Vision-Language Models & Interpretability
Attention Head Probing, Visual Token Redundancy, and Distillation to Small VLMs

12-Week Research Roadmap (Five Milestones; Research-Intern Friendly)

Prepared for: *Harshal Bhat*
Prepared by: *Mentor (Dr. Raj)*
Version: December 10, 2025

Project One-Line: Build a reproducible experimental pipeline to (i) characterize what specific attention heads do in a VLM during multimodal processing, (ii) quantify how much information each visual token carries and where redundancy emerges across layers, and (iii) distill a strong teacher VLM (open or API-based) into a smaller VLM using SFT and preference-style tuning, aiming for measurable quality gains at lower compute.

Primary Research Questions:

- **Attention heads:** Which attention heads consistently implement cross-modal alignment (text→vision and vision→text), grounding, or formatting/position operations, and can we validate these roles with causal interventions and probes?
- **Visual tokens:** What is the *marginal contribution* of each visual token to downstream performance, and do tokens become redundant after certain layers (enabling pruning/merging with minimal loss)?
- **Distillation:** Can we distill stronger multimodal behavior from a teacher (e.g., GPT/Gemini/LLaVA-family) into a compact student via SFT and preference/distillation tuning, and do head/token insights improve the distillation recipe?

Feasibility Note (3 months): All three ideas are achievable in a 12-week window if you (1) standardize on **one primary open VLM family** for mechanistic work (to access attentions and activations) and (2) treat distillation as a **controlled, limited-scope** training run (single student size, fixed dataset, tight ablations). Using GPT/Gemini as a teacher is optional and should be limited to a small, well-chosen dataset due to cost and policy constraints.

Contents

Contents

1 Project Overview

1.1 High-Level Goals (Pattern-Aligned)

Phase	Goal
Discover (W1–2)	Read core VLM interpretability, attention analysis, token pruning/merging, and distillation papers. Lock scope: model(s), tasks, metrics, and a concrete "success band."
Instrument (W3–4)	Build an evaluation harness and add hooks to capture attention maps, head outputs, and hidden states. Establish baselines. Create small probing datasets.
Probe (W5–6)	Run head-level analyses: attention mass patterns, clustering, probes, and causal ablations/patching to validate head function hypotheses.
Compress & Distill (W7–10)	Quantify token information and redundancy; implement pruning/merging policies; run SFT + preference-style distillation to a small student; evaluate tradeoffs.
Conclude (W11–12)	Lock final test runs; do ablations and error analysis; package code and write a short paper-style report (6–8 pages) and a demo notebook.

1.2 Timeline at a Glance

Milestone	Weeks	Key Deliverables
M1: Foundations & Scope	1–2	Literature matrix (10–12 items), scoped research plan (1–2 pages), metrics + target bands, risk/ethics memo.
M2: Models, Data, Instrumentation	3–4	Reproducible eval harness; baseline scores; attention/activation logging; probing datasets; analysis notebook skeletons.
M3: Attention Head Roles + Probing	5–6	Head taxonomy (by behavior), probe results, causal ablation/patching results, top heads list + evidence.
M4: Visual Token Information & Redundancy	7–9	Token contribution curves; layer-wise redundancy analysis; pruning/merging policies; accuracy-vs-tokens plots.
M5: Distillation to Small VLM + Finalization	10–12	Student model results (SFT + preference distillation), ablations, final report, reproducible repo + demo.

1.3 Recommended Model and Benchmark Choices (Keep It Simple)

To keep the project tractable, pick one **primary open** VLM where you can inspect internals (attention maps, head outputs, hidden states). Use a second model only if necessary for comparison.

Component	Recommendation (Practical Defaults)
Primary open VLM	A LLaVA-style model (ViT vision encoder + projector + LLM) or any open VLM with full attention access.
Teacher for distillation	Prefer an open larger teacher for reproducibility; optionally add a small, curated set of GPT/Gemini teacher outputs if permitted and affordable.
Student VLM	A compact variant (e.g., 1–3B LLM backbone) compatible with the same vision encoder or a smaller vision encoder.
Evaluation harness	Use a single evaluation framework (or a consistent script) to run the same prompts across models, log outputs, and compute metrics.
Benchmarks (pick 2–3)	One general VQA set, one text-in-image/OCR style set, and one reasoning set (limit scope). Also include a small in-house "debug" set (50–100 examples) for fast iteration.

