SEMALIGN-VQA: SEMANTIC ALIGNMENT FOR MEDICAL VISUAL QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical visual question answering requires correct answers to persist across changes in question formulation. We study how paired training changes this outcome, separating correct-answer retention from answer agreement. SemAlign-VQA combines original-question answer supervision with SemLoss, a pull–push regularizer on projected pre-answer states, using LoRA and a projector bypassed at inference. Evaluation uses an original question and five fixed context wrappers on PathVQA or automatically screened rewrites on SLAKE-English and VQA-RAD. At seed 42, Moondream2 improves all-six correctness by 1.85 and 2.17 percentage points on PathVQA and SLAKE, with nearly unchanged original accuracy. Matched-question analysis locates these gains in retention, while outcome transitions distinguish sustained correct answers from stabilized errors. The training controls reveal a second distinction: direct rewrite supervision reaches 65.22% all-six correctness on SLAKE versus 57.30% with original-only alignment, and the summary-level answer-aware NCE control nearly matches the margin objective on PathVQA despite lower agreement. Effects also vary across backbones and training seeds. Together, the paired comparisons show where alignment changes correctness, how those changes differ from consistency gains, and why supervision-matched controls matter when choosing a training objective.

1 INTRODUCTION

Medical VQA models can answer an original question correctly and fail under a different formulation. Medical instruction tuning expands their range of answers (Li et al., 2023; Chen et al., 2024a), but accuracy on one prompt leaves this behavior unresolved. Agreement addresses another question: whether answers repeat. A model can become more consistent by preserving correct answers, stabilizing wrong answers, or changing both. Evaluating training interventions therefore requires locating the change within question groups.

SemAlign-VQA couples a paired training design with this group-level analysis. SemLoss regularizes projected pre-answer states while answer cross-entropy (CE) supervises the original question. A same-image rewrite supplies the positive view; LoRA carries the update into ordinary generation, which bypasses the projector. The design builds on question augmentation and consistency learning (Kant et al., 2021; Jiang et al., 2025; Sadanandan & Behzadan, 2026b). Its original-only CE setting holds answer supervision fixed so that the paired representation objective can be examined separately from supervising answers to rewritten questions.

We organize the evidence around three questions: where joint-correctness gains occur, how they change with the training signal, and how they vary across implementations and seeds. The primary Moondream2 gains arise largely from retaining answers that were already correct. On PathVQA, however, the net reduction of inconsistent groups includes both 124 additional all-correct groups and 93 additional identically wrong groups. The Moondream2 controls achieve their highest SLAKE all-six scores with rewrite-answer supervision; on PathVQA, margin and answer-aware NCE have nearly equal mean all-six correctness. We compare the training signals, then examine paired regularization across backbones and seeds.

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.

Table 1: Training signals for four cited consistency methods and SemAlign-VQA.

Work + setting	Aligned signal	Answer supervision	Inference
ConClaT (2021) General VQA	Projected V–L states; rephrase and same-answer positives; different-answer negatives	CE on augmented questions	Classifier; no projector
CCL (2025) Medical VQA	Mean-pooled LLM states; original and three perturbations; other batch questions as negatives	NLL on original and three perturbations	Standard generation
CHIL LoRA (2026) Binary med. VQA	Yes/no distributions via symmetric KL; no negatives	NLL on original and paraphrase	LoRA generation
SEQA (2025) Medical VQA	No pairwise loss; LLM rephrases augment the QA set	Model-native loss on augmented QA	Unchanged model
SemAlign-VQA Generative med. VQA	Projected pre-answer states; pull same-image rewrite; relative-margin push against other-item positives	CE on original only	LoRA; no projector

2 Related Work

Training on related questions. ConClaT combines answer classification with contrastive learning of joint visual–linguistic representations (Kant et al., 2021). CCL combines hidden-state contrastive learning and answer supervision for generative medical VQA (Jiang et al., 2025). Mechanistically guided LoRA aligns binary medical-answer distributions while preserving answer supervision (Sadanandan & Behzadan, 2026b); SEQA trains medical VQA models on generated equivalent questions (Ma et al., 2025). Related approaches include answer-preserving adversarial augmentation, Q3R's modular-reasoning regularization, and SOrT's gradient-attribution alignment (Tang et al., 2020; Whitehead et al., 2020; Dharur et al., 2021). Table 1 identifies the design choices examined here: original-only answer CE, a projected answer-prefix state, and instance-level relative margins.

Group correctness and consistency. VQA-Rephrasings consensus scores and Perfect-Consistency measure joint correctness over related questions (Shah et al., 2019; Ray et al., 2019). Medical consistency work also examines logical relations and correctness conditioned on a main answer (Tascon-Morales et al., 2022). We use the full-group perspective with binary exact match, then inspect common-original-correct subsets and matched outcome transitions. This differs from SEQA's TAR-SC, which averages within-group accuracy. PARC, MM-R^3, and PSF-Med further document formulation sensitivity (Schmalfuss et al., 2025; Chou et al., 2025; Sadanandan & Behzadan, 2026c).

Representation constraints and visual evidence. Pairwise, triplet, and contrastive losses provide attraction and separation mechanisms (Hadsell et al., 2006; Schroff et al., 2015; van den Oord et al., 2018; Chen et al., 2020; Khosla et al., 2020). SemLoss applies a relative-distance constraint in a shared projection space, with gradients passed to LoRA (Hu et al., 2022). Correctness under reformulation is distinct from visual grounding: modality ablations, Consistent but Dangerous, and MedVIGIL examine dependence on visual evidence and other perturbations (Parcalabescu & Frank, 2025; Sadanandan & Behzadan, 2026a; Jiang et al., 2026).

3 SemAlign-VQA

SemAlign-VQA uses the same image in two question views (Figure 1). We call the semantic regularizer $\mathcal{L}_{\text{sem}}$ SemLoss; "SemLoss" in the displays labels joint training with $\mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{sem}}$. CE labels original-answer supervision alone.

