FEINER: PRESERVE, PROJECT, AND LET THE DECODER INTEGRATE VISUAL INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Extending language models to reason over additional modalities is an important step towards general-purpose AI models. Since not all modalities can be efficiently trained natively, integrating multimodal information into pre-trained language decoders is an important task. In this work, we focus on vision, one of the most important non-text modalities, and use LLaVA-style vision-language models to study how information from vision encoders can be preserved and utilized in pre-trained decoders. To this end, we introduce a framework for efficient, controlled training and evaluation. Our resulting design preserves strong modality-specific representations and effectively projects them into the decoder's latent space to increase the availability of vision-specific information. We find that representation preservation and decoder-side information flow are complementary requirements for strong visual understanding. Our *feiner* approach rivals state-of-the-art models at comparable scale while requiring less than half the compute. These results provide an efficient basis for studying how pre-trained decoders can incorporate additional modalities.

1 INTRODUCTION

Modular Vision-Language Model (VLM)s are a dominant paradigm in current AI: images are encoded into semantically meaningful embeddings, projected into the text embedding space, and concatenated with language tokens for decoding. This modular design enables the reuse of powerful pre-trained vision encoders and large language models, dramatically reducing training cost (Alayrac et al., 2022; Liu et al., 2023b). However, despite their strength, their strong overall performance often hides failures on basic tasks such as visual perception and spatial reasoning (Tong et al., 2024b; Wang et al., 2024a; Shi et al., 2026). In fact, VLMs frequently generate responses that *do not causally depend on the vision input*, even when informative visual representations should be available from a strong encoder (Fu et al., 2025; Nikankin et al., 2025; Vo et al., 2025). Interpretability analyses indicate that models predominantly attend to language tokens, rendering vision tokens redundant throughout large parts of the decoder (Zhang et al., 2025; Kaduri et al., 2025). Together, these findings expose a gap between the *perceived availability* of visual information in the network and its *utilization* in the shared decoder.

Image patches and their embeddings are not simply "words in another vocabulary": their spatial and relational structure carries information that may not survive projection into a language model's embedding space or end-to-end multimodal training. It stands to reason that model performance thus depends on preserving the information encoded by the vision encoder, and the decoder being able to organize this information for generation. Improving visual understanding in VLMs can thus be approached from two sides: (1) improving the structure of visual information to facilitate later use together with text and (2) improving utilization within the decoder to learn and reason over a truly multimodal embedding space. While the former has been a primary target in contrastive learning (Zhang et al., 2020; Radford et al., 2021), there has been recent debate if the resulting semantic embeddings lose necessary information (Tong et al., 2024b; Liang et al., 2022). To solve the latter, there has been no consensus so far and instead, recent approaches increasingly focus on natively pre-training decoders from scratch (Qwen Team, 2026; Han et al., 2026; Wu et al., 2026). This seems to alleviate issues with language bias but it comes at a significant training cost that increases proportionally to the number of additional modalities. We contend that better understanding how

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.

to unify pre-trained embedding spaces from different modalities is not only theoretically interesting but will also lead to more efficient and accessible alternatives for scaling multimodal models.

For a simple shared decoder architecture to learn strong multimodal representations and allow for scaling compute, model size and number of modalities, we need to consider its two tasks: (1) *representation preservation* such that multimodal training does not deteriorate the structure and information within each modality, and (2) *information integration* to allow the decoder to bind and retain information from a new modality once it becomes available. In this work, we study these questions in a controlled decoder-focused VLM framework. We first measure how multimodal training changes visual representations and isolate this effect by freezing the encoder. We then target the projector to both preserve relational and global properties of the visual embedding space while transitioning to text space. Finally, we introduce lightweight register tokens that provide the decoder with an additional workspace for integrating visual evidence without requiring targeted interventions at individual components of the decoder. Our main goal is to provide a minimal set of interventions that retain the information of the pre-trained modalities while leaving the decoder free to learn how the knowledge should be used. Because our methods depend on general representation structure, we expect them to transfer to arbitrary pre-trained modalities beyond vision, which we leave to future work. We moreover introduce a systematic framework to efficiently train and evaluate architectural choices to improve reproducibility and directly assess visual understanding. Our work provides evidence for the following approach to modular multimodal learning: *preserve modality-specific information, project it to the decoder's embedding space, and leave its utilization fully to the decoder.* Our paper makes three core contributions:

1. **Structured evaluation of visual understanding.** We define a robust set of evaluation metrics for visual understanding based on pair-level image-caption comparisons that minimize within-modality and cross-modality bias.
2. **Targeted analysis of visual representation degradation.** We separate changes in the pre-trained encoder from information loss at the encoder–decoder interface, evaluating how each affects visual understanding.
3. **Structure-preserving approach to modality alignment.** We show that keeping the encoder fixed and preserving its outputs' relative structure propagates stronger vision tokens into the decoder than unconstrained feature projection.

2 Related Work

Most modern VLMs follow a modular design in which a pre-trained vision encoder is connected to a pre-trained language model through a lightweight projector or connector (Alayrac et al., 2022; Liu et al., 2023b). This recipe has proven highly effective and scalable (Laurençon et al., 2024; McKinzie et al., 2024), with recent improvements stemming from stronger image encoders, better data mixtures and improved training recipes (Steiner et al., 2024; Bai et al., 2025; Clark et al., 2026). Multi-scale and sub-image approaches are often done in accordance with other improvements to reduce the overhead of higher-resolution inputs necessary for fine-grained perception (Lin et al., 2023; Zhang et al., 2024). A special version of common VLMs are encoder-free (Diao et al., 2025; Luo et al., 2024) and image-text generative models (Zhou et al., 2025; Shi et al., 2025; Zhuang et al., 2025). We focus on modular VLMs with vision encoders and a single text-generative decoder to isolate the issue of unifying representation of different modalities.

Approaches for connecting vision encoder outputs to the decoder range from one-to-one projections to resamplers that select and compress vision tokens before integrating them with text (Alayrac et al., 2022; Li et al., 2023a). This compression inherently trades off keeping or discarding local and fine-grained information, motivating approaches that attempt to find an optimal middle ground (Cha et al., 2024; Li et al., 2024). Other approaches focus on increasing language compatibility through earlier pre-alignment (Yu et al., 2026; Chen et al., 2026). Generally, connectors are often used for information selection, be it through query-based selection or text pre-conditioning. We instead focus on preserving information, leaving the decoder responsible for deciding which information is relevant in its multimodal representation for generation.

