MORE THAN SEEING: VISION AS SCALABLE ANCHORS IN NATIVE MULTIMODAL PRE-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Native multimodal pre-training trains a single model from scratch on text and multimodal data, enabling it to understand both language and visual inputs. However, how visual signals shape scaling behavior and support capabilities beyond visual understanding remains unclear. We investigate this question through controlled visual-content ablations across model sizes and training budgets. Specifically, we replace image embeddings with content-free padding token embeddings. Retaining visual content improves text prediction on image-text examples without materially changing the lowest text-only loss at a given compute budget, the optimal balance of model size and training data, or aggregate language scores. The relative multimodal loss benefit grows approximately as a power law of compute. To test benefits beyond understanding images, we evaluate spatial reasoning from text descriptions alone. Both conditions improve with model and data scaling. During training of the largest model, visual gains are most sustained in localization, navigation, and symbolic planning, which involve maintaining object configurations or composing spatial updates; gains in spatial relations, geometry, and physical commonsense are less consistent. These findings identify visual content as a scalable anchor for multimodal learning, with benefits that extend beyond visual inputs while preserving aggregate language performance.

1 INTRODUCTION

Native multimodal pre-training develops visual and language capabilities in a single model trained from scratch on text and multimodal data (Shukor et al., 2025b; Cui et al., 2025; Wu et al., 2026). Considering that images convey information about objects, scenes, and spatial relationships, their contribution may extend beyond tasks that require visual input. This raises two related questions about how visual content affects scaling in native multimodal pre-training and whether it improves capabilities evaluated entirely through text.

Answering these questions requires isolating visual content from other changes introduced by multimodal training. Scaling laws describe how training loss depends on model size, data volume, and training compute, guiding how to allocate training resources between parameters and tokens (Shukor et al., 2025b; Wu et al., 2026). In native multimodal models, these relationships reflect learning from both text and multimodal data. However, changing the multimodal data mixture can also alter language supervision and resource allocation, preventing direct attribution of differences in loss or downstream performance to visual content. Isolating its contribution therefore requires varying visual content while holding language supervision and the training configuration constant.

We conduct this controlled comparison across model sizes and training compute budgets. At each scale, we train matched models using either image patch embeddings or content-free padding, holding the architecture, training text, and training settings constant. We analyze loss-compute frontiers and compute-optimal allocations separately for the text and multimodal objectives (Wu et al., 2026). To quantify the contribution of visual content to multimodal understanding, we define the visual anchor gap as the relative increase in text prediction loss under the padding condition compared with the image condition. We then examine how this gap scales with multimodal compute.

Our analyses reveal complementary scaling patterns. Visual content leaves the text loss frontier and compute-optimal allocations nearly unchanged. Meanwhile, it reduces multimodal loss, and the visual anchor gap follows an approximate power law of multimodal compute. These findings

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.

suggest that the relative contribution of visual content to multimodal text prediction increases with compute, while pure-text scaling remains largely unchanged.

A growing visual anchor gap, however, does not establish a benefit when visual input is absent. We therefore evaluate downstream capabilities using text-only inputs, both across model sizes and throughout training of the largest model. General language benchmarks test whether the stable pure-text loss pattern is accompanied by comparable language performance. Spatial tasks provide a targeted test of benefits from visual learning because images encode spatial relationships that can also be described and queried through language.

The image and padding conditions achieve similar aggregate performance on general language benchmarks across model sizes and training compute budgets. Under both conditions, spatial reasoning performance improves with model size and additional training data. During training of the largest model, visual gains are most sustained in localization, navigation, and symbolic planning, which involve maintaining object configurations or composing spatial updates. Gains in spatial relations, geometry, and physical commonsense are less consistent, suggesting selective transfer rather than a uniform spatial advantage. Long-plan success also emerges at larger scales, although execution constraints remain a bottleneck (Section 5.4). These gains accompany similar aggregate general-language performance. Our main contributions can be summarized as follows:

- We isolate the contribution of visual content through matched image and padding conditions across model sizes and training compute budgets, holding language supervision and the training configuration constant.
- We identify closely matched pure-text loss frontiers and compute-optimal allocations, alongside a multimodal visual anchor gap that grows with compute under multiple allocation rules.
- We examine capability development under text-only evaluation, showing preserved aggregate language performance and selective spatial transfer, with sustained gains in localization, navigation, and symbolic planning during A3B training.

2 PRELIMINARIES

2.1 NATIVE MULTIMODAL PRE-TRAINING

Native multimodal pre-training trains a single model from scratch using interleaved multimodal data (Shukor et al., 2025b; Cui et al., 2025; Wu et al., 2026). This approach enables both modalities to share parameter capacity and integrate fundamentally throughout training rather than being aligned post hoc. Specifically, images are mapped into continuous patch embeddings and interleaved with text tokens within a single decoder-only architecture. The model is trained autoregressively with a next-token prediction objective, with loss evaluated only at text tokens (Shukor et al., 2025b; Wu et al., 2026). Under a fixed compute budget, native multimodal pre-training inherits the trade-off between model size and training data established for LLMs (Kaplan et al., 2020; Hoffmann et al., 2022; Wu et al., 2026). We use the compute-optimal framework and estimators developed for native multimodal pre-training (Shukor et al., 2025b; Wu et al., 2026) to analyze the two streams under matched training conditions.

2.2 COMPUTE-OPTIMAL ALLOCATION

Given a fixed compute budget C, we investigate the optimal allocation between model size N and training token count D. Here N counts activated non-embedding parameters per token, and D counts all input tokens. Data composition is parameterized by $r = D_{\text{mm}}/D_{\text{text}}$, where multimodal tokens include both image-patch embeddings and text. Using the standard approximation $C = 6ND$, compute accounting is

$$D = D_{\text{text}} + D_{\text{mm}} = (1+r)D_{\text{text}},\ C = 6ND = C_{\text{text}} + C_{\text{mm}},\ C_s = 6ND_s,\ s \in \{\text{text}, \text{mm}\}. \tag{1}$$

Let $L_s^q(N, D_s; r)$ denote the smoothed training loss under condition $q \in \{\text{img}, \text{pad}\}$. The two conditions retain or ablate visual content, respectively. Both source losses are evaluated at text tokens and optimized using shared parameters. When training examples are rarely revisited, smoothed training loss serves as a practical proxy for test loss (Kaplan et al., 2020; Hoffmann et al., 2022). We

use it as the loss measure for our scaling analysis. We estimate the preferred allocation separately for each objective using the same jointly trained models (Wu et al., 2026). At fixed (s, q, r), the ideal allocation is

$$\big(N_s^{q,*}(C_s; r), D_s^{q,*}(C_s; r)\big) \in \underset{N, D_s:\, 6ND_s = C_s}{\arg\min} L_s^q(N, D_s; r). \tag{2}$$

Its loss defines the ideal frontier $L_s^{q,*}(C_s; r)$. The definitions below allow general r. Our matched IMG/PAD study fixes $r = 0.3$, and we suppress this fixed argument in Section 3.

2.3 Estimators of the Compute-Optimal Frontier

IsoFLOP profiles provide our primary allocation estimates while the training-curve envelope provides a sensitivity check using the same runs. Let $e \in \{\mathrm{iso}, \mathrm{env}\}$ index the estimator. Hats denote finite-data estimates, in contrast to the ideal optima in Equation (2). We summarize allocations by fitted power laws,

$$\widehat{N}_{s,e}^q(C_s; r) \propto C_s^{a_{s,e}^q(r)},\ \widehat{D}_{s,e}^q(C_s; r) \propto C_s^{b_{s,e}^q(r)},\ a_{s,e}^q(r) + b_{s,e}^q(r) = 1, \tag{3}$$

The exponent constraint follows from $\widehat{D} = C_s/(6\widehat{N})$ when the two log-log regressions use identical budgets and weights. Budget selection and estimator details appear in Sections D.1 and D.2.

IsoFLOP profiles. At a fixed budget C_s, we read each model's loss in log compute, obtaining $\widetilde{L}_s^q(N, C_s; r)$, and fit a quadratic in $\log N$. A convex fit's minimum, clipped to the observed size range, estimates $\widehat{N}_{s,\mathrm{iso}}^q$. The token allocation follows algebraically. Regressing both allocations against compute in log space gives $a_{s,\mathrm{iso}}^q(r)$ and $b_{s,\mathrm{iso}}^q(r)$.

Training-curve envelope. Increasing D_s at fixed N traces a loss-compute trajectory. At each supported budget, the lowest interpolated loss across models defines the equal-FLOPs envelope $\widehat{L}_{s,\mathrm{env}}^q(C_s; r)$. Each envelope point retains its model size and source-token count. Regressing these selected allocations gives the second set of allocation exponents. Agreement between the estimators is a sensitivity check, not independent validation.

The compute frontier. We fit the equal-FLOPs envelope with

$$\mathcal{F}_s^q(C_s; r) = E_s^q(r) + \left(\frac{C_{c,s}^q(r)}{C_s}\right)^{\beta_s^q(r)},\ C_{c,s}^q(r) > 0, \quad \beta_s^q(r) > 0. \tag{4}$$

Here $E_s^q(r)$ is a fitted loss offset, $C_{c,s}^q(r)$ a characteristic compute scale, and $\beta_s^q(r)$ a loss-decay exponent distinct from the token-allocation exponent $b_{s,e}^q(r)$. The curve $\mathcal{F}_s^q$ describes the observed compute range and is distinct from the discrete envelope and ideal frontier. Section D.3 gives the fitting procedure.

2.4 Experimental Setup

Training data. The nominal training corpus contains 250B text-source tokens from web pages, books, academic papers, and other domains. The multimodal corpus comprises web-crawled image-text pairs and interleaved image-text documents. With images represented as continuous patch embeddings, it contains 75B input tokens, including image positions and associated text. Thus, the full-training budget is $D = 325\mathrm{B}$ with $r = 0.3$; both conditions use the same source budgets.

Model architecture. We use a decoder-only MoE Transformer with a single patch-embedding layer that maps images directly into continuous embeddings without a pre-trained vision encoder. The models use auxiliary-loss-free MoE training (Liu et al., 2024). The scaling ladder contains six activated non-embedding parameter scales, $N \in \{71\mathrm{M}, 128\mathrm{M}, 340\mathrm{M}, 590\mathrm{M}, 874\mathrm{M}, 3\mathrm{B}\}$, denoted A71M through A3B. Architecture and optimization details are provided in Section B.

