WHAT IS A VISUAL THOUGHT WORTH? WHEN CONTINUOUS VISUAL STATES HELP VISION-LANGUAGE MODELS REASON

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent vision-language models augment the visual features of a pretrained backbone with continuous intermediate states intended to support reasoning. These "visual thoughts" can be trained to encode additional visual structure, such as segmentation or depth, or can emerge through iterative latent computation. The goal is to provide the model with a more task-relevant representation of the scene than the backbone features alone may offer. But does this added representation actually work as intended? In particular, can the answer be decoded from it, does the model itself use it, can training make it useful, and does it provide value beyond ordinary image features?

We study these questions in two released models, Chain-of-Visual-Thought (CoVT) and Latent Visual Reasoning (LVR). We save each model's visual states from the original image and progressively weaken or remove the image during answering, while keeping the states fixed. This lets us distinguish three properties that are often conflated: whether an answer can be decoded from a state, whether the released model uses its image-specific content, and whether an answer model explicitly trained to read the states becomes more accurate because of them, relative to an otherwise identical model trained with a non-image-specific mean state.

We find that these properties can differ sharply. LVR states become increasingly useful as direct visual evidence disappears, improving counting by 14.2 points and left-or-right accuracy by 12.9 points when the image is removed. CoVT states reveal a different failure mode: the answer is decodable from them even when the answering model initially derives almost no benefit from them, while stronger readout training yields an 8.6-point counting gain without the image. On CLEVR, CoVT states are highly decodable but unused by the released model; after training the answer model to read them, their value grows dramatically as the image degrades, reaching a 71.3-point counting gain with a blank image. Yet at an equal-size interface, ordinary pooled image features outperform the latent states. Because the states are written after the question, part of these gains may reflect an answer already formed, which our runs do not rule out. Thus, adding structured visual states is not sufficient: their value depends on what information is written into them, how that information is exposed to the backbone, and whether subsequent computation learns to use it. Our results motivate latent-reasoning architectures that jointly optimize the representation and its readout, rather than treating the presence of richer intermediate features as evidence that they contribute to reasoning.

1 INTRODUCTION

Vision-language models (VLMs) typically reason about an image through features produced by a pretrained visual encoder. These features are powerful general-purpose representations, but they are not necessarily organized to make every piece of information needed for downstream reasoning explicit. Properties such as object structure, depth, geometry, or task-specific relations may remain implicit in the backbone representation and therefore difficult for subsequent language-model computation to access reliably.

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

A growing class of methods addresses this limitation by introducing *continuous intermediate visual states* between perception and the final answer. The goal is to supplement or restructure the backbone's visual features into a representation that is better suited to reasoning. Chain-of-Visual-Thought (CoVT), for example, trains vectors at special visual-thought positions to match vision experts such as segmentation and depth (Qin et al., 2025). Latent Visual Reasoning (LVR) instead feeds its own hidden vector back as the next input for several steps (Li et al., 2026a). Although the mechanisms differ, both aim to create an intermediate visual representation that makes task-relevant information more accessible to later computation.

This raises a basic question: *when do these added visual states actually help?* Improved end-task accuracy alone does not answer it. A state may encode useful information that the downstream model does not know how to read. Conversely, the model may still obtain the answer primarily from the original image features, while the image-specific content of the added state contributes little. Changes in training, additional computation, or the presence and format of special positions can also affect performance independently of the particular content written into those positions. Prior audits have shown that released models often keep their answers when latent content is replaced (Zhang et al., 2026a; Viveiros et al., 2026; Guo et al., 2026; Li et al., 2026c; Park et al., 2026), and that a representation can be decodable without being used (Huang & Chang, 2025; Li et al., 2026b). Thus, if visual states are meant to become a useful substrate for reasoning, it is important to distinguish what information is written into them from whether subsequent computation actually reads and benefits from that information.

We separate three questions that are often conflated. *Can the state be decoded?* We ask whether a fitted reader can recover the answer from the saved state. *Does the released model use it?* We ask whether changing the state's image-specific content changes the model's answer. *Can the state be made useful?* We train an answer model on each image's own saved state and compare it with an otherwise identical model trained with a fixed mean state that carries no image-specific content. A fourth, practical baseline asks whether the learned state is more useful than the visual representation the model already had: at a common portable interface and equal storage budget, we compare it with pooled features from the frozen image encoder.

Figure 1 shows the protocol. A frozen writer, the released model itself, sees the original image and the question once. Its states, the hidden vectors at the visual-thought positions, are saved and held fixed while the picture available to the answer model is progressively weakened: it is shown in full, blurred, blanked, or removed. The answer model then answers from the question, the saved states, and the remaining visual evidence. This sweep matters because an informative intermediate state may add little when the original image already makes the answer easy to recover. As direct visual evidence weakens, information preserved in the saved state has a greater opportunity to affect the answer. The formal contrasts and training setup are given in Section 2.

Across CoVT and LVR, the same saved state can receive different answers to these questions. Some states expose task-relevant information to fitted readers without being effectively used by the answering model, while additional readout training can make part of that information useful. Other states become increasingly useful as direct visual evidence weakens. At the same time, a learned latent state need not provide a more usable interface to the backbone than ordinary pooled image features, even when the answer is more readily decoded from the latent representation. These observations suggest that latent visual reasoning is not only a representation-learning problem, but also a *writing-and-reading* problem: a useful visual-thought mechanism must both encode task-relevant structure and expose it in a form that subsequent computation can reliably exploit.

The study covers two released checkpoints of one backbone family and two task types, and makes no claim about other models, out-of-distribution data, or iterative reasoning across states (Section 6).

2 Protocol: Three Questions and a Baseline

Notation. Let I be the original image and Q the question. The frozen writer W is the released model's own generation pass, and $Z = W(I, Q)$ is its saved state, the hidden vectors at its K visual-thought positions, or *slots*. In the configuration we test, CoVT writes $K = 12$ slots (eight segmentation, four depth), and LVR writes $K = 8$ (Appendix A). The writer sees the question, so Z may hold a code for the current answer as well as visual content. The answer model sees

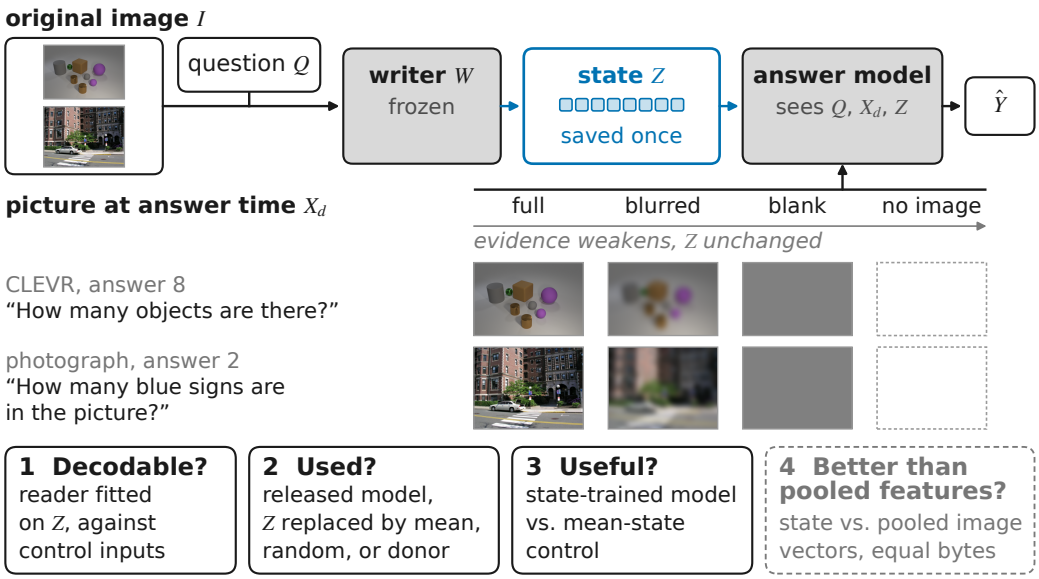

Figure 1: The protocol. The frozen writer W, the released model itself, sees the original image I and the question Q, and its states Z are saved once. The answer model returns the answer $\hat{Y}$ from the same question, the saved states and a picture X_d at answer time. The boxes pose the three questions and, dashed, the baseline question, which compares the state with pooled image vectors of equal size.

$X_d = T_d(I)$, the picture at answer time after degradation T_d at level d. Besides the released model M_0, two answer models are trained: the *state-trained model* M_Z receives each image's own state Z_{own}, and the *mean-state control* M_C is trained the same way on the mean state. We write Y for the answer and $A(\cdot)$ for held-out accuracy. The *all-layer replay cache* holds the input of each of the $L = 28$ decoder layers at the slots, an $L \times K \times D$ tensor with $D = 3{,}584$. In Sections 4.1 and 4.2 a state enters as this cache, written over the answer model's slot inputs at every layer, as an analysis instrument rather than a compact representation. The *portable representation* is K vectors of width D supplied once at the layer-0 input of the slots.

What replaces the states. The *mean state* averages each coordinate over a set of images, separately at each slot and layer and per question type, so it carries no image-specific content. M_C is trained on the *training mean* over the first 128 training images and receives that same tensor at test time. The CLEVR curve of Section 4.2 instead scores M_C under a *reference-pool mean* over images disjoint from the test images. Scoring it under the training mean moves every gain by at most 0.9 pp (Appendix B). The *random state* adds to the mean a random vector as large as a typical reference-pool state's distance from that mean, a *donor state* is another image's state written under the same question, and whichever of them replaces Z_{own} is written Z_{control}.

Three questions and a baseline. Table 1 lists what each question compares and cannot say. *Decodable* is a fitted reader's held-out accuracy on Z against its controls. The other questions use contrasts paired by image:

$$\Delta_{\text{use}}(d) = A(M_0, X_d, Z_{\text{own}}) - A(M_0, X_d, Z_{\text{control}}), \quad (1)$$

$$\Delta_{\text{utility}}(d) = A(M_Z, X_d, Z_{\text{own}}) - A(M_C, X_d, Z_{\text{control}}), \quad (2)$$

$$\Delta_{\text{content}}(d) = A(M_Z, X_d, Z_{\text{own}}) - A(M_Z, X_d, Z_{\text{control}}), \quad (3)$$

$$\Delta_{\text{repr}}(d, K) = A(M_{ZK}, X_d, Z_K) - A(M_{VK}, X_d, V_K), \quad (4)$$

where Z_K is the state's K portable vectors, V_K is K vectors pooled from the frozen image encoder, and M_{ZK} and M_{VK} are new answer models trained on each. *Used* is Δ_{use}, one fixed model with its input changed. *Useful* is the matched-training gain Δ_{utility}, which compares two separately trained models. The content contrast Δ_{content} changes only the input of M_Z, and it carries no utility claim because the mean state can be out of distribution for M_Z. The baseline question is Δ_{repr},

and Section 4.3 reports the pooled advantage $-\Delta_{\mathrm{repr}}$. The use question is answered for CoVT on CLEVR and only piloted for LVR (Appendix F).

Table 1: What each question compares, and what it cannot say.

question	model	compared inputs	shows	cannot say
decodable?	fitted reader	own states against constant, mean, and shuffled inputs and the majority floor	what some reader can extract	that the model reads it
used?	M_0 fixed, Eq. (1)	own against mean, random, and donor states	whether M_0's answer depends on the content	that the content is unread in other settings
useful?	M_Z against M_C, Eq. (2)	own states in M_Z, the training mean in M_C	the gain after matched training	that content rather than input variation drives the gain
better than pooled features?	M_{ZK} against M_{VK}, Eq. (4)	K state vectors against K pooled vectors, one interface, equal bytes	representation advantage at that interface	that the backbone's familiarity with the input plays no part

What is degraded, what is trained. The picture at answer time is the full image, a Gaussian blur of radius 4, 8, 12, 16, or 24 pixels on the 480×320 CLEVR render, a uniform gray picture (*blank*), or no image tokens (*no image*). On photographs the radius is scaled to match the CLEVR blur on the token grid (Appendix C). Blur and blank keep the token grid and every slot, while dropping the image tokens shortens the sequence. The states are the same tensor at every level, and the writer and image encoder stay frozen. Rank-16 adapters (Hu et al., 2022) on the attention projections of every layer act only in the answer pass. They train for $1{,}500$ steps at effective batch 16 on six conditions (full, blur 4, 12, 24, blank, no image). Radii 8 and 16 never enter training.