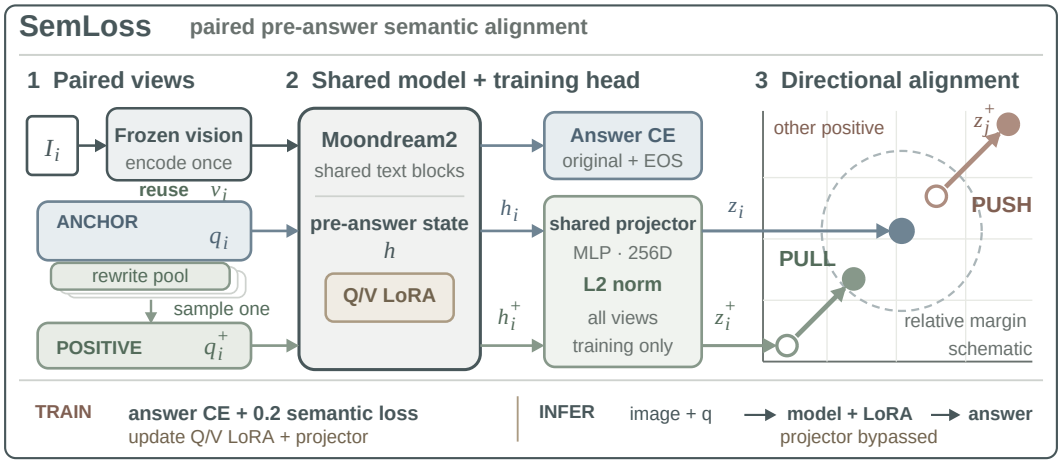

Figure 1: **SemAlign-VQA training and inference.** Every original view is an anchor. Both views share image features, text blocks, and Q/V LoRA. Original answers receive CE; projected pre-answer states receive the pull–push objective. Generation bypasses the projector.

3.1 Paired views and projected pre-answer states

Let $\mathcal{D} = \{(I_i, q_i, y_i)\}$ contain images, original questions, and answers. Every original view (I_i, q_i) is an anchor. Sampling one rewrite q_i^+ from its fixed pool gives the positive (I_i, q_i^+). Positive views from other microbatch items, (I_j, q_j^+) with $j \neq i$, provide instance negatives. This construction uses item identity rather than answer labels; negatives may share an answer or clinical concept with the anchor.

The frozen visual encoder computes each image's features once and reuses them for both question views. Both text forwards share pretrained weights and Q/V LoRA. For Moondream2 and the other causal backbones, $h_\theta(I, q)$ is the last prompt-token state, which conditions the first answer token. BLIP uses its fused encoder CLS state. A shared head applies Linear(d, d), GELU, LayerNorm, and Linear$(d, 256)$, followed by normalization:

$$z_i = \frac{g_\phi(h_\theta(I_i, q_i))}{\|g_\phi(h_\theta(I_i, q_i))\|_2}, \qquad z_i^+ = \frac{g_\phi(h_\theta(I_i, q_i^+))}{\|g_\phi(h_\theta(I_i, q_i^+))\|_2}. \tag{1}$$

The answer-prefix state precedes all answer tokens. The regularizer constrains its projection z and updates the projector and LoRA through the shared state.

3.2 Joint answer supervision and paired regularization

The original view receives teacher-forced answer supervision, averaged over valid answer tokens including EOS:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{\sum_i T_i} \sum_i \sum_{t=1}^{T_i} \log p_\theta(y_{it} \mid I_i, q_i, y_{i,<t}). \tag{2}$$

Prompt and padding positions are masked. With cosine distances $d_{ij} = 1 - z_i^\top z_j^+$, the semantic term is

$$\mathcal{L}_{\text{pull}} = \frac{1}{B} \sum_i d_{ii}, \qquad \mathcal{L}_{\text{push}} = \frac{1}{B(B-1)} \sum_i \sum_{j \neq i} [d_{ii} - d_{ij} + m]_+, \tag{3}$$

$$\mathcal{L}_{\text{sem}} = \tfrac{1}{2}\mathcal{L}_{\text{pull}} + \tfrac{1}{2}\mathcal{L}_{\text{push}}, \qquad \mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{sem}}. \tag{4}$$

Pull attracts matched views. Push compares each negative distance with that anchor's positive distance, contributing zero once $d_{ij} \geq d_{ii} + m$. The margin stops repulsion once the relative-distance constraint is met. We use $m = 0.2$ and $\lambda = 0.2$, with zero push for $B < 2$. CE sets $\lambda = 0$. The

alternative NCE objectives retain pull and use a symmetric contrastive term; answer-aware NCE masks off-diagonal pairs with equal normalized answers. The direct-supervision control, AugCE, averages answer CE over both views, and AugCE+SemLoss adds $\lambda\mathcal{L}_{\text{sem}}$. These controls compare complete training objectives on the available paired records (Section 5.3).

3.3 Parameter-efficient optimization

We adapt `vikhyatk/moondream2`, release `2025-06-21`, using Q/V LoRA with rank 16, alpha 32, and dropout 0.1. Joint training updates 7,870,720 LoRA and projector parameters (0.407% of the training model). CE and SemLoss use the same 1.930B-parameter inference structure (Table 16). Parameter efficiency concerns trainable weights; the additional question view still adds training computation.

Algorithm 1: SemAlign-VQA adaptation

1. Freeze pretrained weights; initialize Q/V LoRA ψ and projector g_ϕ.
2. For each training microbatch, sample one q_i^+ from each item's fixed pool.
3. Encode every image once; reuse its features for the original and positive views.
4. Compute original-answer CE and h_i; obtain h_i^+ from the positive prompt.
5. Project and normalize both states; compute Equations 3–4.
6. Backpropagate $\mathcal{L}/G$; every G microbatches, clip and update (ψ, ϕ).
7. Select the checkpoint with highest validation original-question accuracy.

Moondream2 uses $B = 2$ and accumulation over $G = 4$ microbatches: one negative per anchor and an optimization batch of eight. On PathVQA, BLIP, SmolVLM, and the other backbones use microbatches of 8, 4, and 2, respectively, with the same optimization batch. Accumulation does not enlarge the negative pool. Appendix D gives the backbone-specific preprocessing and objectives.