3 Vision as a Scalable Anchor

Visual content can shape native multimodal learning in two ways. It can provide context for image-text examples and influence shared model parameters that also support language learning. Using

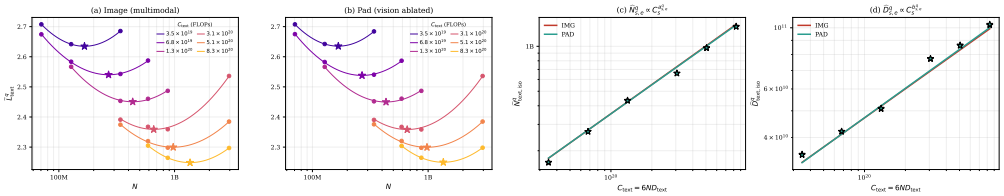

Figure 1: **IsoFLOP curves (language objective).** (a,b) At each target budget C_{text}, we read the $\widetilde{L}^{\text{pad}}_{\text{text}}$ and $\widetilde{L}^{\text{img}}_{\text{text}}$ across various model size at the token count satisfying $6ND_{\text{text}} = C_{\text{text}}$ and plot it against $\log N$. (c,d) Optimal model size $\widehat{N}^{q}_{\text{text,iso}}$ and optimal token count $\widehat{D}^{q}_{\text{text,iso}}$ as a function of the compute budget C_{text}.

a matched visual-content ablation across scales, we investigate whether vision changes the pure-text learning frontier. We also examine whether its contribution to multimodal learning strengthens as compute increases. The pure-text frontier remains nearly unchanged, whereas the relative loss advantage from visual content grows with multimodal compute. We use visual anchoring to describe this contribution to multimodal understanding.

3.1 Isolating the Contribution of Visual Content

A matched visual-content ablation. Removing image-text examples would change both the visual inputs and the language supervision, making their effects difficult to separate. Instead, we retain the examples and replace only their visual content. At each of the six model sizes in Section 2, we train a matched pair of models. The IMG model receives image patch embeddings, whereas PAD replaces each image embedding with a fixed, content-free pad-token embedding. The two models share the same architecture, training text, data order, token budget, and optimization settings. The replacement preserves sequence length, positional structure, and the text-loss mask. Thus, the two conditions differ in access to visual content while retaining the same amount of multimodal data.

Both models are trained from scratch, with losses evaluated only on text tokens. Their comparison therefore captures the combined effect of learning from visual content and using it as context during pre-training. We analyze the text-only and multimodal streams separately, keeping the mixture fixed at $r = 0.3$. As illustrated in Section 2, $C_s = 6ND_s$ denotes compute attributed to stream s within the jointly trained model. Here, D_{mm} includes both image and text tokens.

Measuring visual anchoring. For stream $s \in \{\text{text}, \text{mm}\}$, the anchor gap at allocation (N, D_s) is the relative loss increase

$$\Delta^{\text{anc}}_{s}(N, D_s) = \frac{L^{\text{pad}}_{s}(N, D_s) - L^{\text{img}}_{s}(N, D_s)}{L^{\text{img}}_{s}(N, D_s)}. \tag{5}$$

A positive gap means that visual content lowers the loss. A gap near zero means that the two conditions attain similar losses. To compare allocations at fixed compute, we interpolate each condition's loss in log compute, following Section 2, and obtain the paired relative gap

$$\widetilde{\Delta}^{\text{anc}}_{s}(N, C_s) = \frac{\widetilde{L}^{\text{pad}}_{s}(N, C_s) - \widetilde{L}^{\text{img}}_{s}(N, C_s)}{\widetilde{L}^{\text{img}}_{s}(N, C_s)}. \tag{6}$$

We denote the model sizes supported by both conditions at budget C_s by $\mathcal{N}_s(C_s)$. Gaps are expressed as fractions in the equations and percentages in the figures. This paired measure complements the frontier analysis by comparing the same allocation across conditions, whereas each condition's frontier selects its own loss-minimizing allocation.

3.2 Visual Content Preserves Text Scaling

As shown in Figure 1, the text IsoFLOP profiles closely align under IMG and PAD at matched compute, including the locations of their minima. The training-curve envelopes show the same pattern. As illustrated in Figure 2, their fitted frontiers nearly overlap, with the same decay exponent $\beta_{\text{text}} = 0.198$ and closely matched offsets, 1.829 and 1.831. Thus, visual-content ablation produces

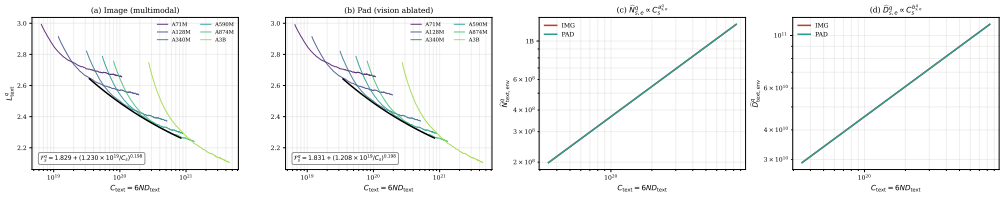

Figure 2: **Training curve envelope (language objective).** (a,b) Training curves span the evaluated model sizes. The lower envelope gives the lowest observed loss at each budget C_{text}. The fitted loss frontier summarizes this envelope. (c,d) Optimal model size $\widehat{N}^q_{\text{text,env}}$ and optimal token count $\widehat{D}^q_{\text{text,env}}$ as a function of the compute budget C_{text}.

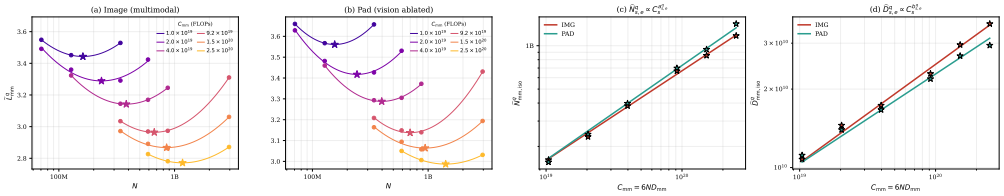

Figure 3: **IsoFLOP curves (multimodal objective).** (a,b) At each target budget C_{mm}, we read the $\widetilde{L}^{\text{pad}}_{\text{mm}}$ and $\widetilde{L}^{\text{img}}_{\text{mm}}$ across various model size at the token count satisfying $6ND_{\text{mm}} = C_{\text{mm}}$ and plot it against $\log N$. (c,d) Optimal model size $\widehat{N}^q_{\text{mm,iso}}$ and optimal token count $\widehat{D}^q_{\text{mm,iso}}$ as a function of the compute budget C_{mm}.

little change in the observed relationship between L^q_{text} and C_{text}. This stability extends to the estimated model-data allocation. The IsoFLOP parameter exponents are 0.649 for IMG and 0.645 for PAD. The envelope estimator gives 0.578 for both. Although the absolute exponents depend on the estimator, the difference between conditions is small within each estimator. Because both estimators reuse the same runs, we interpret this agreement as a sensitivity check on the allocation analysis. The paired anchor gaps provide a direct check that this pattern is not solely a consequence of fitting frontiers. As shown in Figure 5a, $\widetilde{\Delta}^{\text{anc}}_{\text{text}}$ ranges from -0.095% to $+0.128\%$, with a mean of -0.006%. These near-zero differences support preservation of the measured text objective under the fixed mixture and compute budgets. We examine whether this pattern extends to downstream language capabilities separately in Section 4.

3.3 Visual Content Lowers Multimodal Loss

As shown in Figure 3, matched IMG models attain lower loss than PAD models on the text tokens in image-text examples at every evaluated budget. As illustrated in Figure 4, this advantage persists when each condition selects its own allocation. The fitted IMG frontier lies below the PAD frontier throughout $C_{\text{mm}} \in [1.042 \times 10^{19}, 2.504 \times 10^{20}]$ FLOPs. Visual content therefore lowers multimodal loss both at matched allocations and along the estimated loss frontiers.

3.4 Visual Anchoring Strengthens with Compute

A lower multimodal frontier establishes that visual content helps, but it does not show whether its relative contribution persists as the model and data scale. We therefore return to the anchor gap and track its dependence on multimodal compute. As shown in Figure 5, the gaps increase overall while fixed-compute profiles retain a dependence on model size. To characterize the compute trend, we use three complementary summaries.

Separating compute growth from allocation choice. At each budget, $G^{\max}_s$ is the largest paired gap among the supported model sizes $\mathcal{N}_s(C_s)$. A second summary, G^{quad}_s, is the maximum of a quadratic fit to the paired gaps in $\log N$ over the supported size interval. We find this maximum by evaluating both endpoints and any interior stationary point. These summaries describe where the relative effect of ablation is largest. To ask whether the gap also grows along an allocation chosen

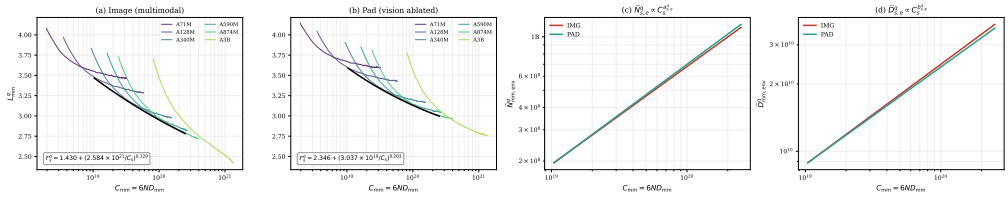

Figure 4: **Training curve envelope (multimodal objective).** (a,b) Training curves span the evaluated model sizes. The lower envelope gives the lowest observed loss at each budget C_{mm}. The fitted loss frontier summarizes this envelope. (c,d) Optimal model size $\widehat{N}^q_{\mathrm{mm,env}}$ and optimal token count $\widehat{D}^q_{\mathrm{mm,env}}$ as a function of the compute budget C_{mm}.