Why degrade the picture. Under log loss, the value a saved state can add given the remaining evidence is $I(Y; Z \mid X_d, Q)$. It is zero when X_d already determines Y and can grow as X_d weakens, though not always, so a state can only show its worth once the picture stops answering the question. Our accuracy contrasts are not estimators of this quantity.

Data and statistics. *Held out* means unseen by our adapters and readers (Section 6). Accuracy is in percent and differences in percentage points (pp), positive when the first named condition is better. Brackets are 95% intervals from $10{,}000$ paired image bootstrap draws with the trained models held fixed, adjusted for multiplicity as stated in Appendix G. Where three pairs were trained, each pair is shown and no training-seed interval is given. The predictions and rules of every study were fixed before its runs were read (Appendices G.1 and G.3).

3 Related Work

Does a released model depend on its latent content? The Token Replacement Test of Zhang et al. (2026a) fills the latent tokens with zero, random, distribution-matched, or oracle content and finds that models (incl. CoVT) keep most of their gain. Viveiros et al. (2026) find accuracy unchanged when dummy tokens replace the latent tokens of LVR and other latent reasoners. Guo et al. (2026) split latent tokens into slot contents, marker tokens, and format, keep most of the gain with the markers alone, and find that another example's states make the model give that example's answer less often than chance. Li et al. (2026c) find that latent tokens barely change with the input, barely move the answer, and hold little visual information under probing. Cui et al. (2026) find LVR's latents bypassed by the answer. The gap between decodable and used has been reported for counting vision transformers (Huang & Chang, 2025), for image tokens (Li et al., 2026b), and for latents that grow richer in training while their effect on the answer is suppressed (Zhang et al., 2026b).

When do latent states become useful? Viveiros et al. (2026) find oracle latents computed from a helper image useful once the relevant image regions are masked and on a synthetic rotation task, while the model's own generated latents stay near chance. Guo et al. (2026) capture a released model's own slot states with the image present, inject them into a run without the image, and recover little, the two ends of our sweep. Park et al. (2026) retrain six latent reasoners so that only the latent

state carries image information, and find that they fail to recover visual competence and that a state from a different-answer image collapses their own model's accuracy. LVR's states recover part of that competence: without the picture its state-trained model counts right on 69.2% of photographs, against 79.5% with it and 55.0% for the mean-state control. We instead hold the states fixed and sweep the picture at answer time, on CoVT and LVR.

Baselines and what a reader can show. Park et al. (2026) compare with raw visual features once, not at equal budget, so we add an equal-byte comparison at one portable interface. Probing controls and usable-information measures separate content from what a bounded reader can extract (Hewitt & Liang, 2019; Voita & Titov, 2020; Xu et al., 2020), so we report what a fitted reader decodes and not what a state contains.

4 Experiments

4.1 Photographs: Useful for LVR, a Readout Gap for CoVT

Setup. Each photograph has one question, a count from TallyQA or a left-or-right relation from GQA (Acharya et al., 2019; Hudson & Manning, 2019). Besides CoVT we test LVR-7B (Li et al., 2026a), whose released checkpoint was fine-tuned without LVR's reinforcement-learning stage and feeds back eight latent steps. Both are built on Qwen2.5-VL (Bai et al., 2025). On photographs the CoVT writer picks which experts' slots to write, so we force it to write eight segmentation and four depth slots on every picture, the *tested configuration*. The writer picks these experts by itself on 71% of counting and 91% of relation pictures (Appendix D). For each model we train three matched pairs under the recipe of Section 2 and score them at all eight levels.

LVR's states become useful as the picture degrades. For LVR the matched-training gain is near zero with the full picture and reaches $+14.2$ $[+12.2, +16.4]$ pp on counting and $+12.9$ $[+10.3, +15.4]$ pp on relations without the picture (Figure 2). Every pair agrees on the direction at every degraded level. The near-zero full-picture gain is not a ceiling effect, since the two models give the right count on 79.5% and 79.4% of pictures there. Pictures LVR may have seen in fine-tuning show no larger gain: without the picture their count gain is $+14.1$ pp, against $+14.7$ on the rest (Appendix C).

Other pictures' states do not give the gain. The mean-state control sees one constant tensor, so it never learns to attend to a varying slot input. A second control therefore trains a third answer model in each LVR pair on the states of other pictures, each written under that picture's own question (Appendix C). This shuffled-state model stays within a point of the mean-state control at each of the four levels scored. The state-trained model beats it by $+14.1$ $[+11.9, +16.3]$ pp on counting and $+13.7$ $[+11.1, +16.2]$ pp on relations without the picture, and every pair agrees. So the gain comes from content specific to each picture and its question, and the state-trained model reads it: given another picture's states its count accuracy changes by -28.0 pp.

CoVT's states are decodable but stay unread under the Section 2 recipe. For CoVT's tested configuration the gain stays within half a point of zero at every level on both tasks, and every interval lies inside the ± 2 pp band fixed before the runs (Table 5). The registered prediction of a gain therefore fails for CoVT, while LVR meets it (Appendix C). Within the state-trained model the states make no difference either: without the picture its content contrast, Equation (3), is 0.0 pp on counting and $+0.1$ on relations. The writer also has perception-feature and edge experts, so three more pairs were trained under the writer's pick for each picture, and three under a forced four-expert configuration. Both give the same near-zero gain (Appendix D). Yet a reader that also sees the question's token embeddings decodes both answers from the last-layer slot vectors that the answer models receive, in both configurations it probed. In the tested configuration, on held-out rows of the training split, it gives the right count on 67.0% and the right relation on 71.6%, above both the majority floor and the question alone (Table 8). Still, no answer model trained by the Section 2 recipe (rank-16 adapters, $1{,}500$ steps) learned to read the states in any of the three configurations, although the same recipe makes CoVT's CLEVR states useful (Section 4.2). We call this a *readout gap*: the states expose answers that the trained answer model does not read.

An answer model trained longer starts to read CoVT's states. A follow-up study trained three more pairs of the tested configuration for 4,500 steps at rank 32 and scored them with the full picture,

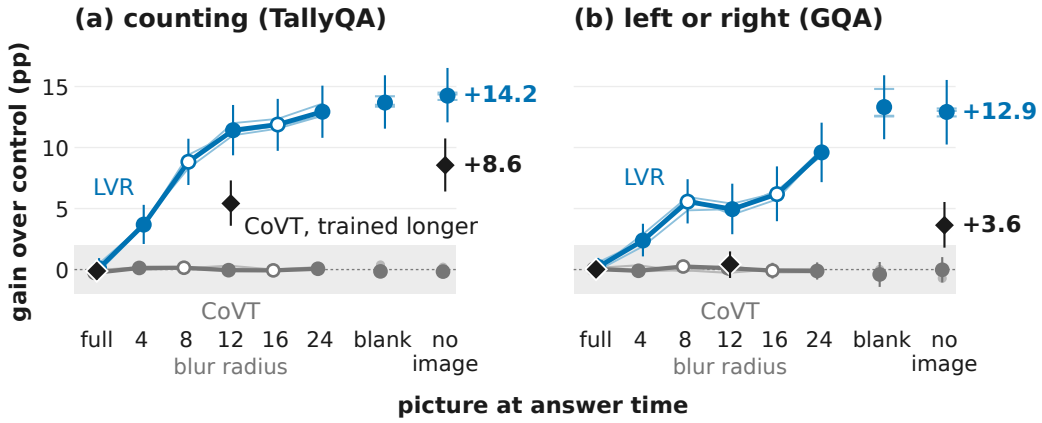

Figure 2: On photographs, LVR's states become useful as the picture degrades, CoVT's do not under the Section 2 recipe, and a CoVT answer model trained longer begins to read them on counting. Each point is the matched-training gain over the mean-state control, Equation (2), on (**a**) TallyQA counting and (**b**) GQA left-or-right questions, with 2,000 held-out pictures per task and three trained pairs per configuration. LVR's three pairs together are blue, each pair light blue. CoVT's tested configuration is gray and its two other configurations light gray. Black diamonds are the tested configuration's answer models trained with three times the steps and twice the rank (4,500 steps, rank 32). Open markers are radii 8 and 16, which training never saw, the gray band is ± 2 pp, and bars are 95% image-bootstrap intervals, adjusted over the two tasks at each level and over six comparisons for the diamonds.

at blur 12, and without the picture (Appendix D). The count gain over the control is now $+5.4$ pp at blur 12 and $+8.6$ $[+6.5, +10.7]$ pp without the picture, every pair above the band (Figure 2). The relation gain without the picture is $+3.6$ $[+1.9, +5.5]$ pp, an interval that excludes zero but reaches into the band, and neither task gains with the full picture. Without the picture the state-trained model's content contrast on counting is now $+11.1$ pp, against 0.0 under the Section 2 recipe, which means its count answers now depend on the states. So under that recipe the near-zero gain came at least partly from the readout, since the same states give a count gain once the answer model is trained longer.

4.2 CLEVR: Decodable, Unused, Then Useful

Setup. The six CLEVR (Johnson et al., 2017) questions are count, count parity (even or odd), shape, color, color with shape, and a relation. CoVT's tested configuration writes one state per image and question. A disjoint pool of 384 reference images supplies the mean and donor states and trains the readers, which are scored on 128 development images (Appendix B).

The count is decodable. A small reader of the last layer's input gives the right count on 95% to 96% of the 128 unseen development images across three reader seeds (Figure 3a). On layer 0 it reaches 9%, the majority floor.

The released model does not use it. On 2,000 held-out confirmation images with the full picture, putting the mean state at all twelve slots and every layer changes the released model's six-question accuracy by $+0.0$ $[-0.1, +0.1]$ pp, inside the ± 2 pp band fixed before the run (count in Figure 3b). The random state does the same, and only 23 of 11,999 parsed answers change, in both directions. Removing the whole span the model writes before its answer changes accuracy by -1.3 pp, almost all on parity, so parity needs that span but not the content of its slots. Swapping in a donor picture's image tokens makes the model give that picture's count on 95% of development examples, so a replacement reaches the answer when the model reads it. Replaying the saved states after a prompt without image tokens drops six-question accuracy from 91% to 27%, and the model writes the same wrong count for every image. Against the mean state, its own states then change six-question accuracy by -0.8 $[-1.1, -0.5]$ pp, a small loss.

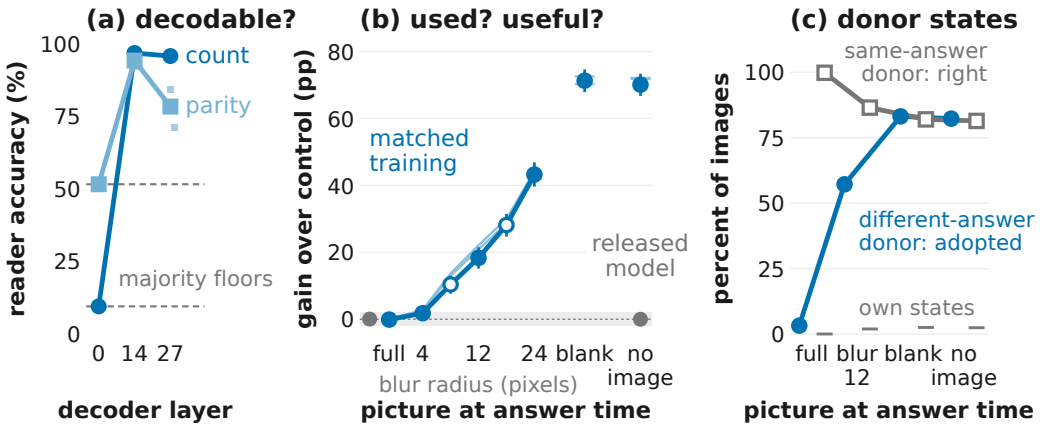

Figure 3: The CLEVR diagnosis on CoVT's states. (**a**) Accuracy on the count and on parity of a reader fitted to the reference images' states, by decoder layer, on 128 development images. Small markers are seeds, dashes are majority floors, and layer 0, whose input is the same placeholder for every image, is the control. (**b**) The matched-training gain on the count, Equation (2), of the first trained pair on 2,000 confirmation images, with intervals adjusted over six questions. Thin traces are the two other pairs, and gray dots Δ_{use} of Equation (1), the released model's own states minus the mean state. Open, unlabeled markers are radii 8 and 16, left out of training, and the band is ± 2 pp. (**c**) Donor states pooled over the three state-trained models: the share of images that give a different-answer donor's count and the share that still give the right count under a same-answer donor. Gray dashes mark the share that already give the different-answer donor's count under their own states.

