachable, because a 70×70 grid at stride 14 covers at most $960{,}400$ pixels. Dividing the area cap by 14^2 therefore overstates the per-image cost by more than a factor of two. The student's forward pass holds the system prompt, the problem image and the question ahead of a completion capped at 1,024 tokens, so it stays under $\approx 6{,}000$ tokens and never reaches the 10,240 budget.

Memory Optimization. Key optimizations to fit within 8×80GB: (1) AdamW state CPU offload during sampling (∼14.5 GiB freed); (2) FSDP precision patch preventing FP32 upcast during parameter unshard (58GB→29GB); (3) Aggressive tensor release between student/teacher forward passes; (4) Sequence length limiting (max 10,240 tokens) with automatic completion truncation.

Training Loop. Algorithm 1 summarizes one optimization step of Visual-OPSD, including on-policy sampling with image-skip injection, dual teacher/student forward passes, the per-token JSD objective, and the EMA teacher update.

Algorithm 1 Visual-OPSD Training Loop

Require: Model M_θ, EMA teacher $M_{\bar{\theta}}$, dataset $\mathcal{D}$
for each training step **do**
  Sample raw data $x \sim \mathcal{D}$ (problem image, question, reference trace with VT)
  On-policy sampling: Generate completion $c \sim M_\theta(\cdot \mid \text{img}, \text{question})$
    On <image_start>, append <image_end> to the context and resume text sampling
    (the injected token is excluded from c, so it carries no loss term)
  Build student batch: $[\text{sys}, \text{ViT}(\text{img}), \text{q}, c]$
  Build teacher batch: $[\text{sys}, \text{ViT}(\text{img}), \text{q}, \text{intro}, (\text{ViT}(\text{VT}_i))^+, \text{trans}, c]$ using $M_{\bar{\theta}}$
  Dual forward: Compute logits_S (with grad), logits_T (no grad)
  Restrict to teacher top-K, renormalize both sides, accumulate clipped per-entry JSD (Eq. 5)
  Backward, optimizer step, EMA update: $\bar{\theta} \leftarrow \alpha\bar{\theta} + (1-\alpha)\theta$
end for

D Initialization Rationale

Visual-OPSD variants (Visual-OPSD, Visual-OPSD-Noise) are initialized from ThinkMorph-7B because the teacher pathway requires interleaved visual generation capability, which is present only in ThinkMorph and not in the base BAGEL model. Visual-OPSD does not optimize a CE loss on the training data; it distills distributional knowledge via JSD on on-policy completions, so it does not directly memorize training examples. Text-only SFT, in contrast, is initialized from BAGEL-7B: it trains with standard CE on text reasoning traces drawn from the same distribution used to train ThinkMorph, so initializing from ThinkMorph would amount to re-fitting on already-seen data, producing overfitting rather than a fair assessment of the text-only training signal. The $+10.28$pp gain of Visual-OPSD over Text-only SFT reflects two factors: (1) the ThinkMorph initialization, which already embeds generation-trained representations, and (2) the Visual-OPSD distillation objective. The Visual-OPSD-Noise control isolates factor (2): it uses the same ThinkMorph initialization and the same objective structure but replaces real VT pixels with semantic-free Gaussian noise, and gains only $+0.40$pp over SFT. The transfer of the generation-pathway signal therefore requires the distillation objective applied to semantically meaningful VT content, not the ThinkMorph initialization on its own. Table 2 additionally reports ThinkMorph in text-only inference mode (69.90) as an untrained reference for that same initialization. Visual-OPSD-Noise sits below it, so distilling against a semantically empty teacher does not merely fail to help but costs accuracy relative to the initialization it starts from, whereas Visual-OPSD exceeds the reference by $+4.13$pp.

E Teacher Context Prompts

The teacher context $\mathcal{C}_T$ (Eq. 1) uses two framing prompts to separate the privileged visual reasoning trace from the student's own completion. Their full text is given below.

