GOGGLES: WHEN OUTPUT SERIALISATION CHANGES THE BENEFITS OF HIGHER RANK IN VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision Language Models (VLMs) can acquire dense visual perception skills by distilling specialized expert targets via Low-Rank Adaptation (LoRA). However, because autoregressive VLMs generate spatial predictions sequentially, a fundamental question remains: how does adapter capacity interact with output serialisation, the way structured targets are formatted into text token sequences? We address this with **GOGGLES**, a framework that evaluates detachable LoRA adapters on frozen VLM backbones. Our design philosophy isolates output serialisation from information content by holding teacher target resolutions strictly fixed while varying output organisation. We uncover three primary insights: i. *Serialisation-Rank Synergy*: lossless output restructuring disproportionately benefits low-rank adaptation, allowing low-rank models to approach the fidelity of high-rank counterparts; ii. *Domain Specificity*: this interaction is highly modality-dependent, holding strongly for continuous depth rasters but vanishing in discrete segmentation, revealing that sequence dynamics depend on spatial target structure; and iii. *Fidelity-Efficiency Trade-offs*: gains in expert-relative fidelity require additional inference prefills and compute, demonstrating that serialisation design must balance fidelity with deployability. Overall, GOGGLES establishes output serialisation as a critical dimension alongside adapter rank, demonstrating how target sequence structure can alter rank requirements in VLM perception.

1 INTRODUCTION

The capacity needed to generate a spatial prediction depends on how that prediction is written. A vision–language model (VLM) must turn an image into a sequence of tokens before a depth map or mask can be reconstructed. LoRA rank limits the trainable update (Hu et al., 2021), whereas serialisation determines the length and conditioning history of each answer. These choices meet at every generated token, yet are usually evaluated separately.

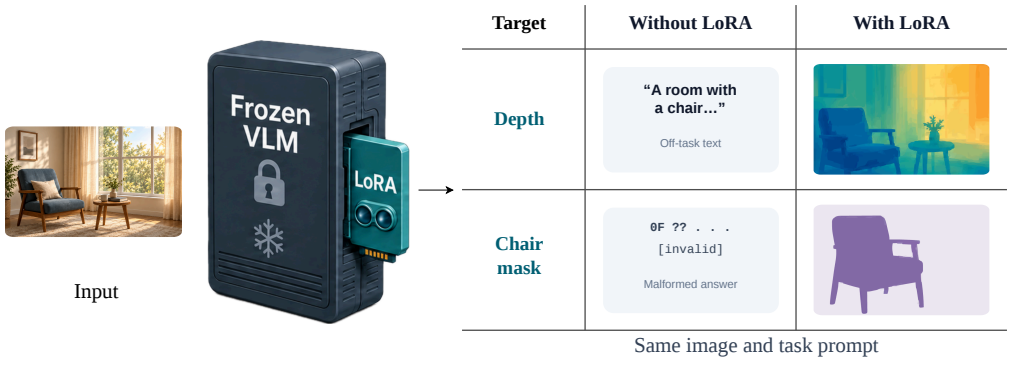

Figure 1: **Detachable perception adapters for spatial readouts.** The same image and task prompt are processed by a frozen VLM with or without an attached LoRA adapter; the right-hand columns depict depth and chair-mask readouts.

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.

Existing work shows that vision models can express spatial outputs through text or special tokens (Chen et al., 2022; Lan et al., 2024; Lai et al., 2023; Cai et al., 2025). Yet a shorter or coarser answer can improve performance simply by removing target values. To ask whether format changes the value of extra LoRA rank, the target must be identical after reconstruction. We compare a single raster answer with four interleaved answers that retain every quantised target cell and use the same backbone, teacher and image set.

We call this controlled protocol **GOGGLES**, a mnemonic for grouping grid outputs when evaluating LoRA skills. Its central statistic is a difference of rank gains: the improvement from rank 16 to 256 with one answer minus the improvement with four answers. A positive value is a *sub-additive interaction*: grouping and higher rank partly substitute for each other. This contrast answers a narrower question than whether either intervention improves absolute fidelity.

The depth experiment yields a positive interaction in all three training seeds on 1,000 fixed confirmation images. The three estimates are $+0.037$, $+0.031$ and $+0.033$, with conditional image-bootstrap intervals above zero. Grouping also reduces low-fidelity outputs. A 2,798-image segmentation replication bounds the mean-IoU interaction within ± 0.02 in every seed, showing that the result does not automatically transfer to a compact binary mask. Coarse-target controls and a gold-prefix diagnostic narrow possible explanations without identifying a single cause.

Together the interaction, mask boundary and four-prefill cost show why adapter rank and output format must be evaluated as a pair when a VLM generates a complete spatial prediction.

2 Related work

Spatial prediction as generation. Pix2Seq expresses detection outputs as token sequences (Chen et al., 2022). Text4Seg and Text4Seg++ design compact textual mask representations (Lan et al., 2024; 2025); DepthLM shows that textual supervision can support metric depth estimation (Cai et al., 2025). LISA instead adds a segmentation token and decoder (Lai et al., 2023). These papers established that the output interface matters. Our question differs: with a fixed quantised target and adaptation recipe, does grouping the answer change the gain from additional LoRA rank?

Expert distillation and adaptation capacity. Depth Anything V2 and SAM 2 provide the frozen expert targets in our experiments (Yang et al., 2024; Ravi et al., 2024). LoRA restricts a weight update to learned low-rank factors (Hu et al., 2021); consequently, rank alone cannot define the difficulty of a generated readout. Work on perception tokens and specialist decoders studies alternative interfaces (Bigverdi et al., 2024; Li et al., 2026). We keep the vocabulary and adapted layers fixed, so the factorial contrast isolates the effect of output grouping within this autoregressive interface. It does not claim that text is the best way to produce dense predictions.

Empirical studies of interacting design choices. Analysis papers such as the attention-head ablation of Michel et al. (2019) and the view-selection study of Tian et al. (2020) use controlled interventions to test a proposed explanation, then state where it stops applying. We follow that evidential pattern: define an estimand, preserve the relevant target, measure the interaction across seeds, and use diagnostic controls and a second task to bound its interpretation. This framing treats GOGGLES as a study protocol, not a new prediction architecture.

