Emergent Vision-Shaped Thinking in Text-Only Reasoning through Multimodal Training

Qi Zhang[1]* **Bohan Wang*** **Qianru Sun**[2] **Hanwang Zhang**[1]
[1]Nanyang Technological University [2]Singapore Management University
qi015@e.ntu.edu.sg, bohan.wang97@gmail.com,
qianrusun@smu.edu.sg, hanwangzhang@ntu.edu.sg

Abstract

Does multimodal training shape how a language model thinks, even when the input is text only? We test this with targeted causal ablations across 10 multimodal configurations from 6 model families, with matched text-only backbones and cross-family transfer controls. We contrast spatial-imagery tasks that invite mental scene construction with symbol-manipulation tasks solvable by symbolic rules alone. We identify a vision–language axis in hidden-state space and ablate its projection at inference. In Chameleon-7B, this ablation drops accuracy on BIG-Bench Hard (BBH) *navigate* by 27.6 pp while leaving *boolean expressions* intact (+5.6 pp), a 33.2 pp double dissociation. The effect is direction-specific and layer-localized, extends to mental-simulation tasks such as mental rotation, paper folding, and perspective taking, and is not explained by task difficulty, reasoning depth, or sequential update load. Together, these results reveal *emergent vision-shaped thinking* (EVT): text-only reasoning that causally recruits representations introduced or strengthened by vision–language training, without assuming an image-like internal format. The measured effect is spatially concentrated at 7–8B and broadens to non-spatial visual attribute binding in the tested 32B configuration. In dual-path unified models (BAGEL, Janus-Pro), the causal handle localizes to the generation pathway as a compact one-dimensional carrier. The understanding pathway also encodes spatial information in a higher-rank, probe-decodable form, but is not the causal handle. To separate decodability from causal use, we analyze concentration, alignment, and semantic location. These diagnostics explain why behavior, probing, and intervention can diverge. Multimodal training thus reshapes the geometry of text-only reasoning, not just how visual input is processed.

1 Introduction

Multimodal training has become standard for large language models, producing diverse architectures for coupling vision with language. These range from shared-vocabulary fusion (Chameleon) [7] and detached-encoder bridging (Qwen3-VL) [2] to vision-fine-tuned backbones (Liquid) [49] and dual-path unified models that jointly train visual generation and understanding (BAGEL, Janus-Pro) [11, 9]. Yet how such training affects the language backbone itself remains poorly understood. Architectural choices describe where visual information enters the model. They say little about what the backbone learns to do with it, or whether the linguistic representations used for text-only reasoning are reshaped in the process. This question is especially sharp in unified models, where visual generation, visual understanding, and language modeling share a single backbone. But it has not been mechanistically tested at scale across modern multimodal architectures.

Behavioral evidence suggests that multimodal training affects text-only reasoning. VLMs outperform LLM backbones on text-only spatial reasoning [47], and visual training can shift internal binding

*Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

even under text-only evaluation on synthetic tasks [6]. Yet existing evidence is behavioral or limited to synthetic mid-scale settings. Whether multimodal models rely on visual-spatial structure during naturalistic text-only reasoning, where this dependency lives, and how it varies across architectures remain unknown. We answer these questions causally.

To make the causal question testable, we contrast two families of text-only tasks. *Spatial-imagery tasks* have textual inputs but naturally invite mental scene construction. Examples include navigating a 2D grid described in words and tracking object locations through textual updates. *Symbol-manipulation tasks* are solvable by formal rule operations alone, such as evaluating boolean expressions or checking logical consistency. Both families are text-only. What differs is the representational form each task calls for. If removing a hidden-state direction selectively collapses spatial-imagery tasks while leaving symbol-manipulation tasks intact, the model uses that direction to carry visual-spatial structure in text-only reasoning. This shows functional use, not mere representational presence.

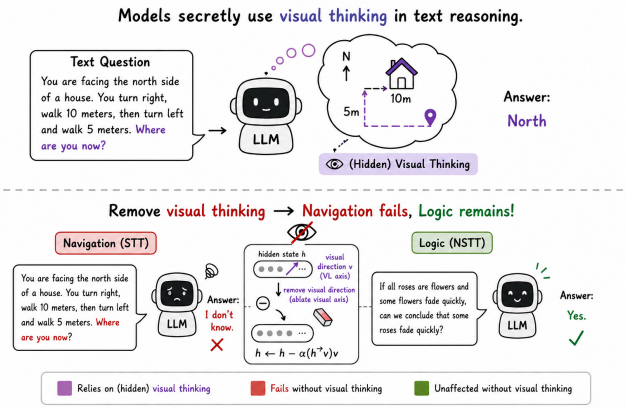


Figure 1: Emergent Vision-Shaped Thinking. Multimodal models use vision-shaped structure even on text-only spatial reasoning. Ablating a vision–language direction selectively damages spatial-imagery tasks (Navigation) while leaving symbol-manipulation tasks (Logic) intact.

For each model, we identify a vision–language axis in hidden-state space and ablate its projection during inference (Figure 1). Inspired by human spatial imagery, we call this collapse *emergent vision-shaped thinking* (EVT), meaning text-only reasoning that causally recruits representations introduced or strengthened by vision–language training. The term deliberately leaves the internal format open: our interventions establish causal use of vision-shaped structure, not the construction of an image-like mental scene (§2, *Terminology and scope*). This definition yields three contributions.

(1) Multimodal training reshapes text-only reasoning geometry. Across 9 VL-axis-positive configurations, VL-axis ablation selectively damages spatial-imagery tasks while leaving symbol-manipulation tasks largely preserved; in the tenth, Chameleon-30B, an input-faithful modality axis is causal at mid/final layers (mid-block DD $= +33.2$ pp; Appendix A.5). The effect is direction-specific, graded across tasks, and localized to a small range of layers. It extends beyond position- and navigation-style tasks to mental-simulation tasks (mental rotation, paper folding, perspective taking), and a controlled battery crossing spatial demand with visual content shows that it is not explained by task difficulty, reasoning depth, or sequential update load (Appendix F). Matched-backbone controls show that multimodal training can either *introduce* a new visual-spatial scaffold or *amplify* one already present in the base LLM. The intervenable effect resides inside the shared LLM hidden state, not in an isolated visual front-end.

(2) In dual-path models, the causal carrier concentrates in the generation pathway. In BAGEL and Janus-Pro, the visual-generation path concentrates spatial structure into a compact one-dimensional causal carrier whose removal selectively disrupts spatial-imagery reasoning. The understanding path also contains spatial information, but in a probe-decodable form that is entangled with general capacity: neither its 1D contrast direction nor any tested rank-k subspace yields a spatially selective handle. In unified models, generation therefore writes a more directly intervenable spatial geometry into the shared language backbone.

(3) What a model encodes is not always what it uses. Behavior, probing, and intervention need not agree. We introduce three diagnostics, concentration, alignment, and semantic location, that explain why a readable feature can be causally weak, why a low-concentration direction can still be behaviorally important, and why a modality axis can fail while a task-contrastive axis succeeds. These diagnostics turn probe-intervention mismatches into evidence about how multimodal training organizes internal representations, so claims about whether a feature is *used* require all three.

2 Terminology, scope, and related work

Terminology and scope. We define the three terms in "emergent vision-shaped thinking" once, and use them in this fixed sense throughout. *Emergent* denotes causal dependence that appears, or becomes substantially stronger, after multimodal training: matched-backbone controls show the dependence is largely *introduced* in Chameleon (the same axis that drops `navigate` by 27.2 pp leaves LLaMA-7B essentially unchanged at -2.0 pp) and *amplified and sharpened* from a scaffold already present in the matched Qwen text backbones (DD $= +18.5$ to $\approx +34$ pp on text-only backbones; §4.3). It does not mean de novo creation of task ability. *Vision-shaped* denotes representations introduced or strengthened by vision–language training and causally recruited during text-only reasoning, without implying an explicitly image-like internal scene; "visual thinking" would overstate the format evidence, while "spatial" would be too narrow for the 32B result below. *Thinking* refers to the internal computation used during text-only reasoning, as measured by interventions on hidden states. *Functional range:* the measured effect is spatially concentrated within the visual domain at 7–8B and extends to non-spatial visual attribute binding in the tested 32B configuration, while the core causal-recruitment result holds at both scales (Appendix F).

Multimodal training and text-only reasoning. Prior work suggests that multimodal training affects more than visual perception. Spatial reasoning remains brittle for both LLMs and VLMs, and visual input does not automatically help [47]. Multimodal neurons and cross-modal structure have been observed in CLIP-style models and text transformers connected to visual encoders [14, 37], and visual training can alter binding strategies and improve text-only generalization [6]. Related work also studies vision–language compatibility and language degradation after multimodal tuning [26, 19, 24, 12, 35]. These studies show that multimodal training can change representations and behavior, but they do not test whether visual-spatial structure is causally recruited during image-absent text-only reasoning. Unlike visual-thought or latent-transfer methods that add, generate, or inject visual/CoT-like intermediates [43, 50, 16, 33, 53], EVT removes an internal VL direction during text-only inference.

Mechanistic interpretability of multimodal models. Mechanistic interpretability separates decodability from causal use [4], motivating causal mediation, activation patching, representation editing, and steering vectors [46, 13, 54, 22]. Prior causal VLM analyses largely focus on image-present settings, including causal tracing, VQA information flow, visual-token processing, hallucination mitigation, and cross-modal corruption [31, 3, 29, 23, 15]. Concurrent work studies image-present spatial mechanisms through linear spatial/temporal IDs [20] and interchange interventions that separate vision-encoder from LM-backbone mechanisms [10]. EVT instead studies the image-absent setting, where a model may still recruit visual-spatial geometry written into the language backbone. This matters because VLMs can rely on linguistic shortcuts even with images [38], and because introduced-versus-amplified scaffolds cannot be inferred from image-present behavior alone. Methodologically, we extend linear concept directions, DiffMean steering, and rank-one editing to architecture-matched VL directions [5, 25, 36, 1, 32], following Othello-GPT in treating latent states as both decodable and causal [21, 28].

Unified multimodal models. EVT also connects to unified models such as Janus-Pro and BAGEL, where understanding and generation share a language backbone but use different routes [9, 11]. Prior work asks whether generation behaviorally improves understanding [44, 17, 8], or whether the two capabilities merely coexist [52, 41, 30, 51]. EVT reframes this debate mechanistically. Rather than asking only whether scores improve, we ask whether understanding and generation write different reusable geometries into the same backbone.

3 Method

Our method has three parts. First, we define a vision–language (VL) direction that separates textual and visual representations. Second, we remove each hidden state's projection onto this direction during text-only inference and measure the performance change. Third, we evaluate the effect with three primary causal checks, namely task selectivity, direction specificity, and layer localization, and two supporting analyses, cross-task gradedness and bidirectional control.

3.1 Identifying Vision–Language directions

We define a vision–language (VL) direction in the representation space where textual and visual signals are comparable. Let μ_{text} and μ_{visual} be the mean textual and visual representations in that space. For hidden dimension d, we define the VL direction as $v = (\mu_{\text{text}} - \mu_{\text{visual}})/\|\mu_{\text{text}} - \mu_{\text{visual}}\|_2 \in \mathbb{R}^d$. We use a standard difference-in-means estimator and refer to the resulting architecture-matched direction as the VL DiffMean axis. The key design choice is where to compute this contrast. For each model, we use the representation space in which textual and visual signals yield an effective modality contrast. For **Chameleon-7B**, we average text-token and visual-token rows in its shared embedding table to obtain a single global VL direction. Although **Emu3** also has a shared discrete vocabulary, its embedding-centroid direction was ineffective in our pilot experiments; for the reported Emu3 results, we instead compute a separate VL DiffMean direction at each layer from hidden states at text positions and synthetic visual-token positions in Emu3-format multimodal sequences. For **continuous-encoder models** (Qwen3-VL), visual features enter the language backbone as continuous hidden states, so we compute per-layer VL directions from hidden states at text and visual positions in multimodal sequences. For **dual-path unified models** (BAGEL, Janus-Pro), visual understanding and visual generation use different visual token streams, so we compute separate VL directions for the two paths.

The construction is therefore heterogeneous by design, following one shared principle: compute the text–visual contrast at the representation site where the two signals are directly comparable within each architecture, using the image source that each encoder can actually process. Every DiffMean axis has split-half stability ≥ 0.9992, and swapping synthetic visual inputs for real COCO images leaves the axes essentially unchanged and reproduces every hub dissociation (Appendix A.4). These checks support qualitative cross-architecture replication, not calibrated effect-size ranking. The canonical VL axis is causal in nine configurations; on Chameleon-30B, only the input-faithful full-sequence construction recovers a causal modality axis (Appendix A.5; §6). Construction details and stability checks are in Appendix A.3.

3.2 Projection intervention

Given hidden state h and normalized axis v, we intervene by $h \leftarrow h - \alpha\,(h^\top v)\,v$. Here $\alpha{=}1$ removes the projection, $\alpha \in (0, 1)$ partially removes it, and $\alpha < 0$ multiplicatively amplifies it. The hook acts on the post-block residual stream of each selected decoder block (Appendix A.1). For $\alpha < 0$, the update scales each token's existing projection on v, whereas additive steering $h \leftarrow h + \beta v$ adds the same offset to all tokens; we therefore use multiplicative amplification for amplification controls.

Dosage selection and held-out validation. The strength α is selected once per configuration and then frozen. Six of ten configurations use $\alpha{=}1.0$ (full projection removal) and require no dosage selection; the remaining positive configurations use a coarse sweep that picks the smallest α yielding a stable selective effect without non-selective collapse. To rule out sample overfitting, we repeated the sweep under a pre-fixed 125/125 calibration/evaluation split: the selected dosages remain effective on held-out samples, with every reported 95% CI excluding 0, and the frozen settings transfer unchanged to the 27-task, external-benchmark, and mental-simulation evaluations. Positive effects persist over broad dosage ranges in the swept configurations; Liquid is the sole cliff-like exception (Appendix G.3). Full dosage choices, sweeps, and the held-out protocol are in Appendices A.6 and A.8.

All evaluation inputs are text-only. Thus, the intervention does not remove visual tokens from the input; it tests whether text-only hidden states *use* a VL direction introduced or amplified by multimodal training.

For higher-rank analyses, we replace $vv^\top$ with $BB^\top$: $h \leftarrow h - \alpha\, BB^\top h$. For each multimodal example i, we compute $\delta_i = \mu_{\text{text},i} - \mu_{\text{visual},i}$, where the two centroids are averaged over text and visual positions, respectively. Here $B \in \mathbb{R}^{d\times k}$ contains the top-k left singular vectors of the matrix formed by $\{\delta_i\}$. This extension is used only for the dual-path geometry analysis in §5.

3.3 Evaluation protocol

We evaluate on two families of text-only tasks. **Spatial-imagery** tasks naturally construct an internal spatial scene, whereas **symbol-manipulation** controls rely on formal symbolic operations. Tasks

Table 1: Cross-architecture double dissociation under VL-axis intervention. For each model, we report the peak causal hub, intervention strength, DD= $\Delta_{\text{control}} - \Delta_{\text{spatial}}$ when available, and direction-specificity control. DD is in percentage points; DD > 0 indicates stronger degradation on spatial-imagery tasks than on symbol-manipulation controls. z reports random-null direction specificity.

Model	Fam.	Axis	Hub	α	DD	z	Specificity
Cham-7B	VQ	Emb.	L5	1.0	$+33.2$ pp	$+12.22$	$0/101$ rand.; CI $[+24.8, +45.4]$ pp
Emu3	VQ	Diff.	L3+F	.50/.50	$+20.8$ pp	–	$0/22$ early/final rand.; CI $[+9.2, +27.0]$ pp
Liquid	FT	Diff.	L27	1.0	$+21.6$ pp	–	$0/22$ rand.; CI $[+12.0, +31.2]$ pp
BAGEL	Dual	Gen-Diff.	L21–27	.5	$+16.8$ pp	$+3.19$	$0/22$ rand.; CI $[+10.7, +23.0]$ pp
Janus	Dual	Gen-Diff. (VQ)	L22–28	.7	$+8.0$ pp	$+3.58$	$0/22$ rand.; CI $[+2.93, +13.00]$ pp
Qwen3-2B	ViT	Diff.	L3	1.0	–	$+27.7$	$0/22$ rand.; CI n.e.
Qwen3-8B	ViT	Diff.	L8	1.0	–	$+38.3$	$0/100$ rand.; CI n.e.
Qwen3-32B	ViT	Diff.	L8	1.0	–	$+262.5$	$0/22$ rand.; CI $[+92.0, +118.0]$ pp
Qwen3-A3B	MoE	Diff.	L6–18	1.0	–	$+48.8$	$0/5$ rand.; CI n.e.

Notes. "0/N rand." means no random direction matched the real-axis effect. Abbrev.: Emb./Diff./Gen-Diff. = embedding-centroid/per-layer/ generation-path DiffMean; the VQ tag denotes VQ-token generation states. F = final block; CI n.e. = not estimated. Dashes mean no paired DD or continuous z is reported. Qwen3-A3B uses a smaller random-null budget; Qwen3-32B's large z reflects small null variance and is used as pass/fail evidence, not cross-model effect size. Chameleon-7B: DD is the single-layer L5 peak (`navigate` -27.6, `boolean` $+5.6$); z and the CI use the matched L0–7 block (block DD $+32.0$). For the Qwen configurations, z is the random-null direction-specificity statistic on the `colored_objects/web_of_lies` pair (at 8B, `navigate` itself is nearly unaffected, so no `navigate`-based DD or CI is reported); for Chameleon/BAGEL/Janus, z is the paired DD significance statistic, with direction specificity carried by the 0/N random-control column.

are drawn from BBH [45] and external spatial-reasoning datasets, including bAbI, SpartQA, and StepGame [48, 27, 42] (§4.1, Appendix A.10).

For each task, we measure $\Delta_{\text{task}} = \text{Acc}_{\text{intervened}} - \text{Acc}_{\text{baseline}}$ and summarize selectivity by DD $= \Delta_{\text{control}} - \Delta_{\text{spatial}}$, where DD > 0 means larger spatial degradation. For random-control summaries, we also report $z = (\text{DD}_{\text{axis}} - \mu_{\text{rand}})/\sigma_{\text{rand}}$, where μ_{rand} and σ_{rand} are computed from the matched random-direction DD distribution. We use z as direction-specificity evidence rather than as a cross-model effect size (Appendix A.13). EVT is identified by three causal checks. **Task selectivity** requires larger degradation on spatial-imagery tasks than on symbol-manipulation controls. **Direction specificity** compares the VL axis against norm-matched random axes, with energy-matched controls in representative cases to match removed projection energy. **Layer localization** requires the effect to peak in a small block range. We also test **cross-task gradedness**, where degradation should increase with spatial-imagery demand, and **bidirectional control**, where amplification or reinjection should reduce the deficit. Each task uses 250 samples; decoding details and dosage sweeps are in Appendices A.1 and A.6.

4 Multimodal training reshapes reasoning geometry

4.1 Cross-Architecture double dissociation

We evaluate 10 multimodal configurations. Table 1 summarizes the 9 configurations where the canonical VL axis is causal; Chameleon-30B is analyzed separately (§6; Appendices G.1 and A.5). Across these 9 VL-axis-positive configurations, VL-axis intervention yields a double dissociation, with spatial-imagery tasks degrading while symbol-manipulation controls are preserved or less affected. We identify the *causal hub* by block-wise sweeps followed by single-layer scans. Hyperparameters, splits, and axis details are in Appendix A.

Hub locations reflect modality integration points. Causal hubs are architecture-dependent and align with where visual information enters the language backbone (Figure 2). Hubs appear early for shared-vocabulary or projection-based models such as Chameleon (L5) and Qwen3-VL (L3–L8), but later for post-hoc visual fine-tuning and dual-path generation models such as Liquid (L27), BAGEL,

and Janus-Pro (L21–L28). This argues against a generic perturbation account. The hub is tied to each architecture's vision–language interface. Emu3 further shows two carrier windows, early (L0–L7) and late (L24–L31), with highly correlated per-task drops ($r = 0.97$), suggesting redundant spatial copies (Appendix E.1). In BAGEL and Janus-Pro, selectivity appears mainly in generation-path layers; the understanding-path VL axis overlaps with dominant hidden-state components, so removal causes broad rather than selective damage (§5).

4.2 Specificity and Cross-Benchmark robustness

We next ask whether the effect is spatial-specific rather than generic damage or a benchmark artifact.

Specificity beyond Table 1. Energy-matched controls test a stricter alternative. The VL axis may look special simply because it removes more residual-stream energy. This alternative fails. On the matched Chameleon-7B L0–L7 block, energy-matched directions yield DD $= -1.9 \pm 3.6$ pp, whereas the VL axis yields $+15.2$ pp ($z = +4.8$, one-sided $p < 10^{-4}$ under a normal approximation; Appendix A.14, Table 12). Thus, the effect reflects direction-specific spatial structure rather than removal magnitude or generic damage.