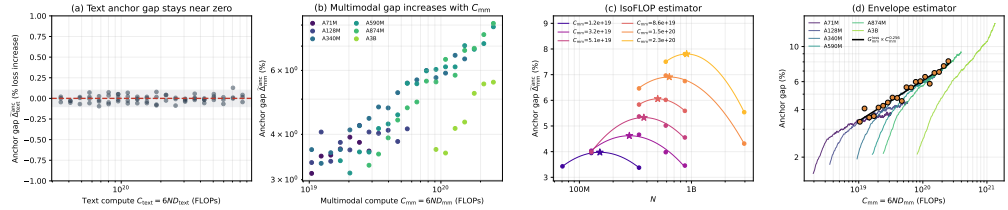

Figure 5: **Visual anchoring strengthens with multimodal compute while the text gap stays near zero.** (a) Paired text gaps, with a shaded $\pm 0.1\%$ reference band. (b) Paired multimodal gaps, colored by model size. (c) Gap profiles at six fixed compute budgets. Stars mark maxima of quadratic fits in $\log N$, not loss minima. (d) Per-model gap trajectories, paired gaps along the IMG envelope, and their power-law fit (black, $\gamma = 0.256$). All gaps are relative loss increases under ablation, expressed as percentages.

for low loss, we evaluate

$$G^{\mathrm{loss}}_s(C_s) = \widetilde{\Delta}^{\mathrm{anc}}_s\left(\widehat{N}^{\mathrm{img}}_{s,\mathrm{env}}(C_s), C_s\right). \tag{7}$$

Here, the IMG envelope selects the model size, and the gap compares IMG and PAD at that same size and compute. This is a paired gap along the IMG envelope, rather than a ratio between independently optimized frontiers. Maximizing the gap and minimizing the loss need not select the same model.

A consistent trend across allocation rules. For each multimodal summary, we fit a power law by linear regression in log gap and log compute:

$$G^j_{\mathrm{mm}}(C_{\mathrm{mm}}) \approx \kappa^j_{\mathrm{mm}} C^{\gamma^j_{\mathrm{mm}}}_{\mathrm{mm}},\ j \in \{\max, \mathrm{quad}, \mathrm{loss}\}, \tag{8}$$

with C_{mm} measured in FLOPs. The three summaries yield closely matched exponents, $\gamma^{\max}_{\mathrm{mm}} = 0.254$, $\gamma^{\mathrm{quad}}_{\mathrm{mm}} = 0.254$, and $\gamma^{\mathrm{loss}}_{\mathrm{mm}} = 0.256$, with log-space R^2 values of 0.978, 0.972, and 0.955, respectively. At $C_{\mathrm{mm}} = 10^{20}$ FLOPs, the fitted gaps are 0.06231, 0.06212, and 0.06099 in the same order. From the smallest to the largest fitted budget, the quadratic maximum rises from 3.63% to 8.09%, while the gap along the IMG envelope rises from 3.35% to 8.09%. This agreement shows that the overall increase is also present along allocations selected for low IMG loss, although local budget-to-budget fluctuations remain.

Compute captures the trend, with allocation still mattering. A descriptive joint fit gives $\widetilde{\Delta}^{\mathrm{anc}}_{\mathrm{mm}} \propto N^{0.212} D^{0.301}_{\mathrm{mm}}$. The unequal exponents, together with the fixed-compute profiles in Figure 5c, show that the gap is not determined by compute alone. The power laws summarize an observed scaling trend across the tested configurations. Because the budgets and summaries reuse correlated observations from the same runs, the quality of these fits does not establish training-seed reproducibility or behavior beyond the fitted range. We provide more details in Section D.

4 General Language Capability

The text scaling law in Section 3.2 remains nearly unchanged when visual content is ablated. We further investigate whether the contribution of vision to multimodal learning is accompanied by a change in general language performance. Using the same matched IMG and PAD runs, we compare

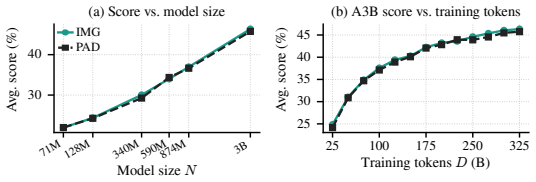

Figure 6: **Aggregate language performance.** Average scores over 15 text benchmarks for IMG and PAD (a) Text performance across model sizes N. (b) A3B text performance over D.

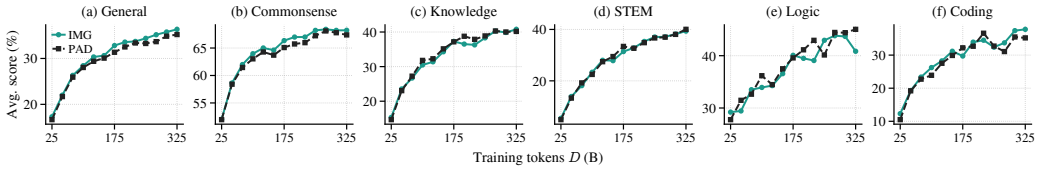

Figure 7: **Per-category language performance.** Average scores evaluated using A3B models for IMG and PAD conditions.

mean scores over 15 text benchmarks spanning general understanding, commonsense, knowledge, mathematics, logical reasoning, and coding. We provide more details in Section C.1.

Aggregate language performance follows a shared scaling trend. As shown in Figure 6, increasing model size produces substantial gains under both conditions, while the IMG and PAD curves remain closely aligned. The difference between conditions stays small relative to the improvement across the scaling ladder. Figure 6 shows the same pattern along the data axis. Specifically, both conditions improve rapidly early in training and more gradually thereafter, with closely matched aggregate scores. The stable pure-text loss pattern is therefore accompanied by preserved aggregate language performance in the evaluated suite.

Category gains accompany matched aggregate performance. As shown in Figure 7, IMG achieves higher mean scores in coding, general understanding, and commonsense reasoning than PAD. These gains accompany closely matched aggregate language performance shown in Figure 6.

5 Spatial Reasoning

Visual anchoring in Section 3 concerns multimodal text prediction. We now investigate spatial reasoning about object locations, relations, and configurations under text-only evaluation. Using the same matched IMG and PAD runs, we evaluate 15 benchmarks across six task groups, averaging benchmarks equally within each group and then weighting groups equally. Evaluation settings and full scores appear in Section C.2 and tables 9 to 12.

5.1 Spatial Reasoning Scales with Compute

At the nominal budget $D = 325$B, aggregate accuracy improves at every successive model size under both conditions. From A71M to A3B, IMG rises from 15.41% to 44.25% and PAD from 15.10% to 39.87% (Figure 8). At fixed A3B size, increasing D from 25B to 325B raises accuracy from 22.56% to 44.25% for IMG and from 20.81% to 39.87% for PAD, with local fluctuations and a further gain at the final checkpoint. Thus, substantial spatial capability develops through scaling even without visual content.

5.2 Visual Content Adds a Measurable Gain

Visual content adds a smaller, scale-dependent benefit. The matched IMG-PAD gap is positive at five of six model sizes and reaches 4.38% at A3B. It is not monotonic. Specifically, the gap is 3.62% at A128M and approximately 1.2% at A590M and A874M. These comparisons distinguish overall capability growth from the additional contribution of visual pre-training.

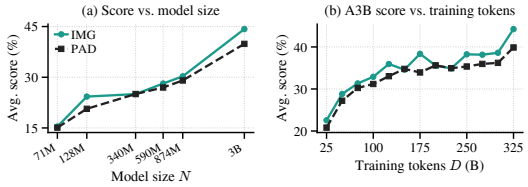

Figure 8: **Aggregate spatial reasoning performance.** Scores for IMG and PAD, averaged equally across six task groups, with equal benchmark weights within each group. (a) Final performance across model sizes N at a nominal budget of $D = 325$B tokens. (b) A3B performance over nominal training tokens D.

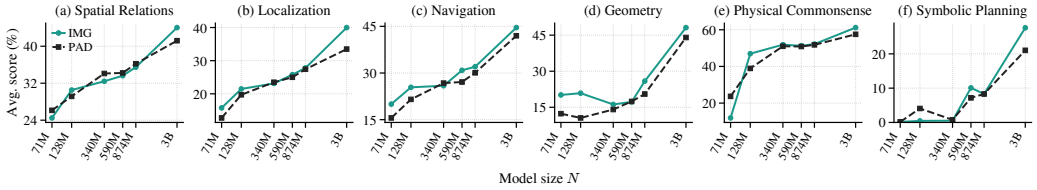

Figure 9: **Scores across model sizes for each task group.**

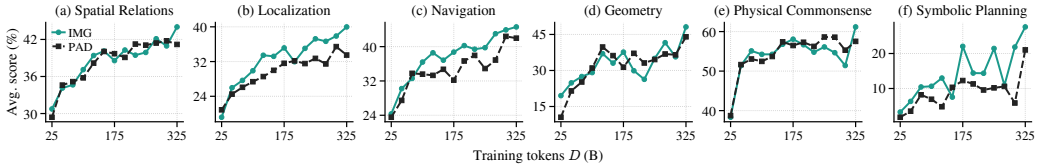

Figure 10: **Scores under data scaling for each task group.**

5.3 Transfer Varies Across Task Groups

Every task group improves from A71M to A3B under both conditions, but visual gains vary across groups and scales (Figure 9). During A3B training, localization, navigation, and symbolic planning have mean IMG-PAD gains of 2.88%, 3.02%, and 5.19%, respectively (Figure 10). The other three groups have mean gaps within 0.6% of zero, although all six favor IMG at the final checkpoint. The distinction is thus the persistence of gains across training, not their presence at a single checkpoint.

The three groups with sustained gains share demands for maintaining or composing spatial configurations: localization binds objects to positions, navigation updates positions along routes, and symbolic planning composes actions toward a target arrangement. The other groups emphasize relation inference, geometric rules, or prior physical knowledge, offering solution routes through language and symbolic constraints. PIQA-spatial, for example, scales strongly with model size despite little average visual advantage during A3B training; geometry gains concentrate in shape identification, while Geometry3K remains difficult. Together, these findings suggest more consistent transfer to organizing configurations and spatial updates. These demands overlap, so this interpretation does not establish a mechanism or strict taxonomy (Section F.1).