The availability of vision information is, however, not sufficient for strong VLM performance. Vision tokens can remain interpretable and directly decodable but with little effect on the final pre-

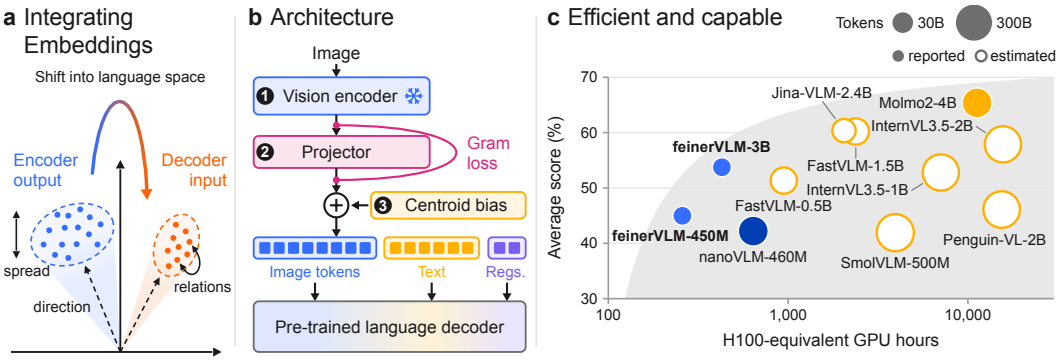

Figure 1: Overview of our method. We preserve embedding structure while projecting into language space and allow the pre-trained decoder to handle both modalities effectively (**a** and **b**). Our resulting model, *feiner*, is Pareto-optimal with efficient training and strong performance on vision benchmarks compared to other open models (**c**).

diction (Fu et al., 2025; Krojer et al., 2026) and their information follows a structured but uneven path through the decoder (Kaduri et al., 2025; Zhang et al., 2025). These findings indicate a distinction between the availability of visual information from its effective use. Prior work addresses this gap through targeted interventions such aligning selected decoder heads with an external vision teacher (Caffagni et al., 2026), adding cross-attention at selected layers (Mahajan et al., 2026), resampling encoder features between decoder blocks (Feng et al., 2025), or steering the model during inference (Guo et al., 2026; Yin et al., 2026). We instead focus on driving multimodal decoder alignment and minimizing targeted interventions, as the decoder should have access to as much information as possible in a readily processable form and with sufficient workspace to learn from it.

Understanding and improving how modality-specific embedding spaces best interact remains intrinsically interesting. Decoder-side interaction between vision and language remains an open problem for improved multimodal understanding. We build on these directions by studying vision in VLMs through the joint lenses of compact visual encoding and effective utilization in the decoder. Our goal is to avoid specialized interventions, instead relying on an efficient, general architecture that scales.

3 Principled Investigation into Vision for VLMs

We study visual representations and their integration in the decoder with a compact modular VLM. This is a deliberately difficult setting: limited model capacity makes it harder to preserve and use information from two pre-trained modalities. At the same time, complete training runs remain cheap enough to test a broad set of interventions and to use the results of early ablations to refine later experiments. This combination makes the setting useful for controlled empirical analysis. We describe the model architecture and establish the experimental baseline in Sec. 4.1. Here, we describe the training data, implementation, and evaluation protocol shared by all comparisons.

3.1 Training Data Curation

We construct *feiner-train* as a small yet reasonable dataset for fast ablations. It is based on FineVision (Wiedmann et al., 2025b) and FineWeb-Edu (Penedo et al., 2024) with fully publicly available data. We filter FineVision based on its image-text correspondence rating, selecting only samples with a score ≥ 4 for learning image-text relations efficiently. Images have substantially more preprocessing and attention cost compared to text sequences, we thus train on single image samples only to reduce and control compute overhead. Furthermore, recent data-centric work shows that single-image training can transfer well to multi-image tasks (DatologyAI et al., 2026). We reject images with a short edge below 80 pixels and resize accepted images to a minimum short edge of 320 pixels and a maximum long edge of 1024 pixels. This further reduces vision overhead while allowing reasonable training, following previous works (Steiner et al., 2024; Laurençon et al., 2024). During training, we use at most two answer turns per image with a maximum of 128 question-and-answer tokens per turn under the SmolLM2 tokenizer (Allal et al., 2025). Answer tokens are the

only tokens in our vision-language pairs that contribute to the language-modeling loss. Thus, this aims at preventing samples with long answers to dominate training and improves packing efficiency and reduce the need to trade image resolution against text length. Samples with long caption answers are trimmed at sentence boundaries before applying this token limit.

This results in 6,030,583 visual question-answer pairs. We including additional data from FineWeb-Edu to have 20% of the total answer-token budget allocated to text-only data. This is done by splitting FineWeb-Edu documents into chunks of 128–256 tokens, which bounds their packing cost and prevents a small number of long documents from dominating the text-only budget. The text-only data helps the decoder to preserve its language capabilities (McKinzie et al., 2024; Marafioti et al., 2025). In total, we end up with $6.2M$ samples and $208M$ loss-generating tokens. All datasets with their preparation scripts will released together with the codebase.

3.2 Efficient Multimodal Training

We use knapsack packing within a bounded look-ahead buffer to maximize decoder-token utilization. The packing constraints bound sequence length and visual attention cost, which we approximate by the sum of squared patch counts of the packed images. Sequence packing follows prior work on separating examples with block-diagonal attention masks (Krell et al., 2022). Our native-resolution encoder follows the image-packing of NaViT (Dehghani et al., 2023). For samples with turns we use tree-structured masks to encode the shared image once and keep the question-answer branches of each turn independent. Position indices restart for each example and each independent turn. Thus, packing changes neither the causal context nor the loss assigned to a sample.

We compile the vision encoder and language decoder, using FlexAttention (Dong et al., 2024) to support the block masks required for sequence packing. Sequences are buffered into pre-determined sequence length buckets to minimize compilation overhead. Fixed-resolution vision encoders use the standard compiled forward pass. Native-resolution encoders use a segment-aware compiled path that keeps attention separate between packed images. Unless stated otherwise, all paper runs use 8 $\times$ H100 80 GB GPUs. A complete *feiner-train* run of our initial nanoVLM-like model takes 2 hours and 52 minutes on 8 $\times$ H100 GPUs, equal to 22.9 allocated H100 GPU-hours. For comparison, nanoVLM reports approximately 640 H100 hours on full FineVision (Wiedmann et al., 2025b).

3.3 Evaluation Protocol

Evaluating VLMs is difficult to their free-form generation and varying influence of both modalities. However, full benchmark suites are expensive and a evaluation signals during training valuable for iterative improvements. We thus construct *feiner-eval* to measure improvements in compositional image discrimination as well as language retention without relying on free-form generation. Following prior likelihood-based image-text scoring (Lin et al., 2024; Tan et al., 2024), we use image-caption compatibility as a binary next-token decision task with *Yes–No* log-odds as the score. For a response, the score r is compute by taking a sample with image x_i and asking the model whether a caption c_j describes the image

$$r(x_i, c_j) = \log \sum_{y \in \mathcal{Y}} p(y \mid x_i, c_j) - \log \sum_{n \in \mathcal{N}} p(n \mid x_i, c_j)\,. \tag{1}$$

Here, $\mathcal{Y}$ and $\mathcal{N}$ contain all valid single-token capitalization and spacing variants of "yes" and "no". To improve evaluation robustness by minimizing cross-modality bias, we calibrate each score with the corresponding text-only response to remove language bias: $s_{ij} = r(x_i, c_j) - r(\varnothing, c_j)$. This procedure reduces generation-dependent variation and caption-specific response priors. It is therefore more stable for small models that have not received extensive instruction tuning.