3 GOGGLES: separating target information from output organisation

The comparison isolates one empirical question: how does grouping a fixed target change the gain from higher adapter rank? We hold the quantised teacher grid fixed, train one adapter for each format–rank cell, and evaluate every cell on the reconstructed complete grid. The grouping changes the number of calls and their answer histories, which are measured rather than treated as free.

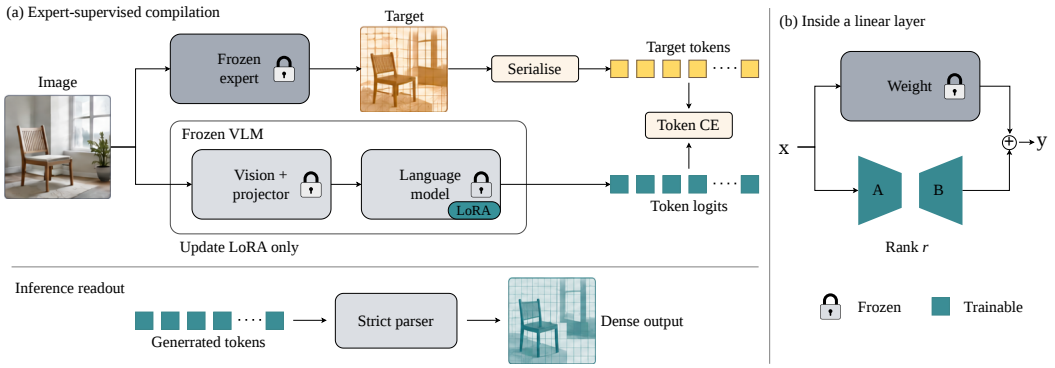

Figure 2: **Distilling an expert target into a detachable adapter.** (a) The expert supervises answer tokens while the base VLM and its visual components remain frozen. (b) LoRA updates the indicated language-model projections, and a strict parser reconstructs the generated answer.

3.1 A TARGET-PRESERVING OUTPUT INTERFACE

Let Y denote the pooled, quantised expert target for image x and task query q. A serialisation $\mathcal{S}_c(Y) = (s^1, \ldots, s^{K_c})$ expresses Y as K_c answer strings using the VLM's existing vocabulary. Exact reconstruction of Y requires retaining every value and its spatial coordinate while changing their grouping or order.

RAW places the entire grid in one row-major answer. CHUNKED partitions an even-sized grid into four interleaved parity phases, each emitted in row-major order:

$$Y^{uv}_{ij} = Y_{2i+u,\,2j+v}, \quad (u,v) \in \{0,1\}^2, \qquad \mathcal{S}_c^{-1}(\mathcal{S}_c(Y)) = Y. \tag{1}$$

Each phase contains one quarter of the cells across the full image; reassembly restores their coordinates (Figure 3). Losslessness refers to the quantised target, not the original expert map.

For a concrete 4×4 example with rows `00 01 02 03`, `04 05 06 07`, `08 09 0A 0B`, and `0C 0D 0E 0F`, the row-major value strings are RAW: `00 01 02 03 / 04 05 06 07 / 08 09 0A 0B / 0C 0D 0E 0F`; the four CHUNKED phase strings are `00 02 / 08 0A`, `01 03 / 09 0B`, `04 06 / 0C 0E`, and `05 07 / 0D 0F`. Here spaces and slashes show cell and row boundaries for readability; the actual answer uses the task template and delimiters specified in Appendix A.

Every call receives the full image and a task–phase prompt $p^k_c(q)$. Phases select target cells, not image crops, and share one adapter per task–format pair. This changes the prediction load and available answer history per call while preserving the complete output's information.

The four answers have no cross-phase context: a later call cannot condition on an earlier phase's generated values. Within each phase, generation is still autoregressive. Reassembly places predictions at their assigned coordinates without interpolation or averaging, so any change in fidelity comes from the predicted values and parsing outcomes. The experiment therefore changes the conditional prediction problem even though its complete supervision target is unchanged.

3.2 LEARNING DETACHABLE PERCEPTION ADAPTERS

We vary capacity through LoRA rank while keeping the VLM backbone and adapted layer family fixed. For a frozen projection $W_\ell \in \mathbb{R}^{d_\ell \times k_\ell}$, LoRA (Hu et al., 2021) learns

$$W'_\ell = W_\ell + \frac{\alpha}{r} B_\ell A_\ell, \qquad \phi = \{A_\ell, B_\ell\}_\ell, \qquad P(r) = r \sum_\ell (d_\ell + k_\ell), \tag{2}$$

where $A_\ell \in \mathbb{R}^{r \times k_\ell}$ and $B_\ell \in \mathbb{R}^{d_\ell \times r}$. $P(r)$ counts trainable parameters across attention and MLP projections. The vision encoder, projector and base language-model weights remain frozen; disabling ϕ recovers the base VLM.

For each image, the chosen codec supplies the supervised answer for each task–phase prompt. Let y be its templated token sequence and $T(y)$ its length. We minimise answer-token cross-entropy on microbatch $\mathcal{B}$:

$$\mathcal{L}_{\mathcal{B}}(\phi) = -\frac{1}{Z_{\mathcal{B}}} \sum_{(x,p,y)\in\mathcal{B}} \sum_{t=1}^{T(y)} \log p_{\theta,\phi}(y_t \mid x, p, y_{<t}), \tag{3}$$

where $Z_{\mathcal{B}} = \sum_{(x,p,y)\in\mathcal{B}} T(y)$ and $p_{\theta,\phi}$ is the next-token distribution with frozen weights θ and adapter ϕ. Image, prompt and padding positions are masked from the loss. Training uses gold answer history; inference must generate that history. This distinction motivates the conditioning diagnostic in Section 5.2.