Reference Introduction (ref_intro)**.** Placed immediately before the privileged VT images. Because the privileged channel is visual-only, the prompt explicitly refers to images rather than to a textual reasoning trace:

> *"The following images are privileged visual references that depict the intermediate visual thoughts on the path to the correct answer. Use them silently as grounding context; do not describe or echo them."*

Transition Prompt (transition)**.** Placed after the privileged VT images and before the completion tokens on which loss is computed:

> *"Now, using your own independent reasoning, answer the problem above. Think step by step."*

Both prompts carry zero loss weight during training; they serve only to structure the teacher's KV cache so that the privileged VT images are absorbed as visual context rather than directly copied or described in the completion.

F Per-Token KL Divergence Analysis

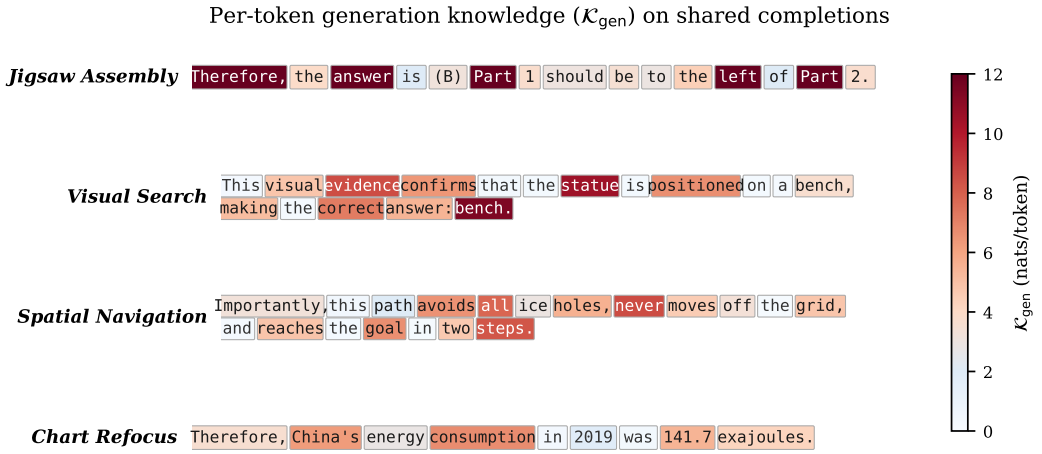


Figure 6: Per-token $\mathcal{K}_{\text{gen}}$ on representative completions drawn from the held-out set of each task category. Generation knowledge concentrates on informationally critical tokens such as spatial labels (`Part`, `left`), object references (`statue`, `bench`), navigation decisions (`goal`, `avoids`), and quantitative values (`141.7`, `2019`), while function words carry near-zero divergence.

The per-token KL analysis reveals that generation knowledge is *non-uniform*: it concentrates on tokens encoding spatial relations, quantities, and visual-grounded answers, while function words and syntactic connectives exhibit near-zero KL. This pattern is distinct from the high-entropy "forking tokens" observed in reinforcement learning (Wang et al., 2025a), where connectives represent trajectory-level decision points. Our cross-context KL instead measures how much additional *perceptual evidence* the VT context contributes at each position: VT informs *what* the answer is (content tokens) rather than *how* to express it (syntactic structure), producing the observed content-specific divergence pattern. Since teacher and student share both the model weights and the completion tokens, the measured divergence is attributable solely to the privileged VT context.