4 Evaluation setup

4.1 Data, variants, and training comparisons

We evaluate PathVQA (He et al., 2021), SLAKE-English (Liu et al., 2021), and VQA-RAD (Lau et al., 2018) on 8,231 fixed six-view groups (Table 2). PathVQA test wrappers preserve each original question verbatim; SLAKE/VQA-RAD use generated rewrites screened by rules and a language model. Their screening uses neither images nor reference answers. Training samples one same-image rewrite from each original's fixed positive pool. Appendix C specifies the constructions and split identities.

Table 2: **Evaluation groups and variant construction.** Test queries comprise an original and five variants. Wrappers preserve the original text; generated rewrites are automatically screened.

Dataset	Train	Val.	Test	Images	Queries	Test variants
PathVQA	17,654	6,259	6,719	858	40,314	Context wrappers
SLAKE-English	4,919	1,053	1,061	96	6,366	Generated rewrites
VQA-RAD	1,618	175	451	203	2,706	Generated rewrites
Total	24,191	7,487	8,231	1,157	49,386	

The original-only CE comparison fixes the backbone, seed, and answer supervision. The augmentation controls additionally supervise rewritten questions with inherited answers. Loss controls change the regularizer within the pre-answer training setup. Complete predictions support ten Moondream2 configurations; other training comparisons use seed-level summaries, with summary-level controls marked †. VQA-RAD augmentation records are not verified against the final variant identities, so their reported scores are listed separately from conclusions about aligned test-set effects.

The backbone study covers Moondream2, BLIP-VQA Base (Li et al., 2022), InternVL2-1B (Chen et al., 2024b), SmolVLM-256M (Marafioti et al., 2025), and Qwen2.5-VL-3B (Bai et al., 2025).

Four backbones have PathVQA seeds 42, 123, and 2026; all five cover three datasets at seed 42. The full inventory contains 72 training configurations (Appendix B).

4.2 Training, scoring, and uncertainty

Moondream2 trains for three epochs using AdamW (Loshchilov & Hutter, 2019), learning rate 2×10^{-4}, weight decay 0.01, linear warmup/decay with warmup fraction 0.06, gradient clipping at 1.0, and BF16. Question/answer limits are 64/24 tokens. Original-question validation accuracy selects the checkpoint, with earliest-epoch tie breaking. Generation is greedy with at most 24 new tokens and model-native image preprocessing.

Let c_{iv} indicate normalized exact match for question i and view v. Original accuracy A, mean variant accuracy P, and all-six correctness R are

$$A = \frac{1}{N}\sum_i c_{i0}, \qquad P = \frac{1}{5N}\sum_i \sum_{v=1}^{5} c_{iv}, \qquad R = \frac{1}{N}\sum_i \prod_{v=0}^{5} c_{iv}. \tag{5}$$

R implements full-group consensus (Shah et al., 2019) for six binary exact-match outcomes. Normalization lowercases, removes punctuation and English articles, and collapses whitespace. We also report original-to-variant agreement C, pairwise agreement PC, $A-P$, and ROUGE-1/L (Lin, 2004). Agreement compares predictions with each other; correctness compares them with the reference.

Paired image-cluster bootstrap samples whole test images and their question groups (2,000 replicates; percentile 95% intervals; question-weighted estimator). These intervals describe test-image variation for fixed checkpoints. Training variation is summarized separately by same-seed SemLoss-minus-CE differences and their sample SD across three seeds. All scoring definitions appear in Appendix E.

5 Results

5.1 The primary gains are gains in correct-answer retention

Table 3: Primary Moondream2 comparison, seed 42. All entries are percentages. A: original accuracy; P: variant accuracy; C: original-to-variant agreement; PC: pairwise agreement; R: all-six correctness.

Dataset	Objective	A	P	C	PC	R
PathVQA	CE	57.21	48.92	60.40	59.38	37.80
	SemLoss	57.24	49.82	63.84	63.88	39.65
SLAKE-English	CE	78.51	70.78	78.32	77.31	55.14
	SemLoss	78.70	71.23	78.40	78.40	57.30
VQA-RAD	CE	45.01	45.01	67.23	68.54	30.16
	SemLoss	44.79	45.01	67.89	69.16	30.60

At seed 42, Moondream2 gains 1.85 points in all-six correctness on PathVQA wrappers and 2.17 on SLAKE rewrites, while original accuracy changes by only +0.03 and +0.19 points (Table 3). The paired image-cluster 95% intervals for ΔR are [1.25, 2.42] and [1.15, 3.15]. VQA-RAD changes by +0.44 points, with an interval of $[-0.42, 1.35]$ (Figure 2). Section 5.3 compares direct rewrite supervision.

For $A > 0$, $R = A\rho$, where $\rho = \Pr(c_{i1} = \cdots = c_{i5} = 1 \mid c_{i0} = 1)$ is retention conditional on a correct original answer. Retention increases from 66.08% to 69.27% on PathVQA and from 70.23% to 72.81% on SLAKE. With nearly unchanged A, retention accounts for most of the gains in R.

The common-original-correct subsets also hold question identity fixed across objectives (Table 4). SemLoss retains 115 additional all-six-correct groups among 3,732 shared correct PathVQA originals, 23 among 828 SLAKE originals, and two among 197 VQA-RAD originals. The analysis therefore locates a benefit on questions both models could already answer correctly.

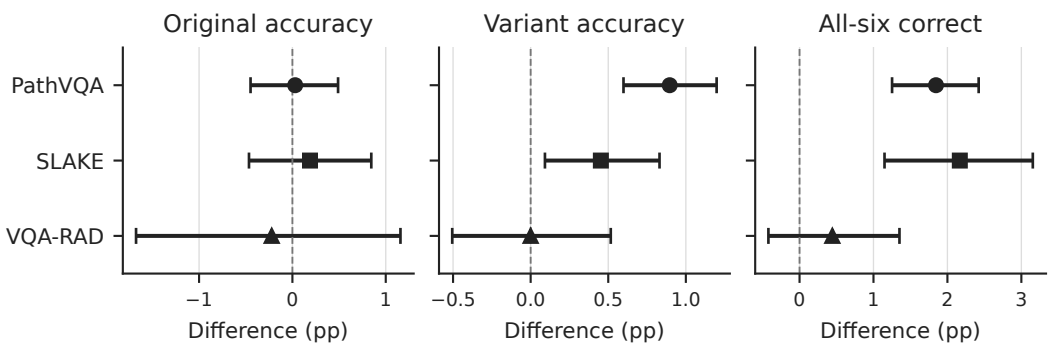

Figure 2: **Moondream2 effects at seed 42.** SemLoss minus CE with paired image-cluster 95% intervals. The larger changes occur in joint correctness across the tested views. These intervals concern fixed checkpoints.