We compute calibrated log-odds for two datasets, Winoground (Thrush et al., 2022) with 400 image-caption groups and BiVLC (Miranda et al., 2024) with 2,933 groups. Each group contains two images and two captions with one correct one-to-one assignment. We use two calibrated scores, with our main visual measure being *relative binding* R. For each group g in the dataset D, decompose its score matrix into a binding contrast b, a caption preference c, and an image preference i:

$$\begin{aligned} b &= s_{00} + s_{11} - s_{01} - s_{10}, \\ c &= s_{00} + s_{10} - s_{01} - s_{11}, \\ i &= s_{00} + s_{01} - s_{10} - s_{11} \end{aligned} \qquad R_D = \frac{100}{|D|} \sum_{g \in D} \frac{b_g}{\sqrt{b_g^2 + c_g^2 + i_g^2}}\,. \tag{2}$$

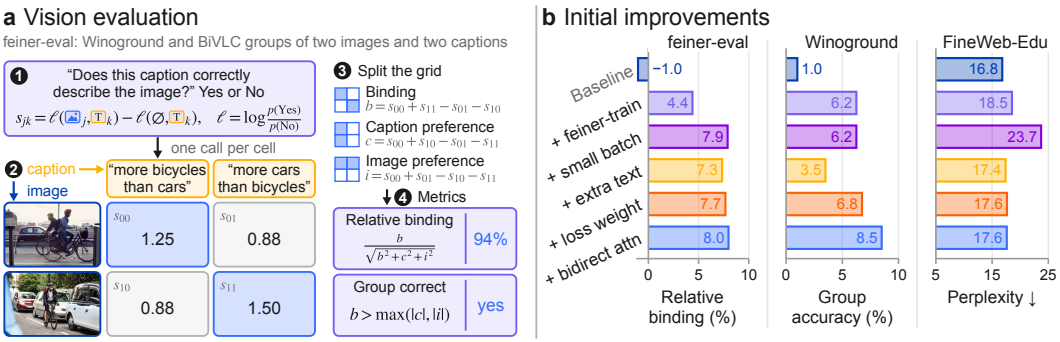

Figure 2: Overview of train-time visual understanding evaluation with *feiner-eval* (**a**) and cumulative improvements over the base architecture through *feiner-train* and additional interventions (**b**).

We compute binding for both datasets individually and then average. Relative binding lies in $[-100, 100]$ with zero being random/neutral. Intuitively, to score highly on binding, the model must match the correct image and caption over two similar but distinct images with different captions. Unlike standard binding $b > 0$ (Zhu et al., 2026), relative binding gives a smaller score when an image or caption modality preference dominates the 2×2 evaluation matrix in Fig. 2**a**. We also report the stricter *group accuracy*, which require $b > |i|$, $b > |c|$, and $b > \max(|c|, |i|)$, respectively. Compared to relative binding, group accuracy is more susceptible to bias from generated negatives samples in BiVLC compared to its natural images and captions. We study this generated bias and its effect on the four metrics in App. A.2. Group accuracy on Winoground remains a demanding secondary measure because all four local choices must be correct. For language retention, we compute perplexity on 1,500 randomly selected, held-out FineWeb-Edu documents. These documents use the same 128–256 token chunking policy as the training text and give 6,114 sequences, or approximately 1.5M next-token predictions. We use *feiner-eval* for experimental comparisons, and the final model is benchmarked on the broader downstream suite described in Sec. 4.5.

4 Experiments

We organize the experiments around three stages of visual information flow. First, we study the features supplied by the vision encoder and how multimodal training changes them. Second, we measure how projection changes patch relations and test whether preserving these relations improves visual understanding. Third, we study how the decoder accesses the available visual states and whether registers with a later image read improve their use.

These stages distinguish the quality of the visual source, the preservation of its structure at the interface, and its use for prediction. We first establish a common training recipe, then pair task comparisons with internal measurements at each stage. A broader benchmark suite then tests the selected system beyond the ablation tasks.

4.1 Baseline Improvements

Before studying the visual pathway, we establish a strong yet efficient baseline that can learn image-text relations while retaining language capability. We start from the nanoVLM model (Wiedmann et al., 2025a) with a pretrained SigLIP2-B/16 encoder (Tschannen et al., 2025), a SmolLM2-360M-Instruct decoder (Allal et al., 2025), and a newly initialized linear projector. Pixel shuffle groups each 4×4 patch neighborhood before projection. The roughly $460M$-parameter model uses fixed-resolution image views and image tiling. Training on *feiner-eval* with the base configuration raises relative binding from random to 4.39.

Training Configuration. Following Marek et al. (2025), we reduce the batch size by adapting the second moment β_2 of Adam. This allows maximizing throughput without gradient accumulation and provides a wider range of close-to-optimal learning rates. In contrast to Marek et al. (2025), we compute β_2 based solely on the number of loss-generating tokens, which is a smaller part of the full sequence in VLM vs. Large Language Model (LLM) pre-training. As common for VLMs, the pre-

trained decoder and vision encoder have a learning rate reduced by $10\times$ and $100\times$ compared to the vision bridge. We use a constant learning rate after warmup and apply layer-wise learning rate decay in the vision encoder following Siméoni et al. (2025). More details in App. A.1. These changes improve performance but increase FineWeb perplexity from 18.53 to 23.72. To retain language capabilities, we interleave text-only tokens from FineWeb-Edu at ∼20% of the output-token budget, which substantially reduces perplexity without negatively affecting visual understanding.

Loss Weighting and Bidirectional Attention. We ablate previously proposed changes to aid integrating vision in the language decoder. Per-token loss averaging lets long-answer examples dominate the gradient over short multiple-choice answers often seen in visual question answering. Weighting the answer loss by $1/\sqrt{n}$ for n answer tokens (Clark et al., 2026) balances contributions across output lengths and improves visual understanding. The attention mask in pre-trained decoders is usually causal across the whole token sequence. Allowing bidirectional attention among vision tokens while keeping causal attention over text (Zhou et al., 2025; Clark et al., 2026) yields further improvements at negligible compute overhead. This results in our final baseline model with *relative binding* of 8.0 and Winoground *group accuracy* of 8.5%. The improvements from the aforementioned interventions are compared in Fig. 2**b**.

4.2 Increasing Availability of Visual Information

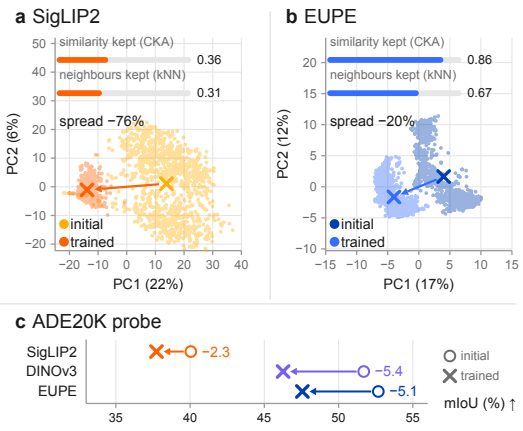

Figure 3: Pre-trained features are vulnerable: multimodal training contracts patch representations. EUPE retains more structure than SigLIP2, while all lose ADE20K probe performance.

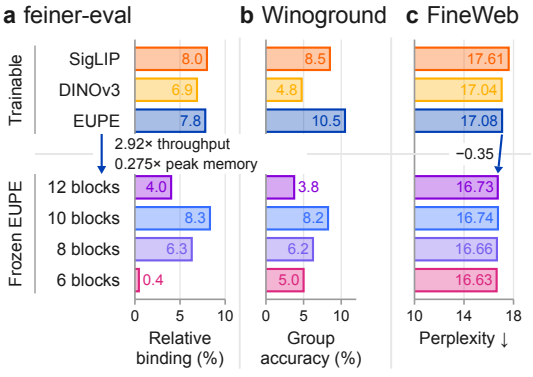

Figure 4: Frozen features need the right readout. Freezing EUPE reduces training cost and perplexity but weakens visual performance. Reading block 10 recovers much of the visual loss.

Contrastive image-text pre-training makes both modalities semantically comparable, but risks losing fine visual understanding (Tong et al., 2024b). We thus compare our SigLIP2 baseline with vision-only DINOv3 and EUPE, which is distilled from vision-only and contrastive pre-trained teachers. We first examine how training the full VLM end-to-end, including the encoder, changes their vision embeddings. We compare each encoder to its initial state on the same image in Fig. 3. We use linear centered kernel alignment (CKA (Kornblith et al., 2019)) to measure overall patch relation preservation and k-NN overlap (k=16) to measure local neighborhood retention. Fixed-resolution SigLIP2 has pre/post CKA of 0.356 and adjusted neighbor retention of 0.308. Its spread, the root-mean-square deviation from the patch centroid within each image view, falls by 76.42%. DINOv3 and EUPE preserve more structure, with EUPE retaining CKA of 0.859 and a reduction in spread by 19.95%.