3.3 Comparing rank gains on complete predictions

Greedy decoding produces each answer in an independent context, followed by reconstruction $\hat{Y} = \mathcal{S}_c^{-1}(\hat{s}^1, \ldots, \hat{s}^{K_c})$. One malformed phase fails the complete prediction and scores zero in the original image denominator; every format is evaluated on the whole target. For N images with teacher agreement $m_i(r, c)$, we report mean fidelity and the fraction of images meeting a task-specific threshold τ:

$$F(r, c) = \frac{1}{N}\sum_{i=1}^{N} m_i(r, c), \qquad F_{\text{tail}}(r, c) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{m_i(r, c) \geq \tau\}. \tag{4}$$

For low and high ranks $r_L < r_H$, the central comparison is

$$I = \big[F(r_H, \textsc{Raw}) - F(r_L, \textsc{Raw})\big] - \big[F(r_H, \textsc{Chunked}) - F(r_L, \textsc{Chunked})\big]. \tag{5}$$

Positive I means that chunking helps low rank more than high rank, even if both improve. We call this *sub-additive interaction* or *substitution*, not synergy. The same contrast applies to F_{tail}. Alongside fidelity, we report adapter parameters $P(r)$, image prefills K_c, answer tokens and measured training time. These quantities distinguish smaller trainable updates from cheaper complete predictions.

3.4 What the interaction identifies

The factorial comparison distinguishes an absolute format gain from a change in the value of rank. Rearranging Equation 5 gives $I = [F(r_L, \textsc{Chunked}) - F(r_L, \textsc{Raw})] - [F(r_H, \textsc{Chunked}) - F(r_H, \textsc{Raw})]$. Thus a positive value requires grouping to help the low-rank adapter more than the high-rank adapter on the same complete output. The fixed target removes the explanation that format helps simply by deleting values; matched image exposure and update counts further focus the comparison. The treatment still includes shorter within-call sequences, phase prompts, four visual prefills and independent answer histories. Section 5 uses controls to probe that bundle. We report image-paired intervals within each trained seed rather than pooling seeds as independent images.

Mean fidelity measures overall expert agreement, while F_{tail} asks whether grouping avoids weak complete maps. A fit-free ordinal companion checks that the depth interaction is not solely an artifact of per-image affine alignment.

4 Where output grouping substitutes for rank

The fixed-target depth comparison exposes a sub-additive rank interaction; the same test on compact masks shows where that pattern stops. These two tasks turn output grouping from an implementation detail into a measurable condition on the value of adapter capacity.

4.1 Experimental setup

Expert targets and serialisations. Both tasks use unlabeled COCO (Lin et al., 2014) training images. Depth Anything V2-Large (Yang et al., 2024) supplies relative depth, which we min–max normalise, mean-pool to 24×24 and quantise to integers in $\{0, \ldots, 127\}$. SAM 2.1 (Ravi

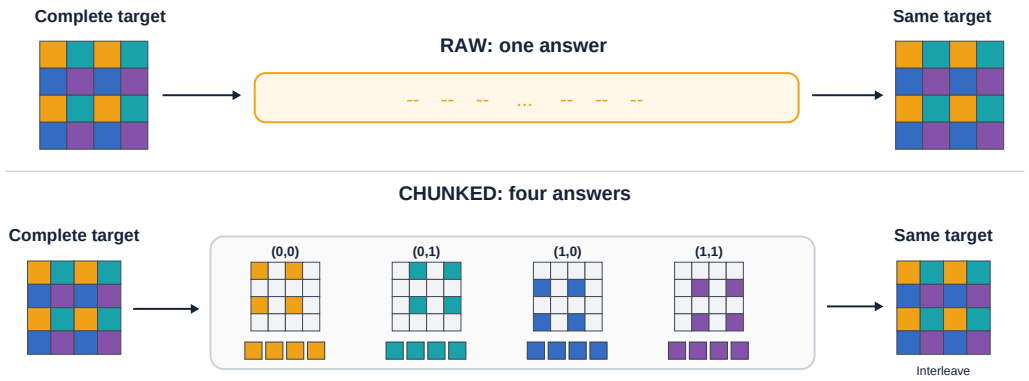

Figure 3: **The same complete target in one or four answers.** RAW emits one raster sequence; CHUNKED emits four interleaved parity phases that reconstruct the identical target. Each phase call sees the complete image.

et al., 2024) supplies point-conditioned masks, majority-pooled to 16×16. Depth values use two hexadecimal digits each; binary masks pack four bits per hexadecimal character. CHUNKED predicts 12×12 depth values or an 8×8 mask per phase, reconstructing the same target as RAW (Figure 3).

Capacity and training. We cross the two formats with ranks 16 and 256 on Qwen3.5-9B, with three seeds per cell. From 60,000 training images, each run presents 30,000 images (half an epoch) in 1,875 updates, using an effective batch of 16 complete images, a cosine learning-rate schedule starting at 10^{-4}, LoRA dropout 0.05, and $\alpha = 2r$. Seven attention/MLP projection families in the language model receive LoRA factors; the vision tower and projector remain frozen. All phases of an image stay together, and image order and update counts match across formats. Supervised depth tokens differ by less than 0.17%. Each chunked answer uses a fresh full-image prefill, so per-call length, available answer history and repeated visual access change together.

Fidelity and confirmation. Depth uses affine δ_1: scale values to $[0, 1]$, fit a per-image least-squares scale and shift with nonnegative scale, then count aligned cells within a factor of 1.25 of the teacher after adding 0.05. Nonpositive shifted predictions count as incorrect. Segmentation uses IoU, with empty teacher masks scoring zero. The tail thresholds in Equation 4 are 0.30 for depth and 0.20 for masks. These metrics measure agreement with pooled expert targets. Strict parsing rejects excess values; sensitivity parsing and a fit-free ordinal depth score provide companion checks (Appendix B.5).

The 1,000-image depth confirmation set was fixed before reading the main outputs. Equation 5 uses $r_L = 16$ and $r_H = 256$; paired image-bootstrap intervals condition on each seed's trained models. Appendix B.1 gives absolute scores and sensitivity checks.

4.2 DEPTH: GROUPING REDUCES THE MARGINAL GAIN FROM RANK

Lossless restructuring reduces the advantage of higher rank on the fixed depth target (Figure 4). Chunking improves rank 16 by $+0.098$ to $+0.111$, compared with $+0.066$ to $+0.074$ at rank 256. The resulting mean-fidelity interaction ranges from $+0.031$ to $+0.037$, with positive intervals in every seed. Mean and tail fidelity show the same direction under both parser readings (Table 1); the ordinal companion agrees.