Table 4: Correct-answer preservation on the common-original-correct subset. SemLoss raises variant accuracy in all five available comparisons and all-six retention in four of five. Retention is the fraction of the subset answered correctly on all six views; Δ is SemLoss minus CE in percentage points.

Dataset	Seed	n	All-six retention (%)			Variant accuracy (%)		
			CE	SemLoss	Δ	CE	SemLoss	Δ
PathVQA	42	3,732	67.87	70.95	+3.08	82.89	84.20	+1.31
PathVQA	123	3,802	55.26	58.47	+3.21	76.18	78.25	+2.08
PathVQA	2026	3,787	58.75	58.49	-0.26	79.01	79.28	+0.27
SLAKE-English	42	828	70.53	73.31	+2.78	85.65	86.14	+0.48
VQA-RAD	42	197	69.04	70.05	+1.02	88.53	88.73	+0.20

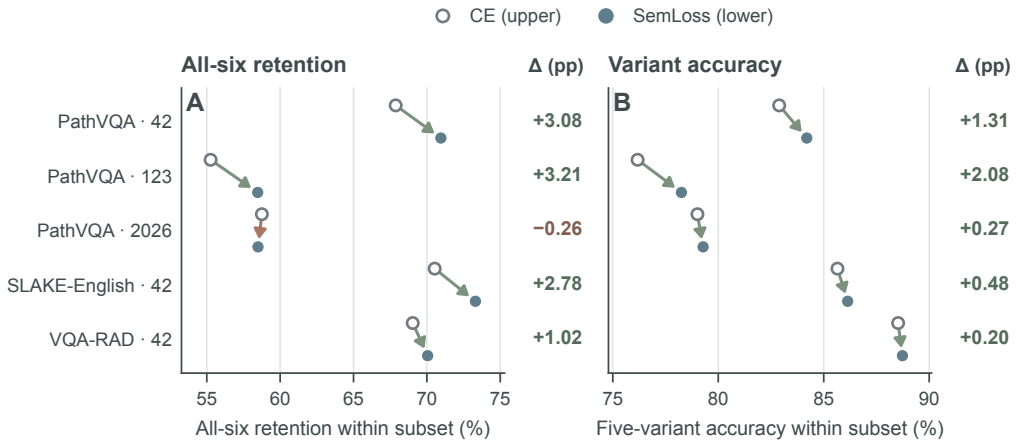

Figure 3: **Retention among shared correct originals.** CE is upper/open and SemLoss lower/filled; horizontal coordinates are the observed scores. Δ denotes SemLoss minus CE in percentage points. Different changes in mean variant accuracy and all-six retention are visible within the same subset.

5.2 GREATER AGREEMENT NEED NOT PRESERVE MORE CORRECT GROUPS

Matched transitions separate correct stabilization from repeated errors (Figure 4). On PathVQA, 290 inconsistent groups become all-correct and 166 reverse. The net reduction of 217 inconsistent groups comprises a net gain of 124 all-correct groups and 93 identically wrong groups. On SLAKE, 27 groups become all-correct and four reverse, while the identically wrong count remains 55. Reducing inconsistency can therefore reflect different mixtures of correct and incorrect stabilization.

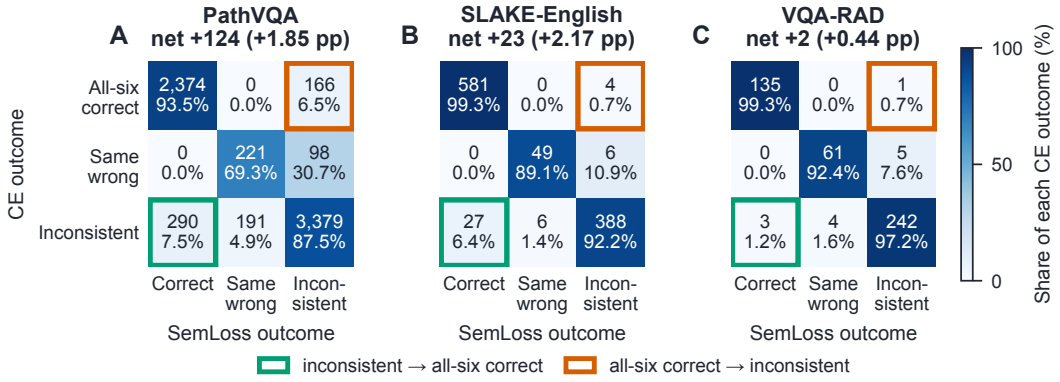

Figure 4: **Correct and incorrect stabilization.** Matched seed-42 transitions show counts and percentages within each CE row. Green marks inconsistent-to-correct transitions; orange marks the reverse. Net changes are interpreted over the full partition, including reverse transitions.

The PathVQA seed-2026 common-correct subset illustrates the difference between average and group correctness (Table 4). Across the same 3,787 items, incorrect variant answers fall from 3,974 to 3,923, but groups with at least one error rise from 1,562 to 1,572. Fewer errors remain spread across more failing groups at these aggregate endpoints.

5.3 ANSWER SUPERVISION AND REGULARIZER CHOICE HAVE DIFFERENT EFFECTS

Direct rewrite supervision. AugCE applies the inherited answer to both views without new manual labels (Table 5). In the recorded Moondream2/SLAKE seed-42 controls, R is 55.14% for CE, 57.30% for SemLoss, 65.22% for AugCE, and 65.50% for their combination. Direct rewrite supervision contributes the larger gain; adding alignment on top changes R by 0.28 points. This comparison places the representation-only intervention alongside the most immediate use of the same rewritten questions.

Table 5: Moondream2 supervision controls (seed 42, %). †: summary-level augmentation result; exact alignment to final VQA-RAD variants is unverified.

