HARNESSING DOMAIN SPECIALISTS IN MULTIMODAL MIXTURE-OF-EXPERTS FOR EFFICIENT ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Mixture-of-Experts (MoE) architectures scale model capacity through sparse computation, routing each token through only a small subset of experts. In this work, we explore whether this sparsity gives rise to emergent intrinsic organization in multimodal MoEs. We find that experts develop strong semantic specialization across modalities and domains despite not being explicitly trained for modularity. Building on this structure, we introduce `ExpertLens`, a data-free method that identifies domain-specialized experts directly from pretrained model weights by decoding router weights into semantically meaningful vocabulary tokens. We leverage this specialization for efficient multimodal adaptation by selectively fine-tuning experts relevant to a target domain. Across math, medical, and remote sensing tasks, `ExpertLens` matches or surpasses full fine-tuning while updating only 21.7–47.0% of model parameters and achieving a $4.0\times$ average training speedup, and outperforms LoRA in both adaptation performance and training efficiency. These results show that sparsity introduced for efficiency can give rise to semantic modularity that is directly useful for efficient adaptation.

1 INTRODUCTION

Mixture-of-Experts (MoE) architectures have become a key ingredient in scaling modern language and multimodal models (Bai et al., 2025; Gemma Team et al., 2026; Kimi Team et al., 2025). By replacing dense feed-forward layers with many experts while activating only a small subset for each token, MoEs decouple model capacity from per-token computation (Shazeer et al., 2017; McGill & Perona, 2017; Fedus et al., 2022). This sparsity has enabled recent models to scale to enormous sizes while remaining computationally tractable (Kimi Team, 2026; Thinking Machines, 2026).

Sparsity in MoEs was introduced primarily for efficiency, not organization (Lepikhin et al., 2020). Experts are learned jointly, with a router selecting a subset per token. Yet consistent routing patterns could induce implicit modularity, with experts specializing to different kinds of information. Such structure has been studied in text-only MoEs (Antoine et al., 2025; Li et al., 2026), but only recently in multimodal models (Xu et al., 2026b; Zhang et al., 2026), where the same expert layers process both visual and textual representations. We therefore ask whether sparse multimodal MoEs are semantically organized: do experts specialize by modality, domain, and visual concept?

In this work, we find that multimodal MoEs develop distinct semantic specialization without explicitly training for it. Across three large multimodal MoEs, experts exhibit strong modality specialization, preferentially processing visual or textual tokens (§3.1). Beyond modality, visual representations and routing organize strongly by semantic domain: medical, math, and remote sensing imagery form distinct structures not explained by low-level properties such as color or spatial position (§3.2). This extends to finer semantics, with visually dissimilar but semantically related object classes recruiting overlapping expert sets (§3.3). Thus, sparsity introduced for computational efficiency gives rise to an emergent semantic decomposition of multimodal MoEs.

This emergent structure raises a practical question: given an MoE, how can we identify which experts specialize in a given domain? To this end, we introduce `ExpertLens`, a data-free method that recovers expert semantics directly from pretrained weights by decoding router weights into semantically meaningful vocabulary tokens (§3.4). `ExpertLens` enables identification of experts relevant to a given domain without requiring domain-specific examples.

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.

We next ask whether this emergent specialization can enable more efficient multimodal adaptation (§4). Using `ExpertLens`, we selectively adapt experts relevant to each domain while freezing the remaining expert MLPs. Across sequential adaptation to math, medical, and remote sensing tasks, this strategy matches or surpasses full fine-tuning while updating only 21.7–47.0% of model parameters and achieving a $4.0\times$ average speedup. `ExpertLens` also outperforms LoRA in both adaptation performance and training speed ($2.9\times$ faster). Thus, the emergent specialization of multimodal MoEs is not only interpretable, but directly useful for efficient adaptation.

In summary, our contributions are:

- We uncover emergent semantic structure in multimodal MoEs, showing specialization across modalities, semantic domains, and fine-grained visual concepts.
- We introduce `ExpertLens`, a data-free method that decodes router weights to identify the semantic specialization of individual experts.
- We demonstrate that emergent modularity enables efficient multimodal adaptation, matching full fine-tuning and outperforming LoRA while reducing training time by $4\times$ on average.

2 Related Work

Mixture-of-Experts. Mixture-of-Experts (MoE) architectures have a long history in conditional computation, scaling model capacity by sparsely routing each input through a subset of expert modules (Shazeer et al., 2017; Jacobs et al., 1991). Sparse MoEs were later incorporated into large-scale Transformers (Lepikhin et al., 2020; Fedus et al., 2022; Jiang et al., 2024; Dai et al., 2024), and are now common for scaling modern language models (Kimi Team, 2026; Xu et al., 2026a) and multimodal foundation models (Bai et al., 2025; Gemma Team et al., 2026; Kimi Team et al., 2025; Thinking Machines, 2026). We provide a brief overview of MoE routing and notation in §A.1.

Interpreting Internal Representations. Mechanistic interpretability seeks to reverse-engineer neural networks into human-understandable mechanisms using tools such as circuit analysis, causal interventions, and representation decoding (Elhage et al., 2021; Meng et al., 2023; Conmy et al., 2023). Logit Lens (nostalgebraist, 2020) applies a model's output head to intermediate residual-stream representations to expose how predictions evolve across layers. This is motivated by the residual stream's approximately linear, additive structure (Elhage et al., 2021) and evidence that transformer MLPs write semantically meaningful, vocabulary-aligned updates into it (Geva et al., 2022; Wu et al., 2025). Subsequent variants have extended this idea to decoding internal representations in LLMs and VLMs (Belrose et al., 2025; Pal et al., 2023; Neo et al., 2025).

Interpretability of Mixture-of-Experts (MoE) models remains comparatively underexplored. Prior work has used attribution and ablation to identify behaviorally important experts (Hu et al., 2026; Li et al., 2025), manipulated routing or expert activation to steer behavior (Fayyaz et al., 2025), and decoded individual expert contributions into output space (Chaudhari et al., 2026). We extend this line of work with `ExpertLens`, a data-free method for decoding expert semantics and identifying domain-specialized experts for efficient adaptation.

Expert Specialization. Prior work on text-only MoEs has found emergent expert structure, including specialization by token type and part of speech (Jiang et al., 2024; Antoine et al., 2025; Li et al., 2026). However, this structure is typically low-level rather than aligned with broader semantic domains. To encourage higher-level organization, EMO (Wang et al., 2026) modifies text-only MoE pretraining to promote domain-level modularity. Expert organization in multimodal MoEs remains less explored, with recent work examining routing across visual, textual, and reasoning stages (Xu et al., 2026b; Zhang et al., 2026). In contrast, we find that pretrained multimodal MoEs develop rich semantic organization spanning modalities, domains, and fine-grained visual concepts.

Parameter Efficient Adaptation. Parameter-efficient fine-tuning methods such as LoRA reduce adaptation cost by introducing a small number of trainable parameters while keeping the pretrained model largely frozen (Hu et al., 2022). Recent work has extended low-rank adaptation to MoE models, applying LoRA updates to shared and/or expert parameters (Wei et al., 2026; Liu et al., 2026). In contrast, our approach exploits the pretrained model's existing expert structure, selectively updating the experts most relevant to the target domain. In direct comparison, `ExpertLens` achieves higher adaptation performance than LoRA while also reducing training time more substantially.

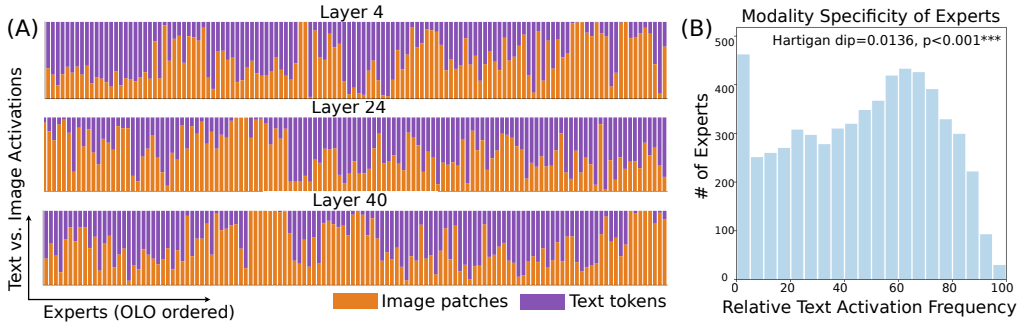

Figure 1: **Experts specialize by modality. (A)**: Fraction of each expert's top-K ($K = 8$) activations from image patches versus text tokens, ordered by optimal leaf ordering (OLO; §A.2.1). Most experts strongly favor one modality. **(B)**: Distribution of each expert's text-token activation share. For QWEN3-VL-30B-A3B, the distribution is bimodal (Hartigan dip $= 0.0136$, $p < 0.001$).

3 EMERGENT STRUCTURE IN MIXTURE-OF-EXPERTS VLMS

Mixture-of-Experts models provide a natural opportunity for functional specialization within their expert modules, but it remains unclear whether such specialization emerges in pretrained models and, if so, how it is organized. We explore this idea via three questions: **(1)** Do experts specialize on different modalities? **(2)** Do experts specialize on different semantic domains within a modality? **(3)** Does this specialization persist at finer-grained, object-level semantics?

Answering these questions can provide insight into how multimodal MoEs organize their internal representations, while also revealing whether their modular structure can be exploited for efficient, targeted adaptation. We study these questions across three frontier multimodal MoEs: QWEN3-VL-30B-A3B (Bai et al., 2025), KIMI-VL-A3B (Kimi Team et al., 2025), and GEMMA-4 26B-A4B (Gemma Team et al., 2026). We conclude by introducing `ExpertLens`, a data-free method for identifying domain-specialized experts from expert router weights.

Throughout this section, we use VQA datasets spanning general, medical, math/scientific, and remote-sensing domains to probe MoE expert behavior across image patches and text tokens. For text-only analyses, we additionally include broader linguistic domains such as legal and abstract reasoning. We include all datasets used and experimental details in §A.2.1.

3.1 EXPERT SPECIALIZATION BY MODALITY

We first ask whether multimodal MoEs exhibit modality-specific routing. In particular, do individual experts preferentially process visual or textual tokens, and is specialization distributed similarly across modalities? Establishing such structure would indicate that experts in pretrained multimodal MoEs are not interchangeable, but develop distinct functional preferences. To this end, we measure modality affinity from each expert's activation counts on modality-specific tokens. We sample 3,840 image patches and equally many text tokens from seven datasets spanning general VQA, medical imaging, and math/science reasoning. Under the expert activation definition (§A.1), we define affinity as the fraction of each expert's top-K activations from image versus text tokens.

Experts develop modality-specific routing. Figure 1A shows, for each expert, the fraction of top-K activations arising from image versus text tokens. The x axis indexes experts within each layer; orange-dominated experts favor image tokens, and purple-dominated experts favor text. Across layers, activations are rarely balanced between modalities. Instead, many experts strongly prefer one, indicating that modality specialization emerges naturally in pretrained multimodal MoEs.