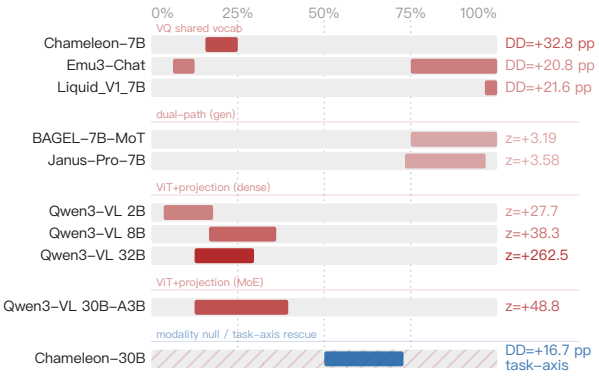
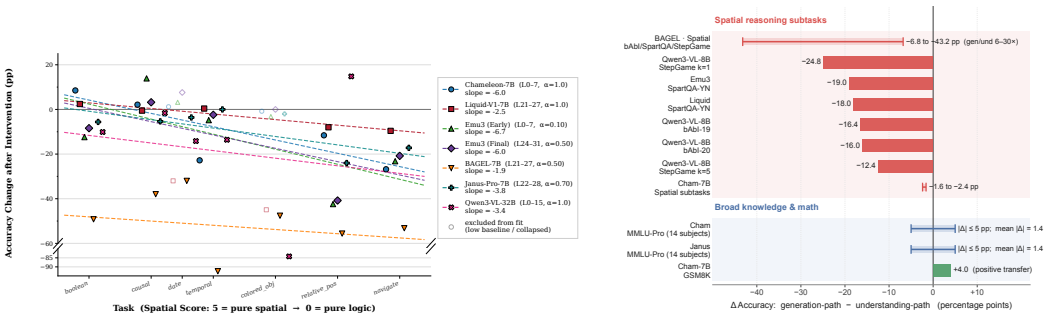


Figure 2: Causal hubs across model architectures.

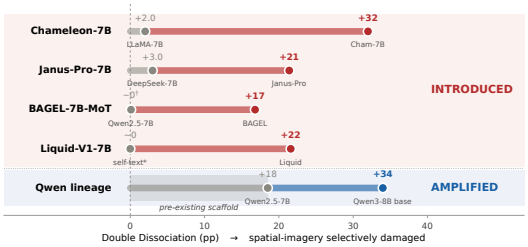
Cross-task gradedness and external benchmarks. Figure 3a reports per-task accuracy changes on a seven-task spectrum ordered by a 0–5 spatial-imagery score annotated by four frontier LLMs and sanity-checked by two AI-background humans. Inter-annotator agreement is high (Kendall's $W \approx 0.98$): six of seven tasks receive unanimous LLM ratings, and the human checks show the same pattern (Appendix A.11). Across configurations, the intervention shows predominantly negative slopes, indicating that degradation increases with spatial demand rather than uniformly across tasks. The negative slope is not an artifact of baseline difficulty: after controlling for baseline accuracy across the seven tasks, the spatial-score slope remains negative in 6 of 7 configurations (-1.6 to -7.3 pp per score point; Chameleon-7B Spearman $\rho = -0.83$, $p = 0.042$; Appendix F). Qwen3-VL-8B is the exception: it passes direction specificity strongly ($z = +38.3$) but shows a positive slope, because its axis also damages low-score control tasks under this protocol (Appendix F.4).



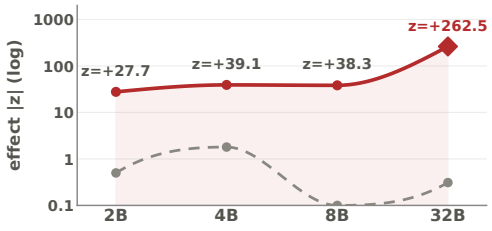
(a) Spatial task spectrum. **(b)** Cross-benchmark transfer.

Figure 3: Task selectivity beyond the main benchmark. (a) VL-axis intervention is increasingly harmful from logic-like to spatial tasks; points with collapsed baselines are excluded from the regression fits. (b) Effects transfer to external spatial benchmarks, while knowledge/math benchmarks stay nearly unchanged.

Figure 3b shows that this selectivity transfers beyond the main benchmark. Drops concentrate on spatial reasoning benchmarks such as bAbI, SpartQA, and StepGame, whereas broad knowledge and math benchmarks such as MMLU-Pro remain nearly unchanged. This pattern argues against generic format disruption and supports a spatial-specific causal effect. One boundary case qualifies this transfer: BAGEL's generation-path intervention at its main dosage also substantially damages GSM8K (-53.6 pp) and LogiQA (-25.2 pp), so knowledge/math preservation holds for the configurations



(a) Cross-architecture attribution.



(b) Qwen3-VL scaling.

Figure 4: Attribution: introduction vs. amplification. (a) Matched-base transfer collapses for four architectures; only Qwen carries a pre-existing scaffold (+18.5 pp on Qwen2.5-7B). (b) In Qwen3-VL, spatial-effect $|z|$ is already large at 2B–8B and becomes strongest at 32B, while control-task $|z|$ stays near zero, indicating scale-strengthened selectivity.

shown in Figure 3b but is not universal across models; we treat the BAGEL external-benchmark cells as a non-selective dosage boundary rather than as transfer evidence.

4.3 Attribution and functional use

We next test where the causal direction comes from and whether the removed signal is functionally *used*, using matched-backbone controls, hub-specific ablation, same-direction rescue, and multiplicative amplification (Appendices G.5 and C.2–C.3).

Attribution: introduction vs. amplification. Multimodal training can either introduce a new causal scaffold or amplify one that already exists in the text-only backbone. For *Chameleon*, the same axis leaves LLaMA-7B essentially unchanged (`navigate` -2.0 pp; every task within ± 3.3 pp, comparable to random directions), versus a -27.2 pp `navigate` drop and DD $= +32.0$ pp in Chameleon-7B, indicating an introduced effect (Figure 4a). For the *Qwen* family, the matched-backbone control reveals an existing scaffold: a Qwen2.5-VL axis yields DD $= +18.5$ pp on the text-only Qwen2.5-7B backbone, and a Qwen3-VL-8B axis transferred to the Qwen3-8B base backbone yields DD $\approx +34$ pp—but with a $\sim$5 pp non-selective damage floor on all tasks, whereas in Qwen3-VL-8B itself the same intervention is cleaner (DD $= +30.0$ pp with controls unaffected). Multimodal training thus amplifies and *purifies* a scaffold already present in the base LLM (Figure 4a).[2] Within Qwen3-VL, direction-specificity z clusters between $+27.7$ and $+39.1$ across 2B–8B and becomes strongest at 32B ($+262.5$). The intermediate 4B run is included only as a scaling sanity check. The affected layer window expands from ≈ 4 to ≈ 11 layers (Figure 4b), supporting amplification.

This amplification trend is family-dependent rather than a universal scaling law: in Chameleon, scaling relocates the carrier deeper (L5 at 7B $\rightarrow$ mid/final blocks at 30B; §6). Cross-family transfer further shows that these carriers are model-specific rather than universal hidden-space directions (Appendix G.5, Table 33).

Functional use: rescue and steering controls. We ask whether the VL direction is functionally used rather than merely diagnostic. Because the L5 ablation causes a layer-specific impairment, we next ask whether the removed information can be recovered. Canonical-axis ablation at the Cham-7B L5 hub gives the headline `navigate` drop of -27.6 pp. In a paired rescue run, L6 reinjection with $\beta = 0.5\sigma$ recovers $+8.0$ pp, while the matched non-hub L20 ablation is null ($+0.8$ pp) and the L20–L21 reinjection control shows no analogous recovery. Rescue is therefore hub-specific rather than a generic residual-stream effect (Appendices C.2–C.5; Tables 18–19).

Additive shifts provide a negative steering control. The update $h \leftarrow h + \beta v$ does not yield a reliable positive rescue regime (Appendix C.4, Table 20). Thus, EVT is carried by a token-wise projection pattern, not by a prompt-independent steering bias.

Finally, BAGEL provides a cross-path rescue test. After removing the compact generation carrier, we ask whether the distributed understanding-path subspace can compensate. We analyze this in §5.

[2] Qwen3 pretraining-data composition is not fully public, so introduction-vs.-amplification cannot be cleanly distinguished for the Qwen3-VL→Qwen3-8B-base transfer in Figure 4a. Qwen2.5-7B, documented as text-only pretrained, provides the cleaner matched-family anchor.

Path	DiffMean (1D, α=0.5) DD (pp)	Top-1 EV	Top-16 EV	PCA rank-k ablation (selectivity)
Generation	$+16.8$ ($z = +3.19$)	19%	69.5%	non-selective at all k; controls drop 56–74 pp
Understanding	-10.9 (n.s., $z = -1.84$)	53%	86.2%	non-selective at all k; k=16: controls -32.3 vs. spatial -19.7 pp

Table 2: Two distinct geometries in BAGEL dual paths. DD $>$ 0 denotes larger spatial than control degradation. DiffMean gives a clean rank-1 carrier in the generation path and is null (slightly reversed) in the understanding path under the same block-level path intervention. PCA rank-k ablation is significant but non-selective in both paths: at every tested k, control tasks degrade at least as much as spatial tasks, indicating entanglement with general capacity rather than a clean carrier (Appendix H).

5 The generation path forms a compact spatial causal carrier

Section 4 shows that EVT appears across architectures and is introduced or amplified by multimodal training. We now locate the carrier in dual-path models. BAGEL-7B-MoT and Janus-Pro-7B provide an internal comparison, because both have a continuous ViT/SigLIP understanding path and a VAE/VQ generation path sharing the same LLM backbone.

One-dimensional carriers are generation-dominant. We construct a path-specific VL axis for each path and intervene over the late LLM blocks used by each model, L21–27 for BAGEL and L22–28 for Janus. Across both models, the one-dimensional generation-path axis yields the stable spatially selective effect, while the understanding-path axis is weak, null, or dominated by broad input-statistics differences. In BAGEL, the generation axis gives a spatial-control double dissociation of $+16.8$ pp ($z = +3.19$), whereas the understanding axis gives -10.9 pp (n.s., slightly reversed). The same pattern is not confined to the BBH-style tasks. On bAbI, SpartQA, and single-hop StepGame, the generation-to-understanding damage ratio ranges from $6\times$ to $31\times$ (Appendix D). On a 27-task suite, Janus-Pro independently reproduces the qualitative effect. `navigate` drops by -28.7 pp, whereas `boolean_expressions` drops by only -7.3 pp (Appendix D.4, Table 23).

This 1D failure does not mean the understanding path lacks spatial information: in Janus-Pro, its DiffMean axis is dominated by a tenfold text/visual norm gap (Appendix G.2), so a null 1D intervention cannot distinguish absent from distributed information, motivating rank-k PCA tests.

DiffMean and PCA reveal complementary geometries. DiffMean and PCA ask different questions. DiffMean tests whether a single vision–text contrast direction forms a clean causal carrier; PCA tests whether related information is distributed over a higher-rank difference subspace. The two can therefore disagree. A 1D contrast can be directly intervenable, while a high-rank subspace can contain stronger but more entangled effects. Figure 5 visualizes this DiffMean–PCA distinction.

For PCA, we center per-sample vision/text mean-difference vectors, compute the top-k basis $\mathbf{B}$, and apply $h \leftarrow h - \alpha \mathbf{B}\mathbf{B}^\top h$, $k \in \{1, 2, 4, 8, 16, 32\}$. In BAGEL's hub block, DiffMean sits off the dominant variance direction in generation (mean $|\cos| = 0.42$ over L21–26; 0.19 averaged over all layers) but is aligned with it in understanding ($|\cos| > 0.85$ at every hub layer; §6; Appendix H, Table 34). Together, they expose complementary path geometries, summarized in Table 2.

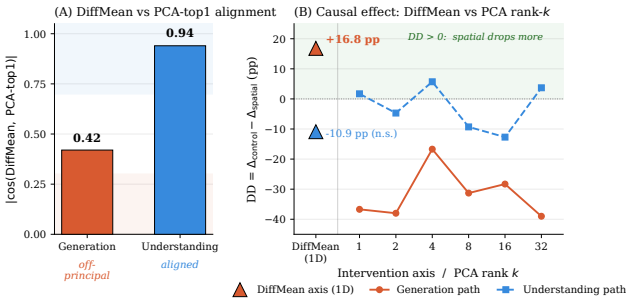


Figure 5: BAGEL dual-path DiffMean vs. PCA geometry. (A) Hub-block DiffMean is off-principal in generation ($|\cos| \approx 0.42$) but top-PC-aligned in understanding (0.94). (B) Positive DD = stronger spatial degradation. Only the 1D generation-path carrier is selective ($+16.8$ pp); the understanding axis is null (-10.9 pp), and rank-k PCA ablation is non-selective in both paths (all DD ≤ 0).

Table 2 summarizes the asymmetry. In generation, DiffMean isolates a compact 1D carrier, while rank-k PCA ablation damages controls more than spatial tasks at every k. In understanding, neither the 1D DiffMean direction nor any tested rank-k subspace yields a spatially selective effect; at k=16 removal is clearly damaging, but controls drop more than spatial tasks (-32.3 vs. -19.7 pp; Appendix H, Table 34). Thus, generation exposes a

compact causal handle, whereas the understanding path's spatial information—though present—is not selectively intervenable at any tested rank.

Generation paths compress; understanding-path information stays entangled. These results support a geometric division of labor. Generation training appears to compress spatial structure into a compact 1D DiffMean signal that can drive visual token generation. The understanding path preserves spatial information in decodable form (§B.4), but neither DiffMean nor PCA isolates a spatial-selective carrier there.

Cross-path rescue is consistent with this reading. After generation-carrier removal, amplifying the rank-16 understanding PCA subspace recovers spatial accuracy (+6.92 pp), but also controls (+5.07 pp; Appendix D.1, Table 21). The recovery is real but not spatially selective: the understanding path carries recoverable task-relevant information in an entangled, non-selective form.

Encoder differences alone are unlikely to explain this, since both paths inject into the same late LLM backbone with the same hidden dimension, and the DiffMean–PCA contrast appears inside the shared backbone rather than only at injection. We interpret this as path-level division of labor, with generation supplying a compact selective carrier and understanding supplying a distributed backup. Controlled training ablations are needed to test whether generation alone causes this compression.

6 From decodability to use: three geometric diagnostics

Sections 4 and 5 show that decodability and causal use need not coincide. In Chameleon-7B, the VL-axis probe peaks at the same layer where VL-axis ablation is most effective, whereas in BAGEL's understanding path, the probe-visible VL-axis direction has little selective causal effect (Appendix B). We therefore use three geometric diagnostics to explain when a readable direction becomes a functional carrier.

First, *concentration* measures how much hidden-state energy lies along an axis: $c_L(\hat{a}) = \mathbb{E}_i[(h_{i,L}^\top \hat{a})^2 / \|h_{i,L}\|^2]$. Second, *alignment* measures whether a readable direction matches the causal direction: $A(\hat{a}) = |\cos(\hat{a}_{\text{readout}}, \hat{u}_1^{\text{causal}})|$. Third, *semantic location* asks which candidate contrast exposes the functional carrier. Operationally, we scan candidate axes or contrasts $a \in \mathcal{C}$ across layers, keep settings that pass the direction-specificity check without causing large control-task damage, and then select the layer–axis pair with the largest double dissociation. For $|\mathcal{C}| > 1$ or task-defined contrasts, we require held-out validation, as in Chameleon-30B (Appendix G.1). This matters when a modality-defined axis is null but a task-contrastive axis succeeds.

Concentration (c) measures how much of the hidden state lies along the causal axis. We report both raw concentration and a centered version that removes the layer mean, because a large raw value can reflect a global mean offset rather than sample-varying signal. This distinction matters for BAGEL. Its generation-path DiffMean axis has very large raw concentration (47–58% across L21–L26), but this is mostly due to the hidden-state mean aligning with the axis ($|\cos(\mu, a)| = 0.71$–0.79). After centering, the concentration drops to 0.43–3.16%, still 15–118× above random directions (Appendix H, Table 35).

Chameleon shows a different geometry. Its L5 VL axis has low centered concentration (0.14%, 5.3× random), yet its ablation produces a spatial-selective effect. Across the L3–L7 hub window, centered concentration ranges from 0.04–0.38%, while the non-hub L20 control falls to the random floor after centering (0.4× random) (Appendix H, Table 35). Thus, concentration separates two mechanisms. Chameleon uses a low-concentration but selective causal signal, whereas BAGEL uses a more concentrated generation-path signal that is mixed with broader task information.

Alignment (A) asks whether the direction that makes a feature linearly readable is also the direction the model causally uses. BAGEL's understanding path shows a clear dissociation. A 1D probe on the DiffMean axis reaches 0.73 accuracy ($+2.65\sigma$ above random), yet removing the same direction is null (DD $= -7.3$ pp; Appendix B, Table 14), and no tested rank-k subspace is selective either (§5): readability and causal alignment dissociate.

We summarize this relation using $|\cos(\text{DiffMean}, \text{PCA-top-1})|$, which yields three regimes. In the *concentric* regime (BAGEL understanding, hub-layer $|\cos| > 0.85$), DiffMean lies inside the dominant variance subspace, but 1D intervention is null. In the *projected-diffuse* regime (BAGEL generation, hub mean $|\cos| \approx 0.42$, all-layer mean 0.19), DiffMean is separated from the dominant

variance direction yet isolates the spatial-selective carrier. In the *mixed* regime (Qwen3-VL), the relation changes across layers, producing non-monotonic rank-k curves that never reach spatial selectivity—the per-layer 1D DiffMean axis is the only selective handle (Appendix H, Table 34).

Semantic location (ℓ^*) asks which contrast reveals the causal carrier. A null result on one contrast does not mean the carrier is absent; it may simply be organized along another semantic axis — or the contrast may have been *estimated* in the wrong space. Chameleon-30B illustrates both. Its modality axis, estimated from the embedding table or synthetic token-ID sequences, is non-causal across 19 block–dosage cells, whereas a `navigate–boolean` task-contrastive axis is causal (held-out DD $= +6.28$ pp, CI $[+0.13, +12.37]$; Appendix G.1). Re-estimating the modality axis input-faithfully, with images through Chameleon's own VQ tokenizer in full multimodal sequences, yields a rotated axis ($\cos \approx 0.4$–0.6) that is strongly causal at mid/final blocks (mid DD $= +33.2$ pp, `navigate` $z = -88.7$; Appendix A.5). The carrier is thus not lost at 30B; it relocates deeper and is exposed by a task contrast or a correctly estimated modality contrast.

Chameleon-7B further shows that the two contrasts are distinct. The modality and task axes are nearly orthogonal ($|\cos| < 0.04$), yet both can be causal at L5 (Appendix G.1, Table 31). Semantic location therefore depends on capacity, architecture, and layer, not simply on model capability.

Diagnostic implication. Together, these cases show why decodability alone is insufficient. Mechanistic claims should pair causal interventions with concentration, alignment, and semantic location. Extended anchor cases are in Appendix H.

7 Discussion and limitations

Our results suggest that multimodal training reshapes text-only reasoning inside the language backbone, not merely via a visual front-end, in an architecture- and pathway-dependent form.

Our study is limited to linear or low-rank interventions and the model families we could access. Our tasks (BBH, bAbI, SpartQA, StepGame, a mental-simulation battery, and a controlled visual$\times$spatial battery; Appendix F) are diagnostic rather than a complete taxonomy. Two boundaries qualify this scope. First, mental-simulation evidence is measurable only in configurations that clear a capability floor: Chameleon-7B is at or below chance on every mental-simulation task, so that evidence comes from the Qwen3-VL configurations. Second, sufficiently long non-spatial state tracking (8–12 sequential updates) shows partial overlap with the same carrier, so the carrier is spatially concentrated rather than exclusively spatial. The analysis may miss nonlinear, highly distributed, or cross-attention-specific mechanisms, and Qwen attribution remains constrained by public uncertainty about pretraining data, although matched-backbone controls reduce this ambiguity. Moreover, axis constructions that bypass the model's real multimodal input pathway can miss true carriers (Chameleon-30B); our conclusions rest on constructions validated against real-image inputs wherever the architecture permits.

Finally, our interventions test necessity rather than the full computation: they do not show whether the model constructs an image-like scene, nor which circuits write and read this signal.

Broader impacts. This work is diagnostic: it introduces no new model, dataset, or generation capability, and improves transparency about when multimodal models rely on visual-spatial representations in text-only reasoning. Similar intervention tools could be misused to manipulate model behavior; we release only analysis code and benchmark scripts.

8 Conclusion

EVT provides causal evidence that multimodal training can make vision-shaped representations usable during text-only reasoning. Across architectures, removing a VL direction selectively harms spatial text tasks while sparing logical controls, with the carrier taking the form of a compact axis, a broader subspace, or an amplified pre-existing geometry. Multimodal training thus changes which internal representations text-only inference causally routes through, not just what models can process.

Acknowledgments and Disclosure of Funding

This research is supported by the National Research Foundation, Singapore, under grant NRF-NRFI10-2024-0004; by the DSO research grant awarded by DSO National Laboratories, Singapore; and by the Ministry of Education, Singapore, under its Tier-1 Academic Research Fund (No. 24-SIS-SMU-040). The authors thank the anonymous reviewers and the area chair for suggestions that led to the mental-simulation battery, the visual×spatial control battery, and the within-task probe analyses added in this version.

References

[1] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction, 2024.

[2] 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, 2025.

[3] Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding information storage and transfer in multi-modal large language models. In *Advances in Neural Information Processing Systems*, 2024.

[4] Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

[5] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357, 2016.

[6] Nicolas Buzeta, Felipe del Rio, Cristian Hinostroza, Denis Parra, Hans Lobel, and Rodrigo Toro Icarte. Seeing to generalize: How visual data corrects binding shortcuts, 2026.