Matched training makes it useful as the picture degrades. The first trained pair's matched-training gain on the count is -0.1 pp with the full picture, where both models are at ceiling. A small gain of $+1.8$ $[+0.8, +2.8]$ pp already appears at blur 4, where the control still gives the right count on 94% of images. The gain rises at every radius to $+71.3$ $[+68.3, +74.3]$ pp with a blank picture and $+70.1$ pp without the picture (Figure 3b). Two more pairs trained the same way gain $+72.6$ and $+70.2$ pp at blank (Appendix B.2). Blending the picture toward gray gives gains of at most $+2.1$ pp while the control still gives the right count on at least 93.7% of images (Appendix B.3). Because the writer saw the question, the gain may also rest on an answer code written under the question, and no run here rules that out (Appendix F).

Donor states show that a donor's count is adopted. For the first trained pair with the full picture, the content contrast of Equation (3) is $+18.1$ pp on the count, yet the gain over the control is zero. Donor states reconcile the two (Figure 3c, Appendix B.4). With a blank picture the three state-trained models give a different-answer donor's count on 83.2% of images, against 3% under their own states. A same-answer donor gives the right count about as often as their own states do, although it changes 28% of answers. With the full picture the different-answer donor's count is given on only 3.2% of images, but the donor changes 32% of answers and every changed answer moves toward its count. Both change rates fail, on every model, the rules fixed before the run: at least 80% of answers unchanged under a same-answer donor and fewer than 10% changed with the full picture. The models read a donor's state as a competing count, and the mean state, which points to a count of 6, competes the same way. The content contrast measures this competition, not utility.

4.3 Pooled Features Beat the State at One Interface

A matched-training gain does not show that the writer beats what the image encoder already offers. We therefore compare LVR's portable representation, its eight input vectors (LVR-8), with eight averages of the frozen image encoder's features over a fixed 2×4 grid of the original image (pooled-8). The grid keeps coarse layout that a left-or-right question can use. Both enter the same eight positions before the first decoder layer of LVR-7B, at 56 KiB per image, with the pooled vectors rescaled once to LVR's vector norm. For each source we train new rank-16 adapters and a mean-state control with the same settings and three seeds. On the CoVT model the same test compares

CoVT-12, the final hidden states at its twelve slots, with pooled-12 from a 3×4 grid, both entering layer 0 at 84 KiB per image (Appendix E.2).

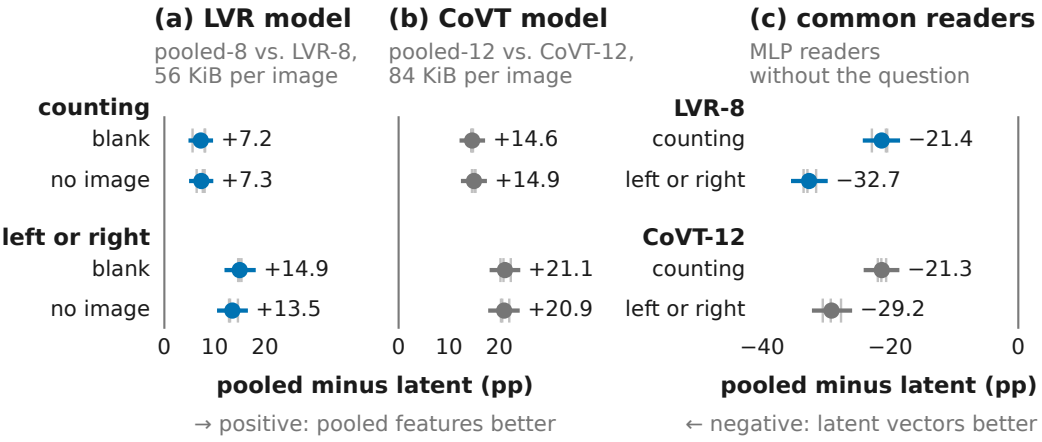

Figure 4: Pooled image features beat the latent vectors at the backbone's portable interface, and common readers reverse the order. Each dot is pooled minus latent accuracy with its 95% image-bootstrap interval, adjusted over eight comparisons, and light gray ticks mark the three seeds. LVR is blue and CoVT is gray, as in Figure 2. (**a**, **b**) At blank and no image, two answer models on one backbone, trained the same way, read either pooled features or latent vectors of the same stored size, given under each panel title. (**c**) The same contrast for common readers, an MLP fitted to each source's standardized vectors without the question, although the latent vectors were written under it. The linear readers are in Appendix E.3.

Pooled features win on both models. Without the picture the pooled features beat LVR-8 by $+7.3$ $[+4.9, +9.7]$ pp on counting and $+13.5$ $[+10.4, +16.6]$ pp on relations, and every seed agrees (Figure 4a). They also lead at blank and at blur 12 (Table 10). At the layer-0 input, where LVR feeds back its own vectors, LVR-8 helps over its mean-state control by $+5.4$ and $+8.0$ pp, less than all-layer replay gives in Section 4.1 with a cache 28 times larger. An answer model trained the same way on pooled features of the same size does better than LVR-8. On the CoVT model without the picture, CoVT-12 does not help over its mean-state control (Table 11), and pooled-12 beats it by $+14.9$ pp on counting and $+20.9$ pp on relations, with every seed agreeing (Figure 4b).

Common readers reverse the order. The answer model may prefer what is familiar, because the backbone saw image-encoder features throughout pretraining but the latent vectors only in LVR's fine-tuning, and never saw CoVT-12 at layer 0. A second test uses common readers, the same linear reader and one-hidden-layer MLP fitted to each source's standardized vectors without the question (Appendix E.3). Without the question every reader decodes the answer from the latent vectors, which were written under it, and none beats the majority floor on the pooled features. The MLP reads the count from LVR-8 on 70.5% of held-out pictures and from pooled-8 on 49.1%, below the majority floor of 50.9%, and CoVT-12 against pooled-12 shows the same pattern (Figure 4c). So the pooled advantage holds at the backbone's interface, and it does not show that the latent vectors hold less of the answer. This test cannot separate familiarity, interface and access to the question (Section 6).

5 Conclusion

Visual thoughts are meant to give a vision-language model a better view of the scene than its backbone features alone. We tested this through three questions that are often mixed up. A state is decodable if a small reader can recover the answer from it, used if the released model's answer depends on it, and useful if a model trained to read it beats one trained on the mean state. We saved each state once from the full picture and then blurred, blanked or removed the picture at answer time, since a state can only matter once the picture stops giving the answer.

On photographs, LVR's states are useful. They add almost nothing with the full picture, but without it they raise counting accuracy by 14.2 points and left-or-right accuracy by 12.9 over the mean-state control. A model trained on other pictures' states does no better than that control, and the state-trained model leads it by 14.1 and 13.7 points. So the gain comes from what each state carries about its own picture and question, not from extra vectors to attend to.

CoVT's states on the same photographs are decodable but not useful under our standard training. A reader that sees the question recovers the count on 67.0% of held-out pictures and the left-or-right answer on 71.6%, yet models trained on these states gain almost nothing in all three configurations we tried. In the tested configuration, training three times longer at twice the adapter rank lifts the counting gain to 8.6 points without the picture. A state with no measured benefit is not necessarily empty, and part of the problem can lie in how it is read.

On CLEVR, a reader recovers the count from CoVT's states on 95 to 96% of images, yet replacing them with the mean state leaves the released model's accuracy unchanged. A model trained to read the same states gains 71.3 points on counting once the picture is blank. Given another image's state, it gives that image's count on 83.2% of blank pictures, and with the full picture 32% of its answers move toward that count. It reads the state as a count and weighs it against the picture.

Even a useful state is not a better input than the image features the model already has. When both are given as a few input vectors of the same size, pooled image features beat LVR's vectors by 7.3 points on counting and 13.5 on left-or-right questions without the picture, and CoVT's by 14.9 and 20.9. Readers without the question decode the answer from the states and not from the pooled features, as expected, since only the states saw the question. The pooled advantage is therefore about what the backbone can use at that input.

Because each state is written after the model reads the question, part of any gain may come from an answer already formed, and our runs do not rule this out. With that caveat, these results explain why one visual thought can look useless in one test and valuable in another. Decodable, used and useful are different properties, and each needs its own test. What a visual thought is worth depends on how it is read, not only on what is written into it. In practice, methods should train a state together with its reader, and evaluations should report a picture sweep, a mean-state control and a pooled baseline of the same size.

6 Limitations and Reproducibility

Limitations. The two checkpoints differ in K, in which slots CoVT writes, in fine-tuning data and in training targets, so the study contrasts two models rather than two mechanisms. They trained on related datasets, so held out refers only to our adapters and readers. The comparison of pictures LVR may have seen is therefore descriptive. The mean state can be out of distribution for the state-trained model, so the utility claim rests on the matched-training gain. The shuffled-state model exists for LVR on photographs only. The near-zero CoVT gain on photographs is shown only for the recipe, configurations, and states tested in Appendix D, and the longer recipe was run on one configuration at three levels only. Because the writer saw the question, the common readers' reversal and the CLEVR and photograph gains can rest on an answer code written under it. The all-layer gains have no pooled-feature baseline, since the equal-byte comparison runs at layer 0 only. Nothing here shows iterative reasoning across states. The intervals cover image sampling with the trained models fixed, not training variability beyond the seeds shown.

Reproducibility. We will release the code, experiment configurations, saved results, fitted-reader and answer-adapter checkpoints, and the scripts that reproduce every analysis, table, and figure. The appendices give the protocol, the statistics, and each registered prediction with its outcome.

AI Use Statement

We used generative AI tools to help write and debug our code and to draft and edit parts of the paper. We did not use them to develop the research questions, design the experiments, set the predictions, or interpret the results. The authors reviewed all code written with AI help and tested it with checks built into the study. For example, replaying CoVT's saved states reproduces its original outputs bit for bit, and swapping in another image's tokens makes the model give that image's count on 95%

of examples (Appendices A and F). The tables and figures are generated by script from the saved results, and every number in the text was checked against those results. The authors revised all drafted text and take responsibility for the content of this work.

REFERENCES

Manoj Acharya, Kushal Kafle, and Christopher Kanan. TallyQA: Answering complex counting questions. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pp. 8076–8084, 2019.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. v1, 19 February 2025.

Jin Cui, Xinyue Long, Xunyong Zhang, Yadong Zhang, Chuanchang Su, Jingye Gan, Boran Zhao, and Pengju Ren. Retrieve, integrate, and synthesize: Spatial-semantic grounded latent visual reasoning. *arXiv preprint arXiv:2605.07106*, 2026. v1, 8 May 2026.

Jiawei Guo, Yu Chen, Xiang Wang, Shuai Li, Xinpei Zhao, Huaxing Liu, Shuai Dong, Feifei Zhai, and Yu Zhou. Beyond visual memory: Mechanistic diagnostics of latent visual reasoning. *arXiv preprint arXiv:2606.01287*, 2026. v2, 26 August 2026.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Lianghuan Huang and Yingshan Chang. Causality $\neq$ decodability, and vice versa: Lessons from interpreting counting ViTs. *arXiv preprint arXiv:2510.09794*, 2025. v1, 10 October 2025.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6700–6709, 2019.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2901–2910, 2017. doi: 10.1109/CVPR.2017.215.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent Visual Reasoning. In *International Conference on Learning Representations (ICLR)*, 2026a.