Modality specialization tends to be bimodal and asymmetric. We pool experts across layers by text-token activation share. Without modality preference, the distribution should peak near 50%; instead, QWEN3-VL-30B-A3B is clearly bimodal (Figure 1B), with an image-leaning mode at 0–5% text share (462 experts) and a text-leaning mode at 60–65% (433 experts). Hartigan's dip test rejects unimodality (dip $= 0.0136$, $p < 0.001$). Image specialization is more extreme, while text specialization spreads across 60–80%. GEMMA-4 26B-A4B is similarly bimodal, while KIMI-VL-A3B is more nearly unimodal, likely due to its smaller expert pool (§A.4.2).

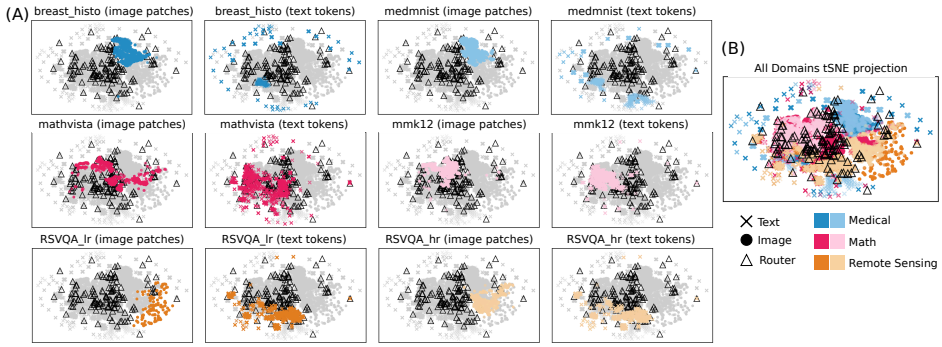

Figure 2: **Visual activations exhibit stronger domain structure than textual activations.** We jointly project layer-18 text-token and image-patch activations with router weights from QWEN3-VL-30B-A3B using t-SNE. **(A)**: Datasets are colored and separated by modality; image activations form tighter domain-specific clusters than text activations. **(B)**: All domains shown together.

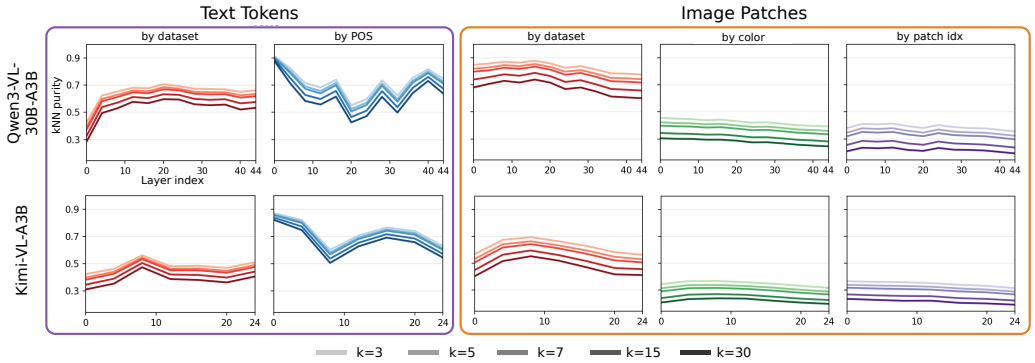

Figure 3: **Visual activations organize by domain; textual activations by linguistic structure.** We measure kNN purity across layers, grouping text activations by dataset or part of speech and image activations by dataset, color, or spatial position. Text activations align most with part of speech, while image activations align with semantic domain rather than low-level visual statistics.

3.2 DOMAIN SPECIFICITY IN VISION AND TEXT EXPERTS

Having established modality preferences, we then ask whether specialization extends to semantic domains. Do some experts preferentially process medical over math inputs? We test whether visual and textual representations organize differently by domain, and whether this structure reflects semantics rather than linguistic factors or low-level visual statistics. We first examine this with t-SNE projections of token activations and router weights, then quantify it using kNN purity.

Visual activations exhibit strong domain structure. We ask whether modality specialization extends to semantic domains by examining the representations that drive expert routing. Because routing depends directly on token activations, domain structure in this space could induce domain-specific expert use. We sample 1.5K image patches and text tokens from each of six VQA datasets spanning medical, math, and remote sensing domains (18K activations total), and jointly project them with all the router weights in QWEN3-VL-30B-A3B using t-SNE. Figure 2 shows layer 18 as an exemplar, with similar patterns across layers and models (§A.4.6).

Figure 2 reveals two trends. First, image-patch activations form tighter domain-specific clusters than text-token activations, with datasets from the same domain overlapping substantially. For example, breast histopathology and MEDMNIST cluster together, as do RSVQALR and RSVQAHR, suggesting that the structure reflects domain rather than dataset identity. Second, router weights concentrate near image activations, while text activations are more diffuse. Together, these observations suggest that visual representations exhibit the domain structure that could induce domain-specific routing.

Quantifying domain structure with kNN purity. The t-SNE analysis suggests strong domain structure in visual activations, but not which factors best explain it. We therefore quantify local activation-space structure using kNN purity: for each activation, we measure the fraction of its k

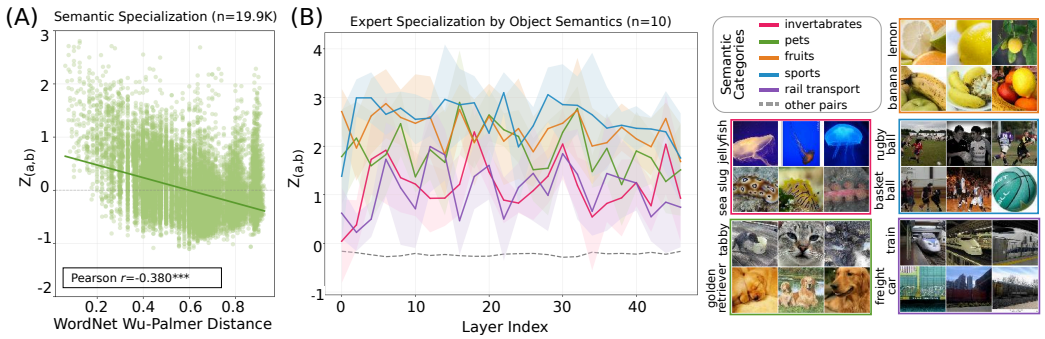

Figure 4: **Semantically related object classes recruit similar experts.** (A) Across all TinyImageNet class pairs, expert-set overlap $Z(a, b)$ decreases with semantic distance (Pearson $r = -0.380$, $p < 0.001$). (B) For 10 classes forming five semantic pairs, related classes show greater expert overlap across layers despite visual variation. Solid lines average thresholds $m \in \{3, 5, 8, 12, 16\}$, shading denotes ± 1 standard deviation, and the gray dashed line shows unrelated pairs.

nearest neighbors sharing the same label, where higher purity indicates better alignment with the local representation geometry. For text tokens, we compare groupings by dataset or part of speech; for image patches, by dataset, mean color, or spatial position. Each grouping contains 12 balanced categories. Further details are in §A.2.2. We report results across several values of k in Figure 3.

Vision and text representations organize differently. Figure 3 shows that text-token activations are better organized by part of speech than by domain, consistent with the findings in prior analyses of the experts of text-only MoEs (Antoine et al., 2025; Jiang et al., 2024). In contrast, image-patch activations achieve highest kNN purity when grouped by dataset, substantially exceeding mean color or spatial patch position. This pattern holds across layers and both QWEN3-VL-30B-A3B and KIMI-VL-A3B. These results support the t-SNE analysis and suggest that visual domain organization is not explained by low-level properties such as color or spatial location.

3.3 VISUAL EXPERT SPECIALIZATION BY OBJECT SEMANTICS

With domain structure in visual representations established, we can now ask if specialization extends to finer-grained semantics. We test whether visually distinct but semantically related classes recruit similar expert sets. To isolate semantics from dataset effects, we conduct the analysis entirely within TinyImageNet (Le et al., 2015), comparing overlap between related and unrelated classes.

Measuring semantic expert overlap. We use all 200 TinyImageNet classes, forming 19,900 object-class pairs spanning a range of semantic similarity, and sample 100 images per class. For each image, we track which experts activate on its patch tokens. An expert counts as active only if it activates on at least m patches, for $m \in \{3, 5, 8, 12, 16\}$. For each class and layer, we retain experts active in at least 60% of that class's images. We then measure Jaccard overlap between each class pair's expert sets and standardize it with a score $Z(a, b)$ relative to all class pairs, where larger values indicate stronger shared specialization. Further details and score calculations are in §A.2.3.

Semantically related classes activate similar experts. Figure 4 shows that related class pairs share more experts than unrelated pairs across layers. In Figure 4A, semantic distance measured by WordNet Wu-Palmer distance (Wu & Palmer, 1994) is inversely correlated with expert-set overlap: larger distance indicates lower semantic relatedness. For the purpose of providing an overview, the semantic overlap scores are shown averaged across all activation thresholds and all model layers. Figure 4B provides a finer-grained view using 10 TinyImageNet classes forming five semantically related pairs, without averaging across layers or patch thresholds. Their specialization scores remain above the gray unrelated-class baseline, with several pairs reaching $Z(a, b) > 3$ in intermediate layers, i.e., more than three standard deviations above baseline. These results show that visual expert specialization extends beyond coarse domains to finer-grained object semantics.

Semantic overlap persists despite visual variation. Importantly, high expert overlap is not limited to visually similar classes. Figure 4B shows that while some related pairs, such as trains and freight cars, share visual structure, others, such as jellyfish and sea slugs or tabbies and golden retrievers,

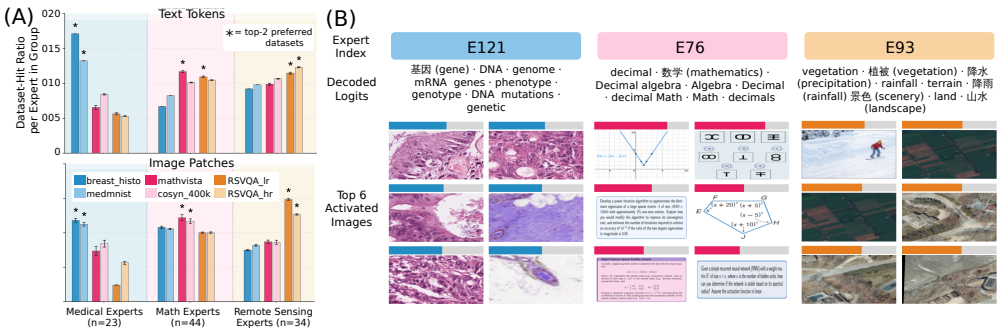

Figure 5: **`ExpertLens` selects domain-specific experts.** (A) `ExpertLens`-derived cohorts preferentially activate on their target domains. Bar colors denote datasets; heights show normalized cohort firing. Across most cohorts, activation is strongest on the corresponding domain. (B) Representative experts with their decoded logits and highest-activating images: E121 is medical, E76 math, and E93 remote sensing. All results use layer 18 of QWEN3-VL-30B-A3B.

differ substantially in appearance and background. Yet they recruit similar expert sets. We observe the same object-level specialization in other MoE-VLMs (§A.4.4). Together, our analyses suggest that visual expert specialization reflects semantics rather than dataset identity or low-level similarity.

3.4 EXPERTLENS: IDENTIFYING DOMAIN-SPECIALIZED EXPERTS

Our analyses show that multimodal MoEs develop domain specialization, but identifying these experts has required domain-specific examples to observe routing behavior. This is limiting for targeted adaptation, where we would prefer to identify experts for a new domain without representative data. We therefore ask: *can we recover an expert's semantic specialization from its routing weights?* We introduce `ExpertLens`, a data-free method that decodes semantic preferences from router weights.