1.4 Ethics, Licensing, and Responsible Use

- Use public datasets with clear licenses. Document dataset sources and intended use via Data Cards.
- If using proprietary teachers (GPT/Gemini), follow their terms for data retention and model training. Store prompts/outputs carefully and avoid sensitive data.
- Avoid unsafe or disallowed content generation. The project is research-only; do not claim deployment readiness.
- Include a brief limitations section: dataset bias, benchmark leakage risk, and interpretability limits of attention-based explanations.

2 Milestone 1 — Foundations & Scope (Weeks 1–2)

Objectives

- Narrow the three ideas into a single coherent project with clear hypotheses and measurable outcomes.
- Choose one primary VLM family, 2–3 benchmarks, and a minimal set of intervention/probing methods.
- Define success criteria for each workstream: heads, tokens, and distillation.

Tasks

1. **Literature sprint (10–12 items):** VLM architecture basics, attention head interpretability, causal tracing/ablation in transformers, token pruning/merging methods, and multimodal distillation.
2. **Scope lock:** Decide (a) primary VLM, (b) student size, (c) benchmarks, (d) compute constraints, (e) what is "core" vs "stretch."
3. **Write hypotheses:**
   - H1: A small subset of heads carry most cross-modal grounding signal; ablating them causes large drops.
   - H2: Visual tokens become redundant after mid-layers; pruning/merging beyond a threshold preserves performance.
   - H3: Distillation improves when training emphasizes grounded answers and/or uses token-efficient representations.
4. **Evaluation plan (1–2 pages):** metrics, acceptance thresholds, and reporting tables/figures.

Deliverables

- `docs/lit-matrix.xlsx` (10–12 entries) and `docs/scope.pdf` (1–2 pages).
- `docs/eval-plan.pdf` with metrics, target bands, and benchmark splits.
- `docs/risk-ethics.pdf` (1 page).

Acceptance Check

A reader can answer: *What exact model(s), datasets, metrics, and ablations will be run, and what would count as success by Week 12?*

3 Milestone 2 — Models, Data, and Instrumentation (Weeks 3–4)

Objectives

- Build a clean evaluation harness and establish baseline performance.
- Add instrumentation to log attention maps, head outputs, and hidden states for multimodal prompts.
- Create probing datasets and a consistent experiment configuration system.

Tasks

1. **Repo skeleton:** Standard folders (`src/`, `configs/`, `notebooks/`, `results/`, `docs/`).
2. **Evaluation harness:** Script to run inference on chosen benchmarks; save (prompt, image id, output, logits optional).
3. **Instrumentation hooks:**
   - Log per-layer attention matrices (or a tractable summary: attention mass to visual tokens).
   - Log per-head outputs (value-weighted output vectors) for selected layers.
   - Log hidden states for a few key tokens (e.g., final answer tokens, special image tokens).
4. **Probing data:** Construct small, controlled tasks:
   - Synthetic grounding (colored shapes, counting, spatial relations).
   - Text-image binding (caption with swapped attributes).
   - "Needle" tasks: hide a small detail in the image and ask targeted questions.
5. **Baselines:** Run baseline metrics for the primary VLM on all chosen benchmarks and on the debug set.

Deliverables

- `results/baseline-metrics.csv` and `results/baseline-outputs.jsonl`.
- `src/instrumentation.py` with hooks and a config to turn logging on/off.
- `data/probes/` (probing datasets) + Data Card describing generation method.
- Notebook: `notebooks/00-baseline-and-logging.ipynb`.

Acceptance Check

Instrumentation produces consistent logs for at least 200 multimodal examples, and baselines reproduce within small variance across runs (fixed seeds, fixed decoding).

4 Milestone 3 — Attention Head Roles and Probing Tests (Weeks 5–6)

Objectives

- Identify and categorize attention head behaviors during multimodal processing.
- Build probing tests to predict head function and validate with causal interventions.
- Produce a ranked list of "important" heads for cross-modal grounding.