5.4 Emergence of Compositional Spatial Skills

Symbolic planning has the largest mean IMG-PAD gain during A3B training, approximately 5.19%. Blocksworld accuracy rises from 3.3% to 27.6% for IMG, versus 1.8% to 21.0% for PAD (Figure 10). Success requires a complete executable action sequence that reaches the goal; any invalid action or unmet goal invalidates the response.

As shown in Figure 11, larger models extend successful composition to longer plans. Here, emergence denotes the appearance of successful long-plan composition at larger scales, without establishing a sharp scaling transition or the acquisition of individual atomic operations. On 3-4-block problems, IMG reaches 46.70% at A590M and 67.45% at A3B. On 5-12-block problems, no IMG response succeeds through A590M; success rises to 0.78% at A874M and 16.62% at A3B, com-

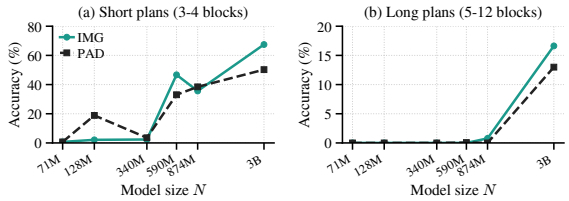

Figure 11: Growth of symbolic tower-building performance in the benchmark Blocksworld. Final-checkpoint scores for (a) short plans with 3-4 blocks (106 problems) and (b) long plans with 5-12 blocks (385 problems) in the tower-building task.

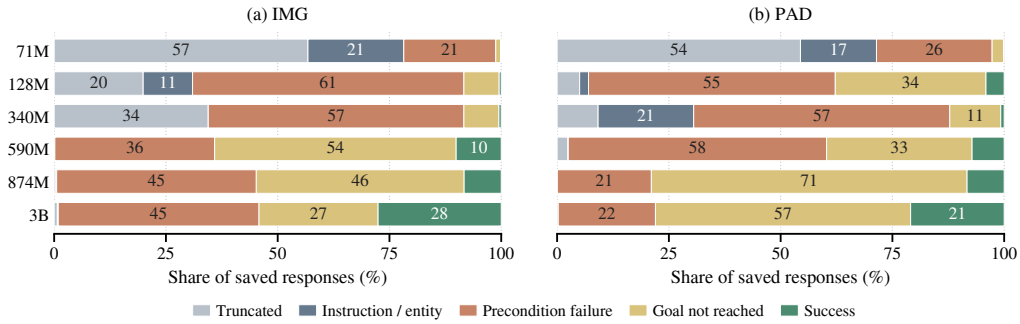

Figure 12: Execution outcomes across scales for (a) IMG and (b) PAD. Length-limited responses are classified first; others are executed until the first invalid instruction or unmet precondition, with a goal check after legal plans. Labels give percentages for segments of at least 10%.

pared with 12.99% for A3B PAD. Section F.1.1 gives the response counts, and the ten-block case in Section F.2 illustrates a complete IMG solution versus a legal but incomplete PAD prefix.

Execution remains a bottleneck. Our final-checkpoint audit distinguishes truncation, instruction/entity errors, precondition violations, and legal plans that miss the goal (Section F.1.1 and fig. 12). Truncation dominates A71M but becomes uncommon at A3B. At A3B, IMG produces fewer legal-but-unsuccessful plans than PAD (523 versus 1,121), but more precondition failures (882 versus 427). Among legal-but-unsuccessful plans, 73.4% under PAD contain fewer actions than minimally required, compared with 44.2% under IMG.

This contrast is consistent with better organization of target configurations or action sequences, as suggested in Section 5.3, while reliable execution remains a bottleneck. Many IMG precondition failures attempt to pick up a covered block: placing a on b too early prevents subsequently placing b on c. The text-prediction objective does not explicitly enforce these constraints. Competing first-failure categories do not isolate improvements in individual operations; Sections F.1.1 and F.2 provide the full audit and cases.

6 Conclusion

Our controlled visual-content ablations show that visual information acts as a scalable anchor in native multimodal pre-training. Its contribution to multimodal text prediction grows with compute, while pure-text loss scaling, compute-optimal allocation, and aggregate general-language performance remain largely unchanged. Spatial reasoning improves with model and data scaling under both conditions, and visual content provides an additional benefit even when evaluation inputs contain only text. During A3B training, gains are most sustained in localization, navigation, and symbolic planning, which share demands for maintaining configurations or composing spatial updates. Gains in spatial relations, geometry, and physical commonsense, with greater emphasis on explicit relations, rules, or prior knowledge, are less consistent. This suggests selective spatial transfer; long-plan success emerges at larger scales, while execution remains a bottleneck.

REFERENCES

Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling Laws for Generative Mixed-Modal Language Models. In *International Conference on Machine Learning (ICML)*, 2023.

Morris Alper, Michael Fiman, and Hadar Averbuch-Elor. Is BERT Blind? Exploring the Effect of Vision-and-Language Pretraining on Visual Language Understanding. *arXiv preprint arXiv:2303.12513*, 2023.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. PIQA: Reasoning about Physical Commonsense in Natural Language. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pp. 7432–7439, 2020.

Chameleon Team. Chameleon: Mixed-Modal Early-Fusion Foundation Models. *arXiv preprint arXiv:2405.09818*, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, et al. Emu3.5: Native Multimodal Models are World Learners. *arXiv preprint arXiv:2510.26583*, 2025.

Mingkang Dong, Hongyi Cai, Jie Li, Sifan Zhou, Bin Ren, Kunyu Peng, and Yuqian Fu. VisNec: Measuring and Leveraging Visual Necessity for Multimodal Instruction Tuning. *arXiv preprint arXiv:2603.01195*, 2026.

Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, et al. SuperGPQA: Scaling LLM Evaluation Across 285 Graduate Disciplines. *arXiv preprint arXiv:2502.14739*, 2025.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training Compute-Optimal Large Language Models. *arXiv preprint arXiv:2203.15556*, 2022.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*, 2020.

Kimi, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi K2.5: Visual Agentic Intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while Reasoning in Space: Multimodal Visualization-of-Thought. *arXiv preprint arXiv:2501.07542*, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*, 2023.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. In *International Conference on Learning Representations (ICLR)*, 2023.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. DeepSeek-V3 Technical Report. *arXiv preprint arXiv:2412.19437*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Annual Conference on Neural Information Processing Systems (NIPS)*, 2023a.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is Your Code Generated by ChatGPT Really Correct? Rigorous Evaluation of Large Language Models for Code Generation. In *Annual Conference on Neural Information Processing Systems (NIPS)*, volume 36, pp. 21558–21572, 2023b.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-GPS: Interpretable Geometry Problem Solving with Formal Language and Symbolic Reasoning. *arXiv preprint arXiv:2105.04165*, 2021.

Sachit Menon, Richard Zemel, and Carl Vondrick. Whiteboard-of-Thought: Thinking Step-by-Step Across Modalities. *arXiv preprint arXiv:2406.14562*, 2024.

Roshanak Mirzaee and Parisa Kordjamshidi. Transfer Learning with Synthetic Corpora for Spatial Role Labeling and Reasoning. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. SPARTQA: A textual question answering benchmark for spatial reasoning. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 4582–4598, 2021.

Tanawan Premsri and Parisa Kordjamshidi. FoREST: Frame of Reference Evaluation in Spatial Reasoning Tasks. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An Adversarial Winograd Schema Challenge at Scale. *Communications of the ACM*, 64(9):99–106, 2021.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. SocialiQA: Commonsense Reasoning about Social Interactions. *arXiv preprint arXiv:1904.09728*, 2019.

Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. StepGame: A New Benchmark for Robust Multi-Hop Spatial Reasoning in Texts. *arXiv preprint arXiv:2204.08292*, 2022.

Mustafa Shukor, Louis Bethune, Dan Busbridge, David Grangier, Enrico Fini, Alaaeldin El-Nouby, and Pierre Ablin. Scaling Laws for Optimal Data Mixtures. In *Annual Conference on Neural Information Processing Systems (NIPS)*, 2025a.

Mustafa Shukor, Enrico Fini, Victor Guilherme Turrisi da Costa, Matthieu Cord, Joshua Susskind, and Alaaeldin El-Nouby. Scaling Laws for Native Multimodal Models. In *IEEE International Conference on Computer Vision (ICCV)*, 2025b.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging Big-Bench Tasks and Whether Chain-of-Thought Can Solve Them. *arXiv preprint arXiv:2210.09261*, 2022.

Hao Tan and Mohit Bansal. Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision. *arXiv preprint arXiv:2010.06775*, 2020.

Shengbang Tong, David Fan, John Nguyen, Ellis Brown, Gaoyue Zhou, Shengyi Qian, Boyang Zheng, Théophane Vallaeys, Junlin Han, Rob Fergus, et al. Beyond Language Modeling: An Exploration of Multimodal Pretraining. *arXiv preprint arXiv:2603.03276*, 2026.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. PlanBench: An Extensible Benchmark for Evaluating Large Language Models on Planning and Reasoning about Change. In *Annual Conference on Neural Information Processing Systems (NIPS)*, 2023.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a Picture Worth a Thousand Words? Delving into Spatial Reasoning for Vision Language Models. In *Annual Conference on Neural Information Processing Systems (NIPS)*, 2024a.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. In *Annual Conference on Neural Information Processing Systems (NIPS)*, 2024b.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. *arXiv preprint arXiv:1502.05698*, 2015.

Haoyuan Wu, Aoqi Wu, Hai Wang, Jiajia Wu, Jinxiang Ou, and Bei Yu. Scaling Native Multimodal Pre-Training From Scratch. *arXiv preprint arXiv:2607.22043*, 2026.

Yutaro Yamada, Yihan Bao, Andrew K. Lampinen, Jungo Kasai, and Ilker Yildirim. Evaluating Spatial Understanding of Large Language Models. *Transactions on Machine Learning Research*, 2024.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can A Machine Really Finish Your Sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Yiming Zhang, Zhonghan Zhao, Wenwei Zhang, Haiteng Zhao, Tianyang Lin, Huanze Tang, Yunhua Zhou, Demin Song, Kuikun Liu, Haochen Ye, Haian Huang, Yuzhe Gu, Haijun Lv, Qipeng Guo, Bin Liu, Gaoang Wang, and Kai Chen. Scalable Visual Pretraining for Language Intelligence. *arXiv preprint arXiv:2607.09657*, 2026.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. *arXiv preprint arXiv:2304.06364*, 2023.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model. *arXiv preprint arXiv:2408.11039*, 2024.