The interaction is supported by the absolute four-cell pattern, not by a single favourable arm. Raising rank helps both formats, while grouping is most valuable at rank 16. In seed 0, for example, the rank gain falls from $.6383 - .5313 = .1070$ with one answer to $.7122 - .6423 = .0699$ with four. The $+.0371$ difference is the interaction, and the other two seeds show the same ordering. This is why the conclusion concerns the *marginal* benefit of rank rather than a claim that high rank is unnecessary.

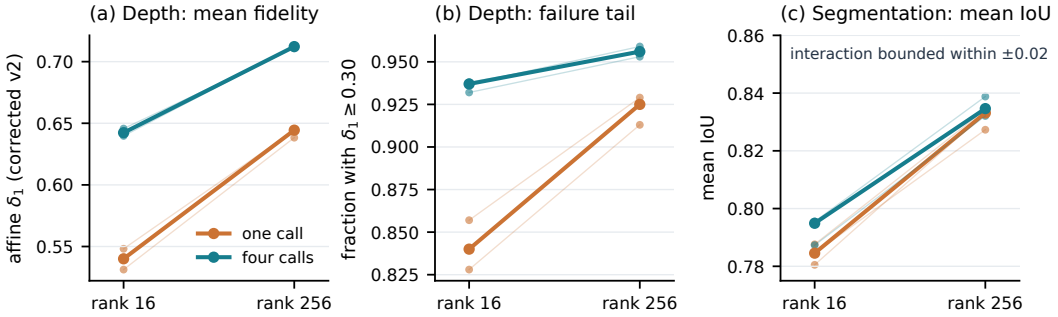

Figure 4: **Depth grouping reduces the gain from higher rank; compact masks bound the interaction.** Faint dots are training seeds and thick lines connect three-seed medians under strict parsing. Orange denotes one answer; teal denotes four. The three panels use different vertical ranges. Depth is evaluated on 1,000 images and masks on 2,798. Table 1 reports the depth interaction and its intervals directly; the corresponding mask estimates appear in Table 2.

Table 1: **The interaction is positive in each training seed.** Strict-parser estimates and 95% paired image-bootstrap intervals. Mean fidelity is affine δ_1; tail fidelity is the fraction of images with $\delta_1 \geq 0.30$. Intervals condition on each trained model; training seeds are not pooled into one confidence interval.

seed	mean-fidelity interaction	tail-fidelity interaction
0	$+0.037$ $[+0.026, +0.048]$	$+0.067$ $[+0.045, +0.090]$
1	$+0.031$ $[+0.020, +0.042]$	$+0.050$ $[+0.030, +0.071]$
2	$+0.033$ $[+0.021, +0.044]$	$+0.064$ $[+0.041, +0.087]$

Higher rank still helps grouped outputs; its smaller marginal gain shows that capacity and serialisation partly substitute for each other. The four-call cost is analysed in Section 6.

4.3 Domain specificity: the segmentation boundary

Increasing rank helps segmentation, but restructuring does not produce the same interaction at 16×16. On 2,798 unseen images, RAW mean IoU improves with rank in every seed. Yet all three mean-IoU interaction intervals lie inside the predefined $(-0.02, +0.02)$ region (Table 2). The tail endpoint meets that bound in two seeds; both parser readings agree exactly.

Table 2: **The depth interaction does not transfer to the tested mask format.** Segmentation RAW rank gaps are positive, but the mean-IoU interaction is bounded within ± 0.02. Intervals are paired image-bootstrap intervals conditional on each training seed.

seed	RAW rank gap	mean-IoU interaction [95% CI]
0	$+0.0427$	-0.0006 $[-0.0100, +0.0087]$
1	$+0.0466$	$+0.0069$ $[-0.0026, +0.0163]$
2	$+0.0526$	$+0.0079$ $[-0.0018, +0.0176]$

The small mask interaction is informative because the task still responds to higher rank. The absolute scores in Appendix B.2 show raw rank gains of .043–.053 across seeds, with grouped gains of similar magnitude. Both formats are well above the image-blind disk score of .298 on the same mask queries. The contrast is therefore a boundary on *how rank and grouping combine*, not a failure of the mask readout or an inability to detect any rank effect.

A packed binary mask occupies just 64 hexadecimal characters. Its bounded interaction is consistent with less per-call prediction load to relieve, although task, target statistics and encoding change together. This domain contrast motivates the following diagnostics of prediction and conditioning demands.

5 WHY DOES GROUPING HELP?

The depth result points to a generation bottleneck: rank 16 improves most when each answer specifies fewer independent spatial values. Coarse-target controls, gold-prefix evaluation and per-image effects show that the gain is connected to both prediction load and free-running answer history.

5.1 INDEPENDENT VALUES MATTER BEYOND ANSWER LENGTH

Two rank-16 controls each reduce the 576-cell target to 144 coarse values. EXPAND12 repeats each coarse value into a full-length raster; POOLED12 emits the short coarse map before deterministic expansion. Both are scored on the original 24×24 teacher target, and their oracle ceiling is 0.873.

Table 3: **Controls separate independently predicted values from answer length.** Strict-parser three-seed median depth fidelity at rank 16. Only RAW and CHUNKED retain all target values. Recorded training wall times are relative to RAW/r16.

arm	free cells/call	tokens/call	calls	fidelity	train wall
RAW	576	1,727	1	0.540	$1.00\times$
CHUNKED	144	431	4	0.642	$1.59\times$
EXPAND12	144	1,727	1	0.634	$1.00\times$
POOLED12	144	431	1	0.648	$0.58\times$

The full-length EXPAND12 answer reaches 0.634, far above RAW/r16 at 0.540. Thus shortening the answer is not necessary for much of the low-rank gain; reducing the number of distinct values per call is consistent with the observed improvement. Because coarse targets also remove detail, the lossless CHUNKED result is the decisive fixed-target evidence.

The ordinal companion shows that answer length still changes some behaviour within the coarse-target controls. Its rank-16 median is .864 for RAW, .903 for full-length EXPAND12, .915 for lossless CHUNKED, and .918 for short POOLED12. The EXPAND12–POOLED12 contrast predicts the same coarse values at different answer lengths, so their separation indicates that generation length can matter even when the primary affine score is near a plateau. The controls jointly support a more precise interpretation: reduced value load is consistent with much of the primary gain, while the way those values are sequenced still influences a fit-free spatial score.