Methods (Core Set)

- **Attention mass diagnostics:** For each head, compute average attention mass from text queries to visual keys, and vice versa; stratify by layer.
- **Behavior clustering:** Cluster heads using features (modality preference, entropy, positional bias, specialization on special tokens).
- **Probing tests:** Train lightweight probes (e.g., linear classifiers) to predict head category or to predict groundedness outcomes from head features.
- **Causal validation:** Head ablation (zero-out head output) and/or activation patching to measure Δ performance on probing set and benchmarks.

Tasks

1. Compute per-head summary statistics and build a head feature table.
2. Define head categories (example taxonomy):
   - Cross-modal grounding heads (text→vision)
   - Vision summarization heads (vision→vision consolidation)
   - Answer formatting / "instruction-following" heads (text→text)
   - Positional / delimiter / special-token routing heads
3. Run ablations:
   - Ablate top-k candidate heads and measure drop on the probing set.
   - Compare to random-k head ablations to show specificity.
4. Write an interpretable "head report" with example heatmaps and short narratives for representative heads.

Deliverables

- `results/head-features.csv` and `results/head-taxonomy.json`.
- `results/head-ablation.csv` with Δ metrics vs baseline.
- Notebook: `notebooks/10-head-analysis.ipynb` (plots + examples).

- 1–2 page writeup: `docs/m3-heads-summary.pdf`.

Acceptance Check

You can name at least 5–10 heads with strong evidence (both diagnostic + causal) that they matter for grounding on the probing set, and the effect is larger than random head removal.

5 Milestone 4 — Visual Token Information and Redundancy (Weeks 7–9)

Objectives

- Quantify how much each visual token contributes to performance and how redundancy evolves across layers.
- Identify a pruning/merging strategy that reduces the number of visual tokens with minimal quality loss.
- Connect token findings to mechanistic head findings when possible (e.g., do "grounding heads" focus on a small subset of tokens?).

Core Analyses

- **Token drop curves:** Drop/zero a fraction of visual tokens and measure performance vs kept tokens (random vs heuristic).
- **Layer-wise sensitivity:** Apply token dropping at different layers (or simulate by masking attention to subsets) and measure sensitivity per layer range.
- **Redundancy metrics:** Token embedding similarity, effective rank, attention entropy, and "marginal utility" of tokens.
- **Pruning policies:** (i) attention-based keep (tokens most attended by answer tokens), (ii) clustering/merging similar tokens, (iii) learned gating (lightweight scorer).

Tasks

1. Implement token-masking hooks and run controlled experiments on the probe set first (fast iteration), then on one benchmark.
2. Produce per-layer plots:
   - Accuracy (or task score) vs number of visual tokens.
   - Sensitivity heatmap: layers vs token fraction kept.
3. Select one "best" token reduction method and run it end-to-end with the base model.
4. Optional stretch: show that the token reduction interacts with specific head categories (e.g., removing tokens changes which heads activate).

Deliverables

- `results/token-drop-curves.csv` and `results/layer-sensitivity.csv`.
- Notebook: `notebooks/20-token-redundancy.ipynb` (plots + key findings).
- A recommended policy: `docs/token-pruning-recipe.pdf` (1 page).

Acceptance Check

You can demonstrate a concrete tradeoff curve (tokens vs score) and identify a regime where you remove a meaningful fraction of tokens with only a small drop on at least one benchmark and the probe set.

6 Milestone 5 — Distillation to a Small VLM (SFT + Preference Distillation) and Finalization (Weeks 10–12)

Objectives

- Distill a teacher VLM into a compact student VLM using SFT and a preference-style objective (DPO/DPT-like).
- Evaluate whether token-efficient representations and/or head-informed signals improve the student.
- Package the project (reproducibility) and write a short report.

Distillation Plan (Practical)

- **Data:** Curate a multimodal instruction set (images + prompts) plus a small held-out test set. Optionally augment with teacher-generated outputs.
- **SFT:** Train student to imitate teacher outputs (or high-quality references) with standard cross-entropy.
- **Preference distillation:** Create preference pairs (chosen vs rejected) from teacher outputs, self-consistency sampling, or rubric-based filtering; train with a preference objective.
- **Token efficiency:** If Milestone 4 yields a safe token reduction method, apply it during student training and inference, and report compute/latency changes.