Introducing ExpertLens. `ExpertLens` builds on a simple property of MoE routing: an expert is selected when its router weight aligns strongly with a token representation. If token representations encode semantics, then a domain-specialized expert's router weight should align with semantically related directions. We therefore map each router weight through the model's LM head to obtain a distribution over vocabulary tokens, analogous to logit-lens decoding (nostalgebraist, 2020). For expert i in layer ℓ, let $w_i^{(\ell)}$ denote its router weight and U the LM head. We compute:

$$z_i^{(\ell)} = U w_i^{(\ell)}, \qquad \mathcal{V}_M^{(\ell,i)} = \mathrm{TopM}\left(z_i^{(\ell)}\right), \qquad S_{\mathrm{EL}}^{(\ell,i)}(d) = \frac{1}{M} \sum_{v \in \mathcal{V}_M^{(\ell,i)}} \mathrm{Rel}(v,d). \tag{1}$$

where $\mathcal{V}_M^{(\ell,i)}$ is the set of top-M tokens decoded from expert i in layer ℓ, and $\mathrm{Rel}(v,d) \in \{0,1\}$ indicates whether token v is semantically related to domain d, as judged by the model (details in §A.3). The `ExpertLens` score is thus the fraction of an expert's top-M decoded tokens associated with the target domain. In all experiments, we use $M = 10$ and select experts with $S_{\mathrm{EL}} \geq 0.9$.

Evaluating ExpertLens. We test whether `ExpertLens` specializations predict expert use on domain-specific inputs. Fig. 5A shows layer-18 routing of `ExpertLens`-identified medical (n=23), math (n=44), and remote sensing (n=34) experts in QWEN3-VL-30B-A3B over text and image tokens from six VQA datasets. For each expert, we compute its activation fraction by dataset and normalize by cohort size. This gives experts equal weight, preventing cohort size or firing rate from dominating. Error bars show SEM; asterisks mark the two highest-scoring datasets.

ExpertLens uncovers domain specificity. Fig. 5A shows that datasets preferentially recruit experts whose `ExpertLens` labels match their domain. Medical, math, and remote-sensing datasets most strongly recruit their corresponding expert cohorts, except for weaker math alignment on text tokens. Thus, experts identified without domain-specific data still align with the domains that preferentially recruit them. Although Fig. 5A shows layer 18, the same pattern holds across layers (§A.3). Fig. 5B shows this correspondence at the level of individual experts. For representative medical (E121), math (E76), and remote sensing (E93) experts, the highest-logit decoded tokens align semantically with the domain of the images that most strongly recruit each expert. This indicates that `ExpertLens`'s decoded preferences closely track the inputs that activate each expert.

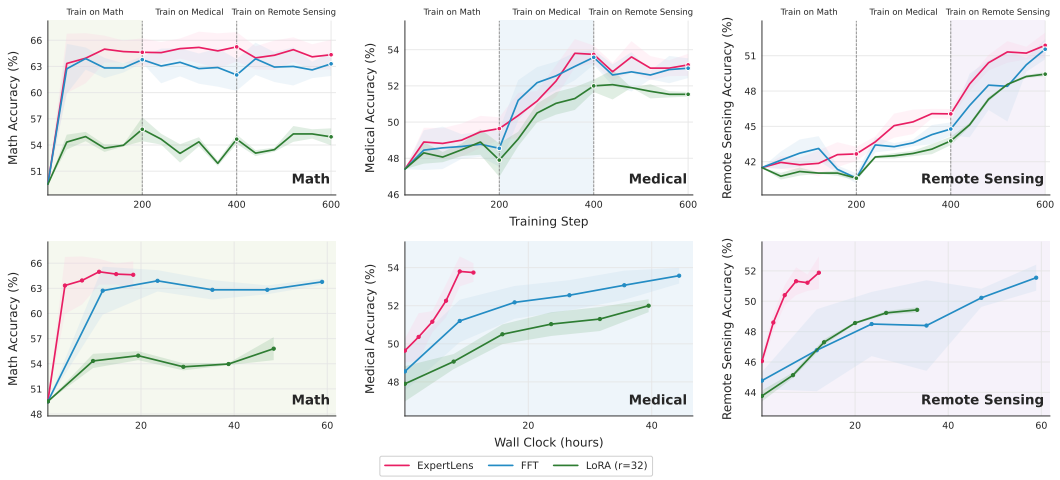

Figure 6: **ExpertLens matches full fine-tuning, outperforms LoRA, and accelerates adaptation.** (A) `ExpertLens` matches or surpasses FFT while outperforming LoRA. (B) Highlighted regions correspond to the same stages in (A), shown against wall-clock time. `ExpertLens` achieves a $4.0\times$ average speedup over FFT while matching performance, and a $2.9\times$ speedup over LoRA.

4 Domain-Specialized Experts for Efficient Adaptation

In §3 we observed that multimodal MoEs develop domain-specialized experts. We now ask whether this structure can enable more efficient adaptation. In particular, can selectively adapting domain-relevant experts match full fine-tuning while reducing post-training parameters and compute?

Sequential Adaptation Protocol. To study this, we adapt a single model sequentially to math, medical, and remote sensing. Rather than resetting to the pretrained checkpoint between domains, this protocol tests whether targeted expert adaptation remains effective as the same model is repeatedly adapted. We evaluate all three domains throughout training and study domain ordering in §4.3.

Selective Adaptation. We identify domain experts with the `ExpertLens` method (§3.4) and select those with $S_{\text{EL}} \geq 0.9$. For Qwen3-VL-30B-A3B, this yields 2,768, 1,103, and 1,290 experts for math, medical, and remote sensing. The sets are largely distinct (Jaccard overlap: 0.13% math–medical, 6.48% math–remote sensing, 13.20% medical–remote sensing). We tune selected expert MLPs, freezing all others and the vision encoder. We also train shared decoder parameters, including QKV projections, router weights, and the LM head (§4.3). As expert MLPs contain most model parameters, 53.0–78.3% of the model is frozen. We call this procedure `ExpertLens` throughout.

Train and Evaluation Setup. We train on $4{,}000$ examples per domain and evaluate on multiple benchmarks spanning math, medical, and remote sensing. Training datasets, evaluation benchmarks, and implementation details are in §A.5. We use the system prompt and rewards from VERO (Sarch et al., 2026), Qwen3-VL-30B-A3B as the base MoE, and GSPO (Zheng et al., 2025) for post-training. We train for 200 steps per domain on 4 H200 GPUs with batch size 32 and group size 5. Results use three seeds, with shaded regions showing 95% confidence intervals.

4.1 ExpertLens for Efficient Multimodal Adaptation

We evaluate whether selectively fine-tuning domain-specialized experts provides an effective and efficient alternative to both full-model and parameter-efficient adaptation. We compare `ExpertLens` against full fine-tuning (FFT) and LoRA (Hu et al., 2022) with the same data, objective, and sequential adaptation protocol. For LoRA, we evaluate ranks $R \in \{32, 64\}$ with $\alpha = 2R$ following prior work (Shuttleworth et al., 2025), and report the best-performing configuration ($R = 32$).

ExpertLens matches FFT and outperforms LoRA. Figure 6A shows that `ExpertLens` matches or surpasses FFT throughout adaptation while updating only 21.7–47.0% of model parameters. At the end of each training stage, `ExpertLens` achieves 64.6% versus 63.8% on math, 53.7% versus 53.6% on medical, and 51.9% versus 51.6% on remote sensing. LoRA reaches 55.8%, 52.0%,

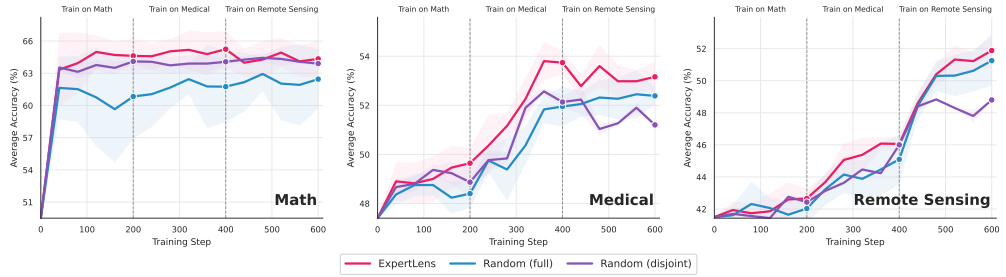

Figure 7: **ExpertLens outperforms matched-budget random expert selection.** `ExpertLens` outperforms matched-budget random selection baselines while updating the same number of experts.

and 49.4%, respectively, underperforming both methods. `ExpertLens` thus identifies sufficient domain-specialized capacity to match full-model adaptation and outperforms LoRA.

ExpertLens accelerates adaptation. Selective adaptation also substantially reduces training time. Figure 6B shows in-domain performance against wall-clock time. Relative to FFT, `ExpertLens` achieves speedups of $3.2\times$ on math, $4.0\times$ on medical, and $4.8\times$ on remote sensing, averaging $4.0\times$. Compared with LoRA, `ExpertLens` trains $2.9\times$ faster and achieves higher performance. Thus, exploiting MoE modularity can yield better performance–efficiency trade-off than FFT or LoRA.

4.2 The Importance of Expert Selection

Next, we ask whether `ExpertLens` gains arise from sparsity alone or from selecting the right experts. In Figure 7, we compare against two random baselines with the same expert budget per domain. *Random (full)* samples from the full expert pool, while *Random (disjoint)* samples only experts not selected by `ExpertLens`. Expert subsets are resampled for each seed. This comparison isolates the benefit of domain-relevant expert selection beyond simply updating fewer parameters.

Sparsity alone does not explain the gains. We first compare `ExpertLens` with *Random (full)*, which updates the same number of experts sampled from the full expert pool. Figure 7 shows that `ExpertLens` consistently outperforms *Random (full)*, with the clearest gains on math and medical. At the end of each training stage, `ExpertLens` achieves 64.6% versus 60.8% on math, 53.7% versus 52.0% on medical, and 51.9% versus 51.3% on remote sensing. Since both methods update the same number of experts, sparsity alone cannot explain the gains; expert identity matters.

ExpertLens-selected experts are particularly effective. We next compare against *Random (disjoint)*, which samples the same number of experts only from those not selected by `ExpertLens`. Figure 7 shows that this baseline matches `ExpertLens` on math but falls behind on medical and remote sensing. Therefore, while alternative expert subsets can suffice for some domains, `ExpertLens`-selected experts are more effective overall. Together with *Random (full)*, this shows that selective fine-tuning depends not only on sparsity, but on selecting domain-relevant experts.

ExpertLens approaches privileged selection. We compare `ExpertLens` with two count-based selectors with the same expert budget. *Frequency Selection* ranks experts by current-domain activation counts, while the *Privileged Oracle* ranks by current-domain counts minus counts from other domains. The oracle requires data from all domains, including future ones, and depends on the datasets used. Despite using no domain examples, ExpertLens achieves comparable final accuracy to both baselines (Table 1). Thus, `ExpertLens` matches count-based selection while remaining usable without target examples or future-domain knowledge.