Jiaang Li, Chengzu Li, Zhaochong An, Yifei Yuan, Xi Liu, Serge Belongie, and Vésteinn Snæbjarnarson. Seeing or knowing? Visual context sensitivity in multimodal large language models. *arXiv preprint arXiv:2607.26326*, 2026b. v2, 25 August 2026.

You Li, Chi Chen, Yanghao Li, Fanhu Zeng, Kaiyu Huang, Jinan Xu, and Maosong Sun. Imagination helps visual reasoning, but not yet in latent space. In *International Conference on Machine Learning (ICML)*, 2026c. arXiv:2602.22766, v3, 3 September 2026.

Suhyeong Park, Junha Jung, and Jaewoo Kang. Reason through the latent! Making latent visual reasoning necessary. *arXiv preprint arXiv:2609.06746*, 2026. v3, 11 September 2026.

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. v3, 5 August 2026.

André G. Viveiros, Nuno Gonçalves, André F. T. Martins, and Matthias Lindemann. What's holding back latent visual reasoning? *arXiv preprint arXiv:2605.18445*, 2026. v2, 19 May 2026.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations (ICLR)*, 2020.

Tianyi Zhang, Mahtab Bigverdi, and Ranjay Krishna. Ablate-to-Validate: Are vision-language models really using continuous thought tokens? *arXiv preprint arXiv:2605.21642*, 2026a. v1, 20 May 2026.

Xin Zhang, Qiqi Tao, Jiawei Du, Moyun Liu, and Joey Tianyi Zhou. Visual latents know more than they say: Unsilencing latent reasoning in MLLMs. *arXiv preprint arXiv:2605.02735*, 2026b. v1, 4 May 2026.

APPENDIX GUIDE

- Appendix A: the two models, the state interfaces with their stored sizes, the all-layer replay cache and the conditions scored with it, the adapters, and checks that a replaced input can reach the answer.
- Appendix B: the CLEVR study behind Figure 3, with the confirmation table, the three trained pairs, the gray blends, and the donor states.
- Appendix C: the photograph data, exposure, scoring, same-question donor states and the shuffled-state control, and two example photographs.
- Appendix D: the CoVT diagnosis, the question-aware readers, and the answer models trained longer.
- Appendix E: the eight-vector and twelve-vector portable comparisons and the common readers on both sources.
- Appendix F: two pilots and the released-model photograph audit.
- Appendix G: estimators and failure rules, every registered prediction with its outcome, and the positioning table.

Accuracy is in percent and differences in percentage points (pp). Brackets give 95% intervals, adjusted where labeled. The follow-up studies are listed with their registered predictions and outcomes in Table 15.

A PROTOCOL, MODELS, AND STATE INTERFACES

Models and states. CoVT-7B-seg_depth_dino_edge (Qin et al., 2025) is based on Qwen2.5-VL-7B (Bai et al., 2025). Its think block, the span the model writes before its answer, places each expert's placeholder tokens between anchor tokens. The tested configuration writes eight segmentation and four depth placeholders, and the checkpoint also writes four perception-feature (DINO) and four edge placeholders on some photographs. LVR-7B (Li et al., 2026a) feeds its final normalized hidden vector back as the next input embedding for eight steps after a start token and then answers in text. The all-layer replay cache of Section 2 is stored in 16-bit form (Table 2). It is recorded during a teacher-forced pass over the model's own think block and written back during a replay, so replaying an image's own states reproduces that pass by construction. The portable representation of Section 4.3 is K vectors written over the layer-0 residual input at the slot positions, and the answer model computes every later layer.

Table 2: Interfaces and stored sizes. Sizes count 16-bit values of width $3{,}584$ per image and exclude the image encoder cache, the weights, the text prefix, and compute. Every study runs one answer pass, and the answer model also sees the picture at every level except the one without it.

study	interface	vectors	layers	bytes
CLEVR, CoVT	all-layer replay	12	28	2,408,448
photographs, CoVT	all-layer replay	12	28	2,408,448
photographs, LVR	all-layer replay	8	28	1,605,632
LVR-8 versus pooled-8	portable, layer 0	8	1	57,344
CoVT-12 versus pooled-12	portable, layer 0	12	1	86,016

Images and questions. The CLEVR study draws from validation images $5{,}000$ to $14{,}999$, disjoint from the $4{,}999$ images of an earlier reader study. A seeded draw fixes the development and confirmation sets. Each image index determines its six scene-graph questions by a fixed rule, so question selection does not depend on the model's answers.

Why a replay reaches everything. CoVT's twelve placeholders are ordinary vocabulary embeddings, and its expert decoders read the last-layer state only during training. A placeholder can therefore reach a later token only through attention, via the keys and values each layer computes from the residual stream. A replay overwrites the residual input at every layer of a replaced slot,

so no trace of the image's own state at that slot survives, while the image tokens, the question, and the placeholder tokens are unchanged. A replay acts only on the pass over the prompt and the think block, and the answer is then decoded from that pass's key-value cache. On CLEVR the writer's think block is the same token string on every image, and on photographs it is forced, so its token ids carry nothing image-specific.

Conditions. With the image: plain generation, replay of the image's own states, reference-pool mean, random state, donor with the same or a different answer, mean on the segmentation or the depth slots only, and no think block. With restricted evidence: the same think block replayed at each degraded level of Section 2, with own states, the mean state, or no think block. Every image and question is scored, and unparsed answers, departures from the fixed think block, and unavailable donors count as wrong and are reported.

Adapters. The rank-16 adapters of Section 2 have scaling $32/16$ and sit on the query, key, value, and output projections, adding 10.1M parameters to the model's 8.30B. They train on CLEVR training images 0 to $9{,}999$ with six question types under cross-entropy on the answer tokens, with AdamW at 10^{-4}, 50 warmup steps, and cosine decay. M_Z and M_C differ only in the slot content. An earlier recipe that trained without the picture left full-picture accuracy below the released model's. The recipe in use, which trains on the six picture levels of Section 2, put it above the released model's by more than the registered 5 pp on two of three pairs, and both are registered failures in Table 13.

Checks and readers. Image-token inputs from a donor picture make the released model give that picture's count on 95% of development images, against zero under its own states. A blank picture of the same size maps to the same token grid and keeps every slot position. With it the released model gives the right count on 0% of development images under its own states or the mean state, as it does without the picture. Each reader of Figure 3a projects the slot vectors to width 128, prepends a learned question embedding, and reads the answer from a two-layer Transformer at the question position. It is fitted on one layer's saved slot inputs over the 384 reference-pool images, with a seeded fifth held out, and scored on the 128 development images. On the confirmation images 841 different-answer donors were unavailable and scored wrong, all in the relation question, whose prompt names an object unique to the image. With these scored wrong, the different-answer donor condition differs from own states by -6.3 pp over all six questions. On the 11,158 image-question cases with a donor that parsed under both conditions, 22 answers change, and the donor's count appears no more often than under own states.

B The CLEVR Study

B.1 Development and confirmation results

On 128 development images the three trained pairs give blank-picture count gains of $+68.8$, $+70.3$ and $+65.6$ pp, with the full picture within a point of zero.

Confirmation table. Table 3 gives each point of Figure 3b with a pointwise interval, under the reference-pool mean, and puts the within-model content contrasts of Equation (3) beside them. In absolute terms the state-trained model gives the right count on 100% of the confirmation images with the full picture, 83% with a blank one and 82% without the picture, so at blank it reads less of the count from the states than the small fitted reader of Section 4.2 does. The mean-state control under its mean gives the right count on 100%, 11% and 12%. Within the state-trained model its own states beat the mean state at every level, while in the control the same swap moves accuracy by less than a point.

Under the training mean the count gain is $+71.3$ $[+69.2, +73.5]$ pp at blank (pointwise interval) and $+71.0$ pp without the picture. At blank the control gives the same answers under both mean states, so the reference-pool and training-mean readings coincide there. The training mean moves no gain by more than 0.9 pp from the reference-pool reading, and no registered verdict changes. For Table 3, M_Z and M_C answered each image on the same type of GPU, and the images run on each of the two GPU types used give similar count results at every level.

Table 3: Confirmation results on CLEVR for the first trained pair (pair 1 of Table 4), on 2,000 confirmation images under the reference-pool mean, in pp and paired by image. Gains carry pointwise 95% image-bootstrap intervals, and Section 4.2 gives the same gains with intervals adjusted over six questions. The content contrast is own minus mean states within one model on the count, Equation (3). Blur 8 and 16 were left out of training.

	matched-training gain		content contrast, count	
level	count	all six questions	state-trained model	mean-state control
full	-0.1 $[-0.3, 0.0]$	-0.1 $[-0.2, 0.0]$	$+18.1$	-0.1
blur 4	$+1.8$ $[+1.1, +2.5]$	$+0.2$ $[-0.1, +0.5]$	$+55.8$	$+0.1$
blur 8	$+10.4$ $[+8.7, +12.0]$	$+2.0$ $[+1.5, +2.5]$	$+67.7$	-0.1
blur 12	$+18.3$ $[+16.3, +20.4]$	$+3.5$ $[+2.9, +4.0]$	$+69.8$	$+0.2$
blur 16	$+28.1$ $[+25.9, +30.4]$	$+5.5$ $[+4.9, +6.1]$	$+71.3$	$+0.1$
blur 24	$+43.2$ $[+40.8, +45.7]$	$+7.6$ $[+7.0, +8.3]$	$+71.2$	$+0.4$
blank	$+71.3$ $[+69.2, +73.5]$	$+14.5$ $[+13.6, +15.3]$	$+70.4$	0.0
no image	$+70.1$ $[+67.9, +72.3]$	$+14.1$ $[+13.2, +14.9]$	$+70.0$	-0.6

B.2 Three trained pairs on the confirmation images

Figure 3b scores the first trained pair. We also score the two other pairs on the same 2,000 confirmation images at all eight levels (Table 15), under the training mean of each pair's control, and score the first pair again the same way. Table 4 gives the count gain per pair and the three-pair mean. The interval of the mean resamples images for the per-image difference averaged over pairs, so it holds the three pairs fixed. Every pair meets the registered criteria except the ± 2 pp band at blur 4, which pairs 2 and 3 exceed at $+2.9$ and $+2.4$ pp. The three agree in direction at every degraded level, and the range across pairs is 2.4 pp at blank and 1.2 pp without the picture, inside the registered 10 pp rule. The within-model content contrast at the full picture varies with the pair ($+18.6$, $+29.1$ and $+0.9$ pp on the count), while the matched-training gain there stays within a point of zero for all three. For half of the images of pair 2 and all of pair 3, M_Z and M_C ran on different GPU types, a departure from the protocol that can affect the gain only through the two answer passes, since the control reads a fixed training mean.