Structural change alone cannot tell us whether visual information remains accessible. We therefore fit separate linear probes to the initial and trained features on ADE20K (Zhou et al., 2019), a dataset with pixel-level semantic labels, following the segmentation protocol in DINOv3. Initial-to-trained mean intersection over union (mIoU) changes from 40.05 to 37.74 for SigLIP2, 51.69 to 46.26 for DINOv3, and 52.67 to 48.15 for EUPE. All encoders thus lose access to dense labels during training. Retention of the initial geometry here does therefore not prevent a loss in visual understanding.

In addition to better probe performance on ADE20K, EUPE also exceeds the common SigLIP2 setup's 8.50% group accuracy as

shown in Fig. 4. SigLIP2 retains higher relative binding in the trainable comparison, but we select EUPE for its combination of strong dense label performance and the highest group accuracy. To further prevent encoder drift, we freeze EUPE, resulting in a constant yet unaligned "reference" vision embedding during training. Surprisingly, freezing reduces perplexity from 17.08 to 16.73 but it comes at the cost of visual understanding, with Winoground group accuracy falling from 10.50% to 3.75%. Freezing thus protects the source embedding but initially weakens visual performance.

As shown by Bolya et al. (2025), embedding quality can peak before the encoder's final layer. We therefore test whether an intermediate readout improves visual performance while allowing us to skip the remaining encoder blocks. We compare frozen EUPE outputs at blocks 12, 10, 8, and 6. Reading out at block 10 instead of block 12 restores binding from 4.01 to 8.31 and Winoground group accuracy from 3.75% to 8.25%. Further reducing the number of encoder blocks performs worse on *feiner-eval*. We thus choose to freeze EUPE with block 10 and use it from now on, as it exhibits the best visual information availability.

4.3 Preserving Vision Information and Improving Integration

Freezing protects the encoder, but the projector can still change the relations between the patches of the input. We hold the visual source fixed and compare the interface on the same 256 images in Fig. 5**a**. The linear projector reduces mean within-image embedding cosine distance by more than 16%. The image patches thus have a larger shared component and are more difficult to separate based on their angular direction even though the encoder itself no longer changes. To preserve features between patches, we test whether a more flexible mapping helps by replacing the linear projector with a two-layer MLP. We follow Liu et al. (2026), using GELU activation and MLP dimensions equal to the decoder hidden dimension, applied after our own pixel-shuffle mapping. The MLP increases mean cosine distance to 0.426 and improves binding from 8.31 to 12.75. However, better patch separation does not necessarily mean structure preservation: adjusted neighbor retention falls from 0.760 to 0.656, and CKA from 0.890 to 0.796. The MLP therefore helps the task while changing more of the encoder output structure.

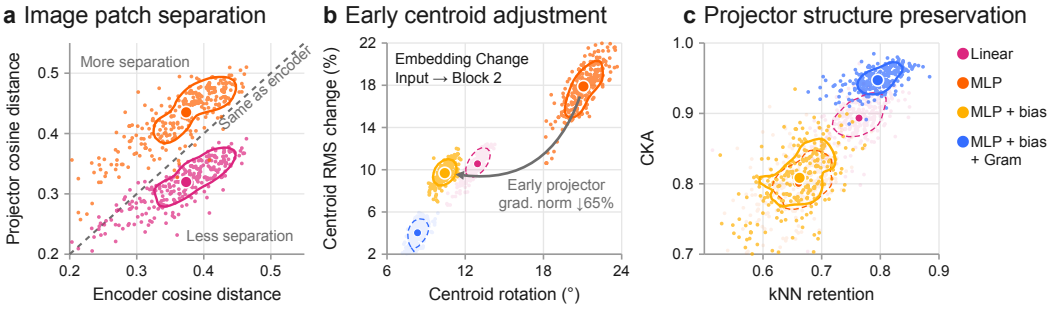

Figure 5: The MLP increases patch separation **(a)**, text-based bias reduces early decoder centroid adjustment **(b)**, and Gram regularization restores patch-level encoder output relations **(c)**.

To determine how the decoder initially transforms these projected states, we measure centroid changes through its first two blocks at the final checkpoints (Fig. 5**b**). MLP image centroids rotate by 21.07°, compared with 12.84° for linear projection, with centroid RMS growing by 17.72% up from 10.46%. Further analysis shows that the MLP's average centroid shift tends toward the same direction of the text embedding centroid. This motivates supplying an offset explicitly to alleviate pressure on the MLP and or early decoder blocks. We add a learned bias to every projected image token to test whether a text-based starting location helps. We initialize the learned bias as the difference between the projector output centroid and the text centroid mid-decoder when the model reasons across modalities. This sets a global reference location without directly changing the relative information in the vision embedding. With this initialization, early centroid rotation falls to 10.43° and RMS increase within the first two decoder blocks to 9.69%. This is also visible in the gradient update. The mean gradient norm for the projector over the first 100 updates falls by 64.9%. Neighbor retention and CKA remain close to the plain MLP, while *feiner-eval* binding rises to 15.69. The bias helps binding and in early decoder layers, but leaves the change in encoder output structure largely unresolved.

Following DINOv3, we use Gram regularization to preserve patch relations during projection. The frozen and shuffled encoder tokens are the Gram teacher and get matched with the projector outputs before adding the learned bias. For each visual example, we use row-normalized encoder output X_E and projected tokens X_P and define $\mathcal{L}_{\text{Gram}} = {}^1\!/_{m^2} \left\| X_P \cdot X_P^\top - \text{sg}(X_E \cdot X_E^\top) \right\|_{\text{F}}^2$, and $\mathcal{L} = \mathcal{L}_{\text{text}} + \lambda \mathcal{L}_{\text{Gram}}$, with $\lambda = 1$ and up to $m = 512$ corresponding tokens per image. The encoder is detached, and the loss acts before the additive bias and before the first decoder layer. This Gram regularization constrains pairwise relations while allowing a change of feature dimension and coordinate orientation. Adding Gram raises neighbor retention to 0.794 and CKA to 0.947, while bringing separation to 0.370, close to the reference value. Early centroid rotation and RMS growth fall further to $8.33°$ and 3.93%. Binding rises from 15.69 to 17.23, and BiVLC group accuracy from 20.83% to 22.43%, with a small Winoground group decrease from 10.50% to 10.25%. Gram thus restores embedding structure and further reduces early centroid change, leaving the decoder free to transform features after projection (Fig. 5**c**). In summary, the aforementioned interventions effectively preserve vision information and aid its intergration.

4.4 Improving Multimodal Information Flow in the Decoder

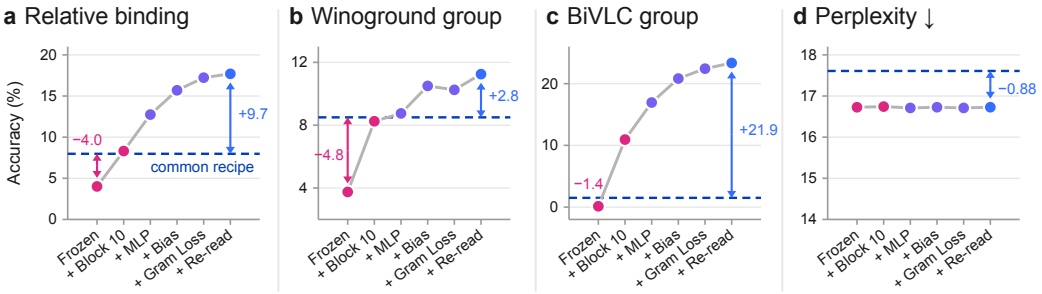

Figure 6: Our targeted interventions recover and further improve visual performance. The dashed line marks the initial trainable common recipe.