Dataset	Objective	Original	Variant	All-six
SLAKE-English	CE	78.51	70.78	55.14
	SemLoss	78.70	71.23	57.30
	Augmented CE†	78.51	76.04	65.22
	Aug. CE + SemLoss†	79.26	75.95	65.50
VQA-RAD	CE	45.01	45.01	30.16
	SemLoss	44.79	45.01	30.60
	Augmented CE†	47.45	46.83	33.26
	Aug. CE + SemLoss†	47.45	46.65	33.04

Margin and contrastive objectives. SemLoss and answer-aware NCE have nearly equal three-seed mean PathVQA R: 35.30% and 35.28% (Table 6). Yet SemLoss has higher agreement in

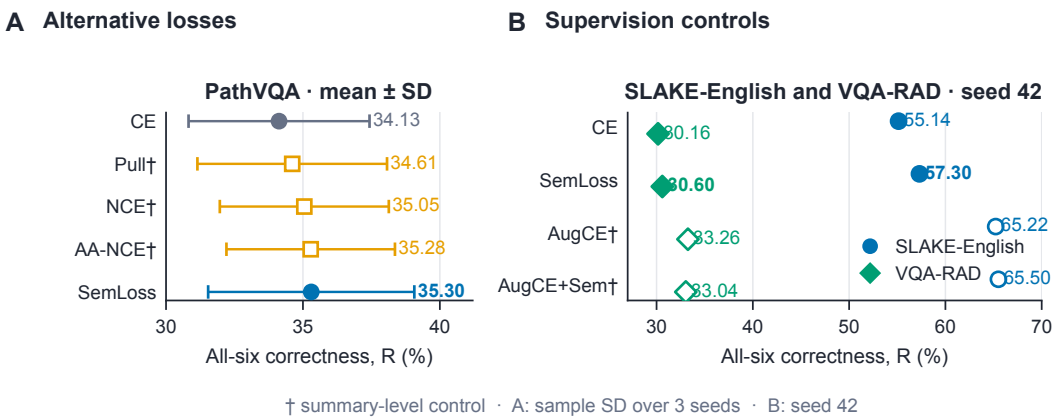

Figure 5: **Regularizer choice and rewrite supervision.** A: PathVQA objective scores, mean and sample SD across three seeds. B: recorded seed-42 augmentation controls. † marks summary-level results; the VQA-RAD control-version caveat is specified in Table 5.

Table 6: PathVQA objectives on Moondream2: mean ± sample SD over three training seeds (%). †: summary-level control. CE and SemLoss use complete predictions.

Objective	Original	Variant	Agreement	All-six
CE	57.35 ± 0.22	47.50 ± 1.34	58.35 ± 2.33	34.13 ± 3.30
Pull only†	57.39 ± 0.25	47.80 ± 1.24	58.70 ± 2.39	34.61 ± 3.46
InfoNCE mix†	57.22 ± 0.16	47.94 ± 1.28	59.15 ± 2.07	35.05 ± 3.09
Answer-aware NCE mix†	57.28 ± 0.15	48.15 ± 1.20	59.21 ± 1.85	35.28 ± 3.08
SemLoss	57.27 ± 0.28	48.33 ± 1.29	60.44 ± 3.08	35.30 ± 3.76

every seed: $\Delta C = +2.71, +0.40, +0.59$ points, compared with $\Delta R = +0.91, +0.37, -1.22$ (Table 7). Thus the regularizers differ consistently in agreement without a corresponding uniform gain in group correctness. These summary comparisons do not identify which question groups produce that difference.

Table 7: Moondream2 PathVQA same-seed differences, SemLoss minus answer-aware NCE (percentage points). SemLoss uses complete predictions; answer-aware NCE uses summary-level results from the same loss study.

Seed	ΔA	ΔP	ΔC	ΔPC	ΔR
42	-0.03	$+0.32$	$+2.71$	$+3.33$	$+0.91$
123	$+0.13$	$+0.30$	$+0.40$	$+1.63$	$+0.37$
2026	-0.13	-0.08	$+0.59$	$+0.21$	-1.22
Mean ± SD	-0.01 ± 0.14	$+0.18 \pm 0.23$	$+1.23 \pm 1.28$	$+1.72 \pm 1.56$	$+0.02 \pm 1.11$

5.4 Observed variation across backbones and seeds

The three-seed PathVQA comparisons provide the clearest view of training variation (Table 8). Moondream2 gains 1.18 points in mean R, with two positive seeds and -0.13 in seed 2026. BLIP declines in all three, from 11.28% to 6.45% in mean R; InternVL2 has one positive seed and SmolVLM two. Agreement improves in eight of these 12 pairs, while R improves in five. BLIP seed 42, Moondream2 seed 2026, and SmolVLM seed 123 each improve agreement while reducing R.

At the common seed 42, ΔR is positive in 13/15 backbone–dataset pairs, with an equal-weight mean of +0.98 points. Dataset means are +1.94 on SLAKE, +0.98 on VQA-RAD, and +0.02 on

Table 8: PathVQA same-seed changes in all-six correctness, SemLoss minus CE (percentage points). Mean ± sample SD is computed across seeds 42, 123, and 2026; † marks summary-level results, while Moondream2 uses complete predictions.

Backbone	42	123	2026	Mean ± SD	$\Delta R > 0$
BLIP-VQA†	−2.53	−4.52	−7.43	−4.83 ± 2.46	0/3
InternVL2-1B†	−1.16	+0.55	−0.77	−0.46 ± 0.90	1/3
Moondream2	+1.85	+1.82	−0.13	+1.18 ± 1.13	2/3
SmolVLM-256M†	+0.42	−0.61	+1.07	+0.29 ± 0.85	2/3