Selection Method	Math	Medical	RS	Mean
Frequency Selection	63.8	52.7	51.6	56.0
Privileged Oracle	63.0	**55.5**	**52.1**	**56.9**
`ExpertLens`	**64.3**	53.2	51.9	56.5

Table 1: `ExpertLens` approaches oracle expert selection without domain examples.

4.3 Ablations and Analyses

We next examine the design choices underlying `ExpertLens` and the robustness of selective expert adaptation. We first ablate which parameter groups are updated alongside the selected experts, then test whether our conclusions depend on the ordering of training domains.

Trainable Parameters	In-domain			Out-of-domain		
	Math	Medical	Remote Sensing	Math	Medical	Remote Sensing
Selected Experts Only	52.2	49.4	48.7	44.7	47.2	51.7
+ Router Weights	63.4 (+11.2)	50.7 (+1.3)	49.3 (+0.6)	45.7 (+1.0)	54.6 (+7.4)	56.2 (+4.5)
+ QKV Projections	65.8 (+2.4)	53.1 (+2.4)	52.3 (+3.0)	45.5 (-0.2)	55.0 (+0.4)	57.7 (+1.5)
+ LM Head (`ExpertLens`)	**66.1** (+0.3)	**54.1** (+1.0)	**52.8** (+0.5)	**46.1** (+0.6)	**56.1** (+1.1)	**59.7** (+2.0)

Table 2: **Shared decoder parameters improve selective adaptation.** Training the router, attention weights, and language-model head alongside selected experts progressively improves performance.

Domain Ordering	Method	Math	Medical	Remote Sensing	Mean
Math → Med. → Rem. Sens.	FFT	62.4	52.6	51.0	55.3
	`ExpertLens`	**65.2** (+2.8)	**53.2** (+0.6)	**52.6** (+1.6)	**57.0** (+1.7)
Med. → Math → Rem. Sens.	FFT	57.1	51.5	51.9	53.5
	`ExpertLens`	**61.3** (+4.2)	**52.3** (+0.8)	**52.7** (+0.8)	**55.5** (+2.0)
Rem. Sens. → Med. → Math	FFT	62.6	**52.9**	**53.1**	56.2
	`ExpertLens`	**67.2** (+4.6)	51.0 (-1.9)	50.9 (-2.2)	**56.4** (+0.2)

Table 3: **ExpertLens is robust to domain ordering.** Across three domain orders, `ExpertLens` achieves higher mean performance than FFT (averaged over 3 seeds) with lower variation.

Shared decoder parameters are important for effective adaptation. Table 2 ablates which parameters are trained alongside `ExpertLens`-selected experts. *In-domain* performance is accuracy after training on a domain, while *out-of-domain* performance averages the other two. Training only selected expert MLPs yields weaker adaptation, especially on math. Adding the router improves math by 11.2 points, with QKV projections and the language-model head providing further gains across domains. These results suggest that selective adaptation benefits from updating the shared parameters that route inputs to the trained experts and shape their representations.

ExpertLens is robust to domain ordering. To test sensitivity to domain sequence, we repeat adaptation under three training orders in Table 3. `ExpertLens` outperforms FFT in mean final accuracy under all three, by 1.7, 2.0, and 0.2 points, respectively. It is also less sensitive to ordering, with the standard deviation across orders decreasing from 1.37 to 0.78 points (43%). Although gains vary across orders, the overall benefit of selective adaptation persists across training sequences.

4.4 Experiments on Kimi-VL

To test whether targeted adaptation generalizes beyond Qwen3-VL-30B-A3B, we repeat our main experiment on Kimi-VL-A3B, comparing `ExpertLens` with full fine-tuning and matched-budget *Random (full)*. Using the same threshold, $S_{\text{EL}} \geq 0.9$, selects 560, 301, and 349 experts for math, medical, and remote sensing, respectively. For the semantic relevance function $\text{Rel}(\cdot)$, we use Qwen3-VL-30B-A3B to classify decoded tokens, as Kimi-VL-A3B was less reliable at this task.

Method	Math	Medical	Remote Sensing	Mean
Random (full)	40.2	46.5	48.4	45.0
FFT	40.0	50.6	**51.5**	47.4
`ExpertLens`	**41.5**	**51.4**	51.0	**48.0**

Table 4: Results for Kimi-VL-A3B.

ExpertLens maintains its advantage on Kimi-VL. Table 4 reports final accuracy after the full sequential adaptation sequence. `ExpertLens` achieves the highest mean performance at 48.0%, compared with 47.4% for FFT and 45.0% for random selection, showing that the benefits of targeted expert adaptation extend beyond Qwen3-VL-30B-A3B. `ExpertLens` also remains faster to train than FFT on Kimi-VL-A3B, providing a 1.53× speedup over the full sequence.

5 Conclusion

In this work, we show that multimodal MoEs develop emergent semantic specialization across modalities, domains, and visual concepts, which can be exploited for efficient and effective adaptation. In this view, `ExpertLens` is a structure-aware adaptation method: instead of introducing new parameters to adapt, it leverages the model's existing organization by identifying and updating experts already specialized for the target domain. To encourage further progress and model development, **we release all data, code, and trained models.**

AI USE STATEMENT

In this work, we used generative AI tools to aid with implementation and editing the manuscript for clarity. We did not use AI for ideation or brainstorming, or to discover relevant literature. The remainder of the disclosure tasks (synthetic data generation, proving mathematical claims, etc.) are not applicable to this work. We have reviewed all AI-assisted writing and implementations. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We include finegrained details for all of the analyses conducted in §3 in the Appendix (§A.2). Similarly, we include details on experiments in §A.5 and all relevant hyperparameters in Table 7. We will release code and trained models upon acceptance.

REFERENCES

Elie Antoine, Frédéric Béchet, and Philippe Langlais. Part-of-speech sensitivity of routers in mixture of experts models. In *ICCL*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens, 2025. URL https://arxiv.org/abs/2303.08112.

BIG-bench authors. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *TMLR*, 2023.

Marmik Chaudhari, Idhant Gulati, Nishkal Hundia, Pranav Karra, and Shivam Raval. MoE Lens – An Expert Is All You Need, 2026. URL https://arxiv.org/abs/2603.05806. Version Number: 1.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R Routledge, et al. Finqa: A dataset of numerical reasoning over financial data. In *EMNLP*, 2021.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

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.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability, 2023. URL https://arxiv.org/abs/2304.14997.

Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models, January 2024. URL http://arxiv.org/abs/2401.06066. arXiv:2401.06066 [cs].

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Mohsen Fayyaz, Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Ryan Rossi, Trung Bui, Hinrich Schütze, and Nanyun Peng. Steering MoE LLMs via Expert (De)Activation, September 2025. URL http://arxiv.org/abs/2509.09660. arXiv:2509.09660 [cs].

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 2022.

Gemma Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space, 2022. URL https://arxiv.org/abs/2203.14680.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *NeurIPS*, 2023.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering, 2020. URL https://arxiv.org/abs/2003.10286.

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

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Guimin Hu, Meng Li, Qiwei Peng, Lijie Hu, Boyan Xu, and Ruichu Cai. What Gets Activated: Uncovering Domain and Driver Experts in MoE Language Models, January 2026. URL http://arxiv.org/abs/2601.10159. arXiv:2601.10159 [cs].

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 1991.

Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 2016.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering, 2019. URL https://arxiv.org/abs/1909.06146.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*, 2017.

Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. Geomverse: A systematic evaluation of large models for geometric reasoning. *arXiv preprint arXiv:2312.12241*, 2023.

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

Kimi Team. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Kimi Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251, 2018.

Yann Le, Xuan Yang, et al. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

Bo Li, Peiyuan Zhang, Kaichen Zhang, Fanyi Pu, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Accelerating the development of large multimoal models, March 2024. URL https://github.com/EvolvingLMMs-Lab/lmms-eval.

Junzhuo Li, Bo Wang, Xiuze Zhou, Peijie Jiang, Jia Liu, and Xuming Hu. Decoding Knowledge Attribution in Mixture-of-Experts: A Framework of Basic-Refinement Collaboration and Efficiency Analysis, May 2025. URL http://arxiv.org/abs/2505.24593. arXiv:2505.24593 [cs] version: 1.

Kun Li, George Vosselman, and Michael Ying Yang. Hrvqa: A visual question answering benchmark for high-resolution aerial images. *arXiv preprint arXiv:2301.09460*, 2023.

Wengang Li, Lingqi Zhang, Toshio Endo, and Mohamed Wahib. Understanding cross-layer contributions to mixture-of-experts routing in llms. In *The Fourteenth International Conference on Learning Representations*, 2026.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pp. 1650–1654. IEEE, 2021.

Yilun Liu, Yunpu Ma, Yuetian Lu, Shuo Chen, Zifeng Ding, and Volker Tresp. Parameter-efficient routed fine-tuning: Mixture-of-experts demands mixture of adaptation modules. In *EACL*, 2026.

Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566, 2020.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022.

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

Mason McGill and Pietro Perona. Deciding how to decide: Dynamic routing in artificial neural networks. In *International Conference on Machine Learning*, pp. 2363–2372. PMLR, 2017.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT, January 2023. URL http://arxiv.org/abs/2202.05262. arXiv:2202.05262 [cs].

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering, 2018. URL https://arxiv.org/abs/1809.02789.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models, 2025. URL https://arxiv.org/abs/2410.07149.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*. IEEE, 2008.

nostalgebraist. interpreting gpt: the logit lens, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, 2022.

Koyena Pal, Jiuding Sun, Andrew Yuan, Byron Wallace, and David Bau. Future lens: Anticipating subsequent tokens from a single hidden state. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pp. 548–560. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.conll-1.37. URL http://dx.doi.org/10.18653/v1/2023.conll-1.37.

Gabriel Sarch, Linrong Cai, Qunzhong Wang, Haoyang Wu, Danqi Chen, and Zhuang Liu. Vero: An open rl recipe for general visual reasoning. *arXiv preprint arXiv:2604.04917*, 2026.

Noam Shazeer, *Azalia Mirhoseini, *Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, 2025.

Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. Lora vs full fine-tuning: An illusion of equivalence. In *NeurIPS*, 2025.

Thinking Machines. Inkling: Our open-weights model, 2026. URL https://thinkingmachines.ai/inkling/.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024a. URL https://openreview.net/forum?id=QWTCcxMpPA.

Ryan Wang, Akshita Bhagia, and Sewon Min. Emo: Pretraining mixture of experts for emergent modularity. *arXiv preprint arXiv:2605.06663*, 2026.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697, 2024b.

Jia Wei, Zhonghao Zhang, Ping Chen, Yancheng Pan, Shaoxun Wang, Ziyi Qiu, Longxiang Wang, et al. Hellora: Hot experts layer-level low-rank adaptation for mixture-of-experts models. *arXiv preprint arXiv:2605.18795*, 2026.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. The semantic hub hypothesis: Language models share semantic representations across languages and modalities, 2025. URL https://arxiv.org/abs/2411.04986.

Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*, pp. 133–138, Las Cruces, New Mexico, USA, June 1994. Association for Computational Linguistics. doi: 10.3115/981732.981751. URL https://aclanthology.org/P94-1019/.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026a.

Haolei Xu, Haiwen Hong, Hongxing Li, Rui Zhou, Yang Zhang, Longtao Huang, Hui Xue, Yongliang Shen, Weiming Lu, and Yueting Zhuang. Seeing but not thinking: Routing distraction in multimodal mixture-of-experts. In *ACL*, 2026b.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific data*, 2023.