Table 4: Matched-training gain on the CLEVR count (pp) under the training mean on the 2,000 confirmation images. Each pair is a state-trained model and its mean-state control, and pair 1 is the pair of Table 3. The three-pair mean carries an image-bootstrap interval adjusted over six questions.

level	pair 1	pair 2	pair 3	three-pair mean
full	-0.1	0.0	$+0.1$	0.0 $[-0.1, +0.2]$
blur 4	$+1.7$	$+2.9$	$+2.4$	$+2.3$ $[+1.5, +3.1]$
blur 8	$+10.5$	$+13.7$	$+13.4$	$+12.5$ $[+10.5, +14.5]$
blur 12	$+18.2$	$+22.2$	$+20.7$	$+20.4$ $[+18.1, +22.8]$
blur 16	$+28.0$	$+30.4$	$+28.2$	$+28.9$ $[+26.2, +31.5]$
blur 24	$+42.8$	$+44.8$	$+42.3$	$+43.3$ $[+40.5, +46.1]$
blank	$+71.3$	$+72.6$	$+70.2$	$+71.4$ $[+69.4, +73.4]$
no image	$+71.0$	$+72.2$	$+71.8$	$+71.7$ $[+69.4, +73.9]$

B.3 Gray blending as a second degradation

The gray-blend check (Table 15) scores the first trained pair under a second degradation that keeps the image-token grid. The picture at answer time is blended toward uniform gray, $I_\alpha = (1 - \alpha)I + 128\,\alpha$ on pixel values from 0 to 255, at α of 25, 50 and 75 percent, with the writer's states unchanged and no model retrained. The blends remove little of the evidence the count needs, since the control still gives the right count on 99.7%, 99.3% and 93.7% of images. The matched-training gain stays small while the control still counts, at $+0.1$ and $+0.2$ pp for the two lighter blends and $+2.1$ $[+1.1, +3.1]$ pp at 75 percent (adjusted over six questions). Under the training mean, the

within-model content contrast rises from $+18.9$ at the full picture to $+41.1$ pp at 75 percent, as expected if the mean state's own count weighs more as the picture fades. The check does not reach a blank picture, α is not a calibrated amount of evidence, and for half of the images M_Z and M_C ran on different GPU types.

B.4 Donor states in the state-trained models

The registered donor test (Table 15) asks what the state-trained model does with another image's state. For every confirmation image and question it compares the image's own states, the training mean, a random state, and two donor states written under the same question on a reference-pool image, one with the same answer and one with a different answer. Each is scored with the full picture, at blur 12, at blank and without the picture, on 2,000 images for each of the three state-trained models (Figure 3c).

Without the picture the models give the different-answer donor's count on 82.3% of images (80.7% to 83.7% across models) against 2% under their own states. The rate is 83.2% with a blank picture and 57.2% at blur 12, where one model reaches only 49.5% and so misses the registered adoption clause (at least 50% give the donor's count). The training mean and the random state each change the answer on about 86% of images at blank and leave the count right on 12%, near the majority floor. Under a same-answer donor the count stays right on 82.0% of images at blank against 83.1% under the image's own states, although 28% of answers change (14% right to wrong, 13% wrong to right, and the rest between wrong answers). So at blank and without the picture the adoption clause holds on every model, and the retention clause (at least 80% of answers unchanged) holds on none.

With the full picture the different-answer donor's count appears on only 3.2% of images, but 32% of answers change and count accuracy falls from 99.7% to 68%, so the registered negative control (fewer than 10% of answers changed) fails on every model. All 1,943 changed answers move toward the donor's count, and a same-answer donor changes only 0.1% of answers. The training mean acts the same way. Under it the models answer 6 most often at blank, and with the full picture 100% of the answers it changes move toward that count. A replacement state therefore contributes a count of its own, which the model weighs against the picture.

C The Photograph Studies

Data and exposure. TallyQA supplies counts from 0 to 10, and GQA supplies left-or-right questions on Visual Genome photographs (Acharya et al., 2019; Hudson & Manning, 2019; Krishna et al., 2017). Each task has 12,000 training, 128 development, and 2,000 confirmation images with one question each. Training and evaluation identifiers are disjoint, but 1 confirmation image of the counting task appears under another identifier in relation training. The most common TallyQA answer is one, and the majority answer is right on 51.3% of counting and 50.8% of relation confirmation targets, where a target is one picture with its question. Held out means held out from our training only. CoVT trained on CLEVR, Visual7W, and TallyQA, including a subset whose split is undocumented, and LVR-7B trained without the reinforcement-learning stage on Visual CoT, which includes GQA and Visual7W training questions (Qin et al., 2025; Li et al., 2026a). Among development and confirmation targets, 1,483 count pictures belong to GQA train and 679 relation pictures to TallyQA train. Which images the released models saw, including in Qwen2.5-VL pretraining, is unknown. Read by stratum, LVR's count gain without the picture is $+14.1$ $[+11.8, +16.3]$ pp on the 1,401 confirmation counting pictures in GQA train, an LVR source, and $+14.7$ $[+11.4, +18.0]$ on the rest. Its relation gain is $+9.7$ on the 637 pictures in TallyQA train, a CoVT source, and $+14.4$ on the rest. The gain is no larger on the pictures in either training split, and CoVT's tested configuration is within a point of zero in every stratum. The strata were not randomized, so this reading is descriptive.

Configurations and states. CoVT uses the tested configuration of eight segmentation and four depth slots on every picture so that slot positions stay the same, and the released writer chooses these experts by itself on 71% of counting and 91% of relation targets (Appendix D). LVR uses eight latent steps. Its writer first generates the fed-back vectors, then a pass with the adapters off records the all-layer replay cache, and every replacement uses that cache at the target's own positions with the rest of its prefix fixed.

Training, blur, and scoring. Each model has three independently trained pairs under the recipe of Section 2, with fixed training means over 128 training images per task. LVR leaves out 283 of 24,000 training rows whose generation holds no usable states, by a rule fixed before fitting. The picture at answer time is blurred at its own size, with the radius scaled by the ratio of the processor's resize factors. The blur on the token grid then equals the blur that radius r produces on the 480×320 CLEVR render, and the processor resizes the picture as usual, so the number of image tokens is the same at every radius. Every confirmation target stays in the gain curves, and missing LVR states (33 count and 12 relation targets) is scored wrong for both models. The two models of a pair ran on the same GPU type and each pair on a different type, so variation across pairs mixes training and device effects. For each contrast we average the three paired correctness differences for each image, then bootstrap images 10,000 times with Bonferroni correction over the two tasks at each level. The registered analysis gave the mean-state control the mean of a separate 384-image reference pool per task, and the training means used here change every gain by less than a point. LVR meets the registered criteria with intervals above zero, and CoVT does not. The criteria ask for count gains of at least 3 pp at blank and 4 pp without the picture, and relation gains of at least 4 and 6 pp. Table 5 gives every level, and Figure 5 shows one counting case where LVR's states help and one left-or-right case where they hurt, at blank. Across the three pairs, LVR's gain without the picture runs from $+13.9$ to $+14.5$ pp on counting and from $+12.6$ to $+13.2$ pp on relations. CoVT's runs from -0.2 to 0.0 and from -0.2 to $+0.3$ pp.

Table 5: Matched-training gain on photographs under the training mean, in pp with 95% image-bootstrap intervals adjusted over the two tasks at each level, three trained pairs averaged on the 2,000 confirmation pictures per task. Count is TallyQA counting, relation is the GQA left-or-right question, and CoVT uses its tested segmentation-and-depth configuration.

level	CoVT count	CoVT relation	LVR count	LVR relation
full	$-0.3\,[-0.6, 0.0]$	$0.0\,[-0.2, +0.3]$	$+0.1\,[-0.6, +0.9]$	$+0.2\,[-0.5, +0.9]$
blur 4	$+0.1\,[-0.2, +0.4]$	$-0.1\,[-0.5, +0.3]$	$+3.7\,[+2.2, +5.2]$	$+2.4\,[+1.2, +3.7]$
blur 8	$+0.1\,[-0.2, +0.5]$	$+0.2\,[-0.2, +0.7]$	$+8.8\,[+7.0, +10.6]$	$+5.6\,[+3.9, +7.3]$
blur 12	$0.0\,[-0.4, +0.3]$	$+0.1\,[-0.4, +0.6]$	$+11.4\,[+9.4, +13.4]$	$+5.0\,[+3.0, +7.0]$
blur 16	$-0.1\,[-0.4, +0.3]$	$-0.1\,[-0.7, +0.5]$	$+11.9\,[+9.8, +13.9]$	$+6.2\,[+4.0, +8.4]$
blur 24	$+0.1\,[-0.2, +0.3]$	$-0.1\,[-0.7, +0.5]$	$+12.9\,[+10.9, +15.0]$	$+9.6\,[+7.2, +11.9]$
blank	$-0.1\,[-0.5, +0.2]$	$-0.4\,[-1.3, +0.5]$	$+13.7\,[+11.6, +15.8]$	$+13.3\,[+10.8, +15.8]$
no image	$-0.2\,[-0.5, +0.1]$	$0.0\,[-1.0, +0.9]$	$+14.2\,[+12.2, +16.4]$	$+12.9\,[+10.3, +15.4]$

Same-question donor states. The frozen LVR writer answers each target's question on a different training image, chosen without labels, and we inject its states into the target's unchanged answering context. A fixed donor assignment uses 384 images per task, each at most six times, and intervals condition on that assignment. The paired sets hold 1,966 count and 1,987 relation images. Without the picture, replacing own states costs 29.3 pp on counting and 12.3 pp on relations, and at blank it costs 29.5 and 11.8 pp. Scoring every picture left out of these sets against own states still leaves large losses without the picture on both tasks. Donors can make the question inapplicable, so this does not identify what the states represent.

Shuffled-state control. The registered shuffled-state control (Table 15) asks whether the state-trained model beats a model trained on equally variable states of other pictures. A random pairing of the training rows, blind to the labels and fixed before any fit, gives every row the states captured for another row of the same task and question source (for counting, TallyQA questions written by crowd workers or imported from Visual Genome), never one from its own picture. A second fixed map does the same for the confirmation targets, so the shuffled-state models receive the state-trained models' own states in permuted order. Each was written under its own picture's question, so these donor states are not the same-question donor states of Section 2. Three shuffled-state models were trained under the LVR recipe of Section 2 with only the input states changed. Every evaluation of the state-trained models, the shuffled-state models, and the mean-state controls ran on one GPU type. The population is 1,967 counting and 1,988 relation pictures with usable own and donor states, and the donor's answer matches the target's on 35.3% of count and 51.1% of relation pairs. Table 6 gives the accuracies and own minus shuffled, the state-trained model under its own states minus the



Figure 5: Two photographs with LVR answers at blank: on counting the states help (the state-trained model is right and the mean-state control wrong), and on left or right they hurt (the reverse). The writer saw the original image shown, while each answer model sees a blank picture, the state-trained model with the picture's own states and the control with the training mean. Each example is the first such case on its task by image identifier for the first of the three trained pairs, a rule fixed before the answers were read. The dataset answer is black, a right answer is green with a tick, and a wrong answer is orange with a cross. For this pair the states help on 374 counting pictures and hurt on 90, and on left or right they help on 495 and hurt on 243.

shuffled-state model under its donor states, with intervals adjusted over the two tasks. Scoring every unavailable answer wrong moves this difference by at most 0.23 pp. Per pair, own minus shuffled without the picture runs from $+13.7$ to $+14.5$ pp on counting and from $+13.1$ to $+14.4$ pp on relations. The shuffled-state model is indifferent to its slot input, since own states in place of donor states move it by -0.1 pp on counting and 0.0 pp on relations without the picture, and it sits within a point of the mean-state control. The state-trained model given donor states moves by -28.0 and -13.1 pp. Scored again on this one GPU type, the state-trained models under their own states stay within 0.2 pp of their accuracies in the main photograph study at every level.

Table 6: The shuffled-state control on LVR photographs. Accuracy in percent of the state-trained model under its own states (own), of the shuffled-state model under its donor states (shuffled), and of the mean-state control under its training mean (mean), then the two differences, own minus shuffled and shuffled minus mean, in pp with 95% image-bootstrap intervals adjusted over the two tasks, three pairs averaged.

		accuracy (%)			difference (pp)	
task	level	own	shuffled	mean	own minus shuffled	shuffled minus mean
count	full	80.9	80.7	80.8	$+0.2\,[-0.6, +0.9]$	$-0.1\,[-0.4, +0.3]$
count	blur 12	71.3	59.7	59.7	$+11.6\,[+9.6, +13.6]$	$0.0\,[-0.4, +0.5]$
count	blank	70.5	56.6	56.4	$+13.9\,[+11.8, +16.1]$	$+0.2\,[-0.2, +0.5]$
count	no image	70.3	56.1	56.0	$+14.1\,[+11.9, +16.3]$	$+0.2\,[-0.2, +0.5]$
relation	full	89.7	89.4	89.7	$+0.3\,[-0.3, +1.0]$	$-0.3\,[-0.7, 0.0]$
relation	blur 12	76.8	72.1	71.8	$+4.7\,[+2.7, +6.7]$	$+0.3\,[-0.3, +0.9]$
relation	blank	71.9	58.0	58.7	$+13.9\,[+11.3, +16.6]$	$-0.7\,[-1.5, +0.1]$
relation	no image	72.4	58.7	59.2	$+13.7\,[+11.1, +16.2]$	$-0.4\,[-1.3, +0.5]$

D THE COVT DIAGNOSIS

Configurations the writer produces. On the CV-Bench photographs of the released-model photograph audit (Appendix F), the released model wrote four configurations: segmentation and depth on a minority of the $2{,}446$ targets, and three mixes of segmentation with depth, perception feature, and edge on the rest. On the Visual Genome development pictures it chose segmentation and depth for most count and relation pictures. The photograph study forces that configuration on every picture so that the slot positions the answer models read stay the same, and the tested configuration gave the same answer as the writer's own choice on nearly every development target. An exploratory reader that received a task index rather than the question decoded little. That bounds the reader and not the states, and it is why the registered reader below sees the question.