[7] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024.

[8] Fengjiao Chen, Minhao Jing, Weitao Lu, Yan Feng, Xiaoyu Li, and Xuezhi Cao. UniHetero: Could generation enhance understanding for vision-language-model at large data scale?, 2025.

[9] 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, 2025.

[10] Kelly Cui, Nikhil Prakash, Ayush Raina, David Bau, Antonio Torralba, and Tamar Rott Shaham. The dual mechanisms of spatial reasoning in vision-language models, 2026.

[11] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining, 2025.

[12] Woody Haosheng Gan, Deqing Fu, Julian Asilis, Ollie Liu, Dani Yogatama, Vatsal Sharan, Robin Jia, and Willie Neiswanger. Textual steering vectors can improve visual understanding in multimodal large language models, 2025.

[13] Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, 2021.

[14] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021. doi: 10.23915/distill.00030.

[15] Michal Golovanevsky, William Rudman, Vedant Palit, Carsten Eickhoff, and Ritambhara Singh. What do VLMs NOTICE? a mechanistic interpretability pipeline for Gaussian-noise-free text-image corruption and evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 11462–11482, 2025.

[16] 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. *arXiv preprint arXiv:2510.27492*, 2025.

[17] Ruiyan Han, Zhen Fang, XinYu Sun, Yuchen Ma, Ziheng Wang, Yu Zeng, Zehui Chen, Lin Chen, Wenxuan Huang, Wei-Jie Xu, Yi Cao, and Feng Zhao. UniCorn: Towards self-improving unified multimodal models through self-generated supervision, 2026.

[18] Mary Hegarty and David Waller. A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence*, 32(2):175–191, 2004.

[19] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis, 2024.

[20] Raphi Kang, Hongqiao Chen, Georgia Gkioxari, and Pietro Perona. Linear mechanisms for spatiotemporal reasoning in vision language models. In *International Conference on Learning Representations*, 2026.

[21] Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations*, 2023.

[22] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, 2023.

[23] Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Weitao Ma, and Xiachong Feng. Causal tracing of object representations in large vision language models: Mechanistic interpretability and hallucination mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 31645–31653, 2026. doi: 10.1609/aaai.v40i37.40431.

[24] Grace Luo, Trevor Darrell, and Amir Bar. Vision-language models create cross-modal task representations. In *International Conference on Machine Learning*, 2025.

[25] Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2023.

[26] Jack Merullo, Louis Castricato, Colin Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. In *International Conference on Learning Representations*, 2023.

[27] Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. Spartqa: A textual question answering benchmark for spatial reasoning. In *Proceedings of NAACL*, 2021.

[28] Andrew Lee Neel Nanda and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models, 2023.

[29] Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *International Conference on Learning Representations*, 2025.

[30] Yuwei Niu, Weiyang Jin, Jiaqi Liao, Chaoran Feng, Peng Jin, Bin Lin, Zongjian Li, Bin Zhu, Weihao Yu, and Li Yuan. Does understanding inform generation in unified multimodal models? from analysis to path forward, 2025.

[31] Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. Towards vision-language mechanistic interpretability: A causal tracing tool for BLIP. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 2856–2861, 2023.

[32] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models, 2023.

[33] Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and Xudong Wang. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. *arXiv preprint arXiv:2511.19418*, 2025.

[34] Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Krähenbühl, and Vladlen Koltun. Does spatial cognition emerge in frontier models? In *International Conference on Learning Representations (ICLR)*, 2025.

[35] Neale Ratzlaff, Man Luo, Xin Su, Vasudev Lal, and Phillip Howard. Training-free mitigation of language reasoning degradation after multimodal instruction tuning. In *AAAI Symposium Series*, 2025.

[36] Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, 2024. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

[37] Sarah Schwettmann, Neil Chowdhury, Samuel Klein, David Bau, and Antonio Torralba. Multimodal neurons in pretrained text-only transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2023.

[38] Haz Sameen Shahgir, Xiaofu Chen, Yu Fu, Erfan Shayegani, Nael Abu-Ghazaleh, Yova Kementchedjhieva, and Yue Dong. VLMs need words: Vision language models ignore visual detail in favor of semantic anchors, 2026.

[39] Roger N. Shepard and Christine Feng. A chronometric study of mental paper folding. *Cognitive Psychology*, 3(2):228–243, 1972.

[40] Roger N. Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171 (3972):701–703, 1971.

[41] 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, Chaoyou Fu, 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, 2025.

[42] Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. Stepgame: A new benchmark for robust multi-hop spatial reasoning in texts. In *Proceedings of AAAI*, 2022.

[43] Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, Linjie Li, Yu Cheng, Heng Ji, Junxian He, and Yi R. Fung. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers, 2025.

[44] Zihan Su, Hongyang Wei, Kangrui Cen, Yong Wang, Guanhua Chen, Chun Yuan, and Xiangxiang Chu. Generation enhances understanding in unified multimodal models via multi-representation generation, 2026.

[45] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL*, 2023.

[46] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, 2020.

[47] Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. In *Advances in Neural Information Processing Systems*, 2024.

[48] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks, 2015.

[49] Junfeng Wu, Yi Jiang, Chuofan Ma, Yuliang Liu, Hengshuang Zhao, Zehuan Yuan, Song Bai, and Xiang Bai. Liquid: Language models are scalable and unified multi-modal generators, 2024.

[50] Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Mind's eye of llms: Visualization-of-thought elicits spatial reasoning in large language models. *arXiv preprint arXiv:2404.03622*, 2024.

[51] Wulin Xie, Yi-Fan Zhang, Chaoyou Fu, Yang Shi, Bingyan Nie, Hongkai Chen, Zhang Zhang, Liang Wang, and Tieniu Tan. MME-Unify: A comprehensive benchmark for unified multimodal understanding and generation models, 2025.

[52] Zhiyuan Yan, Kaiqing Lin, Zongjian Li, Junyan Ye, Hui Han, Zhendong Wang, Hao Liu, Bin Lin, Hao Li, Xue Xu, Xinyan Xiao, Jingdong Wang, Haifeng Wang, and Li Yuan. Can understanding and generation truly benefit together—or just coexist?, 2025.

[53] Yuliang Zhan, Xinyu Tang, Han Wan, Jian Li, Ji-Rong Wen, and Hao Sun. L2v-cot: Cross-modal transfer of chain-of-thought reasoning via latent intervention. *arXiv preprint arXiv:2511.17910*, 2025.

[54] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency, 2023.

Supplementary Material

A Reproducibility

This section consolidates the experimental setup, compute budget, axis-construction provenance, and random-seed accounting for all main-text claims, in the format expected by the NeurIPS reproducibility checklist. We release the intervention code, per-sample correctness JSONs used for bootstrap CI recomputation, and the scripts needed to reproduce the main experiments (§A.12).

A.1 Hardware and software

All experiments use **8× data-center GPUs with ≈140 GB HBM each** (H200-class), with PyTorch in `bfloat16` throughout. A 7B model in `bfloat16` occupies ≈14–16 GB, so most experiments use 1–2 GPUs per model instance. Three lightweight components (axis construction, position-steering pilots) ran on a single 48 GB or 80 GB GPU. Decoding is greedy; `max_new_tokens`=128 for chain-of-thought formats and 16 for direct-answer formats. Bootstrap confidence intervals use n_{boot}=10000 paired-by-index resamples.

A.2 Compute budget

Table 3 reports per-experiment GPU-hours. The total of **≈150 GPU-hours** is a conservative estimate; Emu3, Liquid, and BAGEL main runs lack itemized wall-clock logs and are absorbed into the 7-task spectrum line. Janus random-control numbers are reported at n=100 per task rather than n=250 because the per-example latency on Janus (≈7–8 s) made $22 \times 5 \times 250 \approx 61$ GPU-hours infeasible; we re-computed real-axis DD on the matched n=100 subset for the random-control comparison and retained n=250 as the headline DD.

Table 3: Compute budget. Per-experiment GPU-hours and configurations. Hardware column denotes the GPU count used; Σ denotes total inference forward passes (samples × tasks × seeds/dirs). Per-experiment configurations are in the released codebase under `experiments/<model>/`.

Experiment	Model	n	tasks	seeds/dirs	GPU-h	Hardware
Cham-7B core (Exp1)	Cham-7B	250	2	1+20	∼22	3×GPU
Cham-7B 60-dir random control	Cham-7B	250	2	60	∼50	8×GPU
Cham-7B axis-config sweep	Cham-7B	250	2	3	∼2.5	GPU
Cham-30B Exp1 + Exp2A	Cham-30B	250	2	1+20	∼14	4×GPU
Cham-30B Exp2B (LOO axes)	Cham-30B	250	2	12 pairs	∼6	2×GPU
Janus random control	Janus	100	5	22	∼24	3×GPU
Janus main effect	Janus	250	5	1	∼2.5	GPU
EXP4 phase 0 (axis construction)	3 models	50/64	—	—	<0.5	GPU
EXP4 phase 1 (k-sweep, 6 values)	3 models	100	6	6	∼6	GPU
EXP4 dose-response	BAGEL und	100	6	3	∼0.2	GPU
7-task spectrum (7 configs)	6 models	250	7	1	∼18	GPU
Position steering (additive/multiplicative)	Cham-7B	80–250	7	5	∼3.5	GPU
Cross-family transfer (V.1–V.4)	4 backbones	250	5	—	∼4	GPU
Total (conservative estimate)					**∼ 150**	

A.3 Axis-construction provenance

The 10 model configurations use three axis-construction methods (embedding centroid, per-layer DiffMean, per-layer PCA on per-sample DiffMean vectors). Sample-source heterogeneity is the most important cross-experiment caveat: BAGEL gen-path and Janus axes use real COCO val2017 images, while all other axes use synthetic random-pixel images plus text prompts. We document this asymmetry explicitly; in pilot experiments we compared synthetic-vs-real source on BAGEL und-path and found split-half cosine ≥ 0.99 across both sources, so the choice does not affect axis identity at our N=50 sample size. Appendix A.4 extends this check to all synthetic-axis models: rebuilding each axis from real COCO images reproduces both the axis direction at the causal hub and the hub intervention effect.

Shared vocabulary does not by itself determine the construction space. Chameleon uses the embedding-centroid axis reported in the main experiments, whereas Emu3 uses per-layer hidden-state

DiffMean axes. We tested the analogous global embedding-centroid axis for Emu3 as an initial pilot, but it produced no meaningful intervention effect (all changes within approximately 2 percentage points; all $p \geq 0.5$), so it is not the axis reported for Emu3 in the main results.

Table 4: Axis construction provenance per model. Method, sample source, sample count N, base random seed, and code reference.

Model	Method	Sample source	N	Seed scheme
Chameleon-7B/30B (main)	Embedding centroid	Embedding rows only (no runtime samples)	—	—
Cham-7B (task-contrastive axis)	Per-layer DiffMean	Synthetic random-pixel images	50	`4096+i` for dir i
Emu3-Chat	Per-layer DiffMean	Synthetic visual-token IDs	50	not seeded (fixed prompt loop)
BAGEL gen (main)	Per-layer DiffMean	**Real COCO val2017 images**	50	`4000+i`
BAGEL und (main)	Per-layer DiffMean	Synthetic random images	50	`1000+i`
BAGEL gen v2 (cross-check)	Per-layer DiffMean	Synthetic random images	50	`2000+i`
EXP4 (3 anchor models, phase 0)	Per-layer PCA on δ_i	Synthetic random pixel images	50/64	split1: `1000+i`, split2: `2000+i`
Janus-Pro-7B (gen)	Per-layer DiffMean	**Real COCO val2017 images**	50	not explicitly recorded
Liquid-V1-7B	Per-layer DiffMean	Synthetic sequences (Emu3-style)	50	not recorded
Qwen3-VL-8B (main)	Per-layer DiffMean	Synthetic random pixel images	64	`1000+i`

COCO usage. The 50 images used for BAGEL gen-path and Janus axes are sampled from **COCO val2017** (CC-BY-4.0). To respect the license without distributing image bytes, the code release ships a `download_coco_50.sh` script with a hard-coded list of image IDs.

Split-half stability. For all DiffMean-based axes, we doubled N to 100 and randomly split into two halves; per-layer cosine similarity between the two halves is ≥ 0.9992 across all models and all layers. This justifies N=50 as our default budget. The high split-half stability justifies using N=50 as our default axis-construction budget.

A.4 Real-image source-swap validation of synthetic axes

A natural objection to synthetic axis inputs is that random pixels (or uniformly random visual-token IDs) are out-of-distribution, so the resulting direction might be an artifact of abnormal visual inputs rather than a modality direction that real images would also produce. To test this, we rebuilt the VL DiffMean axis for every synthetic-axis model with the visual source swapped to 50 real COCO val2017 images (the same canonical image set used for the BAGEL gen-path and Janus axes), passed through each model's own visual pathway: the ViT+projector for Qwen3-VL, the Emu3-VisionTokenizer VQ encoder for Emu3, and Liquid's release pipeline (pad-to-square, 512×512, Chameleon VQGAN, code $+256000$ vocabulary offset) for Liquid. Text prompts, sequence templates, position masks, axis estimator, hook mechanics, and the evaluation harness are unchanged; only the visual source varies. We then (i) compare the real-image axis v_{real} to the synthetic axis v_{syn} per layer — this direction comparison covers *every* configuration for which both axes exist, including BAGEL's understanding path, Janus-Pro's generation path, Chameleon, and all five Qwen3-VL sizes — and (ii) re-run the causal-hub intervention with v_{real} on the identical task samples for the four synthetic-axis models with paired references, including a fresh 5-direction random-null control.

Axis identity. Table 5 reports the per-layer cosine between the two axes for every configuration with both sources. At its causal hub, every positive-result model shows $\cos \geq 0.94$ (Qwen3-VL family ≥ 0.98); BAGEL's understanding path, for which the original pilot reported only an aggregate ≥ 0.99, shows hub-window cosines of 0.941–0.997 with a full-stack median of 0.998. Low cosines appear only where the axis is weakly defined: Qwen3-VL layers L0–5 have centroid separation 19–37 versus 786+ from L6 onward, and there the two sources disagree ($\cos$ 0.29–0.60) — a property of an underdetermined direction, not of the source. COCO-axis split-half stability is ≥ 0.98 for the four re-run models (≥ 0.96 for Chameleon), so these cosines are limited by source agreement, not sampling noise. For Liquid we additionally verified that a 256-px encoding (matching the 256-token

Table 5: Axis identity under the source swap: per-layer cosine between the synthetic-input axis and the real-image (COCO) axis, summarized at each model's causal hub and over the full stack. Low minima occur only at layers where the axis itself is weakly defined (centroid separation 1–2 orders of magnitude below hub layers); COCO-axis split-half stability is ≥ 0.96 for all models at all layers, so the reported cosines are not noise-limited. [a]Janus's canonical axis is already built from COCO; the swapped source is random-RGB images through its VQ tokenizer (same comparison, reversed direction). [b]Chameleon rows compare the full-sequence per-layer DiffMean axes of Appendix A.5; Chameleon's main-result axis is the embedding-centroid direction, which uses no input images at all.

Model (path)	Hub layers	cos at hub	Median (all layers)	Min (all layers)
Qwen3-VL-2B	L3	0.999	0.997	0.49 (L0)
Qwen3-VL-4B	L6–8	0.998	0.998	0.38 (L0)
Qwen3-VL-8B	L8	0.999	0.995	0.29 (L0)
Qwen3-VL-32B	L8	0.995	0.995	0.49 (L0)
Qwen3-VL-30B-A3B	L6–18	0.983–0.989	0.987	0.79 (L0)
Emu3-Chat	L0–7	0.956–0.995	0.955	0.61 (L31)
Liquid-V1-7B	L27	0.951	0.641	0.23 (L0)
BAGEL-7B-MoT (und)	L21–27	0.941–0.997	0.998	0.66 (L1)
Janus-Pro-7B (gen)[a]	L22–28	0.849–0.932	0.963	0.81 (L29)
Chameleon-7B[b]	L0–7	0.73–0.79	0.766	0.67
Chameleon-30B[b]	L36–47	0.90–0.96	0.861	0.54

count of the synthetic sequences) agrees with the 512-px axis at $\cos = 0.986$, so token count is not a confound. The Chameleon rows show the weakest direction agreement (0.73–0.79 at the 7B hub); notably, even there both sources' axes produce a causal spatial-selective effect of the same sign (Table 7), and Chameleon's main-result axis is in any case the embedding-centroid direction, computed from embedding rows without any input images, so the synthetic-input objection does not apply to it.

Table 6: Real-image source swap: the same intervention with the synthetic axis and with the real-image (COCO) axis. Each model has two rows — one per axis source — evaluated on identical task samples. If the synthetic axis were an artifact of out-of-distribution inputs, the real-image rows would show no effect; instead, within every model the two rows show the same pattern: the spatial task is damaged and the control task is untouched. Accuracies in %, base → intervened; DD as in Table 1 (pp, DD > 0 = spatial-selective damage). Spatial/control tasks: `colored_objects/web_of_lies` for Qwen3-VL, `navigate/boolean` for Emu3 and Liquid. $\cos(v_{\text{syn}}, v_{\text{real}})$ is measured at the hub layers. [a]The matched synthetic 32B run did not evaluate `web_of_lies`; the paper CI is shown instead of a DD point estimate. [b]Liquid's real-image axis is stronger per unit α (see text), so its dose-matched comparison uses α=0.5.

Model (hub, α)	Axis built from	$\cos(v_{\text{syn}}, v_{\text{real}})$	Spatial acc.	Control acc.	DD [95% CI]
Qwen3-VL-8B	random pixels		81.3 → 52.7	44.0 → 45.3	+30.0
(L0–8, 1.0)	real COCO	0.999	81.3 → 3.3	44.0 → 45.3	+79.3 [+68.0, +90.7]
Qwen3-VL-32B	random pixels		89.3 → 4.0	—[a]	CI [+92, +118][a]
(L0–15, 1.0)	real COCO	0.995	89.3 → 7.3	26.0 → 45.3	+101.3 [+87.3, +114.7]
Emu3-Chat	random token IDs		41.6 → 0.0	39.6 → 18.8	+20.8
(L0–7, 0.5)	real COCO	0.956–0.995	41.6 → 0.0	39.6 → 11.6	+13.6 [+4.0, +23.2]
Liquid-V1-7B	random token IDs (α=1.0)		54.8 → 42.0	54.8 → 63.6	+21.6 [+12.0, +31.2]
(L27)[b]	real COCO (α=0.5)	0.951	54.8 → 42.0	54.8 → 56.4	+14.4 [+5.2, +23.6]

Causal reproduction. Re-running each hub intervention with v_{real} reproduces the spatial-selective dissociation in every configuration (Table 6; all DD CIs exclude zero, and no random direction matches the effect). Magnitudes match the synthetic CIs for Qwen3-VL-32B, Emu3, and Liquid; for Qwen3-VL-8B the real-image axis is *stronger* (+79.3 vs. +30.0 pp), with the control task still untouched (`web_of_lies` +1.3 pp, $z \approx -0.1$ against the random null). Fine-grained patterns replicate point-for-point: Emu3's `navigate` drop under the early-block intervention is −41.6 pp (to 0.0) under both sources, and Qwen3-VL-32B's paradoxical `navigate` *improvement* is +39.3 pp under both sources.

Two source-sensitive components. First, Liquid's L27 hub sits in logit space (tied embeddings; §G.3), and the real-image axis carries a larger projection onto task hidden states there (0.60 vs. 0.46),

shifting the documented dose cliff to lower α: at α=0.5 the COCO axis reproduces the synthetic α=1.0 headline effect almost exactly (`navigate` -12.8 pp, identical to the synthetic drop; `boolean` unharmed at $+1.6$ pp; DD $+14.4$), whereas $\alpha \geq 0.75$ overshoots into nonselective collapse. Second, Emu3's redundant *final* carrier (L24–31; Appendix E.1) does not replicate under the real-image axis (DD -0.0, CI $[-9.2, +8.8]$), consistent with the larger axis divergence in that window ($\cos$ falls to 0.61 by L31); Emu3's primary early carrier is source-robust. Neither caveat affects the headline configurations, which replicate under real images in all four models. We conclude that the synthetic-input axes are not artifacts of out-of-distribution visual inputs.

The swap also runs in the reverse direction for the two dual-path models, whose canonical axes are already built from real COCO images. For Janus-Pro, a generation-path axis rebuilt from seed-fixed random-pixel images (through the VQ-16 encoder) agrees with the COCO axis at $\cos = 0.85$–0.93 over the causal hub L22–28, and in an environment-matched rerun produces the same causal profile (`navigate` -20.8 pp with controls unchanged, versus -25.2 pp for the COCO axis). For BAGEL's understanding path, the COCO-built axis matches the noise-built axis at $\cos \geq 0.994$ (L21–26) and reproduces the same *negative* result under intervention (control-dominated damage, $z = -1.33$, $p = 0.91$): the understanding-path null is a property of the path, not of the visual source. Likewise, the Janus understanding-path axis rebuilt from random images retains the pathological input-source norm gap ($\|\mu_{\text{text}}\|$=14 vs. $\|\mu_{\text{visual}}\|$=131 at L0, projection ratio 0.91) and collapses non-selectively already at α=0.1, matching the real-image failure in Appendix G.2. Positive and negative conclusions are therefore both source-robust.