G Output Token Statistics

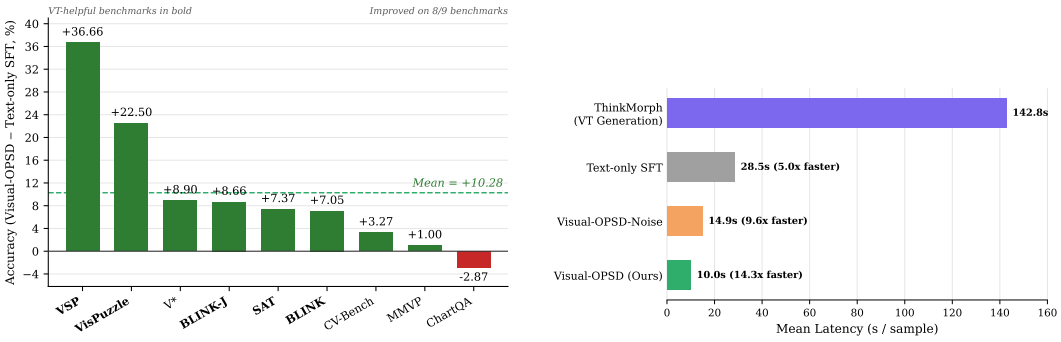
Visual-OPSD produces shorter outputs on every benchmark in Table 5, whose 10 rows split ChartQA and CV-Bench into their two variants. On average Visual-OPSD generates 201.4 tokens per sample, $2.0\times$ fewer than SFT (411.3) and $2.2\times$ fewer than ThinkMorph (451.7, text tokens only). The compression is largest on VisPuzzle (263.4 against 701.5 for SFT, $2.7\times$) and BLINK-J (287.9 against 604.6 for ThinkMorph, $2.1\times$), both complex spatial tasks where the baselines write lengthy "I observe that…" narrations. Even where accuracy is comparable across methods, as on CV-Bench-2D, Visual-OPSD emits $\sim 2\times$ fewer tokens, so the conciseness is a general property of the distillation rather than an artifact of particular tasks.

Table 5: Mean output tokens per sample. All methods use greedy decoding (temperature=0, max 1024 tokens). ThinkMorph token counts exclude diffusion steps (text tokens only).

Benchmark	Visual-OPSD	ThinkMorph	Text-only SFT
BLINK	258.2	522.9	439.4
BLINK-J	287.9	604.6	552.1
ChartQA (h)	181.8	397.7	360.6
ChartQA (v)	194.7	416.8	379.6
CV-Bench-2D	148.3	304.1	307.2
CV-Bench-3D	171.1	383.0	349.8
MMVP	165.7	407.2	371.9
SAT	189.2	404.5	369.4
VisPuzzle	263.4	768.1	701.5
V*	153.9	308.4	281.6
Average	**201.4**	451.7	411.3

H Per-Task Knowledge Transfer and Inference Efficiency

Figure 7 visualizes the per-task gain pattern and the latency profile discussed in Section 3.2. Spatial reasoning benchmarks dominate the mean $+10.28$pp gain over Text-only SFT while ChartQA shows essentially no transfer, and Visual-OPSD at 10.0s per sample is $14.3\times$ faster than the VT teacher and $2.9\times$ faster than text-only SFT, the latter factor following from its $\sim 2\times$ shorter outputs (Appendix G).



(a) (b)

Figure 7: **Task-specific knowledge transfer and inference efficiency.** **(a)** Generation knowledge transfers selectively: spatial reasoning tasks benefit most (mean Δ=+10.28pp over Text-only SFT), while chart understanding shows minimal change. **(b)** Visual-OPSD is faster than both the VT teacher ($14.3\times$) and text-only SFT ($2.9\times$), suggesting distillation produces more concise reasoning.

I Knowledge Transfer Verification: Post-Distillation KL Analysis

The KL diagnostic in Section 2.2 establishes that a distributional gap $\mathcal{K}_{\text{gen}}$ exists between teacher (with VT context) and student (without) *before* any distillation training. A natural question is whether Visual-OPSD successfully closes this gap, which would provide direct evidence that generation knowledge has been internalized into the understanding pathway.

Protocol. We re-run the identical KL diagnostic (Eq. 3) on the same 1,000 held-out samples, replacing the student with each trained checkpoint. These samples never appear in the training data, so the numbers below measure generalization of the distillation objective rather than its fit on the

optimized examples. For each variant, we compute:

$$\Delta\mathcal{K} = 1 - \frac{\mathcal{K}_{\text{gen}}^{\text{post}}}{\mathcal{K}_{\text{gen}}^{\text{pre}}} \in [0, 1] \tag{8}$$

where $\Delta\mathcal{K}$ represents the fraction of the teacher–student distributional gap that has been closed by training. A value of 1 indicates perfect knowledge internalization; 0 indicates no transfer.

Table 6: Post-distillation KL gap analysis on 1,000 held-out samples. We measure the teacher–student distributional gap $\mathcal{K}_{\text{gen}}$ (nats/token) before and after training for each variant (both initialized from ThinkMorph-7B). Visual-OPSD closes 58.4% of the gap on average, with the largest reductions on spatial reasoning tasks, while Visual-OPSD-Noise closes under 4%.