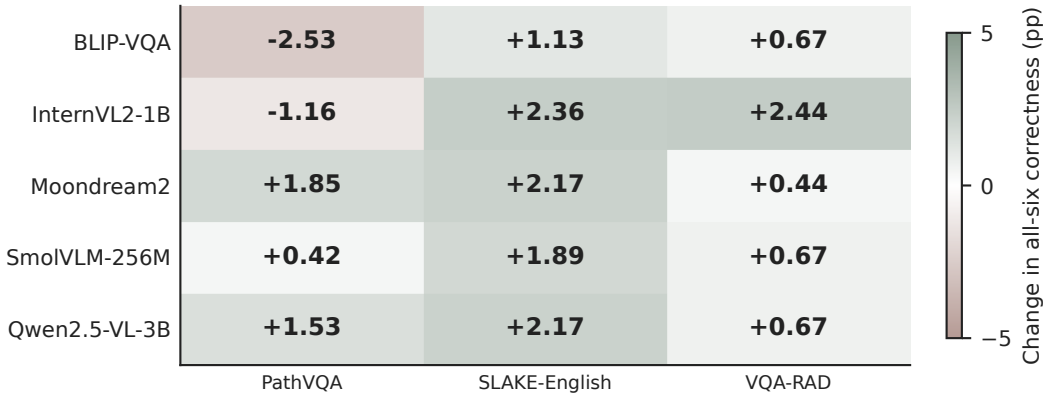

Figure 6: **Coverage at a common training seed.** Each cell is SemLoss minus CE in R at seed 42 (percentage points). The 13 positive cells describe coverage across settings; repeated-training effects appear separately in Table 8.

PathVQA; BLIP and InternVL2 change by -2.53 and -1.16 on PathVQA. State extraction, micro-batch size, and negative-pool size vary with the backbone implementation, so this matrix measures their combined setting.

6 Discussion and Conclusions

Agreement, average variant accuracy, and group correctness respond to different changes in model output. Common-correct subsets and matched transitions distinguish corrected answers from redistributed or stabilized errors. Objective selection therefore requires group-correctness checks alongside agreement and average accuracy.

The available SLAKE controls favor direct rewrite supervision, whereas the PathVQA loss comparison yields similar R from different regularizers. CE–SemLoss predictions support question-group analysis; augmentation and NCE controls provide aggregate objective comparisons within the training design, rather than reproductions of the complete systems in Table 1.

Scope. The evaluation uses fixed wrappers and automatically screened rewrites; human semantic-equivalence and answer-preservation rates were not measured. VQA-RAD follows a question split with 182 train–test and 20 validation–test shared images, and correctness uses single-reference exact match. Repeated training seeds cover PathVQA only; the evidence levels of other controls are specified with their results. The paired analyses identify changed answers, while state choice, negative sampling, and backbone factors were not isolated by component interventions. Visual dependence is a separate evaluation question (Yan et al., 2025; Sadanandan & Behzadan, 2026a).

AI Use Statement

AI tools assisted with language polishing.

Ethics Statement

This study evaluates public medical-VQA benchmarks. The evaluation reports correctness and consistency jointly and documents question-variant construction, image overlap, and reference-answer scoring. Source images and model weights remain governed by their providers' redistribution terms.

Reproducibility Statement

Section 3 and Algorithm 1 define the method. Appendices C–E specify data construction, training, checkpoint selection, inference, and scoring. The accompanying source package provides numerical inputs and scripts for reproducing the figures and tables. Complete predictions support the ten primary and 24 external evaluations; the tables identify summary-level controls.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2425–2433, December 2015. doi: 10.1109/ICCV.2015.279. URL https://openaccess.thecvf.com/content_iccv_2015/html/Antol_VQA_Visual_Question_ICCV_2015_paper.html.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. doi: 10.48550/arXiv.2502.13923. URL https://arxiv.org/abs/2502.13923.

Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. VQA-Med: Overview of the medical visual question answering task at ImageCLEF 2019. In *Working Notes of CLEF 2019 – Conference and Labs of the Evaluation Forum*, volume 2380 of *CEUR Workshop Proceedings*, Lugano, Switzerland, 2019. CEUR-WS.org. URL https://ceur-ws.org/Vol-2380/paper_272.pdf.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the most of text semantics to improve biomedical vision–language processing. In *Computer Vision – ECCV 2022*, volume 13696 of *Lecture Notes in Computer Science*, pp. 1–21, Cham, 2022. Springer Nature Switzerland. doi: 10.1007/978-3-031-20059-5_1. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/html/1557_ECCV_2022_paper.php.

Aofei Chang, Le Huang, Alex James Boyd, Parminder Bhatia, Taha Kass-Hout, Cao Xiao, and Fenglong Ma. Focus on what matters: Enhancing medical vision-language models with automatic attention alignment tuning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9357–9372, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.460. URL https://aclanthology.org/2025.acl-long.460/.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and Benyou Wang. Towards injecting medical visual knowledge into multimodal LLMs at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7346–7370, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.418. URL https://aclanthology.org/2024.emnlp-main.418/.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, June 2024b. doi: 10.1109/CVPR52733.2024.02283. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_InternVL_Scaling_up_Vision_Foundation_Models_and_Aligning_for_Generic_CVPR_2024_paper.html.

Shih-Han Chou, Shivam Chandhok, Jim Little, and Leonid Sigal. MM-r^3: On (in-)consistency of vision-language models (VLMs). In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4762–4788, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.246. URL https://aclanthology.org/2025.findings-acl.246/.

Corentin Dancette, Rémi Cadène, Damien Teney, and Matthieu Cord. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1554–1563, October 2021. doi: 10.1109/ICCV48922.2021.00160. URL https://openaccess.thecvf.com/content/ICCV2021/html/Dancette_Beyond_Question-Based_Biases_Assessing_Multimodal_Shortcut_Learning_in_Visual_Question_ICCV_2021_paper.html.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10088–10115. Curran Associates, Inc., 2023. doi: 10.52202/075280-0441. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html.

Sameer Dharur, Purva Tendulkar, Dhruv Batra, Devi Parikh, and Ramprasaath R. Selvaraju. SOrT-ing VQA models: Contrastive gradient learning for improved consistency. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3103–3111, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.248. URL https://aclanthology.org/2021.naacl-main.248/.

Sedigheh Eslami, Christoph Meinel, and Gerard de Melo. PubMedCLIP: How much does CLIP benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 1181–1193, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.88. URL https://aclanthology.org/2023.findings-eacl.88/.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pp. 1735–1742. IEEE, 2006. doi: 10.1109/CVPR.2006.100. URL https://doi.org/10.1109/CVPR.2006.100.