A.5 Input-faithful full-sequence axes for Chameleon, and the resolution of the Chameleon-30B boundary case

The source-swap protocol above also closes the one construction gap in the Chameleon family. The canonical Chameleon axis is the embedding-table centroid difference, and the per-layer DiffMean variants previously tested on Chameleon (single-token forwards on 7B; 16 random-token-ID synthetic sequences on 30B, method 7 of Table 29) never passed real images through the model's own VQ tokenizer inside a full multimodal sequence — unlike the Qwen3-VL, Emu3, and BAGEL/Janus constructions. We therefore rebuilt per-layer VL DiffMean axes for both Chameleon models under the input-faithful construction used everywhere else: 50 images (either the canonical COCO set or seed-fixed random-pixel images), each encoded by Chameleon's own in-model VQ tokenizer into 1,024 visual tokens inside a full `<image>`+text sequence, with text/visual position masks, estimator, hooks, and evaluation harness unchanged (α=1.0 unless stated; `navigate`/`boolean`, n=250; split-half stability ≥ 0.95).

Table 7: Full-sequence VL axes on Chameleon-7B/30B: the modality axis is causal in both models, under both image sources. Accuracies in % (base $\rightarrow$ intervened); DD in pp (DD > 0 = spatial-selective). Baselines: 7B `navigate`/`boolean` 41.2/38.8; 30B 57.6/55.6. The 30B late block (L24–35) is a clean null under both sources (DD -0.0, CI $[-6.0, +5.6]$), so the effect is layer-localized, not a global degradation. Random null: random directions exceeding the real axis; z and Δ refer to `navigate` (7B z at α=1.0).

Model (block, α)	Image source	Spatial acc.	Control acc.	DD [95% CI]	Random null
Cham-7B (L0–7, 0.75)	random pixels	41.2 $\rightarrow$ 28.4	38.8 $\rightarrow$ 40.8	+14.8 [+4.0, +25.6]	
Cham-7B (L0–7, 0.75)	real COCO	41.2 $\rightarrow$ 4.8	38.8 $\rightarrow$ 39.6	+37.2 [+26.8, +48.4]	0/5; $z = -12.1$
Cham-7B (L5 only, 1.0)	real COCO	41.2 $\rightarrow$ 32.4	38.8 $\rightarrow$ 46.4	+16.4 [+5.6, +27.2]	
Cham-30B (L12–23, 1.0)	random pixels	57.6 $\rightarrow$ 47.2	55.6 $\rightarrow$ 57.2	+12.0 [+4.0, +20.0]	
Cham-30B (L12–23, 1.0)	real COCO	57.6 $\rightarrow$ 24.0	55.6 $\rightarrow$ 55.2	+33.2 [+21.2, +44.8]	0/3; $z = -88.7$
Cham-30B (L36–47, 1.0)	random pixels	57.6 $\rightarrow$ 39.2	55.6 $\rightarrow$ 59.6	+22.4 [+14.0, +30.8]	0/3; all rand. Δ=0
Cham-30B (L36–47, 1.0)	real COCO	57.6 $\rightarrow$ 39.2	55.6 $\rightarrow$ 59.2	+22.0 [+13.6, +30.4]	0/3; all rand. Δ=0

Chameleon-7B: a third, independent causal direction. The full-sequence axis reproduces the canonical double dissociation (early-block DD monotone in dosage: $+12.0/+33.2/+37.2/+30.8$ pp at α=0.25/0.5/0.75/1.0 with COCO images), yet it is nearly orthogonal to both the embedding-centroid axis (L5 $\cos \approx 0.19$–0.25) and the single-token DiffMean axis (L5 $\cos \approx 0$). Real images strengthen it substantially over random pixels (+37.2 vs. +14.8 pp), mirroring the BAGEL noise-vs-COCO gain.

Chameleon-30B: the modality axis is causal after all. The full-sequence axis is rotated away from the synthetic-token axis of method 7 ($\cos \approx 0.43$ against the random-pixel version, 0.56 against the

COCO version, layer mean) and from the embedding axis ($\cos \approx 0.01$), which explains why the original 19-cell sweep was null: it probed a different direction. Along the input-faithful modality axis, Chameleon-30B shows the same spatial-selective, direction-specific, layer-localized dissociation as every other configuration — `navigate` -33.6 pp with `boolean` unharmed (-0.4 pp) at the mid block — under both image sources, with random directions producing exactly zero `navigate` change at the final block. The implications for the Chameleon-30B boundary-case narrative are discussed in Appendix G.1 (resolution paragraph): the spatial carrier at 30B has not decoupled from the modality contrast; it relocates from the early block (7B, L5 hub) to mid/final blocks and is exposed only when the modality contrast is measured on the model's actual multimodal input pathway.

A.6 Intervention dosage choices

Hook location. Across Chameleon, Emu3, Liquid, Qwen3-VL, BAGEL, and Janus-Pro, the main intervention is implemented with a forward hook on the complete decoder block module (e.g., `model.layers[`ℓ`]`), not a pre-hook and not a hook on `self_attn` or the MLP. The hook modifies the first tensor returned by the block, after its MLP residual update. The modified tensor is passed to the next block before that block's input normalization; for the final decoder block it is passed to the model-level final normalization and then the language-model head. BAGEL's implementation also exposes an attention-output hook for diagnostic experiments, but all BAGEL results reported in the main paper use the default full-block-output position.

The intervention strength α controls how much of the projection onto the tested direction is removed: $h \leftarrow h - \alpha(h^\top v)v$. We treat α as an intervention dosage, not as a learned parameter. For each architecture, we choose a dosage from coarse sweeps over layers and α values. The selected dosage is the smallest value that produces a stable task-selective effect at the causal hub while avoiding non-selective collapse of both spatial and control tasks. Once selected, the same dosage is fixed for the corresponding main-text analyses and random-direction controls.

Table 8 summarizes the dosages used in the main experiments. Most models use α=1.0, corresponding to full projection removal. Some architectures require smaller dosages because the tested direction overlaps with a high-variance or broadly used component of the residual stream. In those cases, large α values damage general token prediction rather than selectively affecting spatial-imagery reasoning.

Table 8: Intervention dosages used in the main experiments. α denotes the removal strength in $h \leftarrow h - \alpha(h^\top v)v$.

Model	Axis / path	α	Selection note
Chameleon-7B	Embedding VL axis	1.0	Full removal; strongest layer-localized DD at L5.
Emu3	Early / final VL axes	0.5 / 0.5	Same dosage for both carriers (L0–7, L24–31); α=1.0 also degrades controls.
Liquid	DiffMean axis (L27)	1.0	Single layer. The L21–27 block has a sharp cliff: DD +5.6 pp at α=0.50, +40.0 pp at 0.55, non-selective collapse at 1.0.
BAGEL	Generation-path DiffMean	0.5	α=0.70 over-suppresses control tasks.
Janus-Pro	Generation-path visual-token	0.7	Most significant in the sweep (z=3.86, L22–28); significant for $\alpha \in [0.30, 0.70]$, marginal at 1.0 (z=1.95).
Qwen3-VL (all four)	Per-layer DiffMean axis	1.0	Full removal at the hub; 30B-A3B is a preliminary specificity check.
Chameleon-30B	Task-contrastive axis	–	VL axis null at every tested dosage (Table 29).

We separate primary removal dosages from reinjection dosages used in the rescue experiments. The rescue intervention uses an additive reinjection magnitude β measured in units of the target-layer projection standard deviation σ, and is reported separately in Appendix C.1 and Appendix C.3.

A.7 Chameleon-7B dose–response curve

To characterize dosage sensitivity in the headline configuration, we ran an 11-point α sweep on Chameleon-7B (L0–7 block, embedding VL axis, n=250 per point per task). Table 9 reports absolute accuracy; α=0 is the no-intervention reference within this sweep.

Relative to the α=0 reference, `navigate` falls monotonically to -24.8 pp by α=0.6 and then changes by at most 5 pp more through α=1.0, while `boolean_expressions` improves throughout.

Table 9: Chameleon-7B 11-point dose–response. `navigate` degrades smoothly with α and largely saturates beyond $\alpha \approx 0.6$, while `boolean_expressions` stays flat or improves. There is no cliff-like transition; Liquid remains the only cliff-like case (Appendix G.3).

α	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
`navigate` (%)	41.2	34.4	25.6	22.4	18.0	17.2	16.4	14.8	13.2	11.2	11.6
`boolean` (%)	38.8	42.0	48.4	46.4	47.2	45.6	48.0	48.4	48.4	48.8	50.0

The selective effect is therefore present across a broad dosage range rather than at a tuned operating point.

A.8 Held-out validation of dosage selection

Because the four swept configurations select α from a coarse sweep on 250 evaluation examples, the selected dosage could in principle overfit the sweep sample. To test this, we repeated the original sweep under a pre-fixed 125/125 calibration/evaluation split: α is selected on the 125 calibration examples only and then evaluated once on the 125 disjoint held-out examples.

The dosage selected on the calibration half remains effective on the disjoint evaluation half: held-out DD is $+20.0$ pp for Emu3 early, $+19.2$ pp for Emu3 final, and $+9.6$ pp for Janus, and every 95% CI excludes 0; Liquid remains the dosage-sensitive exception (Appendix G.3). The effect is also not confined to the selected dosage: Emu3 is significant at every swept $\alpha \in [0.30, 1.00]$ and Janus across $[0.30, 0.70]$. Three further checks involve no dosage selection. No external dataset entered α or hub selection, yet the frozen intervention transfers to external benchmarks (Figure 3b) and to the 27-task and mental-simulation evaluations; the Chameleon-30B task-contrastive axis remains effective on held-out tasks (DD $= +6.28$ pp, CI excludes 0; Table 30); and zero of N norm-matched random directions reproduce the effect, while on Chameleon-7B the real axis gives $z = +4.8$ and exceeds all 20 energy-matched controls (Table 12).

Six of ten configurations use α=1.0 and involve no dosage selection at all; Chameleon-30B's seven-dose null sweep was a boundary check, not a selection. Together with the broad significant dosage ranges, these checks indicate that the reported effects are not artifacts of tuning α to a particular sample or task pair.

A.9 Random seeds

Table 10 records all random-seed assignments across experiments. The convention throughout is `base + i` where `base` is the experiment's anchor and `i` indexes the random direction or axis-construction sample.

Table 10: Random seed assignments per experiment. `base+i` convention indexed by random-direction or sample-construction id.

Experiment	Base seed	Use
Cham-7B/30B random-direction control	2026	`seed = 2026 + i`, $i \in \{0, \ldots, 19\}$
Cham-30B axis build	2026	single seed, default
Cham-7B task-contrastive random control	4096	`seed = 4096 + i`, $i \in \{0, \ldots, 59\}$
EXP4 phase 0 split 1	1000	`default_rng(1000+idx)` per random image
EXP4 phase 0 split 2 (stability)	2000	`2000 + idx` per random image
BAGEL gen (main) axis	4000	`4000 + idx` per sample
BAGEL gen v2 axis	2000	`2000 + idx` per sample
BAGEL und axis	1000	`1000 + idx` per sample
Janus random control	0–21	22 independent seeds
Cross-modal leakage analysis	42	`np.random.seed(42)`, PCA visualization only
Cham-30B rescue pilot	1234	`torch.Generator().manual_seed(1234)`
Bootstrap CI (all models)	–	Stochastic in pilot scripts; CIs stable within ± 0.1 pp. Release scripts fix and log the seed.

Bootstrap reproducibility caveat. The reported CIs are computed from per-sample correctness files. In the pilot scripts, bootstrap resampling used stochastic seeds; repeated runs change CIs by

Table 11: Spatial-imagery score annotation. Each task is rated on a 0–5 scale. The final score is the mean of the four LLM ratings. Human annotations are used as a blind sanity check rather than for computing the final score.

Task	GPT	Gemini	Claude	DeepSeek	LLM Mean	Range	Human-A	Human-B
`navigate`	5	5	5	5	5.00	0	5	5
`relative_position`	3	4	4	3	3.50	1	4	3
`colored_objects`	3	3	3	3	3.00	0	3	3
`temporal_sequences`	2	2	2	2	2.00	0	2	2
`causal_judgement`	1	1	1	1	1.00	0	1	1
`boolean_expressions`	0	0	0	0	0.00	0	0	0
`web_of_lies`	0	0	0	0	0.00	0	0	0

less than ± 0.1 pp because the resampling pool is large. The public release scripts fix and log the bootstrap seed to enable bit-for-bit CI recomputation.

A.10 Datasets

Primary benchmark. BBH (BIG-Bench Hard), publicly released under Apache-2.0. We use 7 spatial-imagery tasks plus symbolic-manipulation controls; full task list in §4.1.

External validation. SpartQA (yes/no and multi-choice subsets), StepGame (k-hop variants), bAbI tasks 17/19/20, GSM8K, LogiQA, MMLU-Pro. All publicly available; download URLs and version pins in `data/README.md`.

Axis-construction images. 50 images from COCO val2017 (CC-BY-4.0) for BAGEL gen-path and Janus; synthetic random-pixel images for all other DiffMean-based axes (fully reproducible from seeds).

Models. All models loaded via HuggingFace IDs (no weights distributed): `facebook/chameleon-7b`, `facebook/chameleon-30b`, `BAAI/Emu3-Chat`, `ByteDance-Seed/BAGEL-7B-MoT`, `deepseek-ai/Janus-Pro-7B`, `Qwen/Qwen3-VL-{2B,4B,8B,32B,30B-A3B}-Instruct`, `Qwen/Qwen2.5-7B`, `Qwen/Qwen2.5-VL-7B-Instruct`, `deepseek-ai/deepseek-llm-7b-base`. The Liquid-V1-7B checkpoint is loaded from its public HuggingFace release (`Junfeng5/Liquid_V1_7B`); no model weights are redistributed by us.

A.11 Spatial-Imagery Score Annotation

To test whether the intervention effect varies with spatial demand, we assign each BBH task a spatial-imagery score from 0 to 5. A score of 5 denotes a task whose natural solution strongly relies on constructing or manipulating an internal spatial scene, whereas a score of 0 denotes a logic-like task that can be solved primarily by symbolic manipulation.

The scores were independently annotated by four frontier LLMs: GPT, Gemini, Claude, and DeepSeek. We use the mean LLM rating as the final spatial-imagery score. To reduce the risk that the ordering is an artifact of LLM self-annotation, two AI-background human annotators also performed a blind sanity check. The annotators were highly consistent (Kendall's $W \approx 0.98$). Table 11 reports the full annotations.

Six of the seven tasks receive unanimous ratings from all four LLM annotators. The only disagreement is on `relative_position`: Gemini and Claude assign a score of 4, whereas GPT and DeepSeek assign a score of 3. This disagreement is consistent with the task's mixed nature: it uses a spatial left–right relation, but much of the reasoning can also be implemented symbolically as ordered-list transitivity. The human sanity check shows the same pattern, with Human-A assigning 4 and Human-B assigning 3.

Mapping to the seven-task spectrum. The plotted spectrum (Figure 3a) evaluates `temporal_simple`, a simplified programmatic variant of `temporal_sequences` that inherits its score (2.0), and additionally includes `date_understanding` (scored 1.5 under the same protocol);

web_of_lies (score 0) is used as a symbol-manipulation control task in the selectivity analyses rather than as a spectrum point.

We use the spatial-imagery score only as a coarse proxy for spatial demand, not as ground truth about cognition. Therefore, our prediction is not that every individual task must follow a strictly monotonic damage pattern. Instead, tasks with higher spatial-imagery scores should, on average, exhibit larger degradation under VL-axis ablation.

A.12 Code release

Released layout. The released repository is organised by responsibility rather than by experiment site. src/core/ contains the shared math: DiffMean / embedding-centroid axis construction (axis.py), projection-removal, additive/multiplicative steering, rescue, and rank-k subspace hooks (hooks.py), both BBH evaluation protocols (eval_bbh.py), DD/bootstrap/McNemar statistics (stats.py), linear probes, and IO helpers. src/models/ provides per-architecture loaders (Chameleon, Qwen3-VL dense and MoE, BAGEL, Emu3, Liquid, Janus). src/tasks/ loads BBH and the supplementary task sets, with all dataset paths parameterised through configs/paths.yaml and environment variables. src/pipelines/ exposes one runnable module per experiment type (axis construction, intervention, layer scan, random control, α-sweep, task spectrum, text-only control, activation addition, rescue, linear probes, rank-k subspace construction and intervention, cross-backbone text ablation, and the BAGEL generation-path axis and intervention); every file carries an origin: header naming the original experiment script it was ported from, so results can be audited line by line. experiments/<model>/run.sh provides one annotated end-to-end launcher per model site, and analysis/ ships the figure scripts and the cross-model bootstrap driver. tests/smoke_cpu.py verifies the intervention math (e.g., that $h - \alpha(h^\top v)v$ zeroes the projection at α=1 and preserves the orthogonal component) without a GPU.

Public release package. The public release package contains the following.

1. requirements.txt pinning all Python dependencies (Python $\geq$ 3.10).
2. Top-level README.md with a step-by-step main-experiment reproduction recipe, and reproduce.sh, a one-command reproduction of the headline Chameleon-7B double dissociation (axis construction $\rightarrow$ L0–7 removal $\rightarrow$ DD with paired-bootstrap CI), including a SMOKE=1 small-n mode.
3. configs/paths.yaml (data and checkpoint roots, overridable via env vars) and configs/models.yaml (HuggingFace IDs and per-architecture layer accessors for every model used).
4. Per-model launchers under experiments/<model>/run.sh that drive the canonical sequence end-to-end on a single machine.
5. data/download_bbh.sh and data/download_coco_50.sh for fetching BIG-Bench-Hard and the 50 COCO val2017 images used to build the BAGEL gen-path and Janus axes (the image bytes themselves are not redistributed), plus the small gradient-task jsonl files shipped in-repo.
6. Fixed seeds logged inline in every result JSON's config field; the released bootstrap scripts fix and log the resampling seed so CIs can be recomputed bit-for-bit from the per-sample correctness files.

Items excluded from release. The release does not include model weights (loaded directly from the HuggingFace Hub via the IDs in configs/models.yaml), raw COCO image bytes (replaced by the download script), the SPACE benchmark items (CC-BY-NC-ND; obtainable from the original release [34]), the upstream BAGEL repository that the BAGEL adapter depends on (pointed at via the EVT_BAGEL_REPO environment variable), preliminary pilot runs not used in main claims, and any local session-state artifacts. One component is documented rather than ported: the BAGEL-internal per-sample extraction behind the rank-k phase-0 subspace construction depends on upstream BAGEL internals; the released subspace_construct pipeline covers the Qwen3-VL path, and the PCA step itself is model-agnostic and included.

A.13 Statistical testing standards

For every primary DD claim in the paper, we report:

- **Effect size**: $\mathrm{DD} = \Delta_{\text{control}} - \Delta_{\text{spatial}}$ in percentage points, with paired bootstrap 95% CI (paired-by-index, n_{boot}=10000).
- **Direction specificity vs. random**: $z = (\mathrm{DD}_{\text{axis}} - \mu_{\text{rand}})/\sigma_{\text{rand}}$, with $N_{\text{rand}} \in \{20, 22, 60, 100\}$ independent random unit directions per model. One-sided p is the empirical fraction of random DDs $\geq \mathrm{DD}_{\text{axis}}$.
- **Energy-matched random control** (Chameleon-7B and Cham-30B): 20 random directions where the per-token removal L_2 norm is forced equal to the canonical-axis removal norm, isolating direction effects from magnitude effects (energy-matched protocol).
- **Held-out validation** (Cham-30B task-contrastive axis): 5 tasks not used in axis construction; held-out DD with 95% CI to rule out within-axis circularity (Chameleon-30B task-contrastive analysis).

We report a model configuration as a positive EVT case only if it passes the four causal checks of §3.3; configurations that show a behavioral drop without passing direction specificity or task selectivity are recorded as diagnostic results, not main-claim evidence.

A.14 Energy-matched random controls

To rule out the possibility that the VL-axis effect is stronger simply because it removes more hidden-state energy, we run an energy-matched random-direction control on Chameleon-7B. For each random direction, we match the per-token projection-removal norm to that of the canonical VL-axis intervention. Thus, the control removes the same amount of hidden-state energy, but along a random orientation.

Table 12 reports the result for the Chameleon-7B VL embedding axis at the early block L0–7 with $\alpha = 1.0$. The VL axis remains substantially stronger than the energy-matched random baseline, confirming that the effect is due to direction specificity rather than removal magnitude.

Table 12: Energy-matched random controls on Chameleon-7B. Random directions are rescaled so that their per-token removal norm matches the canonical VL-axis removal norm. The VL axis remains far outside the energy-matched random distribution.

Direction	DD (pp)	Bootstrap 95% CI	Random DD	z
VL axis, L0–7	$\mathbf{+15.22}$	$[+10.0, +20.5]$	-1.88 ± 3.56	$\mathbf{+4.80}$

Note. The random baseline uses 20 energy-matched random directions; none reaches the real-axis effect (empirical $p < 0.05$; one-sided $p < 10^{-4}$ under a normal approximation of the control distribution).

B Decodability vs. Causal Handle: Two Dissociation Patterns

Section 4.1 reports that decodability and causal handle are not interchangeable diagnostics. This appendix presents the per-layer probe data behind that claim for two complementary cases: Chameleon-7B (where decodability is uniform and causality is localized) and BAGEL's understanding path (where decodability is localized but 1D causality is null). The two cases are not redundant: they document the two distinct ways in which probing can mislead the analyst, and together they motivate the alignment property in §6.