Task Category	$\mathcal{K}_{\text{gen}}^{\text{pre}}$	$\mathcal{K}_{\text{gen}}^{\text{post}}$ **(nats/token)** [$\Delta\mathcal{K}$%]	
	(frozen)	**Visual-OPSD-Noise**	**Visual-OPSD (Ours)**
Visual Search	4.23	4.08 [3.5%]	1.72 [**59.3%**]
Spatial Navigation	3.96	3.82 [3.5%]	1.48 [**62.6%**]
Chart Refocus	3.51	3.39 [3.4%]	2.14 [**39.0%**]
Jigsaw Assembly	6.84	6.61 [3.4%]	2.37 [**65.4%**]
Overall	**4.64**	4.48 [3.5%]	**1.93** [**58.4%**]

Results and interpretation. The Visual-OPSD student closes 58.4% of the teacher–student gap on average. On samples it was never trained on, its predictions, made *without* any VT context, move most of the way toward what the model would produce *if* it had observed the full sequence of privileged VT images, which is what transfer of generation-pathway knowledge into the understanding pathway should look like.

The pattern tracks the accuracy gains of Table 2. Jigsaw Assembly closes the most (65.4%) and carries the largest improvement over the teacher (BLINK-J +11.3pp); Chart Refocus closes the least (39.0%) and is the one category where distillation fails to beat text-only SFT (ChartQA −2.87pp). Visual-OPSD-Noise closes 3.5%, which is attributable to mild EMA-based regularization rather than transfer. Aligning with the VT-conditioned teacher therefore requires semantically meaningful VT content; Gaussian noise in the teacher context supplies almost no learning signal.

J JSD Hyperparameter Sensitivity

We ablate the two key JSD loss hyperparameters (top-K vocabulary truncation and per-token clipping threshold) to assess Visual-OPSD's robustness. All variants are trained for 1,000 steps with other hyperparameters held at their default values (Section C). We report average accuracy across 3 representative benchmarks (VSP, VisPuzzle, BLINK-J).

Table 7: JSD hyperparameter sensitivity. **Left**: varying top-K with clip=0.05. **Right**: varying clip threshold with K=256. The default configuration (K=256, clip=0.05) achieves the best overall performance, but Visual-OPSD is robust across a wide range of settings.

Top-K	**VSP**	**VisPuz.**	**BLK-J**	**Avg$_3$**	**Clip**	**VSP**	**VisPuz.**	**BLK-J**	**Avg$_3$**
64	81.67	83.50	73.33	79.50	None	82.50	83.00	74.00	79.83
128	84.17	85.00	76.00	81.72	0.01	83.33	84.50	75.33	81.05
256	**85.83**	**86.00**	**77.33**	**83.05**	**0.05**	**85.83**	**86.00**	**77.33**	**83.05**
512	84.17	85.50	76.00	81.89	0.10	84.17	85.00	75.33	81.50

Table 7 shows that Visual-OPSD is robust to hyperparameter choices: all configurations substantially outperform the Text-only SFT baseline (Avg$_3$=60.45). Performance degrades modestly with very small K (64) due to loss of distributional information in the long tail, or without clipping (None) where noisy style-token gradients introduce variance. The default K=256 and clip=0.05 achieve the best balance between capturing sufficient distributional information and suppressing noise.

K Case Study: VT Interference in Generative Reasoning

We present qualitative examples where ThinkMorph's generated VT images mislead subsequent reasoning, while Visual-OPSD avoids these failure modes. For each case we show: the input with question, the VT image generated by ThinkMorph, key reasoning excerpts from both models, and final answers.



(a) Input (b) Option A (c) Option B ✓ (d) ThinkMorph VT

Question: Which image is the missing part in the first image?

Ground truth: B (the third image).

ThinkMorph ✗ → A

Round 0: "The arm and vest continue naturally from the first image…the contours of the arm line up with the visible body." → commits to A.

VT generation: produces image (d) attempting to verify the completed puzzle.

Round 1 (conditioned on VT): "The second image seamlessly fits into the missing lower right quadrant…the white shirt and black vest connect smoothly."