Preserving patch relations at the interface leaves open how the decoder uses the resulting embedding. Following Kaduri et al. (2025), we therefore measure answer attention in the final decoder block during caption prediction. We use teacher-forcing for both correct image-caption pairs per group with the instruction "Describe the image" for all 400 Winoground groups. Although the fixed instruction supplies little task-specific content, our results in Tab. 1 show that our model without registers assigns similar attention shares to image tokens and the instruction, 9.20% and 9.48%.

Table 1: Decoder attention at final block 32 from all Winoground captions. Chat/formatting tokens are excluded. Re-read uses four registers.

	Registers		Re-read			
Attention (%)	None	4×	B20	B24	B28	B32
Image	9.20	6.44	7.48	8.55	9.09	8.47
Text	9.48	4.71	4.79	6.83	4.13	8.09
Registers	0.00	10.18	9.49	3.75	4.79	6.32
Early answer	81.32	78.67	78.24	80.87	81.99	77.13
Rel. binding	17.23	14.51	11.59	15.55	**17.70**	8.12
WG group (%)	10.25	8.50	7.50	9.75	**11.25**	8.00

Instead of requiring only text tokens to process visual information, we next test dedicated workspace tokens combining image and text information. We insert four learned *register tokens* after the instruction and before the answer. The registers receive more attention than the image and text tokens, resulting in a decrease in relative binding and Winoground group accuracy from 10.25% to 8.50%. The answer thus attends to the workspace, but this alone does not improve the measured visual performance. We thus introduce a *residual re-read*, i.e., a gated image-only attention operation added to the register tokens (see App. A.1). This increases final image-attention share to close to the no-register level and improves relative binding and increases Winoground group accuracy to 17.70 and 11.25%. We finally test whether the full sequence of changes compensates for the initial visual cost of freezing. Fig. 6 compares the cumulative improvements after the common recipe. Relative to the common recipe, the complete design gains 9.72 binding points, 2.75 Winoground group points, and 21.85 BiVLC group points, while perplexity falls by 0.88. We call the final model with all our aforementioned improvements *feinerVLM-450M*.

4.5 Evaluation on Standard VLM Benchmarks

Benchmarks with ground-truth answers provide the practical test of whether our improvements in integrating vision in a pre-trained decoder transfer to task performance. We propose *feiner-bench*, a curated set of established benchmarks spanning general multimodal understanding, vision-focused tasks, and language capability. The suite covers visual question answering, spatial relations, perception, and instruction following. Table 2 summarizes the results, with the six single-image BLINK subsets combined into one score. Evaluation settings and task protocols are given in App. A.2.

We train *feinerVLM-450M* on the broader FineVision mixture used by nanoVLM to properly compare the models outside of *feiner-eval*. Training starts from the pre-trained backbones and random projector initialization, with a budget of 1.67 epochs, similar to nanoVLM but without multi-image samples, which slightly reduces number of samples, especially documents. We also test whether the approach scales by replacing the SmolLM2-360M decoder with SmolLM3-3B for our larger *feinerVLM-3B*. We transfer the frozen EUPE encoder and the interface and register mechanisms directly, without any new ablations or benchmark-specific tuning before the full run. The MLP width is adapted to the decoder's 2,048 hidden dimension, and the decoder base learning rate is reduced from 10^{-4} to 3×10^{-5} (see App. A.2). Benchmarks are run with VLMEvalKit (Duan et al., 2024).

Table 2: Benchmark results on open models with $\leq 4B$ parameters. Scores are percentages (*higher is better*). Bold scores compare feinerVLM-450M with its base nanoVLM.

Benchmark Model size	feinerVLM 450M	nanoVLM 460M	SmolVLM 500M	FastVLM 0.5B	InternVL3.5 1B	FastVLM 1.5B	InternVL3.5 2B	Penguin-VL 2B	Jina-VLM 2.4B	Molmo2 4B	feinerVLM 3B
MMLU	33.40	**35.05**	33.38	30.98	36.03	36.01	41.50	40.11	43.26	47.34	28.70
IFEval	26.53	**27.44**	27.84	18.74	58.03	28.67	67.45	32.73	40.81	80.69	35.39
ChartQA	56.84	**57.72**	58.52	76.00	77.40	80.16	81.36	32.80	81.56	46.56	67.48
MMBench Eng.	**45.98**	45.36	48.76	64.32	67.96	73.37	75.08	46.75	76.32	82.89	61.38
MMMU-Pro	**16.94**	16.07	15.38	17.34	25.90	24.39	32.08	17.57	26.65	36.07	21.21
MathVista	**38.90**	36.80	38.90	44.70	11.40	61.60	22.50	66.50	60.30	57.70	53.70
GQA	**47.15**	46.94	45.31	63.76	50.88	65.00	54.89	43.94	53.68	54.76	53.90
AI2D	**58.42**	47.80	56.74	68.78	70.05	79.24	78.34	43.59	82.16	86.20	70.11
MMVP	**15.33**	**15.33**	16.00	30.67	31.33	44.00	43.33	12.67	49.33	72.67	25.33
SpatialEval	**31.89**	28.37	26.21	31.31	44.27	48.59	41.73	36.68	44.96	60.97	44.57
VLMsAreBiased	**18.79**	16.16	18.46	18.93	17.46	13.90	14.98	18.18	18.25	30.21	16.81
POPE F1	**86.80**	84.43	84.58	84.30	89.30	84.54	89.51	89.73	88.66	88.15	86.74
CV-Bench	**53.44**	51.23	46.44	59.90	67.83	74.36	72.69	60.94	65.03	79.24	60.55
VSR	**61.95**	59.57	55.65	59.57	64.65	75.37	69.31	65.79	76.51	78.48	72.50
BLINK (1 img)	**46.69**	42.90	43.19	46.54	48.66	49.90	54.17	48.93	53.42	62.86	52.17
CRPE Relation	**42.03**	33.21	28.25	43.98	57.07	54.33	65.63	46.52	63.62	64.68	59.19
Vision Mean	**44.37**	41.56	41.60	50.72	51.73	59.20	56.83	45.04	60.03	64.39	53.26

feinerVLM-450M improves on nanoVLM on seven of the eight vision-focused scores and ties on MMVP. Its largest gains include 8.82 points on CRPE Relation and 10.62 points on AI2D. ChartQA remains close, at 56.84% versus 57.72%, while both language scores are slightly lower. The average score for all 14 vision tasks is 44.37%, compared with nanoVLM's 41.56%. Scaling our model to $3B$ raises this mean to 53.26% and improves 13 of the 14 vision-task scores. feinerVLM-3B exceeds SmolVLM-2.2B and Penguin-VL-2B on this mean and is competitive with InternVL3.5-1B (51.73%), whose reported training corpus contains $380B$ stage tokens. These gains support transfer to a larger decoder, although MMLU decreases and models with more data and compute still tend to outperform. Fig. 1 places these results alongside training-token counts and H100 GPU-time measurements or estimates. Our two runs process $22B$ decoder tokens and use 258 and 428 H100 GPU-hours, respectively. The 450M run thus uses about 60% less GPU time than nanoVLM's reported 640 H100 GPU-hours, and even the 3B run stays below that budget. This lower training cost makes repeated experiments more accessible and leaves more compute for further improvements.