Chengxu Zhuang, Evelina Fedorenko, and Jacob Andreas. Visual Grounding Helps Learn Word Meanings in Low-Data Regimes. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.

A Related Work

Native multimodal pre-training. Multimodal systems differ in when and how vision and language enter a shared learning process. Modular approaches connect a pre-trained vision encoder to a language model through bridging modules or instruction tuning (Li et al., 2023; Liu et al., 2023a). Native approaches instead train on mixed-modal sequences within a shared backbone. Chameleon uses discrete image and text tokens (Chameleon Team, 2024), Transfusion combines language modeling with image diffusion (Zhou et al., 2024), and Emu3.5 studies unified autoregressive modeling of large-scale interleaved vision-language data (Cui et al., 2025). In this study, we investigate how the information carried by these inputs affects learning across scales.

Scaling laws and cross-modal interactions. Language-model scaling laws relate loss to model size, data, and compute and formalize the allocation of a fixed training budget (Kaplan et al., 2020; Hoffmann et al., 2022). Mixed-modal scaling extends this analysis to interactions between data sources, including synergy and competition (Aghajanyan et al., 2023). Subsequent work examines native multimodal architectures and sparsity (Shukor et al., 2025b), data-mixture weights (Shukor et al., 2025a), and differences between language and vision scaling requirements (Tong et al., 2026). Our prior study estimates allocations for the text and multimodal objectives separately and analyzes their dependence on the training mixture (Wu et al., 2026). Changing the mixture, however, can also alter language supervision. Tong et al. investigate caption-distribution shifts as a possible source of language degradation (Tong et al., 2026). We hold the text, mixture, sequence structure, and optimization settings fixed while ablating visual content. This comparison targets image information within a common training setup.

Visual context and transfer to language. Visual ablation can probe a trained model's use of image context (Frank et al., 2021). VisNec uses losses with and without visual context to select informative instruction-tuning examples (Dong et al., 2026). Its blind forward pass changes the input to a single model, whereas our anchor gap compares matched models trained with different access to visual content. A complementary line of work studies how visual supervision affects language learning through vokenization (Tan & Bansal, 2020), visual-knowledge probing (Alper et al., 2023), and word learning under limited text data (Zhuang et al., 2024). Work on visual-document pre-training also compares continued learning from scientific pages with learning from parsed text of the same documents (Zhang et al., 2026). These settings motivate examining both the contribution of visual context to training loss and the capabilities retained when visual inputs are absent. We study both within native pre-training while preserving the associated text across conditions.

Spatial reasoning without visual inputs. Textual spatial benchmarks assess relations and operations that can also be expressed visually. StepGame evaluates relation composition and multi-hop reasoning (Shi et al., 2022), while SpatialEval compares textual and visual input conditions (Wang et al., 2024a). Other methods introduce visual computation during inference (Menon et al., 2024; Li et al., 2025). Our evaluation instead uses text-only inputs and outputs after pre-training. Our prior work reports spatial improvements when we add multimodal data to a fixed text budget (Wu et al., 2026). The present matched ablation retains both the associated text and the input budget. This design separates capability growth with scale from the additional contribution of visual content, while leaving the internal mechanism of this transfer unresolved.

B Training settings

For native multimodal pre-training, we employ the Muon optimizer with a weight decay of 0.1 and a gradient clipping threshold of 1.0. Additionally, we utilize a warmup-stable learning rate schedule, a global batch size of 2048, and a maximum sequence length of 4096 (8M input tokens per step). The training loss is computed only on text tokens, with vision tokens masked out from loss computation (Kimi et al., 2026). We provide more details in Table 1.

C Evaluation settings

C.1 General language capability

We evaluate general language capabilities using 15 distinct benchmarks.

Table 1: Pre-training hyperparameters. We detail the hyperparameters used for pre-training different model configurations to derive scaling laws.

	A71M	A128M	A340M	A590M	A874M	A3B
Architecture Configurations						
layers	11	15	19	23	27	48
hidden size	640	768	1152	1280	1536	2048
FFN hidden size	2048	2048	3072	4096	4096	6912
attention heads	8	12	16	20	24	32
query groups	4					
kv channels	128					
total experts	128					
activated experts	8					
expert FFN hidden size	256	256	384	512	512	768
shared-expert hidden size	256	256	384	512	512	768
MoE layer	first layer dense, remaining layers MoE					
Training Settings						
batch size (sequences)	2048					
sequence length	4096					
warmup steps	2000					
learning rate	2.6e-3	1.9e-3	1.1e-3	8.2e-4	6.8e-4	3.0e-4

- **Aggregate:** SuperGPQA(Du et al., 2025)(5-shot), MMLU-Redux(Hendrycks et al., 2020)(5-shot), MMLU-Pro(Wang et al., 2024b)(5-shot), and AGIEval$_{\text{en}}$(Zhong et al., 2023)(3-shot).
- **Coding:** HumanEval+(Liu et al., 2023b)(0-shot) and MBPP+(Liu et al., 2023b)(3-shot).
- **Mathematics:** GSM8K(Cobbe et al., 2021)(4-shot, CoT) and MATH(Lightman et al., 2023)(4-shot, CoT).
- **Logical Reasoning:** BBH(Suzgun et al., 2022)(3-shot, CoT).
- **Knowledge:** NaturalQuestions(Kwiatkowski et al., 2019) (5-shot) and TriviaQA(Joshi et al., 2017)(5-shot).
- **Commonsense Reasoning:** HellaSwag(Zellers et al., 2019)(10-shot), WinoGrande(Sakaguchi et al., 2021)(5-shot), SIQA(Sap et al., 2019)(0-shot), and PIQA(Bisk et al., 2020)(0-shot).

For the evaluation of general language capability, we use greedy generation strategy.

C.2 Spatial reasoning

We evaluate 15 benchmarks in six task groups, all with 5-shot chain-of-thought prompting and text-only inputs. Each entry below combines the group's scope with its benchmarks.

- **Spatial Relations:** Infer or verify relative positions. StepGame (Shi et al., 2022) uses directional statements; SpaRTUN (Mirzaee & Kordjamshidi, 2022) covers direction, distance, containment, and contact; bAbI(positional reasoning) (Weston et al., 2015) tests positional constraints; and FoREST (Premsri & Kordjamshidi, 2025) varies the reference frame.
- **Localization:** Identify, describe, or count objects using their attributes and positions. SpartQA (Mirzaee et al., 2021), SpatialMap (Wang et al., 2024a), and BBH(colored objects) (Suzgun et al., 2022) query objects within described scenes.
- **Navigation:** Follow routes, update positions or orientations, or construct paths. SpatialUnderstanding (Yamada et al., 2024) queries the object reached after successive moves; MazeNav (Wang et al., 2024a) uses text-represented mazes; bAbI(path finding) (Weston et al., 2015) requires a direction sequence; and BBH(navigate) (Suzgun et al., 2022) tests return to the starting point.
- **Geometry:** Recover geometric configurations and apply their constraints. BBH(geometric shapes) (Suzgun et al., 2022) identifies shapes from SVG path coordinates; Geometry3K (Lu et al., 2021) uses textual logic forms for lengths, angles, and areas, without diagram images.
- **Physical Commonsense:** Choose between solutions to everyday physical goals. PIQA-spatial (Bisk et al., 2020) tests knowledge of object use, materials, or support, rather than general physical simulation.

Table 2: Spatial evaluation inventory. The 15 benchmarks cover six task groups.

Capability	Benchmark	Items	Answer type
Spatial Relations	StepGame	3,000	Direction label
	SpaRTUN	3326	Relation label
	bAbI(positional reasoning)	627	Yes/no
	FoREST	1998	Direction label
Localization	SpartQA	1994	Choice label
	SpatialMap	997	Choice label
	BBH(colored objects)	245	Choice label
Navigation	SpatialUnderstanding	3998	Object name
	MazeNav	999	Choice label
	bAbI(path finding)	995	Direction sequence
	BBH(navigate)	245	Yes/no
Geometry	BBH(geometric shapes)	244	Choice label
	Geometry3K	596	Choice label
Physical Commonsense	PIQA-spatial	548	Choice label
Symbolic Planning	Blocksworld	491	Action sequence

- **Symbolic Planning:** Produce a complete executable action sequence satisfying all goal relations. Blocksworld (Valmeekam et al., 2023) asks the model to build a tower from separate blocks, tracking which blocks are clear, held, or placed, and respecting action preconditions.

We generate four responses per item with temperature 0.3 and top-$p = 0.9$, reporting avg@4. Each benchmark belongs to one group; benchmarks are weighted equally within groups, and the six group scores are averaged equally. These groups organize this evaluation rather than define a general taxonomy of spatial or physical reasoning. Table 2 lists item counts and answer types.

D Allocation Estimation and Loss-Compute Frontiers

This section details the estimators in Sections 2 and 3, retaining the source $s \in \{\text{text}, \text{mm}\}$, condition $q \in \{\text{img}, \text{pad}\}$, and estimator $e \in \{\text{iso}, \text{env}\}$ indices. We suppress the fixed mixture argument r and report results at $r = 0.3$. The compute is $C_s = 6ND_s$ in FLOPs. Hats denote finite-data estimates, tildes denote fixed-compute interpolated losses, and $\mathcal{F}_s^q$ denotes a fitted frontier. The loss-decay exponent β_s^q, allocation exponents $a_{s,e}^q, b_{s,e}^q$, and gap-growth exponent γ_s^j have distinct meanings.