B.1 Chameleon-7B: uniform decodability, localized causality

Setup. We trained two probes per layer on 250 samples from `navigate` (spatial) and 250 from `boolean_expressions` (symbolic-control), with 5-fold stratified cross-validation. The first is a full-dimensional logistic probe on the layer-L hidden state at the last input token (d=4096). The second is a 1D probe on the scalar projection $h \cdot \hat{v}$, where $\hat{v}$ is the canonical Chameleon-7B VL embedding-centroid axis used in the main intervention. As a direction-specificity control we trained 22 matched 1D probes on random-direction projections (each direction sampled once and L2-normalized; seeds 0–21).

Full probe is saturated everywhere. The full-dimensional probe achieves 1.000 accuracy at every one of the 32 layers (mean $\pm$ std across folds: 1.000 ± 0.000). This reflects vocabulary divergence between the two tasks (`left/right/forward` vs. `True/False/and/or`) rather than learned spatial structure: any sufficiently expressive probe can separate the tasks at any depth. The full probe therefore provides no localization signal whatsoever.

1D VL-axis probe is informative at the hub. Table 13 summarizes the 1D-axis probe across layer regions. The probe is informative through the early-to-mid layers (L0–L15, where the causal hub L5 sits), with peak accuracy of 1.000 at L5 itself.

Table 13: Chameleon-7B 1D VL-axis probe. Per-region mean accuracy and z-score range against 22 random unit-norm direction probes. The hub at L5 reaches the maximum accuracy of 1.000.

Layer region	Mean acc	z vs. random	Note
L0–L6 (early, includes hub)	0.944	up to $+1.01$	L5 (hub) $= 1.000$
L7–L15 (mid)	0.982	up to $+1.45$	sustained signal
L16–L31 (late)	0.683	near or below baseline	axis no longer aligned

Implication. At Chameleon-7B's causal hub (L5), the VL-axis 1D probe is fully informative (acc=1.000) and intervention along the same axis produces the largest single-layer DD in the model (§4.1). *Decodability and causal handle co-localize at the hub.* The standalone 1D probe is therefore a usable, but not a sufficient, localizer for the hub: it identifies a wide early region (L0–L15) within which the hub falls but does not single out L5 specifically and gives no information about whether intervention will succeed.

B.2 BAGEL understanding path: localized decodability, null 1D causality

The complementary pattern appears in BAGEL's understanding path. Table 14 reports 1D probe accuracy across the hub block; the probe peaks at L25–L27 with statistically reliable separation from the random baseline.

Table 14: BAGEL und-path 1D probe at the hub block. Per-layer accuracy with z-score against 22 random direction probes. The probe peaks in the last three layers; 1D intervention along the same axis is null at every layer in this block, and rank-k ablation is non-selective.

Layer	1D probe acc	z vs. random
L25	0.701	$+1.52$
L26	0.736	$+1.33$
L27	**0.729**	$\mathbf{+2.65}$
1D DiffMean intervention (block L21–27, α=0.3): DD $= -7.3$ pp ($z = -1.25$, null)		
Rank-16 PCA ablation at hub: damaging but control-dominated (spatial -19.7 vs. control -32.3 pp)		

The probe is informative at late layers, but the 1D intervention is null there, and no rank-k subspace exposes a selective handle either: at k=16, removal damages controls more than spatial tasks (§H, anchor 1). This is the geometric opposite of Chameleon-7B: decodability localizes late, while no tested linear handle—1D or rank-k—is spatially selective in this path.

B.3 Synthesis: probing alone is not a sufficient diagnostic

Table 15 contrasts the two cases.

Two cautionary lessons. **(1)** A high probe accuracy does not imply a usable causal handle along the same direction (BAGEL und). **(2)** A uniform full-dimensional probe across all layers does not imply that all layers are causally equivalent (Chameleon-7B). Each lesson rules out one direction of inference from probing to causal structure. The four-check protocol in §3.3 therefore requires intervention evidence in addition to probing; main-text §6 formalizes the geometric properties (concentration, alignment, semantic location) that explain when probing and intervention agree.

Practical heuristic for new architectures. A probe sweep is cheap; an intervention sweep is expensive. From the Chameleon-7B pattern, a usable heuristic is to locate the layers where a 1D

Table 15: Two dissociation patterns between probing and causal handle. Probe accuracy at the hub layer alone does not predict intervention success: Chameleon-7B's hub probe is informative and 1D intervention works; BAGEL und's hub probe is informative but neither 1D nor rank-k intervention yields a spatially selective effect.

Aspect	Chameleon-7B (L5 hub)	BAGEL und (L21–26 hub)
Full-dim probe	1.000 at every layer	≥ 0.999 at every layer
1D probe at the hub	1.000 (informative)	0.729 ($+2.65\sigma$, informative)
1D intervention at the hub	DD $= +33.2$ pp (L5; block $z = +12.22$)	DD $= -7.3$ pp, $z = -1.25$ (null)
Rank-16 intervention	not needed (1D succeeds)	control-dominated (-32.3 vs. -19.7 pp); not selective
Diagnostic shape	Probe and causality co-localize	Probe localizes; no selective handle at 1D or rank-k

probe along the candidate axis is informative for task identification, and intervene in that range first. From the BAGEL und pattern, a null result at the probe peak should not end the inquiry: rank-k subspace intervention is the natural next step—although in BAGEL's understanding path even rank-k removal turns out to be non-selective, which is itself diagnostic (the information is entangled with general capacity rather than absent).

B.4 Within-task probes: spatial state, not task identity or lexical cues

The cross-task probes above (`navigate` vs. `boolean_expressions`) could in principle exploit vocabulary or input format rather than spatial content; indeed, the full-dimensional cross-task probe is saturated at every layer partly because of vocabulary divergence (§B.1). To remove this confound, we probe *only navigate examples*, holding task identity and input format fixed, and predict parametric spatial state variables.

Exact spatial labels. Each of the 250 navigate instructions is parsed by an exact simulator that replays the instruction sequence and computes the final state. The simulated return-to-start answer matches the gold target on **250/250** examples in both the Chameleon and BAGEL sets, so the derived labels are exact rather than heuristic. The probe targets are (i) *final endpoint relation*: whether the endpoint lies ahead of or behind the start (`dy_sign`; n=126 examples with nonzero displacement; majority class 0.67), and (ii) *final displacement magnitude*: whether $|dx| + |dy|$ exceeds the median (n=250; majority 0.50). Both targets integrate the full instruction sequence. Probes are logistic regressions with 5-fold stratified cross-validation on hidden states at the last prompt token.

Table 16: Within-task probes across all 10 configurations. Full-dimensional probe accuracy for the endpoint ahead/behind target (majority baseline 0.67) and the displacement-magnitude target (majority 0.50), best probed layer; for Chameleon-7B and BAGEL-und the causal-hub layer is shown (L5/L27). Because the ahead/behind label is unbalanced, we also report balanced accuracy on pooled out-of-fold predictions with paired-bootstrap 95% CIs (10,000 resamples, fixed seed); for the balanced displacement target, balanced and plain accuracy coincide. Task identity and input format are fixed by construction, so cross-task lexical separation cannot explain these accuracies.

Configuration	Endpoint acc.	Endpoint balanced acc. [95% CI]	Displacement acc. [95% CI]
Chameleon-7B (L5 hub)	0.842	0.807 [0.726, 0.881]	0.664 [0.605, 0.722]
BAGEL-und (L27 hub)	0.825	0.795 [0.714, 0.872]	0.784 [0.733, 0.835]
Qwen3-VL-8B	0.842	0.819 [0.741, 0.891]	0.828 [0.781, 0.873]
Qwen3-VL-2B	0.881	0.849 [0.774, 0.917]	0.764 [0.711, 0.816]
Qwen3-VL-4B	0.872	0.855 [0.784, 0.921]	0.796 [0.745, 0.844]
Qwen3-VL-32B	0.889	0.867 [0.796, 0.931]	0.832 [0.786, 0.877]
Qwen3-VL-30B-A3B	0.881	0.855 [0.782, 0.922]	0.828 [0.780, 0.873]
Emu3-Chat	0.904	0.885 [0.817, 0.945]	0.760 [0.706, 0.811]
Liquid-V1-7B	0.834	0.807 [0.728, 0.881]	0.780 [0.728, 0.832]
Janus-Pro-7B (gen)	0.857	0.825 [0.748, 0.896]	0.728 [0.673, 0.781]
Majority baseline	0.67	0.50 (chance)	0.50

Across all 10 configurations, the endpoint target is decoded at 0.825–0.904 against a 0.67 majority baseline, and displacement magnitude at 0.66–0.83 against 0.50. Balanced accuracy sits 0.02–0.03 below plain accuracy, and its lower CI bound remains well above the majority baseline in every configuration, so the decodability claim survives label-imbalance correction. One caveat: outside Chameleon-7B and BAGEL, the best layer for the endpoint target is the embedding layer (L0),

whereas the displacement target peaks at late layers; probe localization should therefore not be read as hub localization. In contrast, the paper's 1D causal axes are weak readouts of these same variables (mostly $z < 3$ against 22 random directions). The axis is a causal *carrier*, not a content *register*: within-task state information is present in the residual stream, while the direction that causally matters is not the direction that best reads that state out.

Interpreting model failure. This distinction changes how benchmark failures should be read. Emu3 decodes the endpoint state at 0.904 while answering only 12–18% of the same navigate items correctly, so its failure cannot be explained by a missing decodable representation; the bottleneck must lie in causal access, routing, or downstream readout. Benchmark accuracy alone cannot separate a model that lacks a representation from one that encodes it but fails to use it; probing plus intervention can.

B.5 The decodability–causality dissociation across three architectures

To test whether the dissociation generalizes beyond Chameleon-7B and BAGEL, we extended the probe–intervention comparison to Qwen3-VL-8B (hidden states at the last prompt token, n=250 navigate + 250 boolean examples, all 36 layers).

Table 17: Qwen3-VL-8B: the causally strongest direction is not the most decodable one. Task-identity decoding (navigate vs. boolean) at the causal hub L8, against 1D probes on random unit directions.

Direction (at hub L8)	Task-identity acc.	Causal role
Random 1D directions (mean)	0.825	none (0/100 match real axis)
Causal DiffMean axis	0.540 ($z = -1.49$ vs. random)	$z = +38.3$ under intervention
Full-dimensional probe	1.000	—

Task identity is so lexically salient that *random* 1D directions decode it at 0.825 on average, and a full-dimensional probe is saturated at every layer. Yet the causal axis—extremely strong under intervention at the hub ($z = +38.3$ against 100 norm-matched directions)—decodes task identity at only 0.540, *below* the random-direction mean. The direction that best separates the two tasks is therefore not the direction that controls spatial reasoning. Conversely, the 1D causal axes are weak within-task state readouts (mostly $z < 3$; §B.4). Across Chameleon-7B, BAGEL, and Qwen3-VL-8B, decodability and causal importance dissociate in both directions, and causal interpretation requires intervention together with the concentration, alignment, and semantic-location diagnostics of §6.

C Hub-Specific Recovery of Spatial Information

What the rescue and amplification evidence does and does not establish. The rescue results support three distinct conclusions that should not be conflated. *(i) Chameleon-7B shows partial local recoverability.* Hub removal reduces `navigate` from 47.2% to 34.0%, and L6/L7 reinjection restores $+8.0$/$+7.2$ pp (60.6%/54.5% of the loss; Table 18). Because 22 norm-matched random directions recover $+6.8 \pm 6.8$ pp at the same dosage (§C.5), this establishes local recoverability rather than unique direction-specific recovery; direction specificity rests on the removal-side random and energy-matched controls, not on the rescue. Stronger injections collapse (Table 19). *(ii) BAGEL's generation path exposes a compact 1D carrier:* its 1D DiffMean axis is causally selective (DD $= +16.8$ pp, $z = +3.19$; Table 2). *(iii) BAGEL's understanding path is decodable but not selectively intervenable.* Its 1D DiffMean axis is null (DD $= -7.3$ pp, $z = -1.25$), and rank-16 PCA ablation is damaging but control-dominated (spatial -19.7 vs. control -32.3 pp; DD $= -12.7$, CI $[-21.7, -4.0]$); cross-path rank-16 amplification recovers spatial and control performance comparably ($+6.92$/$+5.07$ pp; Table 21). The information is present, decodable, and partially recoverable, but entangled with general capacity rather than organized as a substitute carrier.

Note on the three navigate-drop figures. The -27.6, -20.7, and -13.2 pp `navigate` drops reported for Chameleon-7B correspond to three distinct protocols run in different sessions with slightly different baselines: the single-layer scan at the L5 hub (n=250, baseline 42.0%), the frozen 27-task L0–7 run (n=150, baseline 43.3%; Appendix F.5), and the L5 ablation inside the paired rescue run (n=250, baseline 47.2%; Table 18), respectively.

This appendix expands on the rescue experiment in main-text §4.3: removing the L5 axis component impairs spatial reasoning, and reinjecting that component at L6 restores a sizable fraction of the lost performance. The experiment establishes two facts: (i) spatial information is concentrated in the early hub—L5 is the strongest single-layer site (-27.6 pp in the layer scan, with L6/L7 also substantial at -20.8/-21.2 pp and L0–L4 ≤ 6.0 pp), while the non-hub L20 control is null—and (ii) the lost performance can be partially recovered by reinjection in the immediately downstream layer, indicating that the information is in a recoverable form rather than destructively erased.

C.1 Dosage calibration

Reinjection dosage is reported in units of σ, the standard deviation of the per-token projection of hidden states onto the canonical VL axis at the hub layer. We calibrated σ at L5 using n=200 `navigate` sequences, 13,236 token positions in total, yielding $\sigma = 8.17$ (a previous calibration with n=50 sequences gave $\sigma = 8.16$, providing a consistency check). We use β=0.5σ=4.08 as the canonical reinjection dosage; larger values induce non-selective collapse (§C.3).

C.2 Layer specificity: L5 is the strongest single-layer hub

Table 18 contrasts the L5 hub with a non-hub layer (L20) in the paired rescue run. In this run, L5 ablation drops `navigate` accuracy by 13.2 pp; the same ablation operation at L20 produces no detectable effect (+0.8 pp, within baseline noise). The headline single-layer scan reports a larger 27.6 pp drop, while this paired run is used for the rescue comparison.

Table 18: Hub specificity in the paired rescue experiment. The same canonical-axis ablation at α=1.0 produces a substantial drop at L5 (the causal hub) but is null at L20. Reinjection of the canonical axis at the immediately downstream layer restores spatial performance at L5–L6 but not at L20–L21, confirming that the recoverable information is hub-specific.

Condition	navigate accuracy	Δ vs. baseline
Baseline (no intervention)	47.2%	—
L5 ablation (α=1.0, hub)	**34.0%**	**-13.2 pp**
L5 ablation + L6 reinjection (β=0.5σ)	42.0%	-5.2 pp (recovered +8.0 pp)
L5 ablation + L7 reinjection (β=0.5σ)	41.2%	-6.0 pp (recovered +7.2 pp)
L20 ablation (α=1.0, non-hub)	48.0%	+0.8 pp (null)
L20 ablation + L21 reinjection (β=0.5σ)	45.6%	-1.6 pp (no recovery effect)

Reading the table. The clean comparison is L5 ablation vs. L20 ablation: the same operation at the same dosage produces a 13.2 pp drop at one layer and no effect at another. This is direct evidence that the spatial information targeted by the canonical axis is concentrated in the early hub region (strongest at L5) rather than present throughout the network. The reinjection rows further establish that the information at L5 is in a recoverable form: a β=0.5σ injection at L6 restores 60.6% of the lost performance, and the same restoration is unavailable when neither the ablation nor the reinjection target the hub (the L20–L21 control row).

C.3 Dosage scan

Table 19 reports the β-sweep at the hub. The β=0.5σ choice in the paired rescue experiment lies in the productive regime; larger doses overshoot and induce non-selective collapse, consistent with the general pattern that intervention magnitudes at 1σ or above approach the residual-norm budget at this layer.

Table 19: Reinjection dosage scan at L6 after L5 ablation. The productive regime is at β=0.5σ.

β	navigate accuracy	Status
0.5σ (4.08)	42.0%	productive (recovers +8.0 pp)
1.0σ (8.17)	0.0%	overshoot (token-prediction collapse)

C.4 Additive versus multiplicative steering

We compare two forms of positive-direction intervention along the VL axis. Multiplicative amplification rescales each token's existing projection:

$$h \leftarrow h + \beta\,(h^\top v)\,v, \quad \beta > 0, \tag{1}$$

where the magnitude of the perturbation varies per token according to that token's own projection. Additive steering instead applies a uniform offset independent of each token's current representation:

$$h \leftarrow h + \beta\,v. \tag{2}$$

Both are tested on Chameleon-7B (early block L0–7) and Qwen3-VL-8B (early block L0–8; additionally at the identified causal hub L8 alone), at n=150 samples per task.

Results. Table 20 summarises the outcomes. For reference, the multiplicative *removal* condition ($h \leftarrow h - \alpha(h^\top v)v$, α=1.0) is included in each block.

Table 20: Additive versus multiplicative steering along the VL axis. Δ denotes accuracy change relative to baseline. Spatial task: navigate (Chameleon-7B) or colored-objects (Qwen3-VL-8B). Control task: boolean-expressions (Chameleon-7B) or web-of-lies (Qwen3-VL-8B). DD = $\Delta_{\text{control}} - \Delta_{\text{spatial}}$ (positive DD indicates spatial-selective damage, consistent with the main-text convention). Hub injection uses n=247 at L8 only.

Model / block	Condition	β	Δ_{spatial}	Δ_{control}	DD
Chameleon-7B, early block L0–7 (baseline: nav = 37.3%, bool = 35.3%)					
	Removal ($-\alpha$)	1.0	−22.0 pp	−2.0 pp	+20.0 pp
	Multiplicative ($+\beta$)	0.3	−15.3 pp	+0.7 pp	+16.0 pp
	Multiplicative ($+\beta$)	0.5	−30.7 pp	−14.0 pp	+16.7 pp
	Multiplicative ($+\beta$)	1.0	−37.3 pp	−35.3 pp	+2.0 pp
	Additive shift	0.3	−4.7 pp	−4.7 pp	0.0 pp
	Additive shift	0.5	−24.0 pp	−9.3 pp	+14.7 pp
Qwen3-VL-8B, early block L0–8 (baseline: co = 81.3%, wol = 44.0%)					
	Removal ($-\alpha$)	1.0	−28.7 pp	+1.3 pp	+30.0 pp
	Multiplicative ($+\beta$)	0.1	−0.7 pp	0.0 pp	+0.7 pp
	Multiplicative ($+\beta$)	0.3	**+3.3 pp**	−2.0 pp	−5.3 pp
	Multiplicative ($+\beta$)	0.5	+2.7 pp	+2.0 pp	−0.7 pp
	Multiplicative ($+\beta$)	1.0	−67.3 pp	−44.0 pp	+23.3 pp
Qwen3-VL-8B, hub injection at L8 only (baseline: co = 82.2%, wol = 48.6%, n=247)					
	Removal ($-\alpha$)	1.0	−75.7 pp	+2.0 pp	+77.7 pp
	Multiplicative ($+\beta$)	0.05–0.5	$\leq \lvert-0.8\rvert$ pp	$\leq \lvert-1.2\rvert$ pp	null
	Additive shift	0.5–10	$\leq \lvert-1.2\rvert$ pp	$\leq \lvert+0.8\rvert$ pp	null

Summary of findings. Removal is strongly asymmetric. Across both models and all layer configurations, multiplicative removal selectively disrupts spatial-imagery performance while leaving or improving symbol-manipulation controls, with effects of −20 to −76 pp on the spatial task. Neither positive-direction mode matches this magnitude or selectivity.

No positive regime for Chameleon-7B. On Chameleon-7B, multiplicative amplification damages the spatial task at every tested β (navigate drops −15 to −37 pp). Additive shift at β=0.3 is near-symmetric (−4.7 pp on both spatial and control), while larger magnitudes produce non-selective degradation. Neither mode improves navigation accuracy.

Weak bidirectional signal for Qwen3-VL-8B (early block only). On Qwen3-VL-8B, multiplicative amplification at β=0.3 yields a weak positive signal: colored-objects improves by +3.3 pp while web-of-lies declines by −2.0 pp (DD = −5.3 pp under the main-text convention, i.e., the spatial task improves relative to the control). This is substantially smaller than the removal effect (−28.7 pp; a ratio of approximately 1:9), and the effect disappears at β=1.0 where non-selective collapse occurs.

Hub injection is null for both modes. Targeted injection at the identified causal hub (L8) with both multiplicative amplification (β=0.05–0.5) and additive shift (β=0.5–10) produces changes of at most 1.2 pp on any task—indistinguishable from zero—despite the same hub supporting a −75.7 pp selective disruption under removal.

Interpretation. The hub-injection null result, combined with the large removal effect, is consistent with a saturation account: the model already exploits the VL-axis projection at or near capacity during spatial-imagery reasoning, so amplifying the existing signal yields no further gain. On Chameleon-7B, where the VL axis is derived from embedding centroids rather than per-layer DiffMean activations, the axis may carry broad modality-separation information rather than a precise spatial-imagery carrier; amplifying it then disrupts rather than supports spatial reasoning. The small positive effect observed for Qwen3-VL-8B multiplicative amplification (+3.3 pp) suggests that the DiffMean axis is a more targeted spatial carrier in that model, but the strong removal–amplification asymmetry indicates that the causal role of the axis is primarily revealed by removal rather than by positive steering. Together with the failure of additive shifts, this suggests that the useful signal lies in the existing projection pattern along the axis, not in a uniform global offset.