The diagnosis. Four registered studies ask whether CoVT's near-zero gain on photographs comes from the tested configuration or from its answer model, each with three seeds on the photographs of Appendix C (Table 15). Two train matched answer models, one under the configuration the writer picks for each picture, with one training mean per configuration, and one under the released model's four-expert configuration. The third fits question-aware readers on the tensor the answer model receives. The fourth, run once the first three were read, retrains the tested configuration's answer models with three times the steps and twice the adapter rank. The registered rule reads decodability and utility together. States that are decodable and useful hold content the answer model can reach, and states that are decodable but not useful point to a readout gap in the tested answer model. States that are neither expose too little to these readers, while states that are useful but not decodable mean the reader is too weak.

Gains under the writer's pick and the four experts. When the writer chose for each picture, it picked the segmentation-and-depth configuration on 71% of count and 91% of relation targets, the segmentation-and-perception-feature configuration on most of the rest, and the edge-bearing configurations on 2% to 4%. The writer is frozen, so both models of a pair get the same choice. Every configuration had a training mean of 128 rows except the four-expert configuration, which had 103 relation rows and no count mean. Its 4 to 5 count targets per model are therefore left out of the main analysis, as registered, and scoring them wrong for both models moves no estimate or interval bound by more than 0.02 pp. Table 7 sets both configurations beside the tested one on 2,000 pictures per task (1,995 counting pictures under the writer's pick), with three trained pairs each. The matched-training gain is within a point of zero at every level on both tasks, and no interval leaves the ± 2 pp band. Without the picture the count gain is $+0.2$ $[-0.2, +0.6]$ pp under the writer's pick (pairs -0.1 to $+0.4$) and $+0.1$ $[-0.2, +0.4]$ pp under the four experts (pairs -0.1 to $+0.2$), and the relation gains are $+0.1$ and -0.7 pp. One interval lies above zero, $+0.4$ $[0.0, +0.7]$ pp on the count at blank under the four experts. Its lower bound rounds to zero from above, while the relation interval at blur 4 that prints the same starts just below zero. Per configuration, the count gain under the writer's pick without the picture is within a point of zero on every pair for the segmentation-and-depth and perception-feature strata, and zero on the edge stratum of 30 to 33 pictures. The content contrast is as small, $+0.1$ and 0.0 pp on the count without the picture. Under their own states the count accuracy without the picture is 55.2% and 55.6%, against 55.0% and 55.6% for the controls.

Question-aware readers. The reader receives the question's own token embeddings from the released model's frozen embedding table beside the slot vectors of one layer of the tensor the answer model receives, projected to width 128 and passed through two Transformer layers. It is fitted on the first 4,000 training rows of each task with the last fifth held out, under the tested configuration and under the four-expert configuration, and no confirmation target enters. Only 800 rows are held out, a deviation from the $1{,}000$ the registration named. Two controls are fitted the same way: *shuffled* pairs each question with the slot vectors of another training row of the same task and configuration, so the reader sees the question alone, and *mean* replaces the slot vectors by the training-row mean. Layer 0 holds the placeholders and is the question-only floor of the architecture, where the reader on the picture's own states (own) is within a point of its controls. Table 8 gives the held-out accuracies as means over three reader seeds on one held-out set, with no interval. Decodability was registered as the reader beating its shuffled control by five points. It holds at layer 27 on both tasks and both configurations and for every pairing of seeds, while layer 14 carries the count but not the relation. The question-only readers on the count sit below the 55.1% majority floor because they fit the ques-

Table 7: Matched-training gain (pp) of the three CoVT configurations on the 2,000 confirmation pictures per task, three trained pairs each, with 95% image-bootstrap intervals adjusted over the two tasks at each level. The *tested* column is the segmentation-and-depth configuration of the main study and repeats the CoVT columns of Table 5. In the *writer's pick* columns the writer picks each picture's configuration, and *four experts* forces slots for all four of the checkpoint's experts (segmentation, depth, perception feature and edge) on every picture.

task	level	tested	writer's pick	four experts
count	full	$-0.3\,[-0.6, 0.0]$	$+0.1\,[-0.2, +0.4]$	$0.0\,[-0.3, +0.2]$
count	blur 4	$+0.1\,[-0.2, +0.4]$	$-0.1\,[-0.4, +0.3]$	$+0.2\,[-0.1, +0.5]$
count	blur 8	$+0.1\,[-0.2, +0.5]$	$+0.1\,[-0.3, +0.5]$	$0.0\,[-0.3, +0.4]$
count	blur 12	$0.0\,[-0.4, +0.3]$	$+0.3\,[-0.1, +0.7]$	$0.0\,[-0.4, +0.3]$
count	blur 16	$-0.1\,[-0.4, +0.3]$	$0.0\,[-0.4, +0.4]$	$0.0\,[-0.3, +0.4]$
count	blur 24	$+0.1\,[-0.2, +0.3]$	$+0.2\,[-0.1, +0.5]$	$0.0\,[-0.3, +0.3]$
count	blank	$-0.1\,[-0.5, +0.2]$	$+0.3\,[-0.1, +0.6]$	$+0.4\,[0.0, +0.7]$
count	no image	$-0.2\,[-0.5, +0.1]$	$+0.2\,[-0.2, +0.6]$	$+0.1\,[-0.2, +0.4]$
relation	full	$0.0\,[-0.2, +0.3]$	$0.0\,[-0.2, +0.3]$	$+0.1\,[-0.2, +0.4]$
relation	blur 4	$-0.1\,[-0.5, +0.3]$	$-0.1\,[-0.5, +0.2]$	$+0.4\,[0.0, +0.7]$
relation	blur 8	$+0.2\,[-0.2, +0.7]$	$0.0\,[-0.5, +0.4]$	$-0.1\,[-0.6, +0.3]$
relation	blur 12	$+0.1\,[-0.4, +0.6]$	$-0.3\,[-0.8, +0.3]$	$+0.3\,[-0.2, +0.8]$
relation	blur 16	$-0.1\,[-0.7, +0.5]$	$0.0\,[-0.6, +0.6]$	$-0.2\,[-0.8, +0.3]$
relation	blur 24	$-0.1\,[-0.7, +0.5]$	$0.0\,[-0.6, +0.7]$	$-0.2\,[-0.9, +0.5]$
relation	blank	$-0.4\,[-1.3, +0.5]$	$0.0\,[-0.8, +0.9]$	$-0.3\,[-1.2, +0.7]$
relation	no image	$0.0\,[-1.0, +0.9]$	$+0.1\,[-0.8, +1.0]$	$-0.7\,[-1.6, +0.2]$

tion text of the training rows. On LVR's eight fed-back vectors, states whose answer model does gain, the same reader reaches 75.6% on the count against 44.8% for its shuffled control, and 82.1% on relations against 56.2%. So this reader can decode a useful state, and LVR's vectors expose more of both answers than CoVT's tested states.

Table 8: Question-aware readers on the CoVT tensor the answer model receives and on LVR's eight fed-back vectors: held-out accuracy in percent on 783 to 800 training-split pictures per task, mean over three reader seeds. The columns give the reader on the picture's own states (own), the shuffled control (shuffled, question alone), the mean control (mean), the majority floor (majority), and whether the states are decodable (they beat the shuffled control by five points). Segmentation and depth is the tested configuration, and layer 0 holds the placeholders.

configuration	task	layer	own	shuffled	mean	majority	decodable
segmentation and depth	count	0	48.2	48.2	48.2	55.1	no
	count	14	59.3	45.4	50.0	55.1	yes
	count	27	67.0	44.5	53.8	55.1	yes
	relation	0	57.2	57.2	57.2	49.9	no
	relation	14	59.6	54.7	56.6	49.9	no
	relation	27	71.6	53.4	56.2	49.9	yes
four experts	count	0	49.8	49.8	49.8	55.1	no
	count	14	59.8	45.3	49.4	55.1	yes
	count	27	66.1	44.8	53.9	55.1	yes
	relation	0	57.6	56.8	56.8	49.9	no
	relation	14	57.0	55.6	56.8	49.9	no
	relation	27	69.8	50.9	50.4	49.9	yes
LVR	count	fed back	75.6	44.8	49.8	54.7	yes
	relation	fed back	82.1	56.2	55.2	50.1	yes

What the joint pattern does not settle. Without the picture, the state-trained model of the tested configuration gives the right count on 56.1% of confirmation pictures, below the reader's 67.0% on

held-out training rows. The model trained longer (next paragraph) reaches 66.2%, within a point of the reader on a different population, and on relations it reaches 62.6% against the reader's 71.6%. The states under the writer's pick were not probed separately, and the reader's rows are training-split pictures while the model's are confirmation targets.

The answer models trained longer. The fourth study retrained the tested configuration's state-trained model and mean-state control for 4,500 steps at rank 32 (alpha 64), keeping only the final checkpoint and otherwise following the recipe of Section 2. Both were scored at the full picture, at blur 12, and without the picture on the 2,000 confirmation pictures per task. The registration named one seed, and two more if the count or relation gain without the picture cleared the 2 pp band, which the count did at the first seed. To check that the longer recipe finds a gain where one exists, a pooled-12 pair at the layer-0 interface (Appendix E.2) was trained with the same steps and rank. The four models scored on each picture ran on one GPU, and every CoVT gain whose interval excludes zero keeps its sign on the pictures of each of the four GPU types in every seed. Table 9 gives the accuracies, the matched-training gain with intervals adjusted over the six comparisons of each source, the gain under the recipe of Section 2, and the content contrast Δ_{content}. Per seed, the CoVT count gain without the picture runs from $+7.3$ to $+9.4$ pp and the relation gain from $+2.5$ to $+5.0$ pp. Without the picture, pooled-12 gains more than CoVT by $+4.8$ $[+2.6, +7.1]$ pp on the count and $+19.1$ $[+15.9, +22.3]$ pp on relations. This margin compares a gain at the all-layer interface with a gain at the layer-0 interface, so it is not the state-against-pooled contrast of Appendix E.2. The CoVT count gain clears the band at blur 12 and without the picture in every seed. The relation gain without the picture excludes zero with its lower end inside the band. Every full-picture value is within a point of zero, and pooled-12 stays ahead at both degraded levels. No shuffled-state control exists for CoVT. Neither the answer models under the writer's pick or the four experts nor the CoVT-12 model was retrained, and blank and the other radii were not scored.