D.1 Data Preprocessing

With $T = 2048 \times 4096 \approx 8\text{M}$ input tokens per optimizer step, step k corresponds to

$$D = Tk,\ D_{\text{text}} = \frac{D}{1+r},\ D_{\text{mm}} = \frac{rD}{1+r},\ C_s = 6ND_s. \tag{9}$$

Here D is the total input-token count and D_s is the token count for source s. The 325B total-token budget is nominal, whereas the loss-compute analysis uses actual selected steps. For each source and condition at fixed r, we construct various candidate budgets uniformly in log compute over the union of the model-specific ranges. Writing $C_{s,m}^{q,\min}$ and $C_{s,m}^{q,\max}$ for model m's endpoints, we define the supported model set as

$$\mathcal{M}_s^q(C_s) = \{m : C_{s,m}^{q,\min} \le C_s \le C_{s,m}^{q,\max}\}. \tag{10}$$

We retain budgets with $|\mathcal{M}_s^q(C_s)| \ge 3$ and interpolate losses linearly in $\log C_s$ within each model's observed range, without extrapolation, to obtain $\widetilde{L}_s^q(N_m, C_s)$.

D.2 Estimating Compute-Optimal Allocation

IsoFLOP profiles. At each retained budget and fixed (s, q), let $x = \log N$ and fit the losses with a quadratic. For this source, condition, and budget, we estimate coefficients p_0, p_1, p_2 in the quadratic

$$P_{s,C_s}^q(x) = p_2 x^2 + p_1 x + p_0. \tag{11}$$

For $p_2 > 0$, the allocation estimate is

$$\log \widehat{N}^q_{s,\mathrm{iso}}(C_s) = \mathrm{clip}\left(-\frac{p_1}{2p_2}, x_{\min}, x_{\max}\right), \ \widehat{D}^q_{s,\mathrm{iso}}(C_s) = \frac{C_s}{6\widehat{N}^q_{s,\mathrm{iso}}(C_s)}, \tag{12}$$

where $x_{\min}, x_{\max}$ delimit the model sizes supported at that budget. We regress $\log \widehat{N}^q_{s,\mathrm{iso}}$ and $\log \widehat{D}^q_{s,\mathrm{iso}}$ on $\log C_s$ with equal weights. This yields $a^q_{s,\mathrm{iso}}$ and $b^q_{s,\mathrm{iso}}$, with $a^q_{s,\mathrm{iso}} + b^q_{s,\mathrm{iso}} = 1$ by the compute identity.

Training-curve envelope. The equal-FLOPs envelope is

$$\widehat{L}^q_{s,\mathrm{env}}(C_s) := \min_{m \in \mathcal{M}^q_s(C_s)} \widetilde{L}^q_s(N_m, C_s). \tag{13}$$

Each point retains the minimizing model size $\widehat{N}^q_{s,\mathrm{env}}(C_s)$ and token count $\widehat{D}^q_{s,\mathrm{env}}(C_s) = C_s/(6\widehat{N}^q_{s,\mathrm{env}}(C_s))$. Regressing these allocations in log space gives $a^q_{s,\mathrm{env}}$ and $b^q_{s,\mathrm{env}}$, again summing to one. Fitting the envelope losses gives the distinct frontier parameters below. Both allocation estimators reuse the same runs, so their comparison in Figures 1 to 4 is a sensitivity check rather than independent validation.

D.3 Fitting Loss-Compute Frontiers

At fixed (s, q, r), Equation (4) can be written

$$\mathcal{F}^q_s(C_s) = E^q_s + \left(\frac{C^q_{c,s}}{C_s}\right)^{\beta^q_s} = E^q_s + K^q_s C_s^{-\beta^q_s}, \qquad K^q_s = (C^q_{c,s})^{\beta^q_s}. \tag{14}$$

The amplitude K^q_s is distinct from the characteristic compute scale $C^q_{c,s}$. We search various equally spaced values of β^q_s in $[0.01, 0.60]$. For each candidate, ordinary least squares in loss space estimates (E^q_s, K^q_s) from the various envelope points. Candidates with $K^q_s \le 0$ or any nonpositive prediction at a fitted budget are rejected. Among the remaining candidates, we minimize the mean squared log-loss residual and recover $C^q_{c,s} = (K^q_s)^{1/\beta^q_s}$. This two-stage selection is not joint least squares in log loss; it does not constrain $E^q_s > 0$, although all four reproduced fits have positive offsets. Each source and condition is fitted independently at $r = 0.3$.

D.4 Fitting Visual Anchoring

The anchor analysis applies Equation (6) to paired losses. We read each condition's loss at the same model size and compute budget before taking their relative difference. For the quadratic maximum, we fit paired gaps in centered $\log N$ and check both endpoints and any interior stationary point. The loss-path summary uses the discrete model selected by the IMG envelope at each budget, with the PAD loss evaluated at that same allocation. It is not a ratio of independently optimized or fitted frontiers.

Positive multimodal summaries are regressed in log gap against log compute. Figures express gaps as percentages, whereas Equation (8) defines them as fractions. If the percentage regression has intercept h_j, then

$$\log\big(100 G^j_{\mathrm{mm}}(C_{\mathrm{mm}})\big) \approx h_j + \gamma^j_{\mathrm{mm}} \log C_{\mathrm{mm}}, \qquad \kappa^j_{\mathrm{mm}} = \frac{\exp(h_j)}{100}. \tag{15}$$

Thus κ^j_{mm} is the coefficient of $C_{\mathrm{mm}}^{\gamma^j_{\mathrm{mm}}}$ when compute is expressed in FLOPs. For $j = \max, \mathrm{quad}, \mathrm{loss}$, respectively, it is approximately 5.1874×10^{-7}, 5.1693×10^{-7}, and 4.6420×10^{-7}. The gaps reported at 10^{20} FLOPs in Section 3 are fitted function values, not these coefficients.

E General Language Evaluation Results

We report the benchmark-level scores underlying the general-language comparisons in Section 4.

Table 3: Supplementary execution statistics. The last row is conditional on legal plans that miss the goal; its denominator differs from the outcome counts above.

Scale	Outcome or subset	IMG	PAD
A71M	Truncation	1,115	1,069
A3B	Truncation	14	0
A3B	Legal plan, goal missed	523	1,121
A3B	Precondition violation	882	427
A3B	Too short among legal goal misses	231/523 (44.2%)	823/1,121 (73.4%)

F Spatial Evaluation Results

We report the benchmark-level scores underlying the spatial comparisons in Section 5.

F.1 Spatial Transfer and Planning Analysis

Benchmark-level interpretation. Beyond the selective gains in Section 5.3, performance across model sizes is comparable in the remaining groups, whose averages mask distinct benchmark behaviors. Physical commonsense has the largest end-to-end improvement with model size, but a much smaller visual-content gain at A3B. Its benchmark, PIQA-spatial, selects between two solutions to an everyday physical goal and can draw on physical commonsense and language knowledge; scaling alone therefore does not establish a specifically visual mechanism. Geometry gains concentrate in BBH(geometric shapes), while Geometry3K remains substantially harder. These differences should be considered when interpreting the group averages reported alongside the full benchmark scores in Tables 9 to 12.

F.1.1 Plan Length and Execution Outcomes

Blocksworld requires an executable action sequence satisfying all goal relations, unlike answer selection in PIQA-spatial. This strict criterion helps explain why visual-content gains coexist with limited absolute success. The long-plan results in Figure 11 use 1540 responses to 385 problems with 5-12 blocks. No IMG response succeeds through A590M, while 12 succeed at A874M (0.78%) and 256 at A3B (16.62%), compared with 200 for A3B PAD (12.99%). The short-plan results are reported in Section 5.4.

For the execution analysis in Figure 12, each response receives one outcome. Truncation at the length limit takes precedence; otherwise, execution stops at the first uninterpretable instruction, undeclared or unspecified block, or violated action precondition (e.g., stacking a block that is not held). Goal attainment is checked only when every action is legal. These competing first-failure categories do not isolate improvements in individual operations.

Table 3 shows that truncation becomes uncommon at A3B, although its decline is not monotonic across intermediate sizes. Repetitive action loops in A71M outputs suggest that improved repetition and termination control may contribute. At A3B, IMG has approximately 53% fewer legal-but-unsuccessful plans but 2.1 times as many precondition failures as PAD; its unsuccessful legal plans are also less often shorter than minimally required. Many precondition failures attempt to pick up a covered block, sometimes immediately after placing another block on it. This contrast motivates the distinction between representing target configurations and executing plans in Section 5.4; the ten-block example in Section F.2 illustrates the distinction without establishing its prevalence.

F.2 Showcases: What the Model Actually Does

A complete long plan instead of a plausible prefix. The ten-block example below makes the composition gain concrete. Each consecutive pickup-and-stack pair is abbreviated as $x \to y$, meaning *pick up* x*; stack* x *on* y. Initial states and goals were recovered from the public source instances; all 491 local reference plans execute successfully against their matched instances.

Case 1: Completing a ten-block tower

Problem (condensed). Blocks $a, c, d, e, f, g, h, i, j, k$ are clear and on the table; the hand is empty. Build the tower i–d–c–e–g–h–j–k–a–f (top to bottom).

A590M IMG, first response: $a \to f,\ d \to a,\ c \to d,\ e \to c.$
All eight actions are legal, but only one of nine target relations is met.

A3B PAD, first response: $a \to f,\ i \to a,\ d \to i.$
All six actions are legal, but only one of nine target relations is met.

A3B IMG, first response: $a \to f,\ k \to a,\ j \to k,\ h \to j,\ g \to h,\ e \to g,\ c \to e,\ d \to c,\ i \to d.$
All 18 actions are legal, and all nine target relations are met.

A3B IMG succeeds on all four responses to this item, while every smaller IMG model and A3B PAD fail on all four. The successful plan builds from the bottom upward, preserving earlier placements. Both unsuccessful prefixes shown above terminate normally: their failure concerns goal completion, rather than the output limit.

A correct answer can conceal a broken spatial derivation. The next example asks for the destination after moving right, down, right, up, left from white-headed capuchin.

Case 2: Accepted answer, inconsistent route (`square-3431`)

Correct route (x rightward, y downward): white-headed capuchin $(0, 0) \to$ lacewing $(1, 0) \to$ wallaby $(1, 1) \to$ hare $(2, 1) \to$ cassette $(2, 0) \to$ lacewing $(1, 0)$.

A3B IMG explanation (condensed): replaces the down move with another right move, assigning wallaby to $(2, 0)$, then hare to $(2, 1)$ and cassette to $(3, 1)$; its final coordinate-to-object assignment is also inconsistent.
Final answers: IMG: `lacewing` (accepted); PAD: `hare` (incorrect).