5.2 ANSWER HISTORY CHANGES THE ORDERING

Gold-prefix evaluation reveals that the codecs differ in how they use generated history. On the same 200 development images, the CHUNKED-minus-RAW exact-value accuracy gap is -0.1588 to -0.1575 with correct answer history; under free-running generation it becomes $+0.0110$ to $+0.0175$. The reversal is consistent across all twelve depth adapters (Appendix B.4). Grouping helps complete free generation even though the longer raster can exploit correct preceding values more accurately; this comparison does not isolate prefix errors from the different spatial histories supplied by the two codecs.

Table 4: **Correct prefixes reverse the format ordering.** Three-seed median exact-value accuracy on the same 200 development images. Teacher forcing and free generation use the same target coordinates, but this diagnostic is distinct from affine fidelity.

	gold answer history		free generation	
rank	Raw	Chunked	Raw	Chunked
16	.5444	.3866	.0838	.0963
256	.5685	.4101	.0899	.1068

This reversal offers a sharper reading of the rank–format result than an appeal to shorter outputs alone. With a correct prefix, the long raster contains many nearby teacher values and the model predicts the next value more accurately. Under greedy generation, those values are the model's own outputs; grouping repeatedly begins from a fresh task–phase context. The two conditions therefore

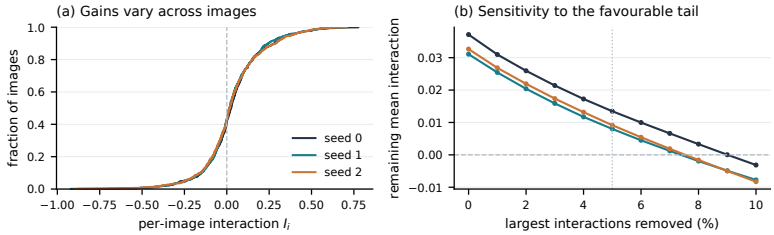

Figure 5: **Depth interaction across images.** Left: empirical per-image distributions for each training seed. Right: remaining mean interaction after removing the largest observed interactions within each seed. The removal is an outcome-selected sensitivity analysis.

rank the same trained adapters differently. The evidence links the benefit to free-running sequence dynamics while leaving repeated visual access and exact spatial conditioning coupled to the format intervention.

5.3 The gain is concentrated where rank 16 fails

Grouping improves the low-fidelity tail as well as the average depth score. At the registered $\delta_1 = 0.30$ threshold, RAW/r16 fails on 143–172 of 1,000 images across seeds, compared with 62– 68 for CHUNKED/r16. The per-image interaction median remains positive in every seed. The practical effect is fewer weak readouts, not merely a small shift in an aggregate metric.

The improvement is paired at the level on which a complete map is used. Table 5 counts the same images before and after grouping at rank 16. Between 88 and 118 images per seed cross the threshold from failure to success, compared with only 8–11 crossing the other way. The net reduction of 80–110 failed maps is more informative for a system that needs every output to clear a minimum fidelity than the mean alone.

Table 5: **Grouping mainly rescues low-fidelity rank-16 maps.** Paired transitions at affine $\delta_1 = 0.30$, strict parsing, 1,000 images per training seed. Each image is counted once in each row.

seed	Raw failures	fail → pass	pass → fail	Chunked failures
0	172	118	8	62
1	143	88	8	63
2	160	103	11	68

The gain is widespread but uneven. Figure 5 shows a positive per-image interaction median in every seed, while the upper tail contributes disproportionately to the mean. Removing each seed's largest 5% observed interactions leaves positive means of +.008 to +.013. This outcome-selected removal is a sensitivity calculation; it supports the direction of the effect while showing that the mean magnitude depends on a subset of strongly helped images. The result therefore combines a common small advantage with a substantial reduction in the failure tail.

6 Cost and interpretation

A smaller adapter is not a cheaper complete prediction. Rank 16 trains 29.1 M adapter parameters, versus 465.6 M at rank 256. But CHUNKED needs four image prefills per complete grid, while RAW needs one; the total depth answer-token count is approximately the same. In recorded runs, training CHUNKED took 1.59× and 1.58× the RAW wall time at ranks 16 and 256. These wall-time ratios include machine changes and restarts. The fidelity gain thus comes with a parameter–prefill trade-off, not a free speedup.

The fixed target makes this trade-off concrete: rank sets trainable parameters, grouping adds image prefills, and both formats emit roughly the same total number of answer tokens. Prefill count is the portable inference-cost descriptor here because recorded decode times vary with hardware and batching.

Table 6: **Complete-target depth operating points.** Three-seed median fidelity; training wall time relative to RAW at the same rank. Wall times are observational, not controlled hardware measurements.

arm	adapter M	prefills	answer tokens	train wall	fidelity
Raw/r16	29.1	1	∼1,728	1.00×	.540
Chunked/r16	29.1	4	∼1,728	1.59×	.642
Raw/r256	465.6	1	∼1,728	1.00×	.644
Chunked/r256	465.6	4	∼1,728	1.58×	.712

Four prefills do not imply four times the latency. The phase answers are shorter and can be batched, whereas a raster has a longer sequential decoding path. Converting this call-count trade-off into latency therefore requires specifying hardware, concurrency and batching for the complete prediction.

The target definition matters to the operating point. When a coarse target suffices, one-call POOLED12 reaches 0.648 at $0.58\times$ recorded training time. Its lower oracle ceiling reflects the loss of fine spatial variation within each 2×2 block. Preserving all 576 target values through CHUNKED incurs the four-call cost. Supplementary costs and baselines appear in Appendix B.

The mask result makes cost target dependent: four calls add 4.4% supervised tokens and recorded training-wall ratios of $5.30\times$ and $5.82\times$, yet its mean-IoU rank interaction stays within $\pm.02$. Grouping is therefore a useful capacity trade-off for this depth readout, without the same evidence of a rank interaction for compact masks.

The choice therefore depends on the resource constraint. If a coarse map is acceptable, POOLED12 offers the strongest observed rank-16 operating point among these controls. If the complete target is required, CHUNKED/r16 and RAW/r256 reach nearby median fidelity with different parameter and prefill requirements. Their proximity is descriptive, not an equivalence test or a cost-matched comparison. Normalising training time within each rank also does not establish which of these two configurations is faster.