C.5 Note on direction-specific recovery

For completeness, we tested whether reinjecting along randomly sampled directions at the hub-downstream layer (L6) reproduces the recovery effect. At β=0.5σ, with 22 norm-matched random unit directions (seeds 0–21), the mean recovery is $+6.8 \pm 6.8$ pp, with a wide spread across seeds. The canonical VL-axis recovery of $+8.0$ pp lies above this distribution mean, but the difference is within the seed-to-seed variance.

We interpret this as a feature of the dosage regime rather than a refutation of the canonical axis: at β=0.5σ in d=4096, partial restoration of residual-stream energy at the hub-downstream layer suffices to relieve some of the L5 ablation damage, and the canonical axis is one of many directions that do this. The principal evidence in this experiment is the layer-specific contrast in Table 18, not a claim of unique direction-specific recovery at this dosage.

Headline claim. The experiment establishes that the spatial impairment is hub-localized (L5 is the strongest single-layer site; the non-hub L20 is null) and that the impairment is recoverable. Direction-specific recovery at lower dosages is a separate question that the present experiment does not adjudicate.

D Cross-Task Path Separation: Extended Evidence

Main-text §5 reports that on BAGEL, the gen-path intervention damages performance on spatial-reasoning datasets several times more than the und-path intervention. This appendix gives the per-task data behind that claim and reports two further findings: a complexity-dependent compression of the path-separation ratio in multi-hop spatial reasoning, and an architecture-dependent variation in path dominance documented in Janus-Pro.

D.1 BAGEL cross-path rescue

Main-text §5 reports an asymmetric rescue experiment on BAGEL. After removing the compact generation-path carrier, we test whether the understanding-path subspace can compensate for the lost spatial information. We amplify the rank-16 PCA subspace from the understanding path rather than reinjecting the removed generation-path direction.

The first pilot used a stronger generation-path ablation ($\alpha_{\text{gen}} = 1.0$) and a 1D understanding-path direction. This setting caused near-complete collapse and did not provide a useful rescue test. We therefore use a milder generation-path ablation ($\alpha_{\text{gen}} = 0.5$) and a rank-16 understanding-path basis.

Table 21: BAGEL cross-path rescue. After generation-path ablation, amplifying the rank-16 understanding-path subspace partially recovers performance. The recovery is real but not spatial-selective, since control tasks recover by a comparable amount.

Condition	α_{gen}	α_{und}	Navigate	Geometric	Control avg.
Baseline	0	0	58.0	36.0	77.6
Gen-path ablation	0.5	0	3.6	18.8	44.0
Rescue	0.5	−0.10	4.8	33.6	49.1

Relative to the generation-path ablation condition, the rescue condition recovers spatial performance by $+6.92$ pp with bootstrap CI $[+4.53, +9.33]$, and recovers control performance by $+5.07$ pp with bootstrap CI $[+1.87, +8.27]$. The two confidence intervals overlap, indicating that the understanding path contains recoverable task-relevant information, but in a broader and less selective form than the compact generation-path carrier.

D.2 BAGEL: gen-path is the dominant carrier on single-hop spatial tasks

Table 22 reports per-task interventions on BAGEL-7B-MoT for both paths at α=0.5, final block (L21–27), n=250 per task. Each row pairs the gen-path and und-path effects on the same dataset; the rightmost column gives the ratio $|\Delta_{\text{gen}}|/|\Delta_{\text{und}}|$.

Table 22: BAGEL gen-path vs. und-path on three external spatial-reasoning benchmarks. On single-hop tasks (bAbI 17/19/20, SpartQA, StepGame k=1), gen-path damage exceeds und-path damage by 6–31×. Multi-hop StepGame (k=5, 9) shows a compressed ratio — discussed in §D.3.

Dataset	Subtask	Δ_{gen}	Δ_{und}	ratio
Single-hop spatial reasoning				
bAbI	task 17 (positional reasoning)	-6.8 pp	$+0.8$ pp	— (sign-reversed)
bAbI	task 19 (path finding)	-12.4 pp	-0.4 pp	**31**×
bAbI	task 20 (agent motivation)	-43.2 pp	-6.8 pp	**6.4**×
SpartQA	yes/no + multi-choice	-12.0 pp	-0.4 pp	**30**×
StepGame	k=1 (single hop)	-18.4 pp	-2.4 pp	**7.7**×
Multi-hop spatial reasoning				
StepGame	k=5 (5 hops)	-16.8 pp	-13.2 pp	1.3×
StepGame	k=9 (9 hops)	-4.8 pp	-4.4 pp	1.1×

Single-hop tasks support the gen-path-dominant view. On all five single-hop variants (the top block), the gen-path intervention damages spatial reasoning 6–31× more than the und-path intervention. The bAbI task 17 row, where und shows a small positive change ($+0.8$ pp) while gen drops by -6.8 pp, has a sign-reversed ratio that we report as "—" rather than ∞; the qualitative direction is consistent with the rest of the single-hop block.

D.3 Multi-hop reasoning recruits both paths

On multi-hop StepGame (k=5, k=9), the gen/und ratio compresses to 1.1–1.3×. Two observations frame this finding.

First, the absolute gen-path damage on multi-hop variants (-16.8 pp at k=5, -4.8 pp at k=9) remains comparable to the single-hop case (-18.4 pp at k=1); the gen-path is still impacted. What changes is that the und-path damage rises in tandem.

Second, multi-hop spatial reasoning is task-structurally different from single-hop: reaching a k=9-step spatial conclusion requires iterated application of intermediate spatial relations rather than a single atomic spatial-mapping operation. This is consistent with iterative reasoning recruiting both representational paths — the gen-path for the underlying spatial scene, the und-path for the symbolic transition rules — in tandem rather than in sequence. Under this view, the path-separation effect we measure is sharpest where a single read-off from a constructed spatial scene suffices, and dilutes when the task has a large symbolic-reasoning component layered on top of the spatial substrate.

Note also that multi-hop StepGame baselines are low (32.4% at k=5, 19.2% at k=9), so absolute Δ values carry larger relative noise than in the single-hop case; this contributes to the compressed ratio. We do not lean on this point: the structural argument in the previous paragraph is the substantive interpretation. We view this as a post-observation interpretation rather than a causal claim; controlled training-objective or path-isolation ablations are needed to test it directly.

D.4 Janus-Pro 27-Task BBH Analysis

Janus-Pro-7B provides an independent test of the generation-path hypothesis on the full BBH task suite. We evaluate 27 BBH tasks with $n = 150$ examples per task, $\alpha = 0.5$, layers L21–29,

Table 23: Janus-Pro-7B gen-path vs. und-path intervention on 27 BBH tasks. The generation-path axis produces a strong drop on `navigate` (-28.7 pp) but only a mild drop on `boolean_expressions` (-7.3 pp), reproducing the qualitative spatial-selective effect on a broader task suite.

Task	Baseline	Gen Δ	Und Δ	Gen–Und	Pattern
`boolean_expressions`	98.7	-7.3	-4.7	-2.7	both
`causal_judgement`	64.0	-4.7	-10.7	$+6.0$	und $>$ gen
`date_understanding`	40.0	-0.7	-6.7	$+6.0$	und $>$ gen
`disambiguation_qa`	50.7	-1.3	-2.7	$+1.3$	$\sim$
`dyck_languages`	12.0	0.0	-3.3	$+3.3$	$\sim$
`formal_fallacies`	24.7	-10.0	-2.7	-7.3	gen $>$ und
`geometric_shapes`	100.0	-16.0	0.0	-16.0	gen $\gg$ und
`hyperbaton`	32.7	$+1.3$	-16.7	$+18.0$	und $\gg$ gen
`logical_deduction_5_obj`	75.3	-24.0	$+1.3$	-25.3	gen $\gg$ und
`logical_deduction_7_obj`	70.0	-3.3	$+3.3$	-6.7	gen $>$ und
`logical_deduction_3_obj`	78.7	-0.7	$+0.7$	-1.3	$\sim$
`movie_recommendation`	32.7	$+5.3$	-4.7	$+10.0$	und $>$ gen
`multistep_arithmetic_two`	7.3	-2.0	-5.3	$+3.3$	$\sim$
`navigate`	81.3	-28.7	-24.0	-4.7	both, similar
`object_counting`	49.3	-6.7	$+1.3$	-8.0	gen $>$ und
`penguins_in_a_table`	31.5	-0.7	-2.1	$+1.4$	$\sim$
`reasoning_about_colored_objects`	25.3	$+0.7$	$+1.3$	-0.7	$\sim$
`ruin_names`	17.3	-0.7	-8.7	$+8.0$	und $>$ gen
`salient_translation_error`	99.3	-34.0	-56.0	$+22.0$	und $\gg$ gen
`snarks`	54.0	-2.7	-20.7	$+18.0$	und $\gg$ gen
`sports_understanding`	90.7	-5.3	$+6.0$	-11.3	gen $>$ und
`temporal_sequences`	68.0	-1.3	-13.3	$+12.0$	und $>$ gen
`tracking_shuffled_5_obj`	52.7	$+6.7$	$+20.7$	-14.0	und helps more
`tracking_shuffled_7_obj`	40.0	-7.3	$+16.7$	-24.0	und helps, gen hurts
`tracking_shuffled_3_obj`	37.3	-11.3	$+4.7$	-16.0	gen $>$ und
`web_of_lies`	70.7	-9.3	0.0	-9.3	gen $>$ und
`word_sorting`	0.0	0.0	0.0	0.0	floor

and 0-shot decoding. This run uses an earlier fixed generation-path configuration rather than the main-table one (L22–28, $\alpha = 0.7$; Table 1); both dosages lie inside Janus's significant range $\alpha \in [0.3, 0.7]$ (Appendix A.6), and neither run is retuned per task. The generation-path axis is constructed from the VQ-16 visual codes, whereas the understanding-path axis is constructed from the SigLIP understanding encoder. Table 23 reports the full per-task results.

Overall, the generation-path intervention causes an average drop of -6.1 pp across the 27 tasks, with 11 of 27 tasks dropping by more than 5 pp. The understanding-path intervention causes an average drop of -4.7 pp, with 9 of 27 tasks dropping by more than 5 pp. Tasks with $|\text{Gen} - \text{Und}| \geq 10$ pp show a balanced split: four are more affected by the generation path (`geometric_shapes`, `logical_deduction_5_obj`, `tracking_shuffled_3_obj`, and `tracking_shuffled_7_obj`), and four are more affected by the understanding path (`salient_translation_error`, `hyperbaton`, `snarks`, and `temporal_sequences`). This supports a functional split between a compact generation-path carrier and a broader understanding-path subspace.

The `tracking_shuffled_objects` family is unstable across object counts: the generation path hurts the 3-object and 7-object variants, while the understanding path improves the 5-object and 7-object variants. This suggests that tracking tasks mix spatial tracking and symbolic ordering strategies, making them less clean as a single functional category.

D.5 Janus-Pro: a more balanced two-path architecture

For comparison, on Janus-Pro the SpartQA intervention shows a different pattern: the gen-path early-block intervention has no net effect ($+2.0$ pp), while the und-path early-block intervention is

mildly performance-enhancing ($+12.0$ pp), and the und-path final-block intervention is null (-1.2 pp). The standalone gen-path-dominance finding seen in BAGEL is not reproduced on Janus.

This is consistent with our 27-task BBH analysis (full per-task data in §D.4), where the across-task gap between gen-path and und-path effects is substantially smaller in Janus-Pro than in BAGEL. We interpret the comparison as evidence that the magnitude of path separation is architecture-dependent: BAGEL's training regime concentrates spatial information on the gen-path more sharply than Janus-Pro's. The qualitative finding that a generation-related axis is causal for spatial reasoning holds in both architectures (§5), and indeed the Janus gen-path DiffMean axis remains the only carrier that passes all four causal checks in that model (§5, §G.2). The Janus result does not refute the gen-path-as-carrier hypothesis; it places a quantitative caveat on cross-architecture extrapolation of the BAGEL ratio.

E Architecture-Specific Main-Effect Details

E.1 Emu3-Chat dual-carrier redundancy

Most positive cases exhibit a single causal hub, but Emu3-Chat is an exception. Its spatial-selective effect appears in two separated layer windows: an early block (L0–7) and a final block (L24–31), both with the main effect at $\alpha = 0.50$ (the task-spectrum runs used $\alpha = 0.10$ for the early carrier). The per-task accuracy drops induced by the two carriers are highly correlated ($r = 0.967$), and their spectrum slopes are nearly matched: -6.7 pp/unit for the early carrier and -6.0 pp/unit for the final carrier.

This suggests that the two blocks carry redundant copies of a similar spatial representation rather than two unrelated task effects. Consistent with this interpretation, ablating either block alone is sufficient to produce a spatial-selective drop (early DD $= +20.8$ pp, final DD $= +17.2$ pp). A complete removal of Emu3's EVT signal may therefore require joint intervention on both layer windows, rather than the single-hub intervention used for most other models.

Source-dependence caveat. The real-image source swap (Appendix A.4) shows that the two carriers are not equally robust. The early carrier replicates under COCO-image axes and even strengthens at the L3 hub (single-layer DD $= +24.4$ pp, CI $[+14.8, +34.4]$, versus $+14.0$ pp with the synthetic axis). The final carrier does not replicate under the real-image axis (DD $= -0.0$, CI $[-9.2, +8.8]$), consistent with the larger axis divergence in the final window ($\cos$ falls to 0.61 by L31). The dual-carrier redundancy described above should therefore be read as a property of the synthetic-token-ID axis; under input-faithful real-image axes, Emu3 presents a single, early causal carrier.

F Task Construct Validity: Mental Simulation, the Visual×Spatial Battery, and Matched Controls

A central alternative explanation is that the spatial/symbolic contrast reflects correlated task properties—baseline difficulty, reasoning depth, reasoning structure, or the particular task pair—rather than spatial demand. This appendix reports four complementary controls: (i) a design-based mental-simulation battery with matched format and frozen interventions (§F.1); (ii) a 2×2 battery crossing spatial demand with visual content (§F.2); (iii) a matched-depth, answer-completion-controlled dissociation (§F.3); and (iv) a statistical control for baseline accuracy across the task spectrum (§F.4), together with the complete 27-task BBH suite under a frozen intervention (§F.5).

F.1 Text-formatted mental-simulation battery

We built text-formatted mental rotation [40], paper folding [39], and perspective-taking tasks [18] with programmatically generated, solver-verified answers and balanced options, and additionally evaluated mental rotation (MRT) and perspective taking (PTT) items *verbatim* from the SPACE benchmark [34]. All tasks are 4-way multiple choice (chance 25%) with the same answer format, prompt template, and scoring. Qwen3-VL-8B/32B are evaluated under the paper's published configuration—per-layer DiffMean axes, early block, α=1.0—with no task-specific retuning (n=150 for 8B, n=100 for 32B).

Interpretability floor. Cells whose baseline falls below a pre-specified chance+10 pp threshold (35%) are marked *floor*, reported without inferential claims, and excluded from inference; † marks a baseline marginally below the threshold, reported for completeness. Every interpretable drop has a paired item-level bootstrap 95% CI (10,000 resamples, fixed seed) and an exact two-sided McNemar test.

Table 24: Mental-simulation battery under the frozen VL-axis intervention. Accuracy is baseline→intervened (%); Δ in pp with paired bootstrap 95% CI and exact McNemar p. Floor cells (baseline $< 35\%$) are excluded from inference.

Model	Task	n	Base→Int.	Δ (pp)	95% CI	McNemar p
Qwen3-VL-32B	Mental rotation (easy)	100	78.0 → 23.0	−55.0	[−66, −44]	1.4e−13
Qwen3-VL-32B	Mental rotation (hard)	100	61.0 → 15.0	−46.0	[−57, −34]	1.2e−10
Qwen3-VL-32B	Paper folding (easy)	100	89.0 → 50.0	−39.0	[−49, −29]	2.2e−10
Qwen3-VL-32B	Paper folding (hard)	100	36.0 → 31.0	−5.0	[−17, +7]	0.52 (n.s.)
Qwen3-VL-32B	Perspective taking (easy)	100	54.0 → 32.0	−22.0	[−31, −13]	2.7e−5
Qwen3-VL-32B	Perspective taking (hard)	100	62.0 → 36.0	−26.0	[−36, −16]	6.2e−6
Qwen3-VL-32B	SPACE MRT (verbatim)	100	43.0 → 26.0	−17.0	[−31, −3]	2.4e−2
Qwen3-VL-32B	SPACE PTT (verbatim)	100	59.0 → 33.0	−26.0	[−39, −13]	3.1e−4
Qwen3-VL-32B	SPACE MPFB (verbatim)	100	floor	—	—	excluded
Qwen3-VL-8B	Mental rotation (easy)†	150	32.0 → 19.3	−12.7	[−20.7, −4.7]	5.4e−3
Qwen3-VL-8B	Mental rotation (hard)	150	floor	—	—	excluded
Qwen3-VL-8B	Paper folding (easy)	150	79.3 → 30.7	−48.7	[−58.0, −39.3]	8.3e−17
Qwen3-VL-8B	Paper folding (hard)	150	floor	—	—	excluded
Qwen3-VL-8B	Perspective taking (easy)	150	59.3 → 27.3	−32.0	[−41.3, −22.7]	1.2e−9
Qwen3-VL-8B	Perspective taking (hard)	150	68.7 → 26.0	−42.7	[−52.0, −33.3]	6.3e−15
Qwen3-VL-8B	SPACE MRT (verbatim)	150	floor	—	—	excluded
Qwen3-VL-8B	SPACE PTT (verbatim)	150	49.3 → 27.3	−22.0	[−31.3, −12.0]	3.8e−5
Qwen3-VL-8B	SPACE MPFB (verbatim)	150	floor	—	—	excluded

Direction specificity. Five norm-matched random directions leave perspective taking within $|\Delta| \leq 4$ pp (real axis: -22 to -43 pp across the two models) and give a mean of -7.2 pp (range -18 to $+6$) on 32B mental rotation, where the real axis causes -55.0 pp. The mental-simulation deficits are therefore axis-specific.

Deficit, not disruption. Perspective taking provides the cleanest reading: under both conditions, both models produce complete, well-formed answers to well-posed questions (answer completion $\approx 100\% \rightarrow 99\%$ under ablation), so the deficit consists of *incorrect answers*, not a failure to answer.

Answer-completion caveat for the other cells. Completion is not fully preserved everywhere: at 32B, mental rotation's completion falls from 77% to 40% and paper folding's from 89% to 50% under ablation, so part of those accuracy drops reflects failure to produce a parseable answer within the generation budget. We therefore treat perspective taking (and the matched-depth battery of §F.3) as the anchor evidence for the mental-simulation claim, with the remaining cells as convergent but completion-confounded support.

Capability floor and architecture coverage. Chameleon-7B is at or below chance on every mental-simulation task, including simplified variants; its outputs are complete and well formed within the generation budget but incorrect, indicating a genuine capability floor rather than a formatting failure. The measurable mental-simulation evidence therefore comes from the Qwen3-VL configurations that clear the interpretability threshold, while the canonical VL-axis phenomenon (Table 1) spans architectures.

F.2 The visual×spatial 2×2 battery

To separate "visual" from "spatial," we cross {visual, non-visual} with {spatial, non-spatial} while matching the 4-way multiple-choice format, prompt template, scoring, and frozen intervention. The visual-spatial and non-visual-spatial quadrants are covered by the mental-simulation battery above and by `navigate`, SpartQA, and StepGame (§4.1–4.2). The critical new cell is *visual non-spatial* attribute binding: recalling colors and materials from a scene description with no spatial relations. Non-visual non-spatial controls (`web_of_lies`, `boolean_expressions`, MMLU-Pro, GSM8K) complete the design.

Table 25: Visual non-spatial attribute binding under the frozen intervention. At 7–8B the visual non-spatial control is largely preserved (Chameleon-7B even improves; CI includes 0), while spatial tasks drop by 22–49 pp; at 32B the same cell collapses, indicating a broader functional range at larger scale.

Model	n	Base→Int.	Δ (pp)	95% CI	McNemar p
Chameleon-7B	250	$38.0 \to 44.0$	$+6.0$	$[-0.4, +12.4]$	0.091 (n.s.)
Qwen3-VL-8B	150	$100.0 \to 96.0$	-4.0	$[-7.3, -1.3]$	3.1e−2
Qwen3-VL-32B	100	$100.0 \to 29.0$	-71.0	$[-80, -62]$	8.5e−22

At 7–8B, the intervention strongly disrupts spatial reasoning while largely preserving non-spatial visual-attribute recall: Chameleon-7B changes by $+6.0$ pp (CI includes 0; all five norm-matched random directions give $|\Delta| \leq 2.8$ pp), and Qwen3-VL-8B changes by only -4.0 pp from ceiling versus 22–49 pp spatial drops. Chameleon-7B starts at only 38%, so preservation of this control is not explained by an easy baseline.

The 32B cell and its robustness. At 32B, attribute binding falls $100.0 \to 29.0$ while `web_of_lies` remains at $58.0 \to 56.0$ (within the random-direction band) and `boolean_expressions` moves mildly ($88.0 \to 77.0$); `navigate` moves moderately ($70.7 \to 57.3$). The attribute-binding drop is axis-specific and robust: a single-layer (hub L8) intervention gives $100.0 \to 9.0$; α=0.5 gives $100.0 \to 20.0$; and a 3-shot short-answer format gives $100.0 \to 32.0$ (with `boolean_expressions` at $88.0 \to 77.0$ under the same 3-shot configuration); three matched-norm random directions all remain at 100.0. Although both attribute-binding cells start at ceiling, their intervention sensitivity differs sharply across scale ($100.0 \to 96.0$ at 8B versus $100.0 \to 29.0$ at 32B).