5 Discussion and Conclusion

Our results support an efficient path to visual-language learning: preserve a strong visual source, adapt its projection to retain patch relations, and give the decoder an explicit route to access information in the image embedding space. The resulting models transfer beyond the diagnostic suite and scale to a larger decoder at modest training cost. Document coverage and language retention remain limits. The framework provides a practical basis for further work on visual reasoning, training mixtures, and how decoders use visual information.

AI USE STATEMENT

In this work, we used generative AI tools to find relevant research we were not aware of, help with ideation and for feedback on hypotheses and content of the paper. Generative AI was used to implement and run experiments, as well as help improve visualization of these experiments. We have not used generative AI tools for creating synthetic data or automated research. We have reviewed all AI-assisted work through regular testing, especially for code written for our experimental setup. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Reproducibility is a key objective of this paper with all training, evaluation and benchmark data publicly available. As a counterweight to the compute demands of multimodal training, we hope that our efficient framework can enable more researchers to contribute to this area. All models are based on publicly available weights and framework. All hyperparameters are in Appendicex 4 and A.1. Due to previous negative experience with desk rejects for including anonymous repositories, the corresponding codebase for training, evaluation and benchmarking will be made available upon publication.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, et al. Smollm2: When smol goes big–data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Abdul Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Dollar, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=INqBOmwIpG.

Davide Caffagni, Alberto Compagnoni, Federico Melis, Sara Sarto, Pier Luigi Dovesi, Mark Granroth-Wilding, Marcella Cornia, and Lorenzo Baraldi. Mind the heads: Topological representation alignment for multimodal llms, 2026. URL https://arxiv.org/abs/2606.23885.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Linqing Chen, Hanmeng Zhong, Wentao Wu, and Peng Zhou. The dominance of text space: Unveiling the asymmetric nature of cross-modal alignment in large language models. In Maria

Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36666–36678, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1699. URL https://aclanthology.org/2026.acl-long.1699/.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *CoRR*, abs/2601.10611, 2026. doi: 10.48550/ARXIV.2601.10611. URL https://doi.org/10.48550/arXiv.2601.10611.

DatologyAI, :, Siddharth Joshi, Haoli Yin, Rishabh Adiga, Haakon Mongstad, Alvin Deng, Aldo Carranza, Alex Fang, Amro Abbas, Anshuman Suri, Brett Larsen, Daniel Zayas, Darren Teh, David Schwab, Diego Kiner, Fan Pan, Jack Urbanek, Jason Lee, Jason Telanoff, Josh Wills, Kaleigh Mentzer, Luke Merrick, Maximilian Böther, Parth Doshi, Paul Burstein, Pratyush Maini, Ties Robroek, Tony Jiang, Vidhi Jain, Vineeth Dorna, Zhengping Wang, Bogdan Gaza, Ari Morcos, and Matthew Leavitt. 20/20 vision language models: A prescription for better vlms through data curation alone, 2026. URL https://arxiv.org/abs/2605.11405.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Peter Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim Alabdulmohsin, Avital Oliver, Piotr Padlewski, Alexey A. Gritsenko, Mario Lucic, and Neil Houlsby. Patch n' pack: Navit, a vision transformer for any aspect ratio and resolution. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=VpGFHmI7e5.

Haiwen Diao, Xiaotong Li, Yufeng Cui, Yueze Wang, Haoge Deng, Ting Pan, Wenxuan Wang, Huchuan Lu, and Xinlong Wang. Evev2: Improved baselines for encoder-free vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21014–21025, October 2025.

Juechu Dong, Boyuan Feng, Driss Guessous, Yanbo Liang, and Horace He. Flex attention: A programming model for generating optimized attention kernels, 2024. URL https://arxiv.org/abs/2412.05496.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11198–11201, 2024.

Ze Feng, Jiang-Jiang Liu, Sen Yang, Lingyu Xiao, Zhibin Quan, Zhenhua Feng, Wankou Yang, and Jingdong Wang. Vision remember: Recovering visual information in efficient LVLM with vision feature resampling, 2025. URL https://openreview.net/forum?id=7Vy620t0gt.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: VLMs overlook their visual representations. 2025. URL https://arxiv.org/abs/2506.08008.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024.

Yichen Guo, Kai Tang, Fenglai Lin, Yiding Sun, Dongxu Zhang, Wenya Wang, Lin William Cong, and Shanghang Zhang. Fade: Mitigating hallucinations by reducing language-prior dominance in large vision-language models, 2026. URL https://arxiv.org/abs/2606.29431.

Junlin Han, Shengbang Tong, David Fan, Minghao Chen, Philip Torr, Filippos Kokkinos, and Mike Lewis. Towards physics of multimodal pretraining: Knowledge flow, modality synergy, early unification, and recipes. *arXiv preprint arXiv:2608.05000*, 2026.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

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

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Kaduri_Whats_in_the_Image_A_Deep-Dive_into_the_Vision_of_CVPR_2025_paper.pdf.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images, 2016.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Andreas Koukounas, Georgios Mastrapas, Florian Hönicke, Sedigheh Eslami, Guillaume Roncari, and Han Xiao. jina-vlm: Small multilingual vision language model, 2026. URL https://arxiv.org/abs/2512.04032.

Mario Michael Krell, Matej Kosec, Sergio P. Perez, and Andrew Fitzgibbon. Efficient sequence packing without cross-contamination: Accelerating large language models without impacting performance, 2022. URL https://arxiv.org/abs/2107.02027.

Benno Krojer, Shravan Nayak, Oscar Mañas, Vaibhav Adlakha, Desmond Elliott, Siva Reddy, and Marius Mosbach. Latentlens: Revealing highly interpretable visual tokens in llms, 2026. URL https://arxiv.org/abs/2602.00462.

Hugo Laurençon, Leo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=dtvJF1Vy2i.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023a.

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. Tokenpacker: Efficient visual projector for multimodal llm, 2024. URL https://arxiv.org/abs/2407.02392.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=S7Evzt9uit3.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024,*

Proceedings, Part IX, pp. 366–384, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72672-9. doi: 10.1007/978-3-031-72673-6_20. URL https://doi.org/10.1007/978-3-031-72673-6_20.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Qiao. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models, 2023. URL https://arxiv.org/abs/2311.07575.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futeral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. Ministral 3, 2026. URL https://arxiv.org/abs/2601.08584.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023a. doi: 10.1162/tacl_a_00566. URL https://aclanthology.org/2023.tacl-1.37/.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023c.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyang Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 2024.

Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jiawen Liu, Jifeng Dai, Yu Qiao, and Xizhou Zhu. Mono-internvl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. *arXiv preprint arXiv:2410.08202*, 2024.

Shweta Mahajan, Hoang Le, Hyojin Park, Farzad Farhadzadeh, Munawar Hayat, and Fatih Porikli. Attention guided alignment in efficient vision-language models, 2026. URL https://openreview.net/forum?id=yneEUGypwJ.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Smolvlm: Redefining small and efficient multimodal models, 2025. URL https://arxiv.org/abs/2504.05299.

Martin Marek, Sanae Lotfi, Aditya Somasundaram, Andrew Gordon Wilson, and Micah Goldblum. Small batch size training for language models: When vanilla sgd works, and why gradient accumulation is wasteful, 2025.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279. Association for Computational Linguistics, 2022.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier Biard, Sam Dodge, Philipp Dufter, Bowen Zhang, Dhruti Shah, Xianzhi Du, Futang Peng, Haotian Zhang, Floris Weers, Anton Belyi, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu He, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. Mm1: Methods, analysis: Insights from multimodal llm pre-training, 2024. URL https://arxiv.org/abs/2403.09611.