⇒ **VT reinforces initial error.**

Visual-OPSD ✓ → B

"The third image, with the blue apron and matching white shirt, fits perfectly into this scene: the apron's straps and the shirt's sleeve connect seamlessly with the upper body."

"If I were to use the second image, the table and objects would not fit as smoothly."

⇒ **Directly compares both options against original input.**

Figure 8: **Case 1: Self-reinforcing VT confirmation.** The VT image (d) is generated *after* ThinkMorph has already committed to Option A. The generated image visually reinforces the initial error, creating a self-reinforcing feedback loop. Visual-OPSD avoids this loop by reasoning directly from the original input.



(a) Input (b) Option A (c) Option B ✓ (d) ThinkMorph VT

Question: Which image is the missing part in the first image?

Ground truth: B (the third image).

ThinkMorph ✗ → A

Round 0: "The left edge of this image aligns with the right edge of the first image, as both show the woman in the tank top at a matching angle." → commits to A.

VT generation: produces image (d) with visible blending artifacts at patch boundaries.

Round 1 (conditioned on VT): Confirms A despite spatial misalignment; pixel-level noise in (d) obscures the boundary discontinuities that would reveal incorrect placement.

⇒ **VT noise masks spatial errors.**

Visual-OPSD ✓ → B

"The third image contains the woman in yellow, the other person, and the officers in perfect alignment. The officers' arms, uniforms, and the sky all connect seamlessly."

"By focusing on the visual flow, especially the alignment of the woman's hands, the officers' arms, and the background, I see that the third image is the only one that fits perfectly."

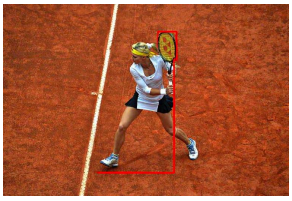
⇒ **Correct spatial continuity without noisy VT.**

Figure 9: **Case 2: Pixel-level VT noise obscures spatial cues.** The VT image (d) introduces blending artifacts at patch boundaries, masking the visual discontinuities that would reveal incorrect placement. Visual-OPSD identifies correct spatial continuity directly from the original.

Summary. The four cases above illustrate complementary VT interference mechanisms: (1) self-reinforcing confirmation of initial errors (Case 1), (2) pixel-level noise masking spatial discontinuities (Case 2), (3) annotation-induced context stripping (Case 3), and (4) reasoning destabilization through ambiguous VT cues (Case 4). In all cases, the Visual-OPSD student avoids these failure modes by reasoning directly from the original input, having internalized the generation pathway's spatial reasoning knowledge at the distribution level without inheriting its pixel-level limitations.



(a) Input image



(b) ThinkMorph VT

Question: Is she about to swing the racket towards her left or right?

Options: (A) right (B) left

Ground truth: A (right).

ThinkMorph ✗ → B (left)

Round 0: "Her right arm is extended with the racket…The racket is currently facing towards her right side." Despite noting rightward cues, concludes "she is preparing to swing towards her left."

VT generation: produces image (b) highlighting racket with bounding box, isolating it from full-body context.

Round 1 (conditioned on VT): "The racket is pointed diagonally towards her left side, aligning with the direction of her body movement." Doubles down on wrong answer.

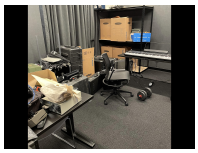
⇒ **VT annotation strips spatial context.**

Visual-OPSD ✓ → A (right)

"Her right arm is bent and gripping the racket…The woman's head and torso are oriented towards the right side…her right leg is extended as if preparing to move."

⇒ **Integrates full-body posture holistically:** foot positioning, torso orientation, and shoulder angle jointly determine the correct swing direction.

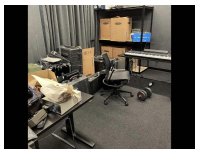
Figure 10: **Case 3: VT annotation strips spatial context.** ThinkMorph's VT image (b) highlights the racket via bounding box but isolates it from full-body context. Post-generation attention shifts to the annotated region (`txt1`→`img1` dominance), causing reasoning from a spatially impoverished representation. Visual-OPSD integrates holistic body cues correctly.