7 DISCUSSION AND CONCLUSION

Output grouping changes the marginal value of LoRA rank for the tested depth readout. With all 24×24 values fixed, four phase answers help rank 16 more than rank 256 across three seeds, both in average fidelity and in the failure tail. The fit-free ordinal score corroborates the direction; the compact 16×16 mask experiment bounds the analogous interaction within ± 0.02.

The controls connect the depth gain to fewer independently predicted values per call, while the gold-prefix reversal implicates free-running answer history. Together they make serialisation part of capacity analysis rather than a cosmetic formatting choice. For this backbone, rank pair and training budget, the useful operating point depends on the full target and the cost of complete prediction: grouping improves depth fidelity but requires four image prefills. Downstream value remains a separate question from agreement with the depth expert.

The remaining uncertainty concerns how this interaction changes with training and with the call structure. A half-epoch comparison measures capacity under a fixed adaptation budget; it does not show whether the raster model catches up after more updates. Likewise, comparing two formats leaves per-call length, visual prefills and context resets coupled. A two-call split and four phase blocks within one call would distinguish some of these explanations. Replicating the complete rank–format design on another backbone would test its scope more directly than showing that another model can learn a depth readout. These comparisons remain open.

AI use statement. Generative AI assisted with literature triage, code and drafting; the authors checked the measurements and take responsibility for the paper.

Reproducibility statement. Appendices A and B give methods and results; a separate archive retains protocol records. Code, prompts, pseudo-labels and adapters will be released.

REFERENCES

Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G. Shapiro, and Ranjay Krishna. Perception tokens enhance visual reasoning in multimodal language models. *arXiv preprint arXiv:2412.03548v2*, 2024.

Zhipeng Cai, Ching-Feng Yeh, Hu Xu, Zhuang Liu, Gregory Meyer, Xinjie Lei, Changsheng Zhao, Shang-Wen Li, Vikas Chandra, and Yangyang Shi. Depthlm: Metric depth from vision language models. *arXiv preprint arXiv:2509.25413v2*, 2025.

Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2109.10852v2.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685v2*, 2021.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692v3*, 2023.

Mengcheng Lan, Chaofeng Chen, Yue Zhou, Jiaxing Xu, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Text4seg: Reimagining image segmentation as text generation. *arXiv preprint arXiv:2410.09855v2*, 2024.

Mengcheng Lan, Chaofeng Chen, Jiaxing Xu, Zongrui Li, Yiping Ke, Xudong Jiang, Yingchen Yu, Yunqing Zhao, and Song Bai. Text4Seg++: Advancing image segmentation via generative language modeling. *arXiv preprint arXiv:2509.06321v1*, 2025. URL https://arxiv.org/abs/2509.06321v1.

Yuchen Li, Amanmeet Garg, Shalini Chaudhuri, Rui Zhao, and Garin Kessler. Perceptio: Perception enhanced vision language models via spatial token generation. *arXiv preprint arXiv:2603.18795v1*, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. *arXiv preprint arXiv:1405.0312v3*, 2014.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/2c601ad9d2ff9bc8b282670cdd54f69f-Abstract.html.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714v2*, 2024.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c2e5eaae9152079b9e95845750bb9ab-Abstract.html.

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

A EXPERIMENTAL DETAILS

A.1 TARGETS, POPULATIONS AND RECONSTRUCTION

The main depth comparison uses unlabeled COCO images and relative-depth maps from frozen Depth Anything V2-Large. Each expert map is min–max normalised per image, mean-pooled to 24×24, and quantised to integers 0–127, with larger values denoting locations nearer the camera. The mask replication uses point-conditioned SAM 2.1 targets majority-pooled to 16×16; the same point query is used for every format of an image. The depth confirmation set has 1,000 held-out images and the mask set has 2,798. These populations are distinct from the original 500-image depth fidelity ledger and the 200-image depth development diagnostic.

Both lossless codecs retain the stored quantised target. RAW lists all cells in row-major order. CHUNKED uses phases $(0,0), (0,1), (1,0), (1,1)$ in that order, selecting cells by Equation 1. Depth answers have spaces between two-digit lowercase hexadecimal values and newlines between rows. Mask answers pack the row-major bits into hexadecimal characters, most significant bit first, without separators. Reconstruction interleaves the four decoded arrays at their original coordinates. There is no interpolation, smoothing or cross-phase correction.

The coarse controls use the same stored 24×24 depth target. EXPAND12 averages each nonoverlapping 2×2 block, rounds its mean to the nearest integer using the implementation's round-to-even rule, and repeats that integer over all four cells. POOLED12 emits the same 12×12 coarse integers once; nearest-neighbour expansion restores a 24×24 grid for scoring. Thus these two controls share recoverable target information but differ in emitted length. Their oracle scores are obtained by applying this pooling and expansion to the teacher target itself, without a model.

A.2 OPTIMISATION AND IMAGE EXPOSURE

The depth and mask factorial designs each contain four format–rank cells and seeds 0, 1 and 2. Every run uses 30,000 of 60,000 available training images under the half-epoch recipe. Table 7 states how microbatches and gradient accumulation preserve complete-image exposure. Sampling permutes image blocks with the same seeded permutation in both formats. A RAW microbatch contains four complete images, whereas a CHUNKED microbatch contains the four phases of one image. Both accumulation windows contain the same 16 images and every target cell is supervised once per image presentation.

Table 7: **Matched training quantities and changed call counts.** The configuration applies to both factorial tasks and both ranks. Rows are separate image–prompt–answer examples.

quantity	Raw	Chunked
available image blocks	60,000	60,000
rows per image block	1	4
rows per microbatch	4	4
accumulated microbatches per update	4	16
complete images per update	16	16
microbatches per run	7,500	30,000
optimiser updates	1,875	1,875
complete-image presentations	30,000	30,000
image prefills during training	30,000	120,000

AdamW uses learning rate 10^{-4}, zero weight decay and gradient norm clipping at 1.0. The cosine schedule advances once per optimiser update over 1,875 updates, with no warm-up. Training uses bfloat16 model weights and gradient checkpointing. LoRA uses $\alpha = 2r$ and dropout 0.05 on the language model's `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj` and `down_proj` linear projections where present. The vision tower, projector and base language weights are frozen. Equation 3 is normalised within each microbatch; its gradient is divided by the accumulation count before the optimiser step. Padding, image and prompt positions carry no training loss. Depth answer payloads contain 1,727 raster tokens or $4 \times 431 = 1,724$ phase tokens; chat-template boundary tokens are additional. Mask tokenisation depends on the bit pattern, and the measured four-phase total contains 4.4% more supervised tokens.