Tasks

1. Define teacher/student configurations and training budget (max steps, batch sizes, resolution).
2. Build distillation dataset and a "data card" (sources, filtering, known risks).
3. Run:
   - Student baseline (no distillation or minimal SFT).
   - Student + SFT.
   - Student + SFT + preference distillation.
4. Ablations (small, high-value):
   - Without token pruning vs with token pruning.
   - Preference distillation on/off.
   - Teacher type: open teacher only vs open + small GPT/Gemini subset (if used).
5. Finalize report and repo:
   - README with 5-minute quickstart.
   - Reproducible configs and scripts.
   - Final tables/figures + error analysis.

Deliverables

- `results/student-metrics.csv` and `results/distill-ablations.csv`.
- Final notebook: `notebooks/30-distillation-results.ipynb`.
- Reproducible repo: configs + scripts + pinned requirements.
- Report (6–8 pages): `report/manuscript.pdf`.

Acceptance Check

A third party can run your evaluation script, reproduce the main table/figure, and observe a clear improvement from distillation (even if modest), with honest limitations and an interpretable analysis of when/why it works.

7 Metrics, Targets, and Reporting (What Results to Obtain)

7.1 Primary Metrics

- **Task performance:** benchmark score(s) you selected (accuracy, exact match, or rubric-based score).
- **Head analysis:** (i) head category separability (probe accuracy), (ii) causal importance (Δ score under ablation), (iii) stability across prompts/images.
- **Token analysis:** (i) performance vs kept tokens curve, (ii) layer-wise sensitivity, (iii) redundancy statistics (e.g., effective rank).
- **Distillation:** student performance gain vs baseline; compute/latency changes if token reduction is used.

7.2 Concrete Target Bands (Educational, Not SOTA)

Set targets relative to your own baseline (avoid chasing leaderboard results):

- **M3 heads:** Top-k heads ablation causes a noticeably larger drop than random-k (clear separation).
- **M4 tokens:** Remove a meaningful fraction of tokens (e.g., 25–50%) with only a small drop on the probe set and at least one benchmark.
- **M5 distillation:** Student improves over its pre-distillation baseline on at least one benchmark and does not regress badly on the probe set.

7.3 Required Figures and Tables

- **Table 1:** Baseline vs ablated-head performance (random-k vs targeted-k).
- **Figure 1:** Token keep fraction vs score (multiple policies on one plot).
- **Figure 2:** Layer-wise sensitivity heatmap (layers vs keep fraction).
- **Table 2:** Student distillation results (baseline, SFT, SFT+preference), plus token-reduction on/off.

8 Milestone Details (Week-by-Week)

Week	Focus	Concrete Outputs
W1	Literature + scope	5 papers summarized; shortlist models/benchmarks; draft hypotheses.
W2	Eval plan + repo plan	Metrics/targets locked; repo skeleton; risk/ethics memo.
W3	Baselines + logging	Baseline runs complete; attention/activation logging works end-to-end.
W4	Probe datasets	Synthetic probing sets ready; fast debug evaluation loop; analysis notebooks started.
W5	Head diagnostics	Per-head attention mass stats; head clustering; initial head categories.
W6	Causal head tests	Head ablations/patching; probe results; top heads + evidence.
W7	Token drop curves	Token masking implemented; random vs heuristic token drop curves on probe set.
W8	Layer sensitivity	Layer-wise token sensitivity experiments; redundancy metrics computed.
W9	Pruning/merging policy	Best token-reduction method selected; end-to-end benchmark run with reduced tokens.
W10	Distillation dataset + SFT	Distillation data built; student baseline + SFT run; initial results.
W11	Preference distillation + ablations	Preference-style tuning run; token pruning on/off; teacher variants if used.
W12	Finalization	Locked test results; error analysis; report draft; repo packaging + demo notebook.

9 Deliverables Summary (Per Milestone)

- **M1:** literature matrix; scoped plan; eval plan; risk/ethics memo.
- **M2:** baseline metrics; logging/instrumentation; probing datasets; initial notebooks.
- **M3:** head taxonomy; probe results; causal ablations; head summary writeup.
- **M4:** token contribution curves; layer sensitivity; pruning/merging recipe.
- **M5:** student distillation results; ablations; reproducible repo; 6–8 page report.