Yue Yang, Ajay Patel, Matt Deitke, Tanmay Gupta, Luca Weihs, Andrew Head, Mark Yatskar, Chris Callison-Burch, Ranjay Krishna, Aniruddha Kembhavi, et al. Scaling text-rich image understanding via code-guided synthetic multimodal data generation. In *ACL*, 2025.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. In *ICLR*, 2024.

Hongyu Zhang, Cheng Yan, Xiang Xia, and Wuyang Zhang. Beyond global routing aggregation: Phase-aware expert merging for moe vision-language models. *arXiv preprint arXiv:2608.04454*, 2026.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025. URL https://arxiv.org/abs/2507.18071.

Xing Zi, Jinghao Xiao, Yunxiao Shi, Xian Tao, Jun Li, Ali Braytee, and Mukesh Prasad. Rsvlm-qa: A benchmark dataset for remote sensing vision language model-based question answering. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 12905–12911, 2025.

A Appendix

Contents

A Appendix **1**

A.1 Preliminaries: Mixture-of-Experts Models 1

A.2 Additional Details on Visualizations 2

A.2.1 Modality Specialization Addendum 2

A.2.2 kNN Purity 3

A.2.3 Semantic Specialization Analysis 3

A.3 ExpertLens Classification 4

A.4 Additional Analyses 7

A.4.1 Routing Sharpness Analysis 7

A.4.2 Additional Modality Specificity Analysis 7

A.4.3 Additional kNN Purity Analysis 8

A.4.4 Additional Semantic Specialization Visualizations. 9

A.4.5 Routing Stability Analysis 10

A.4.6 Additional t-SNE Projections 10

A.5 Experimental Details 13

A.5.1 Training Details 13

A.5.2 Evaluation Details 13

A.5.3 Domain Ordering Experimental Variability 13

A.5.4 Hyperparameters 14

A.6 Prompts 15

A.1 Preliminaries: Mixture-of-Experts Models

Mixture-of-Experts (MoE) models replace the dense feedforward network in a transformer block with a set of expert MLPs (Lepikhin et al., 2020; Fedus et al., 2022). Let N denote the number of routed experts in each MoE layer, and let $\{E_i^{(\ell)}\}_{i=1}^{N}$ denote the experts in layer ℓ. Given the hidden representation $h_t^{(\ell)}$ of token t, the learned router projection $W_r^{(\ell)}$ computes expert probabilities

$$p_t^{(\ell)} = \mathrm{softmax}\left(W_r^{(\ell)} h_t^{(\ell)}\right), \tag{2}$$

and selects the K experts with the largest probabilities,

$$\mathcal{T}_K^{(\ell)}(t) = \mathrm{TopK}\left(p_t^{(\ell)}\right). \tag{3}$$

The layer output is the weighted combination of the selected experts,

$$\mathrm{MoE}^{(\ell)}(h_t^{(\ell)}) = \sum_{i \in \mathcal{T}_K^{(\ell)}(t)} \alpha_{t,i}^{(\ell)} E_i^{(\ell)}(h_t^{(\ell)}), \qquad \alpha_{t,i}^{(\ell)} = \frac{p_{t,i}^{(\ell)}}{\sum_{j \in \mathcal{T}_K^{(\ell)}(t)} p_{t,j}^{(\ell)}}. \tag{4}$$

Some architectures additionally contain shared experts that are evaluated for every token. Sparse routing increases model capacity while improving throughput by evaluating only a small subset of the expert parameters for each token. In multimodal MoE models, both visual and textual tokens are processed by MoE layers, and routing is performed independently for each token. The resulting routing decisions may therefore depend on token modality, semantic content, and input domain. We say that expert i in layer ℓ is *activated* for token t when $i \in \mathcal{T}_K^{(\ell)}(t)$. We distinguish its activation indicator

$$a_{t,i}^{(\ell)} = \mathbb{1}[i \in \mathcal{T}_K^{(\ell)}(t)] \tag{5}$$

from its routing mass

$$m_{t,i}^{(\ell)} = \alpha_{t,i}^{(\ell)} a_{t,i}^{(\ell)}. \tag{6}$$

Aggregating these quantities over tokens from a modality or domain provides a measure of the corresponding expert affinity. In this work, we study whether semantic specialization emerges in multimodal MoE models and whether this structure can be exploited for efficient, targeted adaptation.

A.2 Additional Details on Visualizations

In this section, we provide additional details on the construction of the visualizations in §3. In §A.2.1 we discuss Optimal Leaf Ordering and the datasets used in Figure 1. In §A.2.2 we provide additional details on the KNN purity analysis conducted in Figure 3. Finally, in §A.2.3 we provide details on the semantic specialization analysis in Figure 4.

A.2.1 Modality Specialization Addendum

Experimental Details. Image patches for this experiment were sourced from Breast-Histopathology (Janowczyk & Madabhushi, 2016), Flowers102 (Nilsback & Zisserman, 2008), MedMnist (Yang et al., 2023), MathVista (Lu et al., 2024), ScienceQA (Lu et al., 2022), CoSyn (Yang et al., 2025), PathVqa (He et al., 2020), and text tokens from MedQA (Jin et al., 2020), PubMedQA (Jin et al., 2019), GSM8K (Cobbe et al., 2021), ARC-Challenge (Clark et al., 2018), OpenBookQA (Mihaylov et al., 2018), MMLU (Hendrycks et al., 2021), and LegalBench (Guha et al., 2023).

Optimal Leaf Ordering. In Fig. 1A, we visualize each expert's preference for image versus text tokens. To make structure among experts easier to see, we order experts within each layer according to the similarity of their router weights using *optimal leaf ordering* (OLO). Importantly, this ordering is determined solely by the router weights and does not use the modality-specific activation statistics shown in the figure.

Recall from §A.1 that $W_r^{(\ell)}$ denotes the router projection for layer ℓ. Let $w_i^{(\ell)}$ denote the row of $W_r^{(\ell)}$ corresponding to expert i. For each layer, we compute the pairwise cosine distance between expert router weights,

$$D_{ij}^{(\ell)} = d_{\cos}\left(w_i^{(\ell)}, w_j^{(\ell)}\right).$$

We use these distances to construct a hierarchical clustering dendrogram, i.e., a tree representing the nested grouping of experts by router-weight similarity.

A dendrogram does not uniquely determine the left-to-right ordering of its leaves, since the two children of any internal node can be exchanged without changing the underlying clustering. OLO resolves this ambiguity by selecting, among the orderings consistent with the dendrogram, a permutation π_ℓ that minimizes the total distance between adjacent experts:

$$\pi_\ell = \arg\min_{\pi} \sum_{k=0}^{N-2} d_{\cos}\left(w_{\pi(k)}^{(\ell)}, w_{\pi(k+1)}^{(\ell)}\right).$$

Thus, experts with similar router weights tend to appear near one another along the x-axis while preserving the hierarchical clustering structure.

After obtaining π_ℓ, the expert shown at x-axis position k is $i = \pi_\ell(k)$. We then compute its image-fraction score using the activation indicator $a_{t,i}^{(\ell)}$ defined in Eq. 5. Let $\mathcal{I}_d$ and $\mathcal{X}_d$ denote the sets of image and text tokens, respectively, sampled from domain d. We define

$$f_{\text{img}}^{(\ell)}(i) = \frac{\sum_d \sum_{t\in\mathcal{I}_d} a_{t,i}^{(\ell)}}{\sum_d \left(\sum_{t\in\mathcal{I}_d} a_{t,i}^{(\ell)} + \sum_{t\in\mathcal{X}_d} a_{t,i}^{(\ell)}\right)}.$$

Because image and text tokens are sampled equally, a score near 0.5 indicates no strong modality preference, while values approaching 1 or 0 indicate increasingly image- or text-specific activation, respectively.

OLO therefore affects only the horizontal ordering of experts in Fig. 1A; each expert's modality preference is computed independently from its observed top-K routing behavior.

A.2.2 kNN PURITY

We use k-nearest-neighbor (kNN) purity to quantify which factors best explain the local structure of token representations. Given the same activation vectors, we assign alternative categorical labels to each sample—e.g., semantic dataset or part of speech for text tokens, and semantic dataset, color, or spatial position for image patches—and measure how strongly nearby representations share each label. Higher kNN purity indicates that the corresponding factor better explains the local organization of the activation space.

Data. For image patches, we sample from 12 VQA datasets spanning a diverse set of visual domains: BREASTHISTOPATHOLOGY (Janowczyk & Madabhushi, 2016), FLOWERS102 (Nilsback & Zisserman, 2008), MEDMNIST (Yang et al., 2023), MATHVISTA (Lu et al., 2024), SCIENCEQA (Lu et al., 2022), COSYN (Yang et al., 2025), AI2D (Kembhavi et al., 2016), MMK12 (Meng et al., 2025), GEOMVERSE (Kazemi et al., 2023), PATHVQA (He et al., 2020), VQARAD (Lau et al., 2018), and CHARXIV (Wang et al., 2024b). Each dataset defines one category when evaluating organization by semantic dataset.

For text tokens, we sample from 12 text-based question-answering datasets spanning diverse domains and linguistic content: MEDQA (Jin et al., 2020), PUBMEDQA (Jin et al., 2019), MEDMCQA (Pal et al., 2022), GSM8K (Cobbe et al., 2021), METAMATHQA (Yu et al., 2024), ARC-CHALLENGE (Clark et al., 2018), OPENBOOKQA (Mihaylov et al., 2018), MMLU (Hendrycks et al., 2021), TRIVIAQA (Joshi et al., 2017), BBH (BIG-bench authors, 2023), FINQA (Chen et al., 2021), and LEGALBENCH (Guha et al., 2023). For the part-of-speech analysis, tokens are assigned to one of 12 categories:

$$\{\text{VERB}, \text{NOUN}, \text{PRON}, \text{ADJ}, \text{ADV}, \text{ADP}, \text{CONJ}, \text{DET}, \text{NUM}, \text{PRT}, \text{“.”}, \text{“?”}\}.$$

We balance the number of samples across categories for each comparison so that purity is not driven by differences in category frequency.

kNN Purity. Let $\mathbf{x}_i \in \mathbb{R}^D$ denote the activation vector of token or image patch i, with categorical label y_i. For each sample, we find its k nearest neighbors in the original D-dimensional activation space using Euclidean distance,

$$d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2,$$

excluding the sample itself. Let $\mathcal{N}_k(i)$ denote the indices of these k neighbors. We define the local purity of sample i as

$$p_k(i) = \frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} \mathbb{I}[y_j = y_i],$$

i.e., the fraction of its nearest neighbors that share its categorical label.

For a labeling scheme containing N sampled representations, the reported kNN purity is the mean local purity across samples:

$$P_k = \frac{1}{N} \sum_{i=1}^{N} p_k(i) = \frac{1}{Nk} \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_k(i)} \mathbb{I}[y_j = y_i],$$

where $\mathbb{I}[\cdot]$ is the indicator function. We report this quantity for multiple values of k in Fig. 3.

A.2.3 SEMANTIC SPECIALIZATION ANALYSIS

We quantify whether semantically related object classes recruit similar sets of experts using the TINYIMAGENET experiments described in §3.3. For each MoE layer and patch-count threshold m, we first construct a set of experts associated with each object class. An expert is considered active for an image if it is routed at least m of that image's patch tokens, and is included in the class-level expert set S_c if it is active for at least 60% of images from class c. We evaluate multiple thresholds $m \in \{3, 5, 8, 12, 16\}$ to ensure that the analysis is not sensitive to a particular definition of expert activation.