We read this as a *configuration-dependent broadening of the functional range*: the carrier is spatially concentrated within the visual domain at 7–8B and extends to non-spatial visual attribute binding at 32B, while remaining selective relative to symbolic controls. This broadening is consistent with the widening of the affected layer window at scale (≈ 4 to ≈ 11 layers; Figure 4b) and is why we state the construct (*vision-shaped*) independently of any single task family.

F.3 Matched-depth, answer-completion-controlled dissociation

Spatial tasks might degrade more simply because they require more sequential state updates. We therefore cross spatial demand with update load on Qwen3-VL-8B, using a non-spatial color-tracking task (tracking which color each named object currently has through a sequence of repaint updates) with matched format and frozen intervention.

Table 26: Crossing spatial demand with sequential update load (Qwen3-VL-8B). Completion denotes the fraction of items with a final answer produced within the generation budget (baseline→intervened where both measured).

Task	Update load	Base→Int.	Δ (pp)	Completion
Perspective taking (spatial; easy/hard)	0	$59.3 \to 27.3$ / $68.7 \to 26.0$	-32.0 / -42.7	≥ 0.94
Color tracking (non-spatial)	3	$100.0 \to 79.3$	-20.7	0.98
Color tracking (non-spatial)	5	$100.0 \to 71.3$	-28.7	0.94
Color tracking (non-spatial)	8	$99.3 \to 55.3$	-44.0	0.83
Color tracking (non-spatial)	12	$96.0 \to 44.0$	-52.0	0.81

Three observations separate the two factors. First, mental rotation and perspective taking each present one static scene and one question, with *zero* update load, yet show large drops ($78.0 \to 23.0$ at 32B; $59.3 \to 27.3$ at 8B), so sequential update load cannot be the primary explanation. Second, at matched moderate depth (3–5 updates) the non-spatial cells are less impaired ($100 \to 79/71$) than the zero-update spatial cells, and items are answered within the generation budget on $\geq 94\%$ of trials in these cells, so the spatial deficit is not a failure to produce answers. The non-spatial effect is also axis-specific: five random directions leave color tracking at 100%. Third, at longer horizons (8–12 updates) the non-spatial drop grows to $-44/-52$ pp, revealing partial overlap between the same carrier and long-horizon non-spatial state tracking. Together, these results identify a spatially concentrated carrier with a secondary role in sufficiently long non-spatial state tracking.

F.4 Statistical control for baseline accuracy

Because tasks differ in baseline difficulty, we regress each configuration's per-task intervention effect on the 0–5 spatial-imagery score while controlling for baseline accuracy across the seven-task spectrum (§4.2).

Table 27: Baseline-controlled spatial-score slopes (pp per score point) across the seven-task spectrum. The slope remains negative in 6 of 7 configurations after controlling for baseline accuracy. We report these descriptively and make no pooled-significance claim.

Configuration	Baseline-controlled slope	Note
Qwen3-VL-32B	-7.28	
Chameleon-7B	-5.70	Spearman $\rho = -0.83$, $p = 0.042$
Emu3	-5.25	
Janus-Pro	-3.79	
Liquid	-3.59	
BAGEL	-1.60	
Qwen3-VL-8B	$+1.93$	sole positive slope

F.5 Task selectivity across all 27 BBH tasks

Would the double dissociation survive a broader set of algorithmic and structured-reasoning controls than `boolean_expressions` and `web_of_lies`? We evaluated Chameleon-7B on all 27 BBH tasks under one frozen intervention (embedding VL axis, L0–7, α=1.0, n=150 per task), with no per-task retuning. Table 28 reports the full results; Janus-Pro's independent 27-task run under its own frozen generation-path configuration is in Table 23.

Table 28: Chameleon-7B on all 27 BBH tasks under the frozen L0–7 intervention (α=1.0, n=150). The two largest drops are the two most spatial tasks; algorithmic and structured-logic controls are stable or improve. Tasks with near-floor baselines should be read qualitatively.

Task	Base	Int.	Δ	Task	Base	Int.	Δ
`navigate`	43.3	22.7	−20.7	`date_understanding`	22.0	26.7	+4.7
`tracking_shuffled_3_obj`	29.3	18.7	−10.7	`boolean_expressions`	50.0	54.7	+4.7
`movie_recommendation`	42.7	36.7	−6.0	`ruin_names`	24.0	28.0	+4.0
`snarks`	47.3	41.3	−6.0	`logical_deduction_7_obj`	11.3	15.3	+4.0
`tracking_shuffled_5_obj`	22.7	18.0	−4.7	`formal_fallacies`	84.0	86.7	+2.7
`salient_translation_error`	16.0	12.0	−4.0	`reasoning_colored_objects`	15.3	18.0	+2.7
`logical_deduction_3_obj`	7.3	4.0	−3.3	`geometric_shapes`	2.0	4.0	+2.0
`sports_understanding`	38.7	36.0	−2.7	`temporal_sequences`	8.0	10.0	+2.0
`disambiguation_qa`	24.0	22.0	−2.0	`penguins_in_a_table`	30.8	32.2	+1.4
`hyperbaton`	43.3	42.7	−0.7	`tracking_shuffled_7_obj`	13.3	14.7	+1.3
`multistep_arithmetic_two`	4.7	4.0	−0.7	`dyck_languages`	20.0	20.7	+0.7
`object_counting`	2.0	1.3	−0.7	`word_sorting`	0.0	0.0	0.0
`causal_judgement`	20.0	28.0	+8.0	`logical_deduction_5_obj`	10.0	16.7	+6.7
`web_of_lies`	21.3	46.0	+24.7				

Under the frozen intervention, the two largest drops are `navigate` (−20.7 pp) and the 3-object tracking variant (−10.7 pp), while algorithmic and structured controls— `dyck_languages` (+0.7), `formal_fallacies` (+2.7 at an 84% baseline), `hyperbaton` (−0.7), `word_sorting` (0.0), and `logical_deduction` (−3.3/+6.7/+4.0 for 3/5/7 objects)—remain stable or improve; `web_of_lies` improves by +24.7 pp. Selectivity therefore does not depend on the particular `navigate`/`boolean` pair. On Janus-Pro's independent 27-task run, the spatial-group mean drops −14.5 pp, roughly three times the symbolic-group mean (−4.6 pp), replicating the group-level contrast on a second architecture (Appendix D.4). Two transparency notes bound this table. First, several of the 27 tasks have near-floor baselines for this model (e.g., `word_sorting`, `object_counting`, `geometric_shapes`), so their individual deltas are not individually interpretable, and we do not claim a graded spatial-score trend over the full 27-task suite. Second, for binary yes/no tasks whose Chameleon-7B baselines sit near or below chance (`navigate`, `boolean_expressions`, `web_of_lies`, `causal_judgement`), accuracy changes partly reflect shifts in the output-answer distribution rather than pure reasoning changes—`web_of_lies`' large +24.7 pp "improvement," for example, coincides with a high no-answer rate at baseline. We therefore read this table as a frozen-intervention selectivity check, and rest the strongest causal claims on the answer-completion-controlled batteries of §F.1–F.3.

G When EVT Does Not Appear: Failure Modes and Diagnostic Lessons

EVT is not a uniform phenomenon. Across 10 multimodal configurations and 4 cross-family transfer experiments, we observed several modes in which the canonical 1D VL-axis intervention fails, succeeds only weakly, or requires a non-modality contrast to expose. This appendix consolidates all such cases. We organize them by the diagnostic lesson rather than by model family, because the same model can fail under one method and succeed under another (Chameleon-30B is the clearest example). Each subsection states the observation, the apparent cause, and the methodological correction — if one was found.

Roadmap. §G.1 documents Chameleon-30B, where seven distinct modality-axis methods all return null and a task-contrastive axis recovers a strong effect, demonstrating that the EVT signal can dissociate from the modality axis at scale. §G.2 documents two early failed attempts on Janus-Pro before the generation-path DiffMean axis was identified. §G.3 documents the dosage cliff on Liquid. §G.4 contextualizes the smaller-scale Qwen3-VL-2B effect within the cross-scale narrative. §G.5 documents four cross-family backbone-transfer experiments, three of which produce null or non-selective damage and one of which produces a positive transfer that constrains the attribution argument. Together, these cases sharpen the boundary of the canonical method and motivate the diagnostic protocol of the axis-construction details earlier in this appendix.

G.1 Chameleon-30B: A modality-axis null with a task-contrastive rescue

Chameleon-30B shares architecture and training data with Chameleon-7B; the only difference is capacity (d_{hid}=8192 vs. 4096, 48 vs. 32 layers). The 7B configuration is the cleanest positive case in the paper (single-layer L5 DD $= +33.2$ pp; block-level $z = +12.22$). On the 30B version, seven modality-axis methods produced null effects — all seven, however, construct the modality contrast from the embedding table or from synthetic token-ID inputs; an eighth, input-faithful modality-axis construction (full multimodal sequences through the model's own VQ tokenizer) later succeeds, resolving the boundary case (Appendix A.5, Table 7, and the resolution paragraph at the end of this section).

The seven failed modality-axis methods. Table 29 enumerates each method, its expected effect (extrapolated from the 7B result), and the observed null.

Table 29: Seven modality-axis methods on Chameleon-30B, all null. Each row reports a distinct intervention; all yielded DD ≈ 0 on `navigate/boolean` pair. The eighth row (task-contrastive axis, last row, separated) was the first construction to succeed. A ninth construction — the input-faithful full-sequence *modality* axis — also succeeds (mid-block DD $= +33.2$ pp; Table 7) and supersedes the "modality axis is inert" reading of rows 1–7; see the resolution paragraph. Diagnostic conclusion in the rightmost column.

#	Method	Effect	Apparent cause
1	Embedding axis, α=1.0, 7 layer configs	DD ≈ 0	Axis exists geometrically (separation $0.1065 >$ 7B's 0.0851) but is causally inert
2	Alpha sweep $\alpha \in [1.5, 2.2]$	DD ≈ 0	Stronger removal still inert
3	Alpha sweep $\alpha \geq 2.5$	non-selective collapse	α exceeds residual-norm budget
4	PCA subspace $k \in \{1, \ldots, 200\}$, two depth ranges	DD ≈ 0	Position-PCA is not task-causal-PCA; 2.4% variance subspace contains no causal direction
5	Single-task probe direction	non-selective	Probe fits prompt-template distribution, not spatial content
6	Multi-task LOTO probe (12 layers tested)	0% across all layers	No linearly shared spatial-reasoning feature across tasks
7	Per-layer DiffMean, 4 blocks $\times$ 4 α values plus a 7-dose early-block fine sweep	DD ≈ 0 on all 19 cells	cos(DiffMean axis, embedding axis)≈ 0; modality-position axis $\perp$ task-causal direction
8	**Task-contrastive axis** (`nav` $-$ `bool`, L24–35)	**DD** $= +16.7$ **pp**, $z = +6.4$	Task-contrast direction is causal; modality-position direction is not (§6, semantic location)

Held-out validation of the task-contrastive axis. The task-contrastive axis in Table 29 is built from the `navigate–boolean_expressions` contrast. To rule out circularity, we evaluate the same axis on held-out tasks that were not used for axis construction. The held-out spatial tasks are

relative_position, reasoning_about_colored_objects, and temporal_sequences; the held-out control tasks are web_of_lies and causal_judgement. The intervention uses the late block L24–35 with $\alpha = 1.0$.

Table 30: Held-out validation of the Chameleon-30B task-contrastive axis. The axis is constructed from navigate and boolean_expressions, but evaluated on held-out tasks.

Task	Group	Baseline	Intervened	Δ
relative_position	spatial	62.8	62.8	0.0 pp
reasoning_about_colored_objects	spatial	27.9	26.7	−1.2 pp
temporal_sequences	spatial	28.4	28.8	+0.4 pp
web_of_lies	control	48.8	51.2	+2.4 pp
causal_judgement	control	58.3	67.9	+9.6 pp

The held-out DD remains positive at $+6.28$ pp, with paired-bootstrap CI $[+0.13, +12.37]$ pp. Energy-matched random directions give -1.08 ± 3.28 pp, yielding $z = +2.24$ against the random baseline (one-sided normal-approximation $p \approx 0.013$). Thus, the task-contrastive axis retains a directional double-dissociation signal beyond the task pair used to construct it.

Geometry of modality and task axes. We further compare the canonical VL axis with the task-contrastive axis across layers. The VL axis is the embedding-level text–visual centroid difference. The task axis at layer L is defined as the difference between the mean final-token hidden states of navigate and boolean_expressions examples.

Table 31: VL axis and task axis are nearly orthogonal in Chameleon. The two axes have small cosine similarity across the network, supporting the interpretation that Chameleon-30B decouples modality separation from task-causal spatial reasoning.

Model	Layers	Mean $\lvert\cos\rvert$	Max $\lvert\cos\rvert$
Chameleon-7B	32	0.0118	0.0390 at L9
Chameleon-30B	48	0.0145	0.0360 at L47

In both models, $\lvert\cos\rvert < 0.04$ throughout the network, showing that the VL axis and the task-contrastive axis are almost orthogonal. The important difference is semantic location. In Chameleon-7B, the small cosine peak overlaps the early VL-axis causal hub, whereas in Chameleon-30B the task-axis rescue hub lies in L24–35 and is separated from the VL-axis geometric peak. (Note that the VL axis in this comparison is the embedding-centroid estimate; the input-faithful full-sequence modality axis, which *is* causal at 30B, is itself near-orthogonal to the embedding axis ($\cos \approx 0.01$) — see the resolution paragraph below and Appendix A.5.)

The full data for methods 1–7 ships with the released per-sample result files (§A.12); the rescue evidence is summarized in Table 29 and validated on held-out tasks in Table 30.

Original diagnostic conclusion: capacity-driven decoupling (superseded; retained for the methodological record). Methods 1–7 all parameterize the intervention along a *modality* contrast (text vs. visual). Method 8 parameterizes it along a *task* contrast (navigate vs. boolean_expressions). The fact that the same hidden state, at the same layer block, with the same α, supports a $z = +6.4$ effect under method 8 but DD ≈ 0 under method 7 originally suggested that 30B retains spatial-reasoning structure while *decoupling* it from the modality direction, with the doubled hidden dimension (8192 vs. 4096) relieving the dimensional pressure that, in the 7B model, forces modality and spatial-reasoning circuits to share representational space.

Resolution: the modality axis is causal under an input-faithful construction. The decoupling interpretation is superseded by the full-sequence source-swap experiment (Appendix A.5). Methods 1–7 share a hidden confound: none of them measures the modality contrast on the model's actual multimodal input pathway. Method 7's synthetic axis (16 random token IDs, no image, no VQ encoding) turns out to be rotated $\cos \approx 0.43$–0.56 away from the axis obtained when real or random images are passed through Chameleon's own VQ tokenizer inside a full <image>+text sequence. Along that input-faithful modality axis, Chameleon-30B exhibits the full causal signature: mid-block (L12–23) DD $= +33.2$ pp with boolean unharmed (CI $[+21.2, +44.8]$; navigate $z = -88.7$ against random directions, $0/3$), final-block (L36–47) DD $= +22.0/+22.4$ pp under COCO/random-

pixel sources respectively (random directions produce exactly zero `navigate` change), and a clean null at the late block (L24–35). Three revisions follow. (i) The spatial carrier at 30B has *not* decoupled from the modality contrast; relative to the 7B hub (L5, early), it has *relocated* to mid/final blocks. (ii) What scaling changes is the carrier's depth and its sensitivity to axis construction, not its existence — so the cross-family contrast with Qwen3-VL (§4) is about *where* the carrier lives, not *whether* it survives scale. (iii) The task-contrastive rescue (method 8) and its held-out validation remain valid evidence, but they are no longer the only handle on 30B; the semantic-location lesson of §6 sharpens into a construction lesson: a null along one *estimate* of the modality contrast does not establish that the modality direction itself is inert.

G.2 Janus-Pro-7B: Two failed early stages before generation-path success

Janus-Pro was initially classified as a failure case under two methods before the generation-path DiffMean axis was tried. Both failures were architecture-specific and instructive.

Stage 1 — Embedding centroid axis. We applied the Chameleon recipe directly: $v = \text{normalize}(\mu_{\text{text}} - \mu_{\text{visual}})$ from the embedding table, with α=1.0 on five layer configurations. All cells produced $|z| < 1$. Diagnostic: in Janus-Pro, the `<image_N>` placeholder tokens used to define the visual centroid are *functional dead weights* — during real inference, the SigLIP aligner overwrites those positions, so the embedding rows are never updated by training. The centroid difference therefore reflects initialization noise, not a learned visual direction. This is a generic risk for any architecture where visual tokens enter the residual stream by replacement rather than by table lookup.

Stage 2 — SigLIP understanding-path DiffMean. Following BAGEL precedent, we collected hidden states at `<image_N>` positions from real COCO images passed through the SigLIP aligner, then computed per-layer DiffMean. Axis diagnostic immediately flagged a problem: $\|\mu_{\text{text},L}\| \approx 14$ while $\|\mu_{\text{visual},L}\| \approx 114$ at L=0, with projection ratio 0.83. The axis captured an input-source norm gap, not a clean modality-semantic direction. Applying α=0.5 at L0–7 collapsed the model: `navigate` acc spiked to 0.984 (a false-positive saturation pattern; the model emits `"Yes"` unconditionally), `boolean` dropped to 0.000 with degenerate output (e.g., `"Yes-NoLet'sthinkstepbystep.A:Yes-NoLet'sthinkstepbystep.A:..."` repeating). The intervention destroyed token prediction rather than spatial reasoning specifically.

Stage 3 — generation-path DiffMean (success). The third method — per-layer DiffMean computed on hidden states at VQ-16 generation tokens — passed all four causal checks (DD = +8.0 pp at α=0.7, $z = +3.58$, 0/22 random hits, paired bootstrap CI $[+2.93, +13.00]$ pp). This stage is documented in the main text §5; the failures of stages 1 and 2 are recorded here both as a methodological warning and as evidence that the generation-path axis is not the only construction tried but the only one that survives diagnostic checks.

G.3 Liquid-V1-7B: A narrow dosage cliff at block level

Liquid passes all four causal checks in the single-layer configuration: the headline result is L27-only removal at α=1.0 (DD = +21.6 pp; `navigate` −12.8, `boolean` +8.8; Table 1). The dosage cliff appears in the *block-level* experiment (L21–27), whose full sweep is shown in Table 32.

Table 32: Liquid block-level dosage sweep (L21–27). The transition from a weak effect to `navigate` collapse occurs inside the $\alpha \in [0.50, 0.55]$ window. Large post-cliff DDs are saturation artifacts: `navigate` accuracy is at or near zero, so no usable selectivity remains. Accuracy in %, Δ relative to the `navigate`/`boolean` baselines (54.8/54.8).

α	navigate	boolean	Δnav (pp)	DD (pp)	status
0.40	51.2	51.6	−3.6	+0.4	weak
0.50	44.8	50.4	−10.0	+5.6	block-level sweet spot
0.55	8.8	48.8	−46.0	+40.0	cliff: navigate collapses
0.60	0.0	49.2	−54.8	+49.2	navigate at zero
0.65	0.0	48.4	−54.8	+48.4	plateau (nav dead, bool stable)
0.70	0.0	34.4	−54.8	+34.4	boolean begins to collapse
1.00	0.4	0.0	−54.4	≈ 0	full non-selective collapse

Table 33: Cross-family backbone transfer. Source VLM axes are transferred to target text backbones under the same intervention recipe. Only the matched Qwen family yields positive transfer.

Transfer	Effect	Verdict	Takeaway
Janus → DeepSeek-LLM-7B	DD = +2.4–3.6 pp; $z = +0.32$	null	No task-causal coupling in the unrelated text backbone.
BAGEL → Qwen2.5-7B base	navigate $\Delta = -5.3$ pp; $z = -5.63$; DD ≈ 0	non-selective	Active direction, but no spatial selectivity.
Liquid text-mode self-transfer	navigate $\Delta = +3.6$ pp; $z = +9.04$	sign-reversed	Text-mode axis amplifies rather than suppresses spatial reasoning.
Qwen2.5-VL → Qwen2.5-7B base	**DD = +18.5 pp**; $z = -18.6$	**positive**	**Matched-family transfer supports amplification of a pre-existing Qwen spatial scaffold.**

The cliff is two-staged: `navigate` collapses between α=0.50 and 0.55, then `boolean_expressions` collapses around α=0.70. We emphasize that the post-cliff region is not a usable selective-intervention window: with `navigate` at 0–9% accuracy, the large DD values reflect one task being dead rather than graded selectivity. The best block-level selectivity is only DD = +5.6 pp at α=0.50, far below the single-layer L27 configuration (+21.6 pp), which concentrates the removal at the hub and avoids the cliff entirely.