Xuehai He, Zhuo Cai, Wenlan Wei, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Towards visual question answering on pathology images. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 708–718, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.90. URL https://aclanthology.org/2021.acl-short.90/.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning

for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

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 *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. GLoRIA: A multimodal global–local representation learning framework for label-efficient medical image recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3922–3931. IEEE, October 2021. doi: 10.1109/ICCV48922.2021.00391. URL https://doi.org/10.1109/ICCV48922.2021.00391.

Hanqi Jiang, Junhao Chen, Mingyu Kang, Hyeokjae Kwon, Yi Pan, Lifeng Chen, Weihang You, Haozhen Gong, Ruiyu Yan, Jinglei Lv, Lin Zhao, Hui Ren, Quanzheng Li, Tianming Liu, and Xiang Li. MedVIGIL: Evaluating trustworthy medical VLMs under broken visual evidence. *arXiv preprint arXiv:2605.07919*, 2026. doi: 10.48550/arXiv.2605.07919. URL https://arxiv.org/abs/2605.07919v2. Version 2 preprint.

Songtao Jiang, Yuxi Chen, Sibo Song, Yan Zhang, Yeying Jin, Yang Feng, Jian Wu, and Zuozhu Liu. Knowing or guessing? robust medical visual question answering via joint consistency and contrastive learning. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15970 of *Lecture Notes in Computer Science*, pp. 325–335. Springer Nature Switzerland, September 2025. doi: 10.1007/978-3-032-05141-7_32. URL https://papers.miccai.org/miccai-2025/0471-Paper3893.html.

Shibo Jie, Yehui Tang, Ning Ding, Zhi-Hong Deng, Kai Han, and Yunhe Wang. Memory-space visual prompting for efficient vision-language fine-tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22062–22074. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/jie24a.html.

Yash Kant, Abhinav Moudgil, Dhruv Batra, Devi Parikh, and Harsh Agrawal. Contrast and classify: Training robust VQA models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1584–1593, October 2021. doi: 10.1109/ICCV48922.2021.00163. URL https://openaccess.thecvf.com/content/ICCV2021/html/Kant_Contrast_and_Classify_Training_Robust_VQA_Models_ICCV_2021_paper.html.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

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. doi: 10.1038/sdata.2018.251. URL https://www.nature.com/articles/sdata2018251.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36, pp. 28541–28564. Curran Associates, Inc., 2023. doi: 10.52202/075280-1240. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/5abcdf8ecdcacba028c6662789194572-Paper-Datasets_and_Benchmarks.pdf.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the*

39th International Conference on Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 2022. URL https://proceedings.mlr.press/v162/li22n.html.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

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, April 2021. doi: 10.1109/ISBI48211.2021.9434010. URL https://ieeexplore.ieee.org/document/9434010.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Yongpei Ma, Pengyu Wang, Adam Dunn, Usman Naseem, and Jinman Kim. Bridging the semantic gaps: Improving medical VQA consistency with LLM-augmented question sets. *arXiv preprint arXiv:2504.11777*, 2025. doi: 10.48550/arXiv.2504.11777. URL https://arxiv.org/abs/2504.11777.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. SmolVLM: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025. doi: 10.48550/arXiv.2504.05299. URL https://arxiv.org/abs/2504.05299.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: A multimodal medical few-shot learner. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pp. 353–367. PMLR, 2023. URL https://proceedings.mlr.press/v225/moor23a.html.

Letitia Parcalabescu and Anette Frank. Do vision & language decoders use images and text equally? how self-consistent are their explanations? In *International Conference on Learning Representations*, volume 2025, pp. 21634–21663, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/37294f033582ac0064bf90fa557c2573-Paper-Conference.pdf.

Abhirama Subramanyam Penamakuri, Navlika Singh, Piyush Arora, and Anand Mishra. When big models train small ones: Label-free model parity alignment for efficient visual question answering using small VLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 31644–31661, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1613. URL https://aclanthology.org/2025.emnlp-main.1613/.

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

Arijit Ray, Karan Sikka, Ajay Divakaran, Stefan Lee, and Giedrius Burachas. Sunny and dark outside?! improving answer consistency in VQA through entailed question generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 5860–5865, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1596. URL https://aclanthology.org/D19-1596/.

Shuvendu Roy and Ali Etemad. Consistency-guided prompt learning for vision-language models. In *International Conference on Learning Representations*, volume 2024, pp. 6485–6498, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/1a1f7aeb1f9bd9052cfd11181430b806-Paper-Conference.pdf.

Binesh Sadanandan and Vahid Behzadan. Consistent but dangerous: Per-sample safety classification reveals false reliability in medical vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 6961–6968, June 2026a. URL https://openaccess.thecvf.com/content/CVPR2026W/Med-Reasoner/html/Sadanandan_Consistent_but_Dangerous_Per-Sample_Safety_Classification_Reveals_False_Reliability_in_CVPRW_2026_paper.html.

Binesh Sadanandan and Vahid Behzadan. Mechanistically guided LoRA improves paraphrase consistency in medical vision-language models. In *Proceedings of the 7th Conference on Health, Inference, and Learning*, volume 333 of *Proceedings of Machine Learning Research*, pp. 703–720. PMLR, June 2026b. URL https://proceedings.mlr.press/v333/sadanandan26a.html.

Binesh Sadanandan and Vahid Behzadan. PSF-Med: Measuring and explaining paraphrase sensitivity in medical vision language models. *arXiv preprint arXiv:2602.21428*, 2026c. doi: 10.48550/arXiv.2602.21428. URL https://arxiv.org/abs/2602.21428v2. Version 2, revised 12 April 2026.

Jenny Schmalfuss, Nadine Chang, Vibashan VS, Maying Shen, Andres Bruhn, and Jose M. Alvarez. Parc: A quantitative framework uncovering the symmetries within vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25081–25091, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Schmalfuss_PARC_A_Quantitative_Framework_Uncovering_the_Symmetries_within_Vision_Language_CVPR_2025_paper.html.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823. IEEE, June 2015. doi: 10.1109/CVPR.2015.7298682. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Schroff_FaceNet_A_Unified_2015_CVPR_paper.html.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6642–6651, June 2019. doi: 10.1109/CVPR.2019.00681. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Shah_Cycle-Consistency_for_Robust_Visual_Question_Answering_CVPR_2019_paper.html.