Expert-set overlap. For two classes a and b, we measure the overlap between their associated expert sets using the Jaccard similarity,

$$J(a, b) = \frac{|S_a \cap S_b|}{|S_a \cup S_b|}.$$

A larger $J(a, b)$ therefore indicates that the two classes consistently recruit more of the same experts.

Raw Jaccard overlap can vary across classes because some classes may recruit expert sets that overlap broadly with many others. To account for this, we standardize each pairwise overlap relative to the typical overlap of the two classes with the remainder of the class set. For an analysis containing N classes, we define

$$\mu_{a,b} = \frac{1}{2(N-2)} \left(\sum_{k \neq a,b} J(a,k) + \sum_{k \neq a,b} J(b,k) \right),$$

which measures the average overlap of classes a and b with all other classes. We then compute

$$Z(a,b) = \frac{J(a,b) - \mu_{a,b}}{\sigma},$$

where σ is the standard deviation of all off-diagonal entries in the pairwise Jaccard matrix.

Thus, $Z(a, b) > 0$ indicates that classes a and b share more experts with each other than they typically share with other classes, while larger values indicate increasingly strong shared expert specialization.

A.3 ExpertLens Classification

We provide additional details for `ExpertLens`, our data-free method for identifying domain-specialized experts in multimodal MoE models. For each MoE expert, we project its router weight through the language model head and extract the 10 highest-scoring vocabulary tokens. For each domain, we prompt Qwen3-VL-30B-A3B to classify each token as belonging to that domain or not using deterministic greedy decoding. We include all prompts used in §A.6.

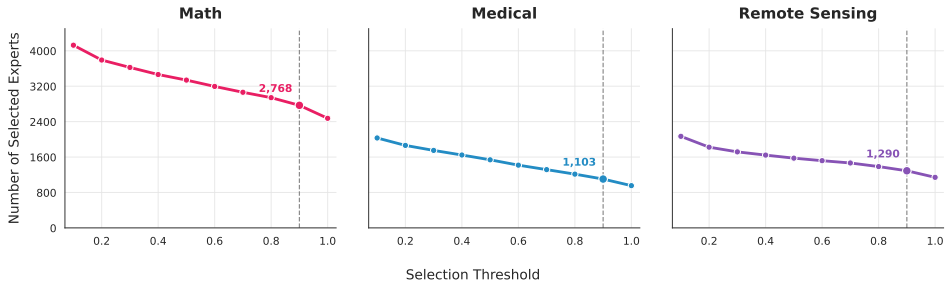

Figure 8: **Sensitivity to score threshold.** We vary the `ExpertLens` selection threshold (S_{EL}) and report the resulting number of selected experts for each domain in Qwen3-VL-30B-A3B. Expert counts decrease smoothly as the threshold becomes more stringent. Our default threshold of $S_{\text{EL}} \geq 0.9$ (dashed line) selects 2,768 math, 1,103 medical, and 1,290 remote-sensing experts.

Expert selection is stable around the chosen score threshold. We examine the sensitivity of `ExpertLens` expert cohorts to the selection threshold S_{EL} in Figure 8. As expected, increasing the threshold progressively reduces the number of experts selected for each domain. Importantly, cohort sizes vary smoothly across thresholds, with no sharp transition around our default value of $S_{\text{EL}} \geq 0.9$. At this threshold, ExpertLens selects 2,768, 1,103, and 1,290 experts for math, medical, and remote sensing, respectively. We use this same threshold across domains rather than tuning it independently for each task.

Limitation: noisy classifications. We now discuss a primary limitation of `ExpertLens`: susceptibility to noise in token-level semantic classification. Since `ExpertLens` determines expert specialization from the semantic relevance of decoded vocabulary tokens, its predictions inherit errors from this token-level classification step. As shown in Table 5, expert readouts can include ambiguous or non-semantic tokens, such as token fragments, punctuation, multilingual fragments, and control tokens. The Qwen3-VL-30B-A3B model used for classification occasionally incorrectly labels these tokens as pertaining to a specific domain. Such errors introduce noise into the inferred expert labels and represent a primary failure mode of our approach.

Medical	Math	Remote Sensing
ly	1	<id 220>
他	<id 220>	...
<id 220>	2	<id 4102>
方面的	0	/or
/or	(	wards
ness	=	\n

Table 5: **Representative noisy ExpertLens token classifications.** Because ExpertLens infers expert specialization by classifying decoded vocabulary tokens, ambiguous tokens, token fragments, punctuation, and control tokens can occasionally be assigned to a semantic domain. We show representative examples for each domain.

ExpertLens specialization persists across layers. Figure 9 extends the routing analysis from Fig. 5A, which visualizes layer 18, to multiple decoder layers. For each layer, we independently identify medical, math, and remote-sensing experts using ExpertLens and measure their routing preferences over the same six datasets. The domain correspondence observed at layer 18 persists broadly throughout the decoder: medical experts preferentially activate on medical datasets, math experts on mathematical datasets, and remote-sensing experts on remote-sensing datasets. Although the strength of this preference varies across layers and modalities, the overall pattern is consistent, indicating that ExpertLens recovers domain-specialized experts throughout the model rather than only at a particular layer.

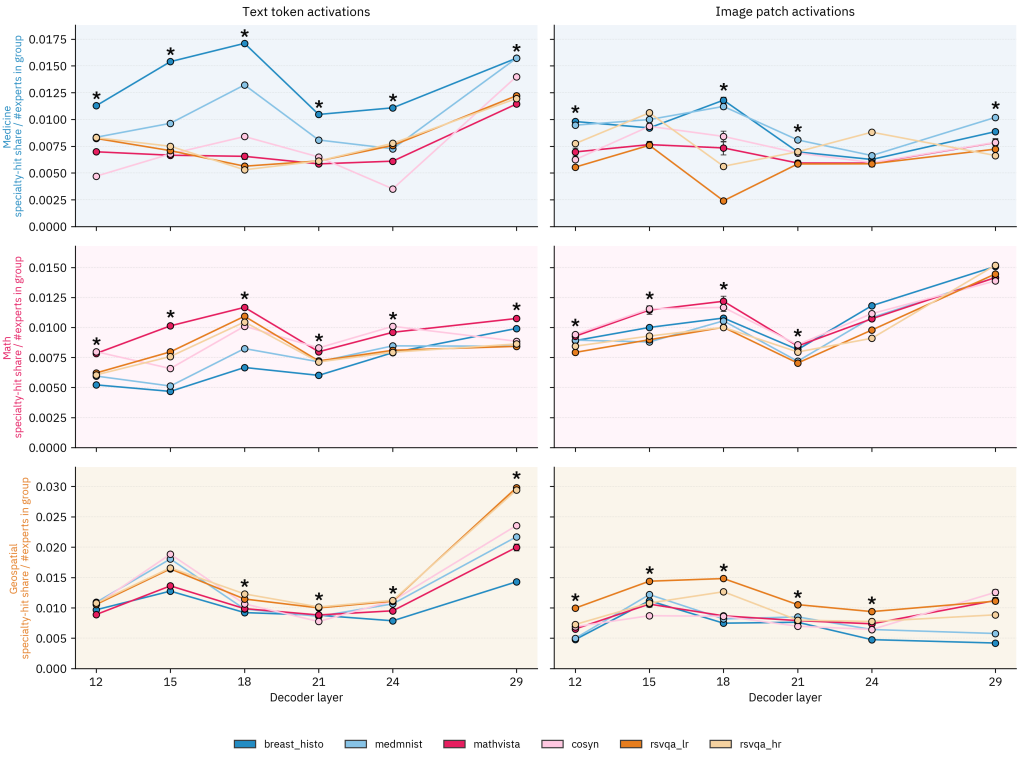

Figure 9: **ExpertLens-derived expert specialization persists across layers.** We extend the analysis from Fig. 5A across decoder layers of QWEN3-VL-30B-A3B. Rows correspond to `ExpertLens`-identified medical, math, and remote-sensing expert cohorts, and columns show routing for text tokens and image patches. Curves report the normalized dataset-hit ratio per expert for each of the six evaluation datasets; asterisks mark layers where at least one of the top-two preferred datasets belongs to the expert cohort's corresponding domain. Across layers, expert cohorts preferentially activate on inputs from their matching domain, showing that the domain alignment observed in Fig. 5A is not specific to a single layer.

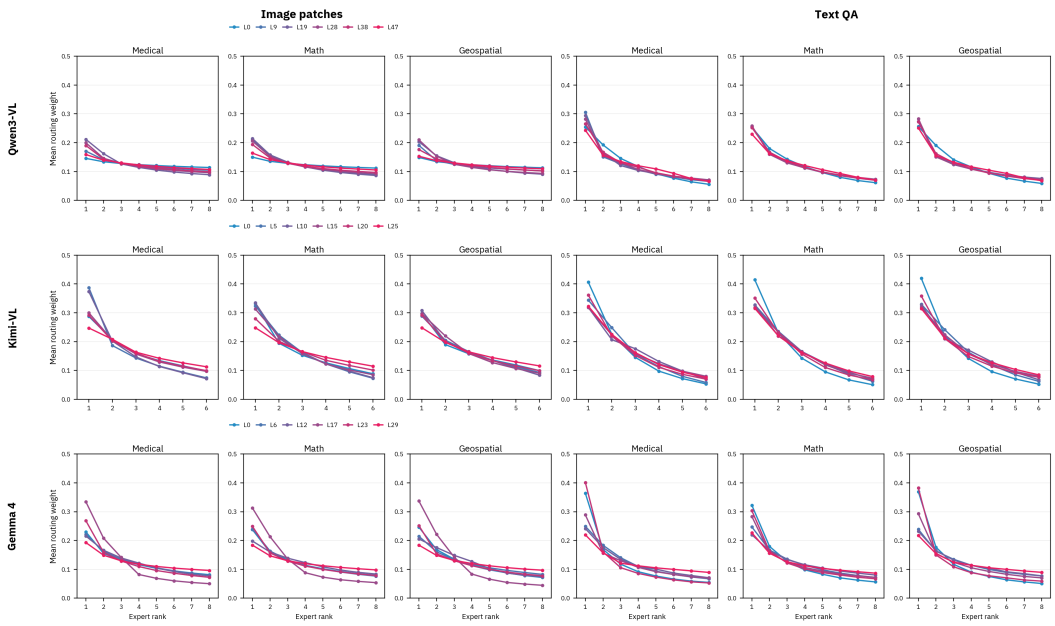

Figure 10: Mean routing weights for the top-k selected experts from each model, on image patches and text tokens. The difference between the top-1 and top-k expert weight is larger for text tokens than image patches in all models. Routing distributions across domains is similar within each modality. Overall, no model exhibits a mean top-1 routing weight over 0.5 or a top-k routing weight below 0.05.

A.4 Additional Analyses

We now provide additional analyses to supplement those in the main paper. We evaluate the distribution of routing scores in §A.4.1, provide additional visualizations for modality preferences (§A.4.2), domain specificity (§A.4.3, §A.4.6), and semantic specialization (§A.4.4). Finally, we explore the stability of routing preferences with `ExpertLens` post-training in §A.4.5.