Table 9: The answer models trained longer, on photographs: accuracy in percent of the answer model trained on each source under that source (own) and of its mean-state control under its training mean (mean), the matched-training gain (pp) with 95% image-bootstrap intervals adjusted over six comparisons per source, the same gain under the recipe of Section 2 ($1{,}500$ steps, rank 16), and the content contrast Δ_{content}, three pairs averaged per picture. CoVT is the tested configuration's all-layer states, and pooled-12 is the pooled-feature pair at layer 0 trained the same way, which checks that the longer recipe finds a gain.

task	level	source	own	mean	gain at 4,500 steps	gain at 1,500 steps	content contrast
count	full	CoVT	81.3	81.4	-0.1 $[-0.7, +0.5]$	-0.3	-0.2
count	full	pooled-12	80.9	81.8	-0.8 $[-1.8, +0.1]$	0.0	$+0.7$
count	blur 12	CoVT	66.6	61.2	$+5.4$ $[+3.7, +7.2]$	0.0	$+6.9$
count	blur 12	pooled-12	72.3	61.2	$+11.2$ $[+8.8, +13.5]$	$+11.6$	$+16.9$
count	no image	CoVT	66.2	57.6	$+8.6$ $[+6.5, +10.7]$	-0.2	$+11.1$
count	no image	pooled-12	71.5	58.2	$+13.4$ $[+10.9, +15.8]$	$+15.0$	$+17.0$
relation	full	CoVT	89.0	89.0	0.0 $[-0.5, +0.5]$	0.0	0.0
relation	full	pooled-12	88.8	88.8	0.0 $[-1.0, +0.9]$	$+0.3$	$+0.5$
relation	blur 12	CoVT	72.3	71.9	$+0.4$ $[-0.6, +1.4]$	$+0.1$	$+0.3$
relation	blur 12	pooled-12	81.8	72.0	$+9.8$ $[+7.3, +12.5]$	$+8.0$	$+16.9$
relation	no image	CoVT	62.6	59.0	$+3.6$ $[+1.9, +5.5]$	0.0	$+2.9$
relation	no image	pooled-12	81.6	58.8	$+22.8$ $[+19.8, +25.7]$	$+20.8$	$+21.3$

E Portable Representation Comparisons

E.1 Eight LVR input vectors versus eight pooled image vectors

This comparison uses the photograph study's splits. LVR supplies the eight hidden vectors it feeds back during its frozen writing pass. The pooled baseline averages the frozen image encoder's features on the original image within the eight cells of a fixed 2×4 grid, without the question or the answer labels. Both sources are eight vectors of width $3{,}584$ in 16-bit form, 56 KiB per image,

while LVR's all-layer replay cache of the photograph study is 28 times larger. One fixed positive scalar (1.30675), fitted on 256 training images, matches the pooled vectors' root mean square norm to LVR's. We train four new answer models per seed, each a set of rank-16 adapters on the LVR-7B backbone: one on each source and one on each source's fixed mean over 128 training images per task, which is its mean-state control. All fits use the same 23,717 training rows, 1,500 steps, batch 16 and the six training image conditions. Each answer pass inserts the eight vectors before the first decoder layer and leaves the target's question and text prefix unchanged. We score each model with the full picture, at blur 12, with a blank picture and without the picture on the 1,967 counting and 1,988 relation images with usable LVR states, shared by both sources and all seeds. Each image's paired differences are averaged over the three seeds before 10,000 bootstrap draws of images, and each column of Table 10 forms its own Bonferroni family over the four levels and two tasks. The design was fixed before this reading, but these images had already been used in the all-layer experiments.

Without the picture the pooled-8 model gives the right count on 68.8% of pictures and the LVR-8 model on 61.4%, against 56.0% for the LVR-8 control (the gains are in Table 10). Across the three seeds the pooled advantage without the picture runs from +6.4 to +8.0 pp on counting and from +12.8 to +14.6 pp on relations. Scoring unavailable states as wrong moves every entry of Table 10 by less than 0.3 pp, and dropping the count picture that also appears in relation training (Appendix C) moves them by less than 0.1 pp. LVR's gains are smaller here than under all-layer replay, but training and interfaces differ, so this does not locate stored information.

Table 10: Portable representation on LVR-7B, in pp: LVR-8 (LVR's eight fed-back vectors) against pooled-8 (eight pooled image-encoder vectors), both at the layer-0 input, with 95% image-bootstrap intervals adjusted over eight comparisons per column. Each gain compares a source's answer model with its separately trained mean-state control, and the last column compares the two sources' answer models directly.

task	level	LVR-8 gain	pooled-8 gain	pooled-8 minus LVR-8
count	full	$+0.2\,[-0.7, +1.1]$	$+0.1\,[-0.8, +1.1]$	$+0.2\,[-0.8, +1.2]$
count	blur 12	$+3.5\,[+1.7, +5.3]$	$+10.0\,[+7.6, +12.4]$	$+6.5\,[+4.2, +8.8]$
count	blank	$+5.2\,[+3.0, +7.4]$	$+12.3\,[+9.7, +14.9]$	$+7.2\,[+4.8, +9.7]$
count	no image	$+5.4\,[+3.4, +7.5]$	$+12.6\,[+10.0, +15.2]$	$+7.3\,[+4.9, +9.7]$
relation	full	$+0.2\,[-0.6, +1.0]$	$+0.3\,[-0.6, +1.1]$	$+0.1\,[-0.9, +1.0]$
relation	blur 12	$+2.7\,[+0.5, +4.9]$	$+8.4\,[+5.9, +10.9]$	$+6.3\,[+3.9, +8.9]$
relation	blank	$+7.3\,[+4.2, +10.4]$	$+22.4\,[+19.2, +25.5]$	$+14.9\,[+11.9, +18.1]$
relation	no image	$+8.0\,[+5.1, +11.1]$	$+21.7\,[+18.7, +24.8]$	$+13.5\,[+10.4, +16.6]$

E.2 Twelve CoVT final-state vectors versus twelve pooled vectors

The same comparison on the CoVT model uses twelve vectors per picture (Table 15). CoVT-12 is the checkpoint's final normalized hidden state at the twelve visual-thought slots of the tested configuration's pass on the original image, the tensor its expert decoders receive. It is therefore a portable final-state readout and not the all-layer replay of the main study. Pooled-12 averages the frozen image encoder's features at language-model width within a fixed 3×4 grid. Both are twelve vectors of width 3,584 in 16-bit form, 84 KiB per picture, and both enter once at the layer-0 residual input of the same twelve slot positions with the target's question and text prefix unchanged. Neither source is written for layer 0 of the CoVT model, so each is scaled by one fixed positive scalar that matches its norm to the layer-0 image tokens (0.393 for CoVT-12 and 1.565 for pooled-12). Both scalars are fitted on the first 128 training pictures per task. The recipe is the photograph study's, with a separate mean-state control per source. Each model is scored with the full picture, at blur 12, with a blank picture and without the picture on the 2,000 confirmation pictures per task, with the estimator and families of Appendix E.1.

With the full picture nothing moves. As the picture degrades the pooled features gain over their control on both tasks, reaching +15.0 and +20.8 pp without the picture. CoVT-12 gains nothing at any level (Table 11), and its content contrast Δ_{content} stays within a point of zero in every point estimate, so the model trained on CoVT's final states does not read them. Without the picture the

pooled-12 model gives the right count on 70.7% of pictures, the CoVT-12 model on 55.8% and its control on 55.6%. Across seeds the pooled advantage runs from $+14.5$ to $+15.2$ pp on counting and from $+20.3$ to $+21.9$ pp on relations. A reader that sees the question decodes both answers from the input of the last layer at these slots (Appendix D), one layer before the vectors used here, and the common readers of Appendix E.3 decode both answers from these vectors themselves.

Table 11: Portable representation on the CoVT model, in pp: CoVT-12 (the final hidden states at its twelve slots) against pooled-12 (twelve pooled image-encoder vectors), both at the layer-0 input, with 95% image-bootstrap intervals adjusted over eight comparisons per column. Each gain compares a source's answer model with its separately trained mean-state control, and the last column compares the two sources' answer models directly.

task	level	CoVT-12 gain	pooled-12 gain	pooled-12 minus CoVT-12
count	full	$0.0\,[-0.3, +0.3]$	$0.0\,[-0.9, +0.9]$	$0.0\,[-0.9, +0.9]$
count	blur 12	$+0.1\,[-0.3, +0.6]$	$+11.6\,[+9.3, +14.1]$	$+11.6\,[+9.3, +14.1]$
count	blank	$-0.1\,[-0.4, +0.2]$	$+14.7\,[+12.2, +17.3]$	$+14.6\,[+12.1, +17.2]$
count	no image	$+0.1\,[-0.2, +0.6]$	$+15.0\,[+12.5, +17.7]$	$+14.9\,[+12.4, +17.5]$
relation	full	$-0.1\,[-0.4, +0.2]$	$+0.3\,[-0.5, +1.1]$	$+0.2\,[-0.6, +1.1]$
relation	blur 12	$+0.4\,[-0.1, +1.0]$	$+8.0\,[+5.4, +10.6]$	$+7.8\,[+5.2, +10.4]$
relation	blank	$-0.1\,[-1.3, +1.0]$	$+20.3\,[+17.2, +23.4]$	$+21.1\,[+18.1, +24.1]$
relation	no image	$-0.3\,[-1.4, +0.9]$	$+20.8\,[+17.7, +23.9]$	$+20.9\,[+17.8, +24.0]$

E.3 Common readers on both sources

The common readers, the same linear reader and MLP fitted to each source without the question, test whether the pooled advantage at the backbone's layer-0 interface persists when both sources are read the same way (Table 15). For LVR-8 against pooled-8 the readers train on the 11,773 counting and 11,944 relation rows that have both sources, and they are tested on the 1,967 and 1,988 confirmation targets. For CoVT-12 against pooled-12 they train on all 12,000 rows and are tested on 2,000 targets per task. Each source's K vectors are concatenated in slot order into $K \times D$ features and standardized per feature with the training rows' mean and standard deviation, so the fixed scales of the two answer-model studies play no role. Each task and source gets two readers, identical in architecture and parameter count across the two sources: a linear reader and an MLP with one hidden layer of 512 units. Both train with AdamW over three reader seeds, with early stopping on a 10-percent validation split drawn once and weight decay chosen from three values by validation accuracy. Two controls are fitted the same way. The mean-input reader sees only the training mean, so it can learn only the class prior, and the shuffled-answer reader trains on shuffled answers. The training-majority answer sets the floor. The readers receive the vectors and no question, while the LVR-8 and CoVT-12 vectors were written under the question and the pooled vectors were not.

Table 12 gives the held-out accuracies and the pooled-minus-latent contrast, with intervals adjusted over the eight comparisons. Under the MLP the latent readers decode the count on 70.5% and 71.3% of targets and the relation on 82.7% and 79.3%, with the linear reader within a few points. The pooled MLP readers stay within two points of the majority floor and of the mean-input control. The linear pooled count readers fall below the floor, to 38.2% and 37.9% against 50.9% and 51.3%. This fitting failure enters the linear count contrasts, and it also puts the linear mean-input count reader on LVR-8 at 41.4%. The shuffled-answer readers sit at chance on relations and below the floor on the count. So common readers without the question decode the answer from the latent vectors, which were written under the question, and not from the pooled features. Under the registered rules this outcome narrows the representation claim of Section 4.3 to the backbone's interface: the pooled features' advantage holds there, and it does not show that the latent vectors hold less of the answer. Because the readers see no question, a pooled reader near the floor is expected by design, and the probe cannot separate familiarity, interface and access to the question.

Table 12: Common readers on both sources: held-out accuracy in percent of the latent and the pooled reader, pooled minus latent with 95% image-bootstrap intervals adjusted over eight comparisons, a control reader trained on shuffled answers (mean of the two sources) and the majority floor, three reader seeds averaged.

sources	task	reader	latent	pooled	pooled minus latent	shuffled answers	majority
LVR-8 versus	count	linear	70.3	38.2	$-32.1\,[-35.0, -29.2]$	24.8	50.9
pooled-8	count	MLP	70.5	49.1	$-21.4\,[-24.3, -18.5]$	42.0	50.9
	relation	linear	83.8	49.6	$-34.2\,[-37.2, -31.0]$	49.3	51.0
	relation	MLP	82.7	50.0	$-32.7\,[-35.5, -29.8]$	51.6	51.0
CoVT-12 versus	count	linear	71.5	37.9	$-33.6\,[-36.6, -30.6]$	24.1	51.3
pooled-12	count	MLP	71.3	50.0	$-21.3\,[-24.2, -18.6]$	43.6	51.3
	relation	linear	82.6	50.1	$-32.6\,[-35.8, -29.4]$	50.2	50.8
	relation	MLP	79.3	50.1	$-29.2\,[-32.2, -26.0]$	49.5	50.8

F PILOTS

LVR on CLEVR. On 128 development images (pointwise intervals throughout this appendix), replacing all eight latent positions of the released LVR model by their mean changes its count accuracy by $+3.9$ $[0.0, +7.8]$ pp. This does not establish equivalence. Donor states and removing the think block give negative point estimates with intervals that include zero, so whether the released model uses its state (Section 2) is answered for CoVT and only piloted for LVR.