Multiple movement and entity errors make the written derivation invalid despite the accepted answer. They cannot be repaired by correcting one entity name, but also do not establish that the answer was guessed randomly.

Separating verbal slips from failed operations. In `bw-hf-4`, A590M IMG's first response executes the correct plan $f \to b,\ j \to f,\ g \to j$, despite once writing *a clear on top* where the block is f. This local prose slip leaves the solution intact. By contrast, the following reversal changes the actual path:

Case 3: Reversing only one of two relations (bAbI task 19, item 0)

Relevant facts (condensed): the bathroom is north of the bedroom; the bedroom is north of the hallway. Find a path from the bathroom to the hallway.
Correct path: south, south (`s,s`).
A3B IMG, first response: south, north (`s,n`).

Only the first north-of relation is inverted correctly; executing the predicted path returns to the bathroom. Checking every edge in the travel direction would expose the error. These cases illustrate mechanisms; the exhaustive counts in Figure 12 remain specific to Blocksworld.

Counting explanation quality within a fixed case panel. We revisit eight fixed problems across all 12 conditions, inspecting each first response (96 traces). Of 27 accepted answers, 13 have consistent derivations, two contain local verbal errors, nine contain substantive errors, and three remain borderline (Table 4). The panel covers StepGame, three SpatialUnderstanding layouts, bAbI pathfinding, and Blocksworld. It supports same-problem case comparisons, but neither population frequency estimates nor a monotonic scaling trend in explanation reliability.

The analyses connect capability growth to both scaling axes, selective visual-content gains, and longer successful action compositions. These results concern single training runs under the archived evaluator. Executable plans and explanation checks complement final-answer accuracy.

	IMG		PAD	
Scale	Accepted	C / L / S / U	Accepted	C / L / S / U
A71M	3	0 / 0 / 3 / 0	1	1 / 0 / 0 / 0
A128M	1	1 / 0 / 0 / 0	3	1 / 0 / 2 / 0
A340M	3	1 / 0 / 1 / 1	2	0 / 0 / 1 / 1
A590M	2	1 / 0 / 0 / 1	2	1 / 1 / 0 / 0
A874M	1	1 / 0 / 0 / 0	3	2 / 0 / 1 / 0
A3B	4	2 / 1 / 1 / 0	2	2 / 0 / 0 / 0

Table 4: Explanation quality among accepted answers in eight fixed problems per condition. C: consistent derivation; L: local verbal error with the solution intact; S: substantive reasoning error; U: unresolved boundary between local and substantive error. The four counts sum to Accepted, which is out of eight first responses. This manual case audit is separate from the all-response planning analysis and does not estimate population frequencies.

G Declaration of LLM Usage

We acknowledge the use of LLMs to improve the clarity and grammar of this paper. All research ideas, experimental design, implementation, analysis, and conclusions are the work of the authors.

Table 5: **A3B IMG language scores across various D.**

	# Pre-Trained Tokens	25B	50B	75B	100B	125B	150B	175B	200B	225B	250B	275B	300B	325B
Aggregate	MMLU-Redux (Acc.)	27.05	38.81	44.74	48.35	51.19	51.37	54.49	54.21	55.05	56.00	57.17	57.09	57.98
	MMLU-Pro (Acc.)	10.85	12.99	17.73	19.19	19.86	20.21	22.03	23.84	24.61	24.92	25.33	27.53	27.68
	$\text{AGIEval}_{\text{en}}$ (Acc.)	20.49	22.71	27.88	29.27	32.13	32.66	36.03	36.58	35.02	35.99	37.10	36.92	37.73
	SuperGPQA (Acc.)	10.78	13.48	14.92	16.88	18.11	18.27	18.69	19.63	20.17	20.63	21.08	21.53	22.00
Coding	HumanEval+ (Pass@1)	6.10	11.59	15.24	14.63	15.24	15.24	13.41	18.90	19.51	15.85	17.68	21.34	21.34
	MBPP+ (Pass@1)	18.52	26.98	31.48	37.83	41.27	47.09	46.03	48.94	49.47	48.94	49.73	53.44	54.23
Mathematics	GSM8K (Pass@1)	4.85	17.21	22.21	30.63	37.00	36.01	42.15	43.82	48.07	50.19	49.58	51.40	52.39
	MATH (Pass@1)	5.95	10.60	14.05	16.05	19.00	19.70	20.65	21.80	22.60	23.90	24.65	25.00	26.30
Logic Reasoning	BBH (EM)	29.21	29.44	33.50	33.93	34.28	36.56	40.07	39.49	39.08	42.94	43.82	43.68	40.84
Knowledge	NaturalQuestions (EM)	10.08	14.99	16.32	18.17	19.31	21.77	24.13	22.05	19.64	22.02	23.35	21.55	23.63
	TriviaQA (EM)	20.56	32.08	37.08	42.78	43.47	46.81	50.42	50.97	52.92	54.58	57.22	58.19	57.92
Commonsense Reasoning	HellaSwag (Acc.)	51.23	59.40	63.20	66.47	66.97	67.87	68.27	69.37	69.53	70.60	71.23	71.37	71.57
	SIQA (Acc.)	34.85	45.14	52.71	54.15	56.24	55.48	59.72	59.67	60.90	62.28	62.08	61.87	62.95
	PIQA (Acc.)	71.16	74.59	75.63	76.11	77.31	76.88	77.42	77.31	78.29	77.86	79.05	78.45	77.80
	WinoGrande (Acc.)	50.83	55.41	56.43	59.12	59.51	58.25	59.98	61.64	59.27	62.12	61.25	61.17	60.62
Average		24.83	31.03	34.88	37.57	39.39	40.28	42.23	43.22	43.61	44.59	45.36	46.04	46.33

Table 6: **IMG language scores across various N.**

	Model Sizes (N)	A71M	A128M	A340M	A590M	A874M	A3B
Aggregate	MMLU-Redux (Acc.)	25.40	32.10	38.91	44.95	49.03	57.98
	MMLU-Pro (Acc.)	10.68	11.63	13.87	14.97	17.29	27.68
	$\text{AGIEval}_{\text{en}}$ (Acc.)	20.64	20.45	25.54	26.29	29.67	37.73
	SuperGPQA (Acc.)	8.00	10.31	11.81	14.16	14.90	22.00
Coding	HumanEval+ (Pass@1)	4.88	7.93	12.20	16.46	15.85	21.34
	MBPP+ (Pass@1)	7.94	13.23	26.45	30.95	37.57	54.23
Mathematics	GSM8K (Pass@1)	2.27	5.08	14.25	23.27	28.05	52.39
	MATH (Pass@1)	3.70	4.95	9.40	11.95	15.60	26.30
Logic Reasoning	BBH (EM)	26.87	28.06	30.31	33.03	35.94	40.84
Knowledge	NaturalQuestions (EM)	7.15	10.03	13.85	14.65	18.86	23.63
	TriviaQA (EM)	17.08	17.92	25.97	36.94	39.86	57.92
Commonsense Reasoning	HellaSwag (Acc.)	42.13	48.13	57.57	62.10	65.60	71.57
	SIQA (Acc.)	32.45	33.88	43.04	49.90	52.05	62.95
	PIQA (Acc.)	68.50	69.64	73.45	74.05	75.68	77.80
	WinoGrande (Acc.)	51.54	51.70	53.83	56.35	58.41	60.62
Average		21.95	24.34	30.03	34.00	36.96	46.33

Table 7: **A3B PAD language scores across various D.**

	# Pre-Trained Tokens	25B	50B	75B	100B	125B	150B	175B	200B	225B	250B	275B	300B	325B
Aggregate	MMLU-Redux (Acc.)	25.89	38.30	44.72	48.49	50.44	51.63	52.74	54.63	54.40	55.74	56.72	56.32	57.75
	MMLU-Pro (Acc.)	11.17	12.20	15.96	16.94	18.53	18.59	20.05	21.95	21.62	22.29	22.20	24.36	25.49
	$\text{AGIEval}_{\text{en}}$ (Acc.)	19.16	23.64	28.36	31.37	33.51	33.49	34.94	35.63	37.57	35.60	36.13	37.86	37.42
	SuperGPQA (Acc.)	10.43	12.63	14.66	15.33	15.04	16.57	17.68	17.84	19.89	19.44	19.61	20.87	20.17
Coding	HumanEval+ (Pass@1)	6.71	10.37	9.76	13.41	18.29	14.02	18.90	20.12	19.51	12.80	15.85	16.46	18.29
	MBPP+ (Pass@1)	14.29	28.04	35.71	34.39	36.77	45.77	45.50	45.24	53.70	52.65	46.30	54.50	52.12
Mathematics	GSM8K (Pass@1)	4.40	16.07	23.50	29.49	38.59	39.50	45.19	43.44	46.70	50.72	50.04	49.81	53.83
	MATH (Pass@1)	5.40	10.50	14.90	15.50	16.30	19.40	21.60	21.90	23.00	22.65	23.95	26.35	26.25
Logic Reasoning	BBH (EM)	27.77	31.47	32.64	36.17	34.47	37.50	39.58	41.12	42.94	40.14	44.44	44.40	45.04
Knowledge	NaturalQuestions (EM)	8.50	15.12	17.40	19.28	21.88	22.22	22.66	24.35	22.71	23.43	24.77	24.18	23.68
	TriviaQA (EM)	20.83	30.97	37.08	44.31	42.64	48.06	51.67	53.19	53.06	54.31	55.83	55.83	56.67
Commonsense Reasoning	HellaSwag (Acc.)	50.90	59.70	64.03	66.00	67.20	67.90	68.47	70.07	69.83	70.50	71.13	71.60	71.20
	SIQA (Acc.)	33.47	44.37	48.82	51.38	52.30	50.97	53.94	54.04	55.53	58.65	59.52	58.91	58.65
	PIQA (Acc.)	71.06	73.56	75.90	76.17	78.02	77.15	77.09	78.07	77.53	76.88	78.89	77.42	78.62
	WinoGrande (Acc.)	52.41	56.04	57.06	58.56	59.59	58.80	60.85	60.77	61.01	62.90	62.90	63.22	61.09
Average		24.16	30.87	34.70	37.12	38.91	40.10	42.06	42.82	43.93	43.91	44.55	45.47	45.75