A.4.1 Routing Sharpness Analysis

We additionally examine how routing mass is distributed among the top-K selected experts. Figure 10 plots the mean routing weight assigned to experts at each rank, separately for image patches and text tokens across medical, math, and remote sensing domains. Across all three models, routing weights decrease smoothly with expert rank rather than collapsing onto a single expert, indicating that routing remains distributed across several selected experts. Text tokens generally exhibit sharper routing than image patches, with a larger gap between the top-ranked expert and the remaining experts. In contrast, routing profiles are broadly similar across domains within each modality, suggesting that routing sharpness is driven more strongly by modality than by domain. Overall, no model assigns more than 0.5 mean routing weight to its top-ranked expert, and even the lowest-ranked selected experts retain non-negligible routing mass, indicating that the observed expert specialization does not arise from near-deterministic routing to a single expert.

A.4.2 Additional Modality Specificity Analysis

We supplement our analysis from §3.1 with similar plots for Kimi-VL-A3B and Gemma-4 26B-A4B in Figure 11. On the left, we show the fraction of top-K activations arising from image patches versus text tokens, organized by OLO (§A.2.1). On the right, we show the number of experts per distinct frequency bins of text-to-image hits ratio. Gemma-4 26B-A4B displays a clear bimodal distribution of expert specialization, with a Hartigan dip value of 0.0214 ($p < 0.001$), whereas Kimi-VL-A3B appears to consist of experts without strong modality preference. Here, note that Gemma-4 26B-A4B and Qwen3-VL-30B-A3B have 128 experts per layer, while Kimi-VL-

A3B has only 64, possibly explaining its lack of modality-specific experts. These results build on our findings from §3.1 and suggest modality preferences are not specific to QWEN3-VL-30B-A3B.

Modality preferences are robust to image scale. We test whether the modality specialization observed in § 3.1 is sensitive to image scale. For each model, we repeat the analysis while varying image width across the original resolution distribution, the dataset mean resolution, and the maximum resolution, while keeping the total number of sampled image patches per domain fixed. We then measure, for each pair of settings, the fraction of experts that remain on the same side of the 50% image/text activation boundary. Figure 12 shows high agreement across scale settings for all three models, indicating that expert modality preference is largely unchanged by image resolution. Thus, the modality specialization we observe is unlikely to be an artifact of image scale or the number of patches contributed by individual images.

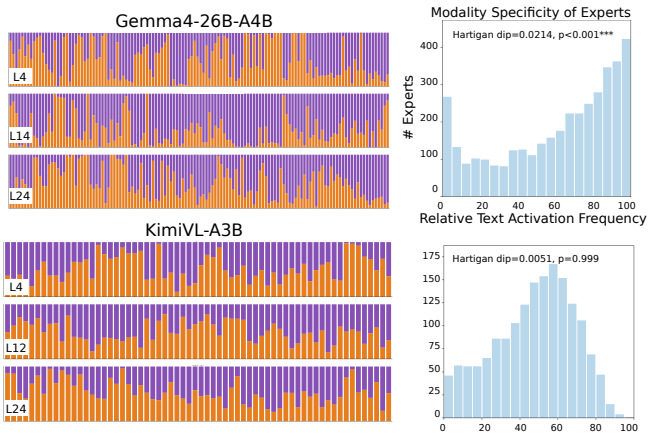

Figure 11: Expanding on Fig.1, GEMMA-4 26B-A4B shows a strong modality-preference with a bimodal frequency histogram, while KIMI-VL-A3B with fewer net experts appears to consist of mostly neutral experts.

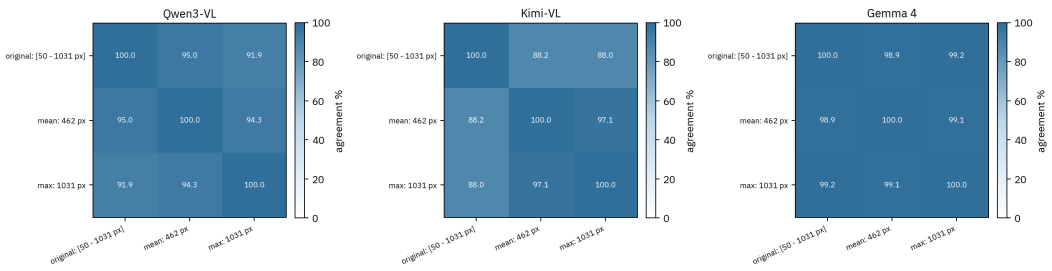

Figure 12: Robustness to image scale. We repeat the experiment in §3.1 but vary the scale of images. Specifically, we fix the total number of patches per domain and vary only image width (`original`: $[50, 1031]$ px; `mean`: 462 px; `max`: 1031 px), so smaller images contribute fewer patches each and we sample from more images. We then count the number of image-leaning versus text-leaning experts in the different experimental runs. Cells show the fraction of experts that retain the same $\geq 50\%$ image/text side across a pair of settings. High agreement across scale changes implies modality specialization is stable under image scale change.

A.4.3 ADDITIONAL KNN PURITY ANALYSIS

We now expand on our domain-specificity analysis in §3.2. Specifically, Figure 3 showed that grouping by dataset led to the highest purity for image activations, while part-of-speech showed the highest purity for text tokens. In Figure 13 we repeat this analysis with clustering by domain. Specifically, we consider our three domains (math, medical, remote sensing) and sample from 2 datasets per domain. Grouping by domain is still the most kNN-pure method in image activations,

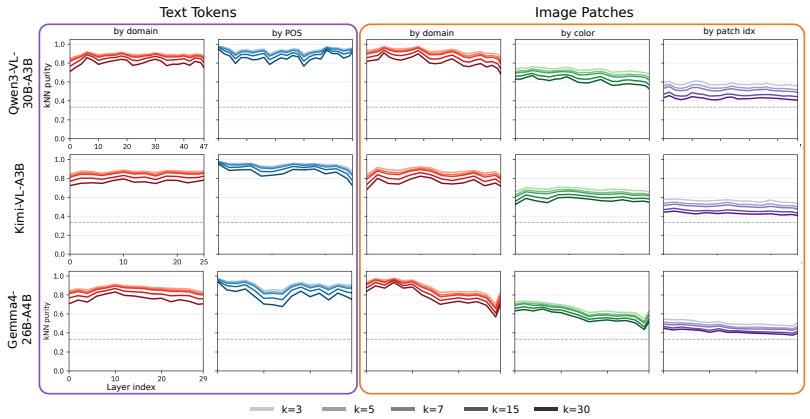

Figure 13: Expanding on Fig.3, which showed that grouping by dataset showed highest kNN purity for image activations, while grouping by POS showed highest kNN purity for text tokens. Clustering by 3 domains (medical, math, remote sensing) with images from 2 datasets per domain, we see that grouping by domain is still the most kNN-pure method in image activations, while POS narrowly exceeds grouping by domain in text tokens. All kNN purity values are higher due to the fewer number of classes. The gray dotted line indicates chance kNN purity.

while POS narrowly exceeds grouping by domain in text tokens. All kNN purity values are higher due to the fewer number of classes. Additionally, we find that both groupings are higher than chance-grouping (shown by the gray dotted line).

A.4.4 Additional Semantic Specialization Visualizations.

We extend our analysis in §3.3 to Kimi-VL-A3B and Gemma-4 26B-A4B. Specifically, in Figure 14 we find that both Gemma-4 26B-A4B and Kimi-VL-A3B illustrate the same behavior as Qwen3-VL-30B-A3B (with Pearson coefficients of -0.343 and -0.349 respectively, comparable to Qwen3-VL-30B-A3B's -0.380 (left). Moreover, the right column in Figure 14 shows that semantically related classes elicit similar experts, with all pairs of related classes above the gray baseline.

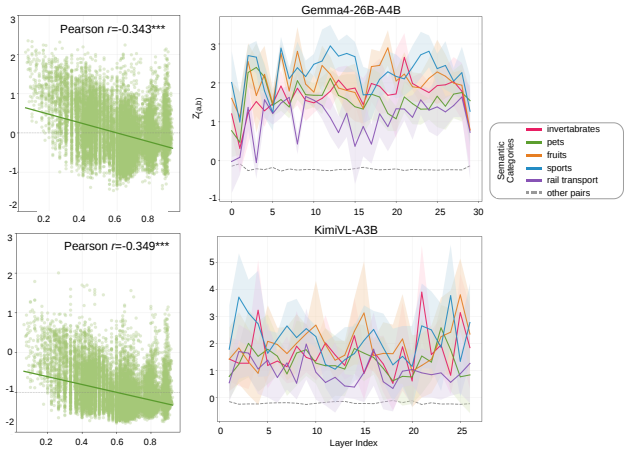

Figure 14: Expanding on Fig. 4, object level categorization can be observed in Kimi-VL-A3B and Gemma-4 26B-A4B. The left hand for each plot shows Z(a, b) plotted against the semantic distance between object class pairs (x axis). Semantically unrelated object classes share fewer experts, as with Qwen3-VL-30B-A3B. The right hand side for each plot shows the category level expert specialization compared to unrelated category pairs. Images from objects in the same semantic category share more experts than any two unrelated object classes.

A.4.5 Routing Stability Analysis

Although ExpertLens updates the router throughout continual adaptation (Table 2), its routing structure remains highly stable. Consecutive checkpoints preserve $92.0 - 94.0\%$ top-8 expert-set overlap, measured as the Jaccard similarity between the highest-probability experts selected for each input and layer. After completing all three training stages, the final model retains 90.1% overlap with the pretrained router, with similarly high retention across Math (90.2%), Medical (90.1%), and Remote Sensing (89.9%) inputs. Thus, training the router enables adaptation without substantially disrupting the model's pretrained routing structure or causing routing collapse.

A.4.6 Additional t-SNE Projections

To supplement our analysis in §3.2, we provide additional t-SNE projections. Specifically, we visualize additional layers for Qwen3-VL-30B-A3B in Figure 15, and provide visualizations for Kimi-VL-A3B in Figure 17 and Gemma-4 26B-A4B in Figure 16. Across all visualizations, we find that projections of image patch activations are highly clustered by domain, while textual activations are comparatively more diffuse.

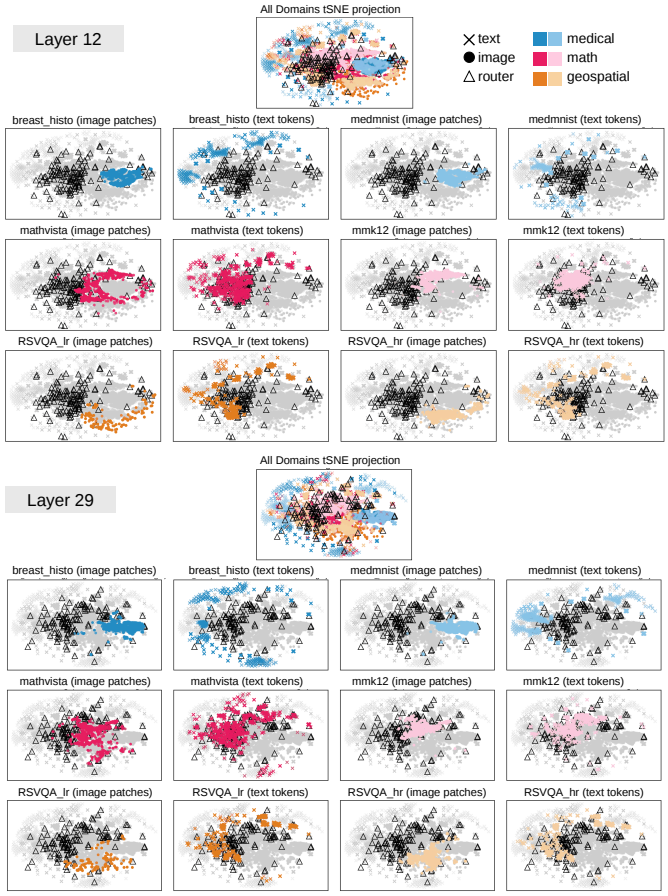

Figure 15: **t-SNE projections for additional layers of Qwen3-VL-30B-A3B**. We visualize layers 12 and 29 of Qwen3-VL-30B-A3B, supplementing the plot included in the main paper. Across various layers we find that visual activations exhibit stronger domain structure than textual activations.