One state per image. On 128 development images, we reuse the CoVT states written under the count prompt to answer all six questions. For the state-trained model, reused minus question-specific states is $+1.2$ $[-0.3, +2.6]$ pp at blank and -1.0 $[-2.7, +0.7]$ pp without the picture, and evidence of useful content is limited to count and parity. On the count the reused and question-specific states are the same, so this pilot cannot tell whether the count gain rests on an answer code written under the question, and no run here rules that out. It does not confirm general reuse.

Released-model photograph audit. CV-Bench supplies 2,446 targets on 1,972 photographs, and 48 other examples per task supply the control states. CoVT chooses different expert configurations across images, so controls are computed within configurations. On 26 count targets none of these examples shares the configuration and the unavailable control is scored wrong, which accounts for the registered no-image difference of $+0.6$ pp on all targets. On targets with both conditions available the difference computed after the run is $+0.2$ $[-0.4, +0.7]$ pp. Replaying the model's own states reproduces its outputs bit for bit on every target, and this audit is why the photograph study fixes CoVT's configuration.

G STATISTICS, REGISTRATIONS, FAILURES, AND POSITIONING

Estimator, failures, and families. Every contrast is a paired difference of per-image correctness between two conditions, averaged over the images each study fixed in advance, with a percentile interval from 10,000 bootstrap draws of images (seed fixed) that resample every question and condition of an image together. Where three trained pairs exist, the per-image differences are averaged over pairs before the bootstrap. The interval therefore conditions on those pairs, so each pair's estimate or range is shown beside it and no interval over training seeds is claimed. Unparsed answers, unavailable controls and operational failures score as wrong in the main CLEVR and all-layer photograph analyses, with complete-case values beside them where they differ. Three studies registered a narrower set of targets and report beside it the value with every unavailable target scored wrong: the CoVT study under the configuration the writer picks keeps the targets whose configuration has a training mean, and the eight-vector comparison and the same-question donor test on photographs keep the targets with usable LVR states. The count intervals quoted in Section 4.2 and Table 4 are Bonferroni-adjusted over six questions, while the count curve of Table 3 carries pointwise intervals. Photograph intervals and the shuffled-state control are adjusted over the two tasks at each level. The portable comparisons are adjusted over the eight comparisons in each column of Tables 10 and 11,

the common readers over their eight comparisons, the answer models trained longer over six per source, and the gray blends over six questions. The training-mean gains of Appendix B.1, the other outcomes of Table 13, the exposure strata and the pilots carry pointwise intervals. The donor rates of Appendix B.4 are means over the three models, shown with their range and no interval. Intervals at different levels are separate readings, so statements about the shape of a curve describe point estimates. Intervals that include zero are never read as equivalence, since a no-effect claim needs the point estimate and its interval inside the registered 2 pp band. Confirmation images are never used for tuning, and all-layer replay and layer-0 readings are reported separately (Table 2).

G.1 Registered predictions and outcomes

Table 13 lists the predictions of the CLEVR study that decide the paper's claims and their outcomes, including failures. The rules and adapters were fixed before the confirmation run, and fixes to the analysis code and one clarified criterion for the untrained blur radii came before its outputs were read.

Table 13: Registered predictions of the CLEVR study and their outcomes. The released-model tests and the gain test use $2{,}000$ confirmation images, while adapter development and state reuse use 128 development images. State reuse answers all six questions from the states written under the count prompt. Accuracy in percent, differences and tolerances in pp, positive differences favor the first condition named, brackets give 95% intervals.

test	registered criterion	outcome
Released model, full picture	Mean state minus own states within ± 2 pp on six-question accuracy	Met: $+0.0$ $[-0.1, +0.1]$ pp
Released model, no image	Own states minus mean state, allowing either sign	Reported: -0.8 pp, with the count wrong on every image
Earlier recipe, trained without the picture	Preserve the released model's full-picture accuracy	Failed: 67.2% versus 90.4%
Recipe in use, six picture conditions	Full-picture accuracy within 5 pp of the released model (development)	Not met on two of three pairs, upward: 98.4% and 98.6% versus 90.4%. Pair 2 within tolerance at 92.8%
Gain as the picture degrades	Count gain within ± 2 pp at the full picture and blur 4. Interval above zero at blur 12, 24, blank, and no image. Endpoint gains at least 20 pp on count and 10 pp over six questions	Met on the point estimates (Table 3). Not on the interval at blur 4, where the count interval adjusted over six questions $[+0.8, +2.8]$ excludes zero and reaches above the band
Untrained blur radii	Count gain ordered between trained neighbors and within their intervals' union, clarified before reading as the envelope from the lower neighbor's lower bound to the upper neighbor's upper bound	Ordered and inside the envelope. Both radii fail the literal union criterion
Full-picture accuracy after training to read states	At least the released model's accuracy minus 2 pp	Met: $+8.1$ pp against the released model
State reuse, released model	Reused minus question-specific states with the full picture, and reused minus mean without the picture, within ± 2 pp	Met with the full picture. Not established without the picture: -1.3 $[-2.3, -0.4]$ pp
State reuse, state-trained model	Reused within 5 pp of question-specific states on six-question accuracy	Met at blank and without the picture

G.2 Positioning against the closest papers

Table 14 places the closest papers on factual axes, checked against the arXiv versions cited in the bibliography. Yes means the paper does it, part means a restricted or different form, and no means it is absent. The table ranks nothing. This paper adds the sweep of the picture at answer time with fixed states, matched answer models against a mean-state control at every level, readers on

the tensor the answer model receives, same-answer retention beside adoption, and the equal-budget pooled comparison. Its no-image level is a bottleneck for the answer model only, and its marker analysis covers only the condition without a think block.

Table 14: Factual axes of the closest papers. Replacement: inference-time replacement of latent content in a fixed model. Reader: a fitted reader on the states. Retrain: a matched answer model trained with and without image-specific states. Sweep: a sweep of the picture at answer time with states fixed. Pooled: an equal-budget pooled-feature baseline at the same interface. Bottleneck: an interface where the state is the only visual path. Markers: a decomposition into content, markers, and format. Donor: a semantic donor adoption or retention test. Cross: the same protocol on two or more latent mechanisms.

paper	replacement	reader	retrain	sweep	pooled	bottleneck	markers	donor	cross
Zhang et al. (2026a)	yes	no	part	no	no	no	part	no	yes
Viveiros et al. (2026)	yes	no	part	no	no	part	no	part	yes
Guo et al. (2026)	yes	no	no	part	no	yes	yes	yes	yes
Zhang et al. (2026b)	yes	no	no	no	no	no	no	no	part
Li et al. (2026c)	yes	part	part	no	no	part	no	no	yes
Park et al. (2026)	yes	no	part	part	part	yes	no	part	yes
Cui et al. (2026)	no	no	no	part	no	part	part	no	part
CoVT, Qin et al. (2025)	no	part	part	no	no	no	no	no	no
LVR, Li et al. (2026a)	no	no	part	no	no	no	no	no	no
this paper	yes	yes	yes	yes	yes	part	part	yes	yes

G.3 Registrations of the follow-up studies

Table 15 lists the predictions of the follow-up studies, fixed before any of their runs were read, with their outcomes. Two later amendments changed no analysis rule, and one of them let the two extra seeds of the answer models trained longer start before the first seed's evaluation had finished.

Table 15: Predictions of the follow-up studies, fixed before any of their runs were read, and their outcomes. Parts marked not run are never used as evidence.

study	registered analysis	outcome
CLEVR replication	The two other trained pairs, evaluated on the 2,000 confirmation images at every level under the training mean, each pair's curve and the three-pair mean reported, and direction agreement at blank and no image	every pair meets the criteria except the blur-4 band, which pairs 2 and 3 exceed at $+2.9$ and $+2.4$ pp, blank count gains $+71.3$, $+72.6$ and $+70.2$ pp, three-pair mean $+71.4$ pp (Appendix B.2)
Donor test	Adopted answer of a different-answer donor and retention under a same-answer donor at blank, no image, and blur 12, against the training mean and a random state, with the full picture as negative control and unavailable donors scored against the claim in a worst case	on three answer models: adoption 83.2% at blank and 82.3% without the picture against 3% and 2% under own states, same-answer donors keeping the count right about as often as own states, the retention clause and the full-picture control failing on every model (32% of answers changed at the full picture), and every changed answer moving toward the donor's count (Appendix B.4)
CoVT, configuration picked by the writer	Matched CoVT answer models trained and evaluated under the configuration the writer picks for each picture, with one training mean per configuration, the no-image gain per task as the main outcome	no gain of more than two points at any level on three pairs, $+0.2$ $[-0.2, +0.6]$ pp on the count without the picture (Appendix D)
CoVT, other fixed configurations	The same under the checkpoint's other configurations, the four-expert configuration first	the four-expert configuration gives no gain of more than two points at any level on three pairs, $+0.1$ $[-0.2, +0.4]$ pp on the count without the picture (Appendix D)
Question-aware readers	Held-out accuracy of a reader that sees the question text and the slot states of one layer, against shuffled and mean controls, decodable when it exceeds the shuffled control by five points	decodable at layer 27 on both tasks and both configurations (Appendix D)
CoVT-12 against pooled-12	Twelve final slot states against twelve pooled image vectors at one layer-0 interface on the CoVT model, 84 KiB each, newly trained answer models, three seeds, reported whichever way it falls	pooled-12 beats CoVT-12 by $+14.9$ and $+20.9$ pp without the picture, CoVT-12 gains $+0.1$ and -0.3 pp over its control, and the registered question-aware probe was replaced by the common readers, which do not see the question (Appendix E.3)
Gray blending	The first trained CLEVR pair and three LVR photograph pairs under blends of 25, 50, and 75 percent gray against the training mean	count gains $+0.1$, $+0.2$ and $+2.1$ pp at 25, 50 and 75 percent gray while the control still gives the right count on 99.7%, 99.3% and 93.7% of images (Appendix B.3), the LVR photograph pairs not run
Donor test without the picture	Only if the donor test is ambiguous without the picture	conditional, not run
Shuffled-state control	Three LVR answer models trained on other pictures' states under a fixed pairing chosen without labels, scored under those donor states against the state-trained model and the mean-state control at full, blur 12, blank, and no image, with the state-trained model minus the shuffled-state model per task read against the 2 pp band	the state-trained model beats the shuffled-state model by $+14.1$ $[+11.9, +16.3]$ pp on counting and $+13.7$ $[+11.1, +16.2]$ pp on relations without the picture, and the shuffled-state model sits within a point of the mean-state control (Appendix C)
Common readers	The same linear and MLP readers on the standardized vectors of both sources for each model, without the question, pooled minus latent over eight comparisons	under common readers without the question the latent vectors win for both models, both tasks and both readers, pooled minus latent from -34.2 to -21.3 pp, and no pooled reader beats the majority floor (Appendix E.3)
CoVT answer models trained longer	One seed at 4,500 steps and rank 32 for the tested configuration's all-layer states and for pooled-12 at layer 0, scored at no image and blur 12 (the full picture added), with two more seeds only if the CoVT gain without the picture at the first seed clears the band	three seeds, since the first seed's count gain without the picture cleared the band: count gain $+5.4$ pp at blur 12 and $+8.6$ $[+6.5, +10.7]$ pp without the picture, every seed above the band, relation $+3.6$ $[+1.9, +5.5]$ pp with its lower end inside the band, full picture within a point, pooled-12 ahead at both degraded levels (Appendix D)