(a) Frame 1



(b) Frame 2



(c) ThinkMorph VT

Question: Were any objects moved from their original positions between frames?

Options:
(A) chair moved right & towards camera
(B) chair moved left & away from camera

Ground truth: A.

ThinkMorph ✗ → B

Round 0: Self-contradictory reasoning; first correctly observes "moved left and slightly towards the camera," then states "the answer is (A)…*However*…the correct answer is (B)."

VT generation: produces image (c) with spatial distortions in chair position.

Round 1 (conditioned on VT): "The chair is now more centrally placed…its legs further from the camera…the answer is (B)." VT resolves internal ambiguity toward wrong direction.

⇒ **VT destabilizes uncertain reasoning.**

Visual-OPSD ✓ → A

"The chair, which was central in the first image, is now positioned slightly to the left and closer to the camera. Its backrest and seat are more directly facing the viewer, and the wheels are aligned with the carpet in a way that suggests it has been moved."

⇒ **Consistent reasoning without self-contradiction.** No VT annotation to second-guess the correct initial spatial assessment.

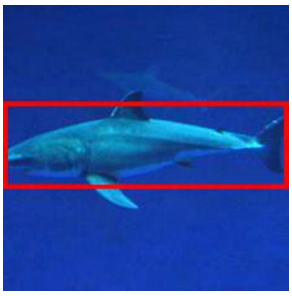
Figure 11: **Case 4: VT-induced reasoning self-contradiction.** ThinkMorph's Round 0 reasoning explicitly contradicts itself, and the VT image (c) resolves this ambiguity toward the wrong answer. Without VT, Visual-OPSD reasons consistently and correctly. This illustrates how VT conditioning can *actively degrade* reasoning quality.

L FAILURE CASES: WHEN HIGH-QUALITY VT HELP

While Visual-OPSD outperforms ThinkMorph on complex spatial reasoning tasks where VT quality is low (Section 3.4), ThinkMorph retains an advantage on a subset of simpler benchmarks where its diffusion model produces high-quality VT images. Below we present representative failure cases from V* and MMVP. In these examples, the generated VT provide genuinely informative visual cues, e.g., magnified object regions or highlighted spatial details, that aid ThinkMorph's reasoning, whereas Visual-OPSD, lacking access to explicit VT images, misses these fine-grained cues despite having internalized the generation pathway's distributional knowledge.



(a) Input



(b) ThinkMorph VT

Question: Is the shark's belly visible in this image?

Options:
(A) Yes
(B) No

Ground truth: B (No).

ThinkMorph ✓ → B (No)

Round 0: "The lower part of the shark, including its belly, is cut off by the edge of the frame... there is no visible tissue continuing along the belly."

VT generation: produces image (b) with a bounding box around the shark's main body, explicitly delineating the region where the belly terminates at the frame boundary.

Round 1 (conditioned on VT): "The area beneath the shark within this box does not show any continuation of its body... the viewer is unable to see the shark's belly." Confirms B.

⇒ **VT bounding box clarifies the frame-cropping boundary.**

Visual-OPSD ✗ → A (Yes)

"The area beneath it is clearly visible... the lighting and angle highlight the belly. The shark's fins, tail, and upper body are visible, but the belly is also distinctly seen, with no major visual blockages."

⇒ **Without the VT bounding box, Visual-OPSD conflates the visible underside with the cropped belly region.**

Figure 12: **Failure Case 1 (MMVP): VT bounding box clarifies spatial occlusion.** The task requires determining whether the shark's belly is visible despite the image being cropped at the bottom. ThinkMorph's high-quality VT (b) draws a bounding box delineating the visible body region, making the cropping boundary explicit and leading to the correct answer. Visual-OPSD, lacking this visual annotation, mistakes the partially visible underside for the full belly.

Discussion. These failure cases reveal the boundary of distribution-level distillation. In both cases, the tasks are perceptually straightforward (requiring visibility judgment for Case 1 or fine-grained color discrimination for Case 2), and ThinkMorph's diffusion model produces high-quality VT annotations that genuinely aid reasoning. Visual-OPSD acquires the generation pathway's reasoning *patterns* but cannot replicate the explicit visual operations (bounding-box delineation, region isolation) that VT provides at the pixel level. When VT quality is high, the benefit of filtering pixel-level noise is outweighed by the loss of these informative visual cues, resulting in the small net accuracy deficits observed on MMVP (-1.0pp) and V* (-2.1pp). Future work could explore hybrid strategies that selectively invoke VT generation when the expected quality is high and the task demands fine-grained perception.