A.3 PROMPTS AND DECODING

Each call supplies the full image followed by the corresponding user prompt below. Braced fields are substituted with the query or phase coordinates. Line wrapping here is typographical; each prompt is a single text string. The generation prompt uses the same chat template as training and is checked to be a token-exact prefix of the full supervised sequence.

Depth, Raw.

> Output the 24x24 relative depth grid of this image: 24 rows, each with 24 two-digit lowercase hex values in [00,7f] separated by spaces, where 7f is nearest to the camera. Output the grid only.

Depth, Chunked.

> Output part of the 24x24 relative depth grid of this image: the values at every other row and every other column, starting from row {u} and column {v} (0-indexed). Give 12 rows, each with 12 two-digit lowercase hex values in [00,7f] separated by spaces, where 7f is nearest to the camera. Output the grid only.

Segmentation, Raw.

> Segment the object at point (x={x}, y={y}) on a 0-999 grid of this image. Output the mask as a 16x16 binary grid packed as 64 lowercase hex characters, row-major, most significant bit first. Output the hex string only.

Segmentation, Chunked.

> Segment the object at point (x={x}, y={y}) on a 0-999 grid of this image. Output part of the 16x16 mask: the cells at every other row and every other column, starting from row {u} and column {v} (0-indexed). Give the resulting 8x8 binary grid packed as 16 lowercase hex characters, row-major, most significant bit first. Output the hex string only.

EXPAND12 uses the RAW depth prompt. POOLED12 uses that prompt with both grid dimensions and both row/value counts changed from 24 to 12. Greedy decoding has no sampling. Maximum new-token budgets are 2,400 for RAW/EXPAND12 depth, 900 for each CHUNKED depth phase or POOLED12 answer, 120 for the full mask and 30 for each mask phase. Generation stops at the model's end-of-turn or end-of-sequence token, or at the cap. Independent phase calls share adapter weights but no generated answer history.

A.4 PARSING, METRICS AND UNCERTAINTY

Complete-output parsing. The primary depth parser extracts two-digit hexadecimal values at word boundaries and requires exactly 576 values for a raster or 144 per phase, all in $[0, 127]$. Its sensitivity reading truncates excess values. The primary mask parser requires exactly one maximal hexadecimal run of length 64 for a raster or 16 per phase; the sensitivity reading searches for a run of the requested length. Missing, short, illegal or over-length primary answers score zero. If any phase fails, the complete image scores zero on every endpoint and stays in the denominator.

Affine depth agreement. Let p_j and z_j be predicted and teacher cell values divided by 127, with $M = 576$ cells. We align the prediction using

$$(a^*, b^*) = \arg\min_{a \geq 0,\, b} \sum_{j=1}^{M} (a p_j + b - z_j)^2, \qquad \widetilde{p}_j = a^* p_j + b^*. \tag{6}$$

A constant prediction or a negative unconstrained fitted scale gives the teacher-mean constant fit. Define $\rho_j = \infty$ when $\widetilde{p}_j + \epsilon \leq 0$, and otherwise $\rho_j = \max\{(\widetilde{p}_j + \epsilon)/(z_j + \epsilon), (z_j + \epsilon)/(\widetilde{p}_j + \epsilon)\}$, where $\epsilon = 0.05$. The image score is

$$m_{\text{depth}} = \frac{1}{M} \sum_{j=1}^{M} \mathbf{1}\{\rho_j < 1.25\}. \tag{7}$$

Nonfinite ratios also fail. This corrected-v2 score enforces positive shifted values and a nonnegative fitted scale. It measures relative agreement with the quantised teacher; the fitted scale and shift remove global calibration differences.

Fit-free ordinal companion. We sample 8,192 distinct unordered pairs of the 576 cell coordinates once with NumPy random seed 20260906, sharing the pairs across formats and training seeds. For each image we retain pairs whose normalised teacher values differ by more than 0.05. A matching predicted order scores 1, a predicted tie scores $1/2$, and a reversed order scores 0. The image score is the mean over retained pairs, followed by a mean over images including parse failures. This metric uses neither affine fitting nor a ratio test.

Masks, tail endpoints and intervals. Mask fidelity is per-image intersection over union, with zero assigned to an empty teacher mask. Consequently, perfect target reconstruction has ceiling $1 - 15/2798 = 0.994639$ on the mask confirmation set. Tail fidelity counts images with depth agreement at least 0.30 or mask IoU at least 0.20. For each seed, bootstrap resampling draws images with replacement and retains the four arm scores for every sampled image together. We compute F and I on each of 10,000 resamples and report the 2.5th and 97.5th percentiles. Intervals quantify image sampling for fixed trained adapters; training seeds are reported separately. The segmentation equivalence region is $(-0.02, +0.02)$ in absolute score units: containing zero alone is insufficient to bound the interaction within that region.

B Supplementary results

B.1 Depth interaction and absolute scores

The complete-target depth comparison shows a positive mean interaction in all three seeds. Table 8 includes all four arm scores so the effect can be checked from absolute performance as well as the contrast.

Table 8: Strict-parser depth mean fidelity on the 1,000-image confirmation set. Each row is one training seed; I is the difference of rank gains and the interval is a conditional paired image bootstrap.

seed	Raw/16	Raw/256	Chunked/16	Chunked/256	I [95% CI]
0	.5313	.6383	.6423	.7122	+.0371 [.0260,.0485]
1	.5482	.6444	.6457	.7108	+.0311 [.0200,.0422]
2	.5400	.6452	.6397	.7123	+.0327 [.0210,.0441]

The tail endpoint also supports the interaction. Tail-interaction estimates are +.067 [.045,.090], +.050 [.030,.071] and +.064 [.041,.087] for seeds 0–2. The fit-free ordinal companion yields interactions of $+0.021\,[+0.014, +0.028]$, $+0.021\,[+0.015, +0.028]$ and $+0.026\,[+0.019, +0.034]$ for seeds 0–2, respectively. A parser that truncates excess values changes only one depth factorial cell's mean estimate, by +.00064, and does not change the interaction conclusion. Removing each seed's largest 5% observed interactions leaves positive means of +.008 to +.013; this outcome-selected sensitivity is descriptive rather than an independent test.