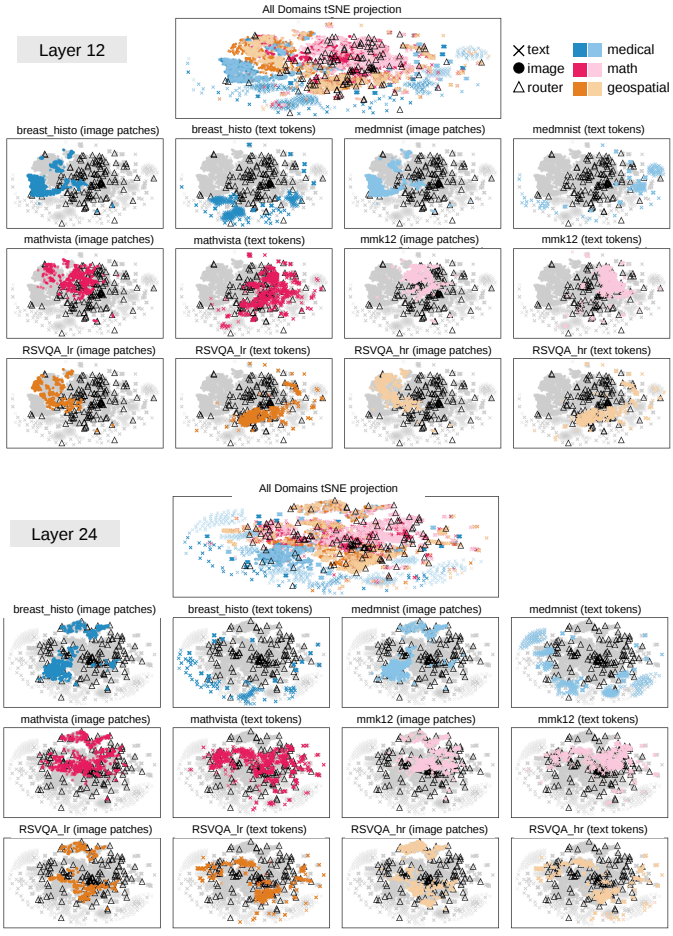

Figure 16: **t-SNE projections for GEMMA-4 26B-A4B**. We visualize Layers 12 and 24 and find image patches are more tightly clustered than text tokens.



Figure 17: **t-SNE projections for KIMI-VL-A3B**. We visualize Layers 12 and 24 and find image patches are more tightly clustered than text tokens.

A.5 Experimental Details

A.5.1 Training Details

Training Data and Rewards. We train on 4,000 examples per domain, sampled from the corresponding training splits. For math, all samples are from MMK12 (Meng et al., 2025). For medical, we use all 1,793 samples from VQARAD (Lau et al., 2018) and 2,207 examples from PATHVQA (He et al., 2020). For remote sensing, we sample 2,000 examples each from RSVQALR (Lobry et al., 2020) and RSVQAHR (Lobry et al., 2020). We adopt the system prompt and reward functions from VERO (Sarch et al., 2026).

Implementation. We use QWEN3-VL-30B-A3B as our base multimodal MoE and GSPO (Zheng et al., 2025) as our post-training algorithm. All experiments are conducted on 4 H200 GPUs for 200 steps per domain, with a batch size of 32, group size of 5, and learning rate of 1×10^{-5}. We use verl (Sheng et al., 2025) for training. Additional hyperparameters are provided in Table 7. We show results for three seeds, with shaded regions indicating 95% confidence intervals.

A.5.2 Evaluation Details

We evaluate all models using lmms-eval (Li et al., 2024) with the vLLM inference backend. We use zero-shot evaluation and sample responses with recommended sampling parameters for QWEN3-VL-30B-A3B: temperature 0.7, top-p 0.8, top-k 20, a presence penalty of 1.5, and a repetition penalty of 1.0, allowing up to 4,096 generated tokens. We use identical prompts, image preprocessing, decoding parameters, and evaluation seeds for all methods.

Evaluation Benchmarks. We evaluate on benchmarks spanning all three domains. For medical, PATHVQA (He et al., 2020) pairs pathology images with open-ended and yes/no questions about clinical findings. VQARAD (Lau et al., 2018) contains clinician-authored open- and closed-ended questions over radiographs, CT scans, and MRI images, covering anatomy, imaging modality, and abnormalities. SLAKE (Liu et al., 2021) spans multiple imaging modalities and body regions, with questions requiring both visual recognition and medical knowledge.

For mathematical reasoning, MATHVISTA (Lu et al., 2024) covers diverse visual mathematics tasks, including arithmetic, geometry, logical reasoning, and interpretation of charts, tables, and scientific figures. MATHVISION (Wang et al., 2024a) contains challenging visual problems drawn from real mathematics competitions, spanning 16 mathematical disciplines and five difficulty levels. MMK12 (Meng et al., 2025) comprises manually collected K–12 multiple-choice problems.

For remote sensing, HRVQA (Li et al., 2023) uses high-resolution aerial imagery to test counting, scene recognition, spatial understanding, and fine-grained object attributes such as size, color, and shape. RSVLMQA (Zi et al., 2025) combines remote-sensing imagery with questions about captions, semantic content, spatial relationships, and object counting. RSVQALR (Lobry et al., 2020) uses lower-resolution Sentinel-2 imagery to evaluate object presence, counting, comparison, and rural-versus-urban classification, whereas RSVQAHR (Lobry et al., 2020) uses very-high-resolution aerial orthophotos to test fine-grained object presence, counting, comparison, and area estimation.

A.5.3 Domain Ordering Experimental Variability

In Table 6 we report standard deviations across three seeds for the domain ordering results shown in Table 3. We find that `ExpertLens` generally exhibits lower intra-seed variation, as well as lower variability per-ordering.

Domain Ordering	Method	Math	Medical	Remote Sensing	Mean
Math → Med. → Rem. Sens.	FFT	1.3	**0.7**	0.4	0.8
	`ExpertLens`	**0.4**	**0.7**	**0.2**	**0.4**
Med. → Math → Rem. Sens.	FFT	**0.7**	0.9	**0.2**	**0.6**
	`ExpertLens`	2.0	**0.2**	0.5	0.9
Rem. Sens. → Med. → Math	FFT	1.6	1.0	**0.2**	0.9
	`ExpertLens`	**0.1**	**0.1**	0.3	**0.2**

Table 6: **Domain ordering experimental variability.** We report standard deviations across three seeds for our domain ordering experiments.

A.5.4 HYPERPARAMETERS

Hyperparameter	**Value**
Training objective	GSPO
Training steps per domain	200
Batch size	32
Group size	5
Learning rate	1×10^{-5}
Optimizer	AdamW
Weight decay	0
LR schedule	Constant
Maximum prompt length	1,024
Maximum response length	2,000
KL coefficient	0
Clipping range	$[0.20, 0.28]$
Temperature	1.0
Top-p	1.0
Top-k	Disabled
Repetition / presence penalty	1.0 / 1.5
Maximum image pixels	401,408
ExpertLens selection threshold	≥ 0.9
Training and rollout dtype	bfloat16
Rollout engine	vLLM

Table 7: Hyperparameters for the Qwen3-VL experiments.

A.6 PROMPTS

```
Classify each token list below (one MoE expert per item).
Tokens may be English, Chinese, or other languages.
Include partial/subword matches (e.g. physio, histo, sat, geo, raster, ndvi, geom, digits, math symbols).

Return JSON only with one entry per expert below (same expert_id values, none omitted):
{ "items": [ { "expert_id": <int>, "labels": [[<specialist>, ...], ...] } ] }
Each inner list is one token's specialist tags (subset of medicine, geospatial, math).
Use [] for generalist-only tokens. A token may appear in multiple inner lists' domains.

Experts:
  expert_id=<expert ID> tokens=[<10 decoded tokens>]
  expert_id=<expert ID> tokens=[<10 decoded tokens>]
  ...
```

Figure 18: **ExpertLens system prompt.**

```
You classify vocabulary tokens from a MoE router logit lens (top predicted tokens per expert).

These tokens are often BPE or byte-level fragments, NOT only complete dictionary words.
Judge semantic association generously using morphology, substrings, symbols, and context.

For EACH token, return a list of specialist domains that apply (zero or more):
  medicine, geospatial, math

A token may match MULTIPLE specialists (e.g. a coordinate token can be both geospatial and math, or a
biomedical measurement can be both medicine and math).
Use an empty list [] when the token is generalist only (no specialist signal).

NUMBERS, UNITS, AND SYMBOLS:
  -- Digit / numeric tokens (0, 1, 2, 12, 0.5, ...): tag math when they indicate quantities, formulas,
  coordinates, measurements, or other numeric reasoning.
  -- Percent sign (\%) and percent-related fragments: tag math.
  -- Latitude/longitude, coordinate, map-scale, band-index, raster-size, pixel-size, ground-sampling-distance/
  GSD, resolution, and geodesic notation may be tagged both geospatial and math.

Do NOT use "generalist" inside per-token lists; use [] instead.

medicine---clinical / biomedical / healthcare (patient, disease, diagnosis, treatment, anatomy, physiology,
pathology, radiology, cells, DNA, gene, symptom, drug, hospital, ...).
geospatial---geospatial / remote-sensing / satellite, aerial, overhead, drone, Earth-observation, GIS, mapping
, terrain/topography, land cover/use, agriculture/crops, urban features, roads/streets/highways/intersections/
bridges/railways/runways/airports, buildings/roofs/parking lots, water/lakes/ponds/rivers/streams/canals/
reservoirs/wetlands/coastlines/beaches/ocean/sea, floods, wildfire/burn scars, clouds/shadows, elevation DEM/
DSM/DTM, LiDAR/point clouds, SAR/InSAR/radar/backscatter, panchromatic/RGB/false-color/NIR/SWIR/multispectral/
hyperspectral bands, Sentinel/Landsat/MODIS/Planet/Maxar/WorldView/NAIP, NDVI/EVI, raster/vector/tile/chip/
scene/polygon/bounding box, GeoTIFF/GeoJSON/shapefile, latitude/longitude/projection/CRS/UTM/WGS84/
georeferencing, and related fragments.
math---equations, operators, geometry/algebra fragments (geom, trig, calc, ^, ...).

Tokens may be English, Chinese, or other languages.
Respond with valid JSON only.
```

Figure 19: **ExpertLens classification prompt.**