Imanol Miranda, Ander Salaberria, Eneko Agirre, and Gorka Azkune. Bivlc: Extending vision-language compositionality evaluation with text-to-image retrieval, 2024.

Yaniv Nikankin, Dana Arad, Yossi Gandelsman, and Yonatan Belinkov. Same task, different circuits: Disentangling modality-specific mechanisms in VLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=AD8ksC9bw1.

Chun-Liang Li Cem Koc Nate True Albert Antony Gokul Santhanam James Gabriel Peter Grasch Oncel Tuzel Hadi Pouransari Pavan Kumar Anasosalu Vasu, Fartash Faghri. Fastvlm: Efficient vision encoding for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Chufan Shi, Cheng Yang, Yaokang Wu, Linghao Jin, Bo Shui, Taylor Berg-Kirkpatrick, and Xuezhe Ma. Are VLMs seeing or just saying? uncovering the illusion of visual re-examination. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=DdU1o2ZvWi.

Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and LILI YU. LMFusion: Adapting pretrained language models for multimodal generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=Kc1WTxZbrP.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. 2025. URL https://arxiv.org/abs/2508.10104.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. PaliGemma 2: A family of versatile VLMs for transfer, 2024. URL https://arxiv.org/abs/2412.03555.

Alvin W. M. Tan, Sunny Yu, Bria Long, Wanjing Anya, Tonya Murray, Rebecca D. Silverman, Jason D. Yeatman, and Michael C. Frank. Devbench: A multimodal developmental benchmark for language learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 77445–77467. Curran Associates, Inc., 2024. doi: 10.52202/079017-2462. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/8d987b2981388c99c7eab6095d1d29fd-Paper-Datasets_and_Benchmarks_Track.pdf.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Shengbang Tong, Ellis L Brown II, Penghao Wu, Sanghyun Woo, ADITHYA JAIRAM IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, Xichen Pan, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=Vi8AepAXGy.

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

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased. 2025. URL https://arxiv.org/abs/2505.23941.

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 *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*, 2024a.

Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, et al. The all-seeing project v2: Towards general relation comprehension of the open world. *arXiv preprint arXiv:2402.19474*, 2024b.

Luis Wiedmann, Aritra Roy Gosthipaty, and Andrés Marafioti. nanovlm. https://github.com/huggingface/nanoVLM, 2025a.

Luis Wiedmann, Orr Zohar, Amir Mahla, Xiaohan Wang, Rui Li, Thibaud Frere, Leandro von Werra, Aritra Roy Gosthipaty, and Andrés Marafioti. Finevision: Open data is all you need. *arXiv preprint arXiv:2510.17269*, 2025b.

Haoyuan Wu, Aoqi Wu, Hai Wang, Jiajia Wu, Jinxiang Ou, and Bei Yu. Scaling native multimodal pre-training from scratch, 2026. URL https://arxiv.org/abs/2607.22043.

Jianghao Yin, Qin Chen, Kedi Chen, Jie Zhou, Xingjiao Wu, and Liang He. Dynamic multimodal activation steering for hallucination mitigation in large vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=YtWZdwEG5K.

Tianyu Yu, Kechen Fang, Zihao Wan, Kaidong Zhang, Yicheng Zhang, Jun Song, Bo Zheng, and Yuan Yao. Deep pre-alignment for vlms, 2026. URL https://arxiv.org/abs/2605.15300.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024.

Boqiang Zhang, Lei Ke, Ruihan Yang, Qi Gao, Tianyuan Qu, Rossell Chen, Dong Yu, and Leoweiliang. Penguin-VL: Exploring the efficiency limits of VLM with LLM-based vision encoders, 2026. URL https://arxiv.org/abs/2603.06569.

Yi-Fan Zhang, Qingsong Wen, Chaoyou Fu, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. Beyond LLaVA-HD: Diving into high-resolution large multimodal models. 2024. URL https://arxiv.org/abs/2406.08487.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text. *CoRR*, abs/2010.00747, 2020. URL https://arxiv.org/abs/2010.00747.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Zhang_Cross-modal_Information_Flow_in_Multimodal_Large_Language_Models_CVPR_2025_paper.pdf.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.

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. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=SI2hI0frk6.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Yinglun Zhu, Jiancheng Zhang, and Fuzhi Tang. Test-time matching: Unlocking compositional reasoning in multimodal models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=wWxdT6LB2D.

Weiming Zhuang, Chen Chen, Zhizhong Li, Sina Sajadmanesh, Jingtao Li, Jiabo Huang, Vikash Sehwag, Vivek Sharma, Hirotaka Shinozaki, Felan Carlo Garcia, Yihao Zhan, Naohiro Adachi, Ryoji Eki, Michael Spranger, Peter Stone, and Lingjuan Lyu. Argus: A compact and versatile foundation model for vision. In *Proceedings of The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2025*, 2025.

A Appendix

A.1 Additional Experimental Information

A.1.1 Improvement Ladder

We used the hyperparameters reported in Tab. 3 to train our models. The β_2 value for Adam is scaled according to Marek et al. (2025) with the VLM adaptation descriped in Sec. 3.2. Learning rates η are 10^{-4} for the decoder and 10^{-3} for the projector scaled both by the packed-token capacity per optimizer update $\eta_{\text{peak}} = \eta_{\text{base}} \left(T/10{,}482\right)^{\log_2(3)/10}$, with T = global packed rows $\times$ maximum row length. The value 10,482 here is the reference sequence length from the common-recipe baseline model. For the larger *feinerVLM-3B* we adapt the learning rate to account for the larger model size and set η_{base} to 3×10^{-5} for the decoder but keep the projector at 10^{-3}.

Table 3: Training hyperparameters. Learning rates are peak values after token-budget scaling. Epoch counts are configured targets.

Hyperparameter	feinerVLM-450M feiner-train	feinerVLM-450M FineVision	feinerVLM-3B FineVision
Epochs	1.00	1.67	1.67
Global batch size (packed rows)	8	8	8
Maximum tokens per packed row	20,965	20,965	8,192
Optimizer	Adam	Adam	Adam
β_1	0.9	0.9	0.9
β_2	0.997112	0.993703	0.997646
Gradient norm limit	1.0	1.0	1.0
Decoder peak learning rate	1.55186×10^{-4}	1.55186×10^{-4}	4.01134×10^{-5}
Projector peak learning rate	1.55186×10^{-3}	1.55186×10^{-3}	1.33711×10^{-3}
Decoder linear warmup (steps)	520	7,985	17,896
Projector linear warmup (steps)	52	799	1,790
Schedule after warmup	Constant	Constant, then cosine	Constant, then cosine
Cosine decay start (step)	–	79,853	178,956
Cosine decay end (step)	–	133,355	298,857
Final learning rate / peak	1.0	0.1	0.1

For our targeted improvements, Tab. 4 shows the exact numbers for relative binding and group accuracy after the common-recipe. For the larger *feinerVLM-3B* in Sec. 4.5 we use the same recipe but change the reference text centroid bias to the corresponding middle layer (18 of 36) of the larger SmolLM3 decoder.

Table 4: Cumulative improvements from baseline. Every row after frozen EUPE block-12 retains the preceding changes. For perplexity (PPL) lower is better, for the other metrics higher is better.