B.2 Segmentation boundary

The mask comparison has a rank gain in both formats, but its mean-IoU interaction intervals all lie within $\pm.02$ (Table 9). The two parser readings agree on every factorial prediction. Because 15 teacher masks are empty and receive zero by convention, exact reconstruction has mean-IoU ceiling .994639. The same-set image-blind disk control scores .2980. These reference values should not be carried to the separate depth or fidelity populations.

The mask tail interactions are $-0.0004\,[-0.0118, +0.0107]$, $+0.0139\,[+0.0018, +0.0257]$ and $+0.0064\,[-0.0050, +0.0179]$ for seeds 0–2. The seed-1 upper endpoint exceeds $+0.02$, so the tail intervals lie wholly inside the equivalence region in two seeds, compared with all three for mean IoU. A bounded mean interaction does not require the formats to have identical absolute fidelity.

Table 9: Strict-parser mask mean IoU on 2,798 confirmation images. Intervals condition on each trained seed.

seed	Raw/16	Raw/256	Chunked/16	Chunked/256	I [95% CI]
0	.7846	.8273	.7955	.8388	-.0006 [-.0100,.0087]
1	.7876	.8342	.7949	.8346	+.0069 [-.0026,.0163]
2	.7805	.8331	.7873	.8321	+.0079 [-.0018,.0176]

B.3 MECHANISM AND IMAGE-LEVEL DIAGNOSTICS

The rank-16 coarse controls test whether answer length alone explains the gain. EXPAND12 predicts 144 coarse values but repeats each into a 576-cell answer, whereas POOLED12 emits the 144 values once. Their three-seed median primary scores are .634 and .648; RAW/r16 and CHUNKED/r16 score .540 and .642. Both coarse controls have an oracle ceiling of .873 rather than 1, and smoothing changes the target. They support per-call prediction load as a candidate, not an identified mechanism. The comparison of POOLED12 with EXPAND12 also shows an ordinal length effect in two seeds.

The interaction is uneven across images. Per-image interaction medians are +.0243, +.0156 and +.0156 across seeds 0–2, below their corresponding means. At the 0.30 fidelity threshold, RAW/r16 fails on 172, 143 and 160 images, while CHUNKED/r16 fails on 62, 63 and 68. The paired transitions are informative: for seed 0, 118 images move from fail to pass and 8 move from pass to fail, out of 1,000. This asymmetry explains why the tail endpoint changes alongside the mean. Figure 5 shows the full distributions.

B.4 GOLD-PREFIX READOUT

Teacher forcing changes the ordering between codecs. On 200 development images, all twelve trained depth adapters are evaluated with gold answer history and compared with their stored free-running predictions using exact-value accuracy, not affine fidelity. For rank 16, the gold-prefix accuracy is .544–.545 for RAW and .386–.387 for CHUNKED; under free generation it is .082–.086 and .096–.097, respectively. At rank 256 the same reversal occurs. Correct prefixes provide different neighbouring target values to the two codecs, so this diagnostic does not isolate error accumulation from spatial conditioning or context reset.

Exact-value measurement. For each two-digit teacher value, correctness requires every token overlapping either digit to have the correct argmax prediction under its own gold prefix. These token decisions are combined at the value level; they are not a joint greedy prediction of the entire map. All phase values are returned to the same 576 coordinates before aggregation. Free-running accuracy uses strict parsing of the saved development answers, with failed complete maps scoring zero. Table 10 includes all six rank–seed pairs and mean negative log-likelihood (NLL) over value tokens. No retraining or new free-running decoding is used for this comparison.

Table 10: **Complete gold-prefix diagnostic.** Exact-value accuracy on 200 development images, with all three seeds shown. TF denotes teacher forcing. NLL is in nats per value token; lower is better. Accuracy uses neither affine fitting nor a tolerance band.

		Raw			Chunked		
rank	seed	TF acc.	free acc.	NLL	TF acc.	free acc.	NLL
16	0	0.5442	0.0822	0.6332	0.3866	0.0962	0.9179
16	1	0.5444	0.0838	0.6332	0.3865	0.0963	0.9180
16	2	0.5450	0.0856	0.6331	0.3870	0.0967	0.9173
256	0	0.5687	0.0897	0.5821	0.4101	0.1066	0.8539
256	1	0.5683	0.0917	0.5822	0.4108	0.1068	0.8539
256	2	0.5685	0.0899	0.5832	0.4097	0.1074	0.8536

B.5 METRIC AND DEPLOYMENT CHECKS

The affine depth score removes global scale and shift differences, so the main claim is also checked with a fit-free ordinal score. The original fidelity ledger used a different depth metric implementation and a separate 500-image population; it must not be numerically pooled with the confirmation set. On the separate fidelity set, a rank-16 depth adapter scores .540 versus an image-blind mean-grid score .404 and a four-shot prompt score .222. On NYUv2, the rank-16 adapter obtains median sparse Spearman correlation .889 with sensor depth, versus .952 for the expert and .579 for a vertical prior; these are eight queries per image, not dense metric-depth estimates.

Other deployment probes further limit interpretation. An attached adapter can interfere with general VLM answers; self-rehearsal reduces the measured interference while lowering depth fidelity. In the tested question-answering pipeline, generated-map input scores 67.2% on BLINK and 80.9% on CV-Bench, compared with 68.4% and 86.5% for supplied expert-map input. These are distinct populations and pipelines from the rank–format study; the comparisons do not establish that changing serialisation improves downstream reasoning.

C PROTOCOL ARCHIVE AND REPORTING BOUNDARY

The separate `supplementary_protocol_archive.zip` preserves the source manuscript's full run tables, dated registrations, amendments and execution notes. It is supplied as an audit record rather than interleaved with the scientific argument. The main text and the two appendix sections above use one ordinary section/table numbering system. The archive includes superseded measurements and run failures, so its historical labels should not be read as additional independent confirmation. The central scientific boundary is unchanged: the positive interaction is established for the tested depth teacher, Qwen3.5-9B, two ranks and half-epoch schedule, with four-call inference as part of the treatment.