Apparent cause. Liquid was created by VQGAN fine-tuning on top of a Gemma-7B text backbone. Visual information is injected only at the final layer block, with no joint training of earlier layers. This produces a maximally localized causal hub: spatial information sits in a thin sliver of representational space at L27, and the block-level effect appears or disappears together as α crosses a threshold. This is the geometric opposite of Emu3, where the same causal pathway is duplicated at early and final blocks (§E.1). The single-layer headline configuration does not exhibit the cliff at the synthetic-axis dosage (α=1.0); the real-image axis, which projects more strongly onto L27 hidden states, shifts the same cliff to lower α (§A.4). The block-level cliff is a methodological warning for block-scale interventions on late-fusion architectures.

G.4 Qwen3-VL-2B: A weaker but coherent positive case

Qwen3-VL-2B passes all four causal checks with $z = +27.7$, showing that the Qwen3-VL effect is already present at the smallest dense scale we test. Together with Qwen3-VL-4B/8B ($z = +39.1/+38.3$) and Qwen3-VL-32B ($z = +262.5$), this supports an overall amplification trend across scale, while keeping the small-scale result as a clean positive case rather than a mixed scaling point.

Note on 30B-A3B (MoE). The MoE configuration achieves $z = +48.8$ (stronger than 2B). Its random-direction control was run at $N = 5$ rather than the $N = 22$ used for other models, and as the only MoE in our suite (n=1), conclusions about MoE-specific effects carry the standard sample-size caveat.

G.5 Cross-family backbone transfer: Three failures and one positive

To test whether VL-axis effects are intrinsic to the source VLM or merely reflect a generic property of the underlying backbone, we transferred axes from VLMs to their text-only counterparts and to unrelated text backbones. Four experiments were run; three produced null or non-selective damage, and one produced a positive transfer that constrains the attribution argument.

The positive transfer constrains the attribution claim. The Qwen2.5-VL → Qwen2.5-7B result (last row) is what allows §4.3 to distinguish *introduction* (Chameleon: 13× magnitude gap between VLM and matched text backbone) from *amplification* (Qwen family: aligned scaffold pre-exists, multimodal training strengthens it). Without this transfer experiment, the Qwen3-VL → Qwen3-8B base result alone would be ambiguous because Qwen3 pretraining data is not fully documented; the Qwen2.5 variant has fully published training documentation and shows the same DD≈ +18pp pattern, consistent with amplification rather than introduction (§4.3).

The three nulls bound the universality claim. Janus → DeepSeek and BAGEL → Qwen2.5 both rule out a generic-direction hypothesis: there is no single "spatial reasoning direction" portable across pretrained text backbones. An axis must either (i) be extracted from the same family ancestor or (ii) target a feature already present in the backbone. The Liquid self-transfer further shows that even the same-backbone embedding-space axis behaves qualitatively differently in text-only mode — a methodological warning that axis construction is recipe-specific.

G.6 Summary: failure modes and the boundary of the canonical method

Across the cases above, three distinct failure modes account for the observed nulls:

(M1) Wrong axis semantics. Cham-30B (modality axis null but task-contrastive succeeds), Janus stage 1 (embedding axis on dead-weight visual tokens), Janus stage 2 (SigLIP path axis dominated by norm gap). In each case the candidate axis exists geometrically but does not align with the causal direction. Diagnostic: report axis projection ratio (proj/h), centroid separation, and split-half cosine before claiming a null.

(M2) Wrong dosage window. Liquid cliff. The axis and layer are correct; only a narrow α band exposes the effect. Diagnostic: scan α at fine resolution near any apparent transition before concluding null.

(M3) Wrong target backbone. Cross-family transfer (3 of 4 cases). The axis is correct in its source VLM but does not couple causally to spatial reasoning in an unrelated text backbone. Diagnostic: same-family transfer is the strongest control; cross-family transfer measures something different (geometric portability) and is null by default rather than informative.

The Qwen3-VL-2B cases are not failures under any of M1–M3; they are quantitative cases of the canonical method at smaller scale, consistent with the introduction-vs-amplification gradient documented in §4.3.

The methodological correction implied by these cases is the diagnostic protocol of axis-construction validation: before reporting a null on a new architecture, validate axis construction (axis projection ratio, separation, split-half cosine), scan dosage at the cliff resolution, and, for dual-path or cross-family cases, attempt the task-contrastive axis as a fallback (Cham-30B precedent).

H Three Regime Anchor Cases: Extended Evidence

§6 introduces three geometric regimes for the alignment between the readout (DiffMean) axis and the dominant variance subspace: *concentric*, *projected-diffuse*, and *mixed*. Each regime is anchored by one configuration in our experiments. This appendix consolidates the underlying data — per-layer $|\cos(\text{DiffMean}, \text{PCA-top-1})|$, cumulative explained variance (EV), and rank-k DD curves — for the three anchors. All data come from EXP4 (released as `src/pipelines/subspace_intervention.py`). One provenance note for auditability: the raw EXP4 logs store a `dd` field computed as $\Delta_{\text{spatial}} - \Delta_{\text{control}}$, the *opposite* sign of the Methods convention $\text{DD} = \Delta_{\text{control}} - \Delta_{\text{spatial}}$; all rank-$k$ DD values in this paper are reported in the Methods convention (sign-corrected from the logs), and the released pipeline documents the discrepancy in its docstring. A full classification of the remaining 7 model configurations into regimes is deferred to future work, since systematic rank-k ablation requires a phase-0 PCA construction that is beyond the scope of this work (§G).

Anchor summary. Table 34 gives the headline numbers for the three regimes; Figures 6–8 visualize the underlying curves.

Concentration diagnostic. Table 35 reports the unified concentration diagnostic used in the main text. For a unit axis $\hat{a}$, raw concentration is computed as

$$c_{\text{raw}} = \mathbb{E}_{x,t}\left[\frac{(h_\ell(x,t)^\top \hat{a})^2}{\|h_\ell(x,t)\|_2^2}\right].$$

Table 34: Three regime anchor cases. Hub-block mean $|\cos(\text{DiffMean}, \text{PCA-top-1})|$, cumulative explained variance at k=1 and k=16, and the most informative rank-k DD value with 95% CI. "ctrl crash" marks DD values that are driven by control collapse rather than spatial selectivity (see text).

	BAGEL und (L21–26, hub)	**BAGEL gen** (L21–27, hub)	**Qwen3-VL-8B** (L0–8, hub)
Regime	concentric	projected-diffuse	mixed
Mean $\lvert\cos\rvert$ in hub	0.94	0.42	0.44 (layer-bimodal)
Cumulative EV at k=1	52.9%	18.9%	47.1%
Cumulative EV at k=16	86.2%	69.5%	77.4%
DD @ k=1 (pp)	$+1.7$ $[-5.0,+8.3]$	-36.7 ctrl crash	-24.0 ctrl crash
DD @ k=16 (pp)	-12.7 $[-21.7,-4.0]$ ctrl-dom.	-28.3 ctrl crash	-15.3 $[-24.0,-6.3]$ ctrl-dom.
1D causal carrier	none (all ranks non-sel.)	DiffMean (1D, z=+3.19)	per-layer DiffMean (1D)
Ctrl.-adj. slope (pp/unit)	—	-1.6	$+1.9$

We also report a centered version, c_{cen}, computed after removing the layer mean from the hidden states. The centered diagnostic separates sample-varying signal from a global mean offset along the axis.

Table 35: Unified concentration diagnostic. c_{raw} measures total axis-aligned hidden-state energy, while c_{cen} removes the layer mean before computing the same quantity. Chameleon has low centered concentration but a selective causal effect. BAGEL has very high raw concentration due to mean alignment, and still remains above random after centering.

Model / axis	**Layer**	$\mathbf{c}_{\text{raw}}$	$\mathbf{c}_{\text{cen}}$	$\lvert\cos(\mu, \mathbf{a})\rvert$	**Random**	$\mathbf{c}_{\text{cen}}/\mathbf{rand}$
Chameleon-7B / VL axis	L3	0.104%	0.380%	0.032	0.018%	20.9×
Chameleon-7B / VL axis	L4	0.174%	0.091%	0.042	0.022%	4.1×
Chameleon-7B / VL axis	L5	0.131%	0.137%	0.036	0.026%	5.3×
Chameleon-7B / VL axis	L6	0.215%	0.045%	0.047	0.024%	1.9×
Chameleon-7B / VL axis	L7	0.249%	0.038%	0.051	0.024%	1.6×
Chameleon-7B / VL axis	L20 ctrl	0.123%	0.007%	0.038	0.019%	0.4×
BAGEL-7B / gen DiffMean	L21	49.2%	0.426%	0.718	0.029%	14.5×
BAGEL-7B / gen DiffMean	L22	57.4%	0.929%	0.778	0.029%	32.5×
BAGEL-7B / gen DiffMean	L23	49.6%	1.522%	0.725	0.028%	55.1×
BAGEL-7B / gen DiffMean	L24	58.3%	3.163%	0.786	0.027%	118×
BAGEL-7B / gen DiffMean	L25	47.3%	2.298%	0.709	0.026%	87.7×
BAGEL-7B / gen DiffMean	L26	58.4%	0.936%	0.785	0.026%	36.4×

For Chameleon, $|\cos(\mu, a)|$ is small, so raw and centered concentrations are of the same order; the signal is not dominated by a global mean offset. For BAGEL, $|\cos(\mu, a)|$ is large, so the raw concentration is dominated by mean alignment. After centering, BAGEL still has a stronger axis-aligned signal than Chameleon, but the gap is moderate rather than several hundred-fold.

H.1 Anchor 1 (concentric): BAGEL understanding path

The understanding path of BAGEL exhibits the cleanest *concentric* alignment: the DiffMean axis lies inside the dominant variance subspace ($|\cos|$ peaks at 0.978 at L21 and stays > 0.85 through L26 before collapsing at L27). Despite this near-alignment with the top variance direction, the rank-1 PCA intervention is null (DD= $+1.7$ pp, CI $[-5.0, +8.3]$), and expanding the rank does not expose a selective handle: at k=16 (cumulative EV 86.2%), removal is clearly damaging but *control-dominated* — control tasks drop by 32.3 pp versus 19.7 pp for spatial tasks (DD= -12.7 pp, CI $[-21.7, -4.0]$). No tested rank yields spatial selectivity in this path.

Diagnostic implication. Probing accuracy is high at L27 (1D probe 0.729 ± 0.017, $+2.65\sigma$ above norm-matched random; full per-layer probe data in Table 14; cross-architecture comparison with Chameleon-7B's complementary pattern in §B), yet 1D intervention along the same direction is null and rank-k removal is non-selective. This is the canonical *probe–intervention dissociation* in its strongest form: linear decodability does not imply any linearly intervenable causal handle. The spatial information carried by this path is entangled with general representational capacity rather than organized as a removable carrier.

H.2 Anchor 2 (projected-diffuse): BAGEL generation path

The generation path of BAGEL exhibits the *projected-diffuse* pattern: the DiffMean axis sits well off the top PCA direction across the hub (per-layer $|\cos|$ between 0.26 and 0.51 over L21–L26, mean 0.42; the all-layer mean is 0.19, and L27 is a known LayerNorm-dominated outlier at 0.819). The 1D DiffMean axis is nonetheless causal (DD $= +16.8$ pp, $z = +3.19$; main text §5, Table 2). The PCA subspace, in contrast, is broadly entangled with general representational capacity: rank-k ablation collapses controls by 56–74 pp at every k, more than the spatial tasks (DD -16.7 to -39.0), so removing variance-dominant subspaces produces broad damage rather than a selective effect.

Diagnostic implication. The two interventions probe orthogonal axes. The DiffMean intervention recovers a compact one-dimensional spatial carrier; the PCA intervention recovers a broader general-purpose subspace whose removal damages everything. The qualitative contrast between BAGEL's two paths — concentric on und (1D null, rank-k effect) vs. projected-diffuse on gen (1D causal, rank-k non-selective) — is the geometric core of the *compression vs. distribution* hypothesis in main text §5.

H.3 Anchor 3 (mixed): Qwen3-VL-8B

Qwen3-VL-8B is the *mixed* case: $|\cos|$ is layer-dependent and bimodal (Figure 6). At L0–L5 the DiffMean axis is nearly orthogonal to the top PCA direction ($|\cos| \in [0.015, 0.465]$); from L6 onwards $|\cos|$ jumps to ≈ 0.999 and remains saturated through L33 before collapsing at L34. This bimodal profile is consistent with the architectural fact that the ViT projection feeds visual information into the LM stream around L6, after which the dominant variance direction and the modality-contrast direction converge.

The rank-k DD curve is correspondingly non-monotonic (Figure 8, panel C), but at no rank does it become spatially selective. At k=1, controls collapse ($\Delta_{\text{control}} = -65$ pp; DD$= -24$). At k=4 and k=8, effects are small and near-symmetric (DD $+0.7$/$+2.3$). At k=16 and k=32, controls again drop more than spatial tasks (DD -15.3 and -21.0). The selective handle in Qwen3-VL-8B is therefore its per-layer 1D DiffMean axis (Table 1), not any variance-dominant PCA subspace.

Diagnostic implication. A model can simultaneously satisfy "concentric" alignment in some layers and "projected-diffuse" in others. When evaluating a new architecture, neither the rank-1 result nor any single rank-k value is decisive; the non-monotonic, never-selective DD-vs-k curve is itself diagnostic of the mixed regime, and rank-k removal should always be checked against control damage before any selectivity claim.

H.4 Figures

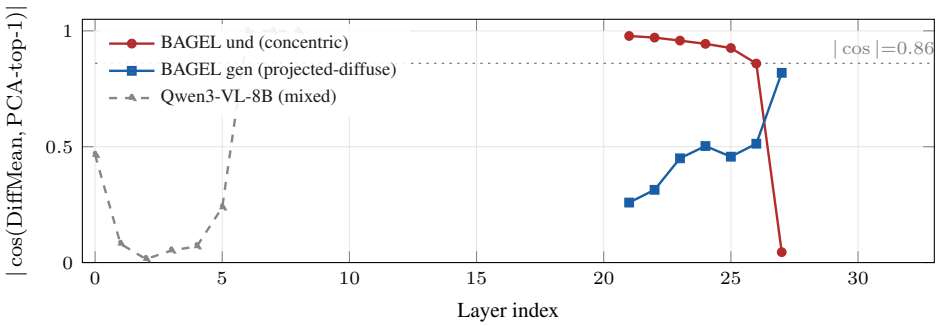


Figure 6: Per-layer $|\cos(\text{DiffMean}, \text{PCA-top-1})|$ for the three anchor cases. BAGEL und stays near alignment ($|\cos| > 0.85$) across the hub block before a layer-norm-dominated drop at L27 (excluded from hub mean). BAGEL gen remains well off the top PC across the hub (per-layer 0.26–0.51, mean 0.42; the all-layer mean is 0.19). Qwen3-VL-8B is bimodal: near-orthogonal at L0–L5, saturated at L6–L8 corresponding to ViT injection. Dotted reference at $|\cos|$=0.86.

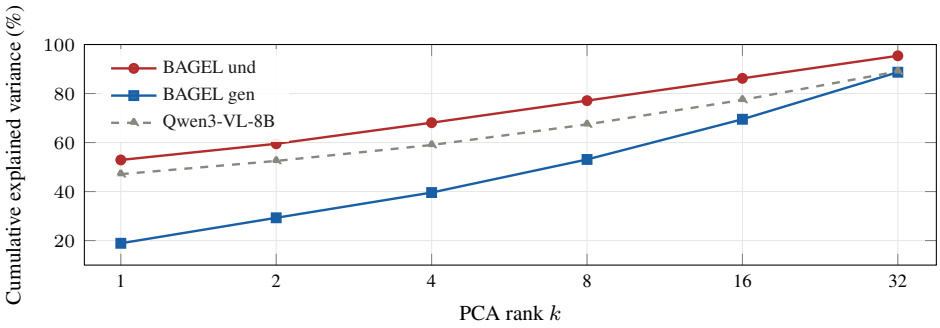


Figure 7: Cumulative explained variance of the per-sample DiffMean PCA subspace, hub-block average. BAGEL und reaches 86% by k=16, indicating a relatively concentrated dominant variance structure. BAGEL gen requires k=32 to reach the same coverage — consistent with the projected-diffuse pattern. Qwen3-VL-8B is intermediate.

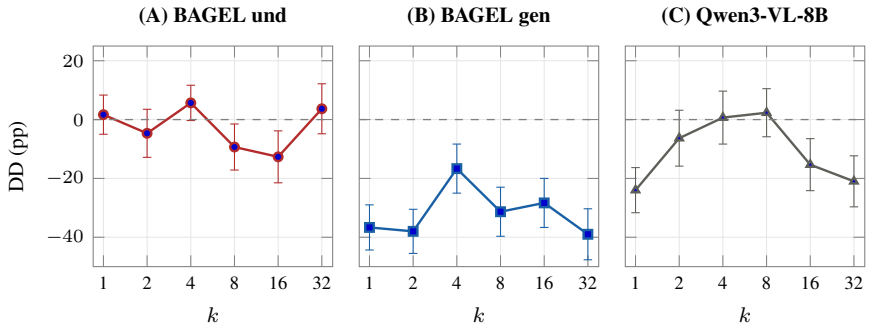


Figure 8: DD vs. rank-k subspace ablation, three anchor cases (DD $= \Delta_{\text{control}} - \Delta_{\text{spatial}}$; positive would mean spatial-selective damage). Error bars are paired-bootstrap 95% CIs. **(A)** BAGEL und: null at k=1; at k=16 removal is damaging but control-dominated (-12.7 pp). **(B)** BAGEL gen: $\Delta_{\text{control}} \leq -56$ pp at every k — broad collapse, not a selective effect (the selective handle is the 1D DiffMean axis, main text §5). **(C)** Qwen3-VL-8B: non-monotonic but never spatially selective; the k=1 dip is driven by control collapse and $k \in \{16, 32\}$ remain control-dominated.

H.5 Scope of this analysis

The three anchor cases establish that the alignment regimes named in §6 are realized geometries with testable diagnostic signatures, not abstract possibilities. The methodological scope of this appendix is to characterize the regimes in detail through these anchors; a complete taxonomy of all 10 configurations into regimes is a separate question and lies outside the present paper's scope. Within our framework, each new architecture can be classified by running the EXP4 phase-0 and phase-1 protocols.

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification: The abstract and introduction state the main causal claims, the evaluated model families, and the spatial-imagery versus symbol-manipulation contrast. The claims are supported by the main experiments and further scoped by the limitations section.
2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [Yes]
   Justification: Section 7 discusses limitations in intervention class, task coverage, model coverage, and attribution uncertainty. We explicitly state that the evaluated task families are a diagnostic contrast rather than a complete taxonomy of text-only reasoning.
3. **Theory assumptions and proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: [N/A]
   Justification: The paper does not present formal theoretical results, theorems, or proofs. The equations define intervention operators, evaluation metrics, and diagnostic quantities used in the experiments.
4. **Experimental result reproducibility**
   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?
   Answer: [Yes]
   Justification: Appendix A provides the experimental setup, hardware and software details, compute budget, axis-construction provenance, datasets, model identifiers, random seeds, intervention dosages, and statistical testing standards. We also release code, intervention scripts, and per-sample correctness JSONs for reproducing the reported results.
5. **Open access to data and code**
   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
   Answer: [Yes]
   Justification: We release a code archive with the intervention framework, task-contrastive axis pipeline, subspace-intervention scripts, per-model experiment scripts, per-sample correctness JSONs, and BBH task data. The release also includes instructions for dataset access and preparation, including scripts for obtaining COCO images rather than redistributing image bytes.
6. **Experimental setting/details**
   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?
   Answer: [Yes]
   Justification: Sections 3–4 describe the intervention, axis construction, task setup, and evaluation protocol. Appendix A gives additional details on model identifiers, datasets, intervention dosages, sample counts, seeds, decoding settings, and statistical testing.
7. **Experiment statistical significance**
   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
   Answer: [Yes]
   Justification: The paper reports double-difference effect sizes, paired bootstrap confidence intervals, random-direction specificity tests, empirical one-sided p-values, and energy-matched random controls for the primary claims. Appendix A.13 defines the statistical testing protocol.
8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Appendix A.1 and Appendix A.2 report the GPU types, memory, software precision, decoding settings, per-experiment GPU-hour estimates, and total estimated compute budget.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The study analyzes existing public or permissioned models and public benchmark datasets, does not involve human-subject experiments, and does not release model weights or scraped sensitive data. We document asset provenance and licenses in the appendix.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The discussion includes a broader-impact paragraph describing the diagnostic value of the work, its non-deployment nature, and the potential risk that intervention tools could be misused to manipulate model behavior without sufficient understanding.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a new pretrained language model, image generator, scraped dataset, or model checkpoint with high misuse risk. The released assets are analysis code, intervention scripts, benchmark-processing utilities, and per-sample correctness metadata.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the original model and dataset sources, and Appendix A.10 lists dataset licenses or access details where applicable. We use public model identifiers, do not redistribute model weights, and provide a COCO download script rather than redistributing image bytes.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new released assets are the analysis code, intervention scripts, task-processing scripts, and per-sample correctness JSONs. Appendix A.12 documents the released directories, required reproduction files, excluded preliminary artifacts, and dataset/model access procedure.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, or human-subject experiments. The two AI-background human annotations are used only as a blind sanity check for task-level spatial-imagery scores and do not constitute a behavioral user study.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, or human-subject experiments. Therefore IRB approval or equivalent human-subject review is not applicable.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs and VLMs are the primary experimental subjects of the paper, and the evaluated model families are described in the main text and appendix. We also use frontier LLMs to annotate task-level spatial-imagery scores, with human sanity checks reported in Appendix A.11.