M EVALUATION PROTOCOL

Internal models. All internal models (BAGEL-7B, ThinkMorph, Text-only SFT, Visual-OPSD-Noise, Visual-OPSD) are evaluated using identical infrastructure: greedy decoding (temperature=0), max output tokens=1024, single H800 GPU, batch size 1. Images are preprocessed with NaViT-style patching (stride 14, max 980px, min 378px) identically across all systems. For ThinkMorph, VT generation uses 50 DDPM denoising steps with classifier-free guidance scale 3.5. Latency measurements include the full pipeline (image preprocessing, ViT encoding, LLM decoding, and VT generation where applicable) but exclude data loading I/O.



(a) Input



(b) ThinkMorph VT

Question: What is the color of the scarf?

Options:
(A) red (B) green
(C) white (D) yellow

Ground truth: B (green).

ThinkMorph ✓ → B (green)

Round 0: "This person is wearing a yellow jacket and a green scarf, which is visible around their neck." Correctly distinguishes scarf from jacket.

VT generation: produces image (b) with bounding boxes around both the ATV rider and the lead vehicle passenger, isolating the person with the scarf.

Round 1 (conditioned on VT): "The scarf stands out distinctly and is green in color, visible against the yellow jacket and the white snow." Confirms B.

⇒ **VT bounding boxes isolate the target person and scarf.**

Visual-OPSD ✗ → D (yellow)

"The person in the foreground, riding an ATV, is wearing a dark jacket and a scarf. The scarf is visible around the neck and is distinctly colored… it is a bright yellow."

⇒ **Without VT isolation, Visual-OPSD confuses the scarf's color with the adjacent yellow jacket.**

Figure 13: **Failure Case 2 (V*): VT bounding boxes disambiguate adjacent colors.** The scene contains a person in a yellow jacket with a green scarf, two colors in close proximity on a small, distant figure. ThinkMorph's VT (b) draws bounding boxes around the relevant persons, enabling precise color discrimination between the jacket and scarf. Visual-OPSD, lacking this visual isolation, conflates the scarf's green with the jacket's yellow.

External VLMs. InternVL3.5-8B/38B (Wang et al., 2025b) and Qwen3-VL-8B/32B (Bai et al., 2025) are evaluated locally by the authors using official model weights and default inference configurations from their respective repositories. Janus-Pro-7B (Chen et al., 2025) and Chameleon-7B (Chameleon Team, 2024) are similarly evaluated locally. All open models use greedy decoding with max 1024 output tokens and the same benchmark prompts as internal models. For V* (Wu & Xie, 2024), we disable any external visual-search tools (e.g., the SEAL/V* visual search pipeline) for all evaluated VLMs, so every model is scored solely on its native single-pass multimodal reasoning capability, ensuring a fair comparison to our text-only Visual-OPSD student that has no tool access.

API models. GPT-4o and GPT-5 results are obtained via the OpenAI API (May 2025 snapshot for GPT-4o, January 2026 for GPT-5) with temperature=0 and max_tokens=1024. All API evaluations use the same prompt templates and image formats as internal evaluations.

Prompting. Each benchmark uses a standardized prompt format: the input image(s) followed by the question text, with answer options listed when applicable. For multiple-choice benchmarks (VSP, BLINK, BLINK-J, MMVP, SAT, CV-Bench, V*), the prompt instructs the model to respond with the option letter. For open-ended benchmarks (ChartQA, VisPuzzle), the prompt requests a direct answer.

Visual-OPSD-Noise details. In Visual-OPSD-Noise, each real VT image in the teacher context is replaced with Gaussian noise $\mathcal{N}(0, 1)$ of the same spatial dimensions. The noise tensor is processed through the same ViT encoder as real VT images, producing tokens of the same shape and position encoding. All other elements of the teacher context (system prompt, problem image, question, reference intro, and transition prompt) are preserved unchanged.