	Relative binding			Group accuracy (%)		FineWeb
Setup	Winoground	BiVLC	Mean	Winoground	BiVLC	PPL
Common recipe, trainable SigLIP2	6.92	9.04	7.98	8.50	1.50	17.609
Frozen native EUPE, block 12	2.79	5.23	4.01	3.75	0.14	16.727
+ Block-10 readout	6.17	10.45	8.31	8.25	10.94	16.743
+ MLP projector	5.92	19.58	12.75	8.75	16.95	16.710
+ Block-16 text bias	7.50	23.88	15.69	10.50	20.83	16.728
+ Standard Gram	9.21	25.25	17.23	10.25	22.43	16.711
+ Four registers and block-28 re-read	7.38	28.02	17.70	11.25	23.35	16.728

A.1.2 Text-based Centroid Bias

Our projector experiments show that text-based centroid bias initialization improves visual binding and reduces shifts in the vision embedding through the first decoder blocks. As the text representations change across decoder depth, we evaluate using text centroid from different decoder blocks

as reference. Text centroids are computed from $n = 500$ text-only samples at decoder blocks $b \in 0, 4, 8, 12, 16, 20$. Each centroid is the coordinate-wise median of the text embedding after removing special tokens and high-norm outliers. We use each text centroid reference to initialize the projector bias and compare performance after training. Using the text centroid from decoder block 16 gives the highest relative binding and Winoground group accuracy in this ablation. This is exactly halfway through the 32-block decoder. This preference is qualitatively consistent with previous interpretability work such as Kaduri et al. (2025) that identifies the middle decoder blocks as the main site of information transfer from vision to text.

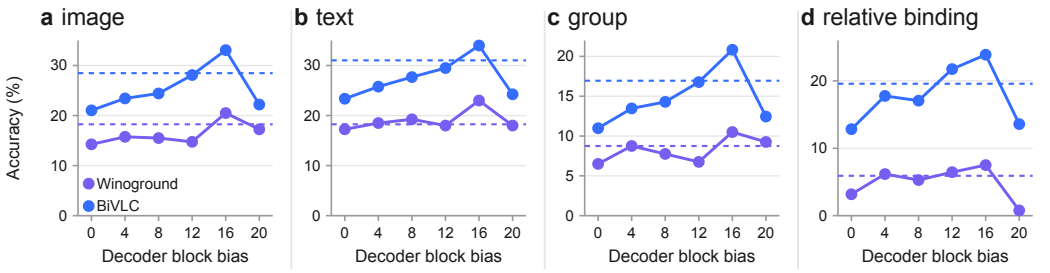

Figure 7: The bias reference matters. Text states from block 16 give the strongest task scores among the tested initialization targets without Gram regularization.

A.1.3 Residual Re-Read Register Tokens

Attention of ordinary register tokens mixes image and text sources within one softmax. To give registers a separate route to visual information for improved information flow, we add an image-only attention read at one decoder block. For each attention head, Q_R contains the queries at register positions, while K_I and V_I contain the keys and values at image positions. We reuse these quantities from the block's ordinary attention computation, so the image states are those inside the decoder. The additional register output is $\Delta_R = \alpha\ \mathrm{Attn}(Q_R, K_I, V_I)$. Here, Attn is standard scaled dot-product attention, with the softmax normalized over image positions only. Each head has its own learned scalar gate $\alpha \in (-1, 1)$, parameterized with $\tanh$ and initialized to 0.1. We add Δ_R to the ordinary attention output at register positions before concatenating heads and applying the shared output projection. Subsequent decoder blocks can then use the updated registers when predicting the answer. The separate image-only read is excluded from the ordinary attention shares in the table. To test when this additional read is useful, we compare late routes at blocks from 20 to the last decoder block (32). Among these late routes, block 28 gives the strongest results (see Tab. 1).

A.2 Evaluation and Benchmark Details

A.2.1 Generative Bias in BiVLC

In BiVLC each pair consists of a natural and generated image, which introduces, as already noted in the original paper by Miranda et al. (2024), a systematic difference in image source and dimensions. We examine the final baseline improvements model from Sec. 4 using its saved, calibrated score matrices. We measure the fraction of groups in which both captions select the first image, the natural one for BiVLC, versus the second image, which is generated. The model selects the natural image for both captions in 96.01% of BiVLC groups, leaving group accuracy at 1.50% (Tab. 5). Winoground also has image preferences within pairs, but its first- and second-image rates are balanced at 37.25% and 37.00%.

Table 5: Image-source audit of the common recipe on 400 Winoground and 2,933 BiVLC groups. Dimensions are measured before model preprocessing. BiVLC places the natural image first and the generated image second. Relative binding is a score with zero as its neutral value.

Measurement	Winoground	BiVLC
Both captions select first image (%)	37.25	96.01
Both captions select second image (%)	37.00	1.47
Positive binding $b > 0$ (%)	58.00	74.91
Relative binding (points)	6.92	9.04
Group accuracy (%)	8.50	1.50

Standard binding $b > 0$ counts even a small positive difference in scores as a success and reaches 74.91% on BiVLC. The actual underlying image preference cancels for b, so this test does not reveal when that preference dominates the scores. Relative binding addresses this measurement problem through by normalizing by $\sqrt{b^2 + c^2 + i^2}$. As a result, a large image preference reduces the magnitude of the normalized binding contrast, giving 9.04 points for this model on BiVLC. We therefore use relative binding on both datasets and retain group accuracy only for Winoground in the main comparisons, since it avoids the fixed natural/generated source distinction and is a more challenging task compared to relative binding. The BiVLC group score is retained in this table as a diagnostic.

A.2.2 Benchmark Tasks and Compute Estimates

The general multimodal tasks in *feiner-bench* are ChartQA (Masry et al., 2022), MMBench English v1.1 (Liu et al., 2023c), MMMU-Pro with ten answer choices (Yue et al., 2024), MathVista MINI (Lu et al., 2024), GQA balanced test-dev (Hudson & Manning, 2019), and AI2D test (Kembhavi et al., 2016). The vision-focused tasks include MMVP (Tong et al., 2024b), SpatialEval (Wang et al., 2024a), VLMsAreBiased (Vo et al., 2025), POPE with F1 scoring (Li et al., 2023b), CV-Bench 2D and 3D (Tong et al., 2024a), VSR on its zero-shot test split (Liu et al., 2023a), and CRPE Relation (Wang et al., 2024b), which requires correct answers under all four option rotations. We also include BLINK (Fu et al., 2024) as the unweighted mean of all six single-image validation subsets: Spatial Relation, Relative Depth, Counting, Object Localization, Relative Reflectance, and IQ Test. The language tasks are MMLU (Hendrycks et al., 2021), evaluated by cloze log-likelihood, and IFEval (Zhou et al., 2023), evaluated by the mean of strict prompt-level and instruction-level accuracy.

We use published token budgets in Fig. 1 where available and reconstruct packed budgets as number of steps $\times$ batch size $\times$ sequence length (Pavan Kumar Anasosalu Vasu, 2025; Luo et al., 2024; Zhang et al., 2026). For models without a token budget, we multiply the documented training samples by $2{,}532.33$ tokens per sample, derived from Jina-VLM's 47B tokens over $25{,}000 \times 128 + 60{,}000 \times 256$ sample exposures, and assume one pass when repeats are not reported (Koukounas et al., 2026). Missing H100 GPU-hours are estimated as $H = 17.54164 \cdot P_B \cdot T_B$, where P_B is model size in billions and T_B is the token budget in billions. The coefficient is the geometric mean of $H/(P_B T_B)$ for feinerVLM-450M, feinerVLM-3B, nanoVLM, and Molmo2, whose measured or reported hours are retained.