Table 8: **PAD language scores across various N.**

	Model Sizes (N)	A71M	A128M	A340M	A590M	A874M	A3B
Aggregate	MMLU-Redux (Acc.)	25.44	30.68	38.51	45.79	47.60	57.75
	MMLU-Pro (Acc.)	11.18	11.63	13.22	16.48	16.70	25.49
	$\text{AGIEval}_{\text{en}}$ (Acc.)	21.42	21.73	24.94	27.60	30.52	37.42
	SuperGPQA (Acc.)	8.39	9.90	12.65	13.55	13.83	20.17
Coding	HumanEval+ (Pass@1)	7.32	7.93	10.98	16.46	13.41	18.29
	MBPP+ (Pass@1)	8.99	12.17	22.75	31.48	35.98	52.12
Mathematics	GSM8K (Pass@1)	1.82	5.16	13.34	22.52	27.37	53.83
	MATH (Pass@1)	3.60	4.20	6.65	12.80	16.90	26.25
Logic Reasoning	BBH (EM)	25.55	28.09	30.68	31.66	35.26	45.04
Knowledge	NaturalQuestions (EM)	7.76	9.06	11.55	16.76	18.23	23.68
	TriviaQA (EM)	14.72	18.75	26.39	35.00	39.31	56.67
Commonsense Reasoning	HellaSwag (Acc.)	41.87	47.43	56.40	61.83	65.00	71.20
	SIQA (Acc.)	32.34	35.82	42.78	51.33	54.81	58.65
	PIQA (Acc.)	68.28	69.04	73.18	74.43	75.52	78.62
	WinoGrande (Acc.)	50.75	52.17	54.77	57.70	59.20	61.09
Average		21.96	24.25	29.25	34.36	36.64	45.75

Table 9: **A3B IMG spatial scores across various D.**

	# Pre-Trained Tokens	25B	50B	75B	100B	125B	150B	175B	200B	225B	250B	275B	300B	325B
Spatial Relations	StepGame	16.34	19.31	21.07	24.62	27.58	26.99	27.02	29.59	29.45	29.04	29.68	31.56	33.11
	SpaRTUN	19.97	26.72	24.69	29.19	27.83	31.30	27.44	30.37	27.64	32.22	32.34	32.25	32.17
	bAbI(positional reasoning)	47.49	47.85	47.05	48.52	52.07	54.19	53.79	54.27	51.20	52.31	59.41	51.36	58.41
	FoREST	39.18	42.62	45.78	46.11	50.10	48.01	46.03	46.91	49.54	46.01	47.00	48.65	52.41
Localization	SpartQA	22.93	33.09	32.35	31.86	31.81	33.55	29.61	29.41	27.62	33.38	31.01	33.35	31.78
	SpatialMap	20.91	26.45	29.01	28.91	33.45	33.30	33.27	33.10	31.47	34.25	35.21	32.22	33.25
	BBH(colored objects)	13.67	18.37	21.63	28.88	35.10	32.76	42.55	33.37	46.02	44.08	43.78	48.06	54.90
Navigation	SpatialUnderstanding	23.36	23.90	23.82	25.77	27.68	30.15	31.77	31.72	35.86	34.43	35.14	38.52	38.54
	MazeNav	16.32	36.99	40.42	43.52	49.80	38.36	42.94	45.15	42.44	41.59	45.70	51.45	50.73
	bAbI(path finding)	1.48	6.81	5.20	7.14	5.75	8.44	10.20	8.99	6.66	8.42	12.01	7.51	11.36
	BBH(navigate)	56.12	53.16	60.92	69.18	71.02	70.41	70.00	75.10	72.65	74.59	79.18	78.16	77.76
Geometry	BBH(geometric shapes)	24.49	27.15	30.53	34.84	49.18	40.88	51.13	35.66	30.74	41.70	56.25	42.01	70.80
	Geometry3K	14.56	22.53	24.41	23.41	24.83	25.25	24.16	24.12	21.90	28.31	26.89	29.36	25.46
Physical Commonsense	PIQA-spatial	38.32	51.41	55.11	54.24	54.24	56.80	58.07	56.66	54.79	56.11	54.65	51.41	61.18
Symbolic Planning	Blocksworld	3.26	6.31	10.44	10.64	12.98	7.54	22.05	14.41	14.36	21.44	10.90	21.84	27.60
Average		22.56	28.81	31.32	32.90	35.94	34.59	38.37	35.57	34.89	38.24	38.15	38.61	44.25

Table 10: **IMG spatial scores across various N.**

	Model Sizes (N)	A71M	A128M	A340M	A590M	A874M	A3B
Spatial Relations	StepGame	15.46	14.20	16.97	20.64	21.67	33.11
	SpaRTUN	19.76	24.11	24.95	25.17	24.68	32.17
	bAbI(positional reasoning)	28.51	48.09	47.73	47.77	47.73	58.41
	FoREST	34.23	35.70	40.04	40.90	47.75	52.41
Localization	SpartQA	25.93	33.35	32.72	31.01	30.58	31.78
	SpatialMap	8.78	18.61	22.19	24.35	27.63	33.25
	BBH(colored objects)	12.55	12.45	14.49	21.94	25.20	54.90
Navigation	SpatialUnderstanding	14.64	19.05	17.52	27.15	22.37	38.54
	MazeNav	11.59	22.35	24.12	26.70	32.73	50.73
	bAbI(path finding)	7.64	6.93	8.52	9.52	6.83	11.36
	BBH(navigate)	46.12	53.27	53.47	60.10	66.12	77.76
Geometry	BBH(geometric shapes)	22.23	29.00	15.47	12.70	30.23	70.80
	Geometry3K	18.04	12.75	16.90	21.77	21.52	25.46
Physical Commonsense	PIQA-spatial	11.91	46.94	51.87	51.23	52.19	61.18
Symbolic Planning	Blocksworld	0.15	0.46	0.51	10.08	8.30	27.60
Average		15.41	24.28	25.01	28.13	30.27	44.25

Table 11: **A3B PAD spatial scores across various D.**

	# Pre-Trained Tokens	25B	50B	75B	100B	125B	150B	175B	200B	225B	250B	275B	300B	325B
Spatial Relations	StepGame	15.72	19.23	20.70	21.46	26.37	27.28	26.95	27.72	29.63	31.77	31.73	29.41	30.02
	SpaRTUN	21.41	25.37	25.14	26.45	28.83	29.83	30.60	30.49	31.64	31.95	32.33	31.20	33.32
	bAbI(positional reasoning)	46.53	47.81	47.89	50.32	48.41	51.59	52.43	49.96	52.07	50.48	52.19	57.50	52.75
	FoREST	33.95	46.01	46.95	44.89	49.06	51.50	48.79	48.21	51.64	50.10	49.41	49.00	48.69
Localization	SpartQA	31.23	31.96	29.11	30.20	31.68	29.00	30.20	29.53	30.12	29.35	28.84	30.03	28.97
	SpatialMap	16.22	22.77	28.18	25.95	30.07	31.04	30.62	30.44	27.76	31.19	30.74	34.05	31.92
	BBH(colored objects)	15.20	18.78	20.92	25.92	23.88	29.90	33.98	36.53	36.73	37.65	34.80	42.24	39.59
Navigation	SpatialUnderstanding	20.88	23.08	27.38	26.79	23.77	26.86	28.47	31.98	33.45	30.87	32.02	33.69	32.40
	MazeNav	15.54	31.13	44.87	38.94	38.61	35.24	23.05	38.46	34.91	35.66	32.36	46.40	46.35
	bAbI(path finding)	6.51	4.12	2.96	4.12	7.36	7.89	10.60	8.54	10.13	7.54	15.88	12.96	12.41
	BBH(navigate)	51.22	51.53	59.90	64.49	63.47	68.88	66.63	67.65	73.37	65.51	67.24	76.53	76.73
Geometry	BBH(geometric shapes)	8.30	27.87	32.38	39.75	57.58	48.05	42.11	52.15	44.47	42.93	45.49	48.67	60.35
	Geometry3K	12.75	15.10	18.04	22.27	21.94	24.33	20.64	22.15	21.77	26.17	28.23	24.37	27.68
Physical Commonsense	PIQA-spatial	38.73	51.64	53.06	52.46	53.70	57.39	56.48	57.30	56.25	58.62	58.58	55.29	57.53
Symbolic Planning	Blocksworld	1.78	3.51	8.20	6.92	4.79	10.34	12.27	11.30	9.57	10.18	10.59	5.86	21.03
Average		20.81	27.20	30.25	31.19	33.04	34.78	33.93	35.61	34.95	35.34	35.96	36.21	39.87

Table 12: **PAD spatial scores across various N.**

	Model Sizes (N)	A71M	A128M	A340M	A590M	A874M	A3B
Spatial Relations	StepGame	13.50	13.51	17.14	20.94	22.34	30.02
	SpaRTUN	16.03	19.69	25.82	24.59	25.98	33.32
	bAbI(positional reasoning)	47.33	47.05	48.37	48.41	48.92	52.75
	FoREST	27.69	36.37	45.08	42.93	47.61	48.69
Localization	SpartQA	22.81	26.62	31.80	31.03	32.33	28.97
	SpatialMap	8.32	18.33	18.88	24.85	26.63	31.92
	BBH(colored objects)	7.04	14.18	19.80	19.08	23.27	39.59
Navigation	SpatialUnderstanding	13.07	17.67	20.42	21.84	25.41	32.40
	MazeNav	11.71	17.77	21.25	21.30	24.47	46.35
	bAbI(path finding)	4.40	5.15	6.98	6.73	11.11	12.41
	BBH(navigate)	33.06	45.51	58.37	58.57	59.29	76.73
Geometry	BBH(geometric shapes)	18.44	11.68	15.27	14.96	19.26	60.35
	Geometry3K	6.25	9.48	12.88	19.92	21.77	27.68
Physical Commonsense	PIQA-spatial	23.68	38.91	51.00	50.87	51.78	57.53
Symbolic Planning	Blocksworld	0.15	4.07	0.76	7.18	8.30	21.03
Average		15.10	20.66	25.03	26.97	29.05	39.87