Guohao Sun, Can Qin, Huazhu Fu, Linwei Wang, and Zhiqiang Tao. Self-training large language and vision assistant for medical question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20052–20060, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1119. URL https://aclanthology.org/2024.emnlp-main.1119/.

Ruixue Tang, Chao Ma, Wei Emma Zhang, Qi Wu, and Xiaokang Yang. Semantic equivalent adversarial data augmentation for visual question answering. In *Computer Vision – ECCV 2020*, volume 12364 of *Lecture Notes in Computer Science*, pp. 437–453. Springer International Publishing, 2020. doi: 10.1007/978-3-030-58529-7_26. URL https://doi.org/10.1007/978-3-030-58529-7_26.

Sergio Tascon-Morales, Pablo Márquez-Neila, and Raphael Sznitman. Consistency-preserving visual question answering in medical imaging. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, volume 13438 of *Lecture Notes in Computer Science*, pp. 386–395. Springer Nature Switzerland, 2022. doi: 10.1007/978-3-031-16452-1_37. URL https://doi.org/10.1007/978-3-031-16452-1_37.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, Anil Palepu, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Semturs, S. Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Natarajan. Towards generalist biomedical AI. *NEJM AI*, 1(3):AIoa2300138, 2024. doi: 10.1056/AIoa2300138. URL https://doi.org/10.1056/AIoa2300138.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. doi: 10.48550/arXiv.1807.03748. URL https://arxiv.org/abs/1807.03748.

Tom van Sonsbeek, Mohammad Mahdi Derakhshani, Ivona Najdenkoska, Cees G. M. Snoek, and Marcel Worring. Open-ended medical visual question answering through prefix tuning of language models. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14224 of *Lecture Notes in Computer Science*, pp. 726–736, Cham, 2023. Springer Nature Switzerland. doi: 10.1007/978-3-031-43904-9_70. URL https://doi.org/10.1007/978-3-031-43904-9_70.

Pavan Kumar Anasosalu Vasu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokula Santhanam, James Gabriel, Peter Grasch, Oncel Tuzel, and Hadi Pouransari. Fastvlm: Efficient vision encoding for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19769–19780, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Vasu_FastVLM_Efficient_Vision_Encoding_for_Vision_Language_Models_CVPR_2025_paper.html.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3876–3887, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.256. URL https://aclanthology.org/2022.emnlp-main.256/.

Spencer Whitehead, Hui Wu, Yi Ren Fung, Heng Ji, Rogerio Feris, and Kate Saenko. Learning from lexical perturbations for consistent visual question answering. *arXiv preprint arXiv:2011.13406*, 2020. doi: 10.48550/arXiv.2011.13406. URL https://arxiv.org/abs/2011.13406.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. *Nature Communications*, 16(1):7866, 2025. doi: 10.1038/s41467-025-62385-7. URL https://www.nature.com/articles/s41467-025-62385-7.

Qianqi Yan, Xuehai He, Xiang Yue, and Xin Eric Wang. Worse than random? an embarrassingly simple probing evaluation of large multimodal models in medical VQA. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19188–19205, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.981. URL https://aclanthology.org/2025.findings-acl.981/.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D. Davison, Hui Ren, Jing Huang, Chen Chen, Yuyin Zhou, Sunyang Fu, Wei Liu, Tianming Liu, Xiang Li, Yong Chen, Lifang He, James Zou, Quanzheng Li, Hongfang Liu, and Lichao Sun. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 30(11):3129–3141, 2024. doi: 10.1038/s41591-024-03185-2. URL https://www.nature.com/articles/s41591-024-03185-2.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1):AIoa2400640, 2025. doi: 10.1056/AIoa2400640. URL https://doi.org/10.1056/AIoa2400640.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-VQA: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. doi: 10.48550/arXiv.2305.10415. URL https://arxiv.org/abs/2305.10415.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pp. 2–25. PMLR, 2022. URL https://proceedings.mlr.press/v182/zhang22a.html.

Yang Zhou, Tan Li Hui Faith, Yanyu Xu, Sicong Leng, Xinxing Xu, Yong Liu, and Rick Siow Mong Goh. Benchx: A unified benchmark framework for medical vision-language pretraining on chest x-rays. In *Advances in Neural Information Processing Systems*, volume 37, pp. 6625–6647. Curran Associates, Inc., 2024. doi: 10.52202/079017-0213. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/0cb35e10bf7bb73d10c12414edbd63fd-Paper-Datasets_and_Benchmarks_Track.pdf.

Kangyu Zhu, Ziyuan Qin, Huahui Yi, Zekun Jiang, Qicheng Lao, Shaoting Zhang, and Kang Li. Guiding medical vision-language models with diverse visual prompts: Framework design and comprehensive exploration of prompt variations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11726–11739, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.587. URL https://aclanthology.org/2025.naacl-long.587/.

A Medical VQA and efficient adaptation

Visual question answering couples image understanding with language-conditioned response prediction, and medical benchmarks extend this setting to clinical images and questions (Antol et al., 2015; Ben Abacha et al., 2019). Image–text pretraining spans CLIP, ConVIRT, GLoRIA, BioViL, PubMedCLIP, MedCLIP, and BiomedCLIP, with BenchX providing a unified medical evaluation framework (Radford et al., 2021; Zhang et al., 2022; Huang et al., 2021; Boecking et al., 2022; Eslami et al., 2023; Wang et al., 2022; Zhang et al., 2025; Zhou et al., 2024). Generative medical systems and data resources include Med-Flamingo, open-ended prefix tuning, Med-PaLM M, BioMedGPT, RadFM, PMC-VQA, LLaVA-Med, and self-training, alongside visual prompting and attention alignment (Moor et al., 2023; van Sonsbeek et al., 2023; Tu et al., 2024; Zhang et al., 2024; Wu et al., 2025; Zhang et al., 2023; Li et al., 2023; Sun et al., 2024; Zhu et al., 2025; Chang et al., 2025). Adapters, QLoRA, MemVP, FastVLM, and model-parity alignment motivate parameter-conscious adaptation, while CoPrompt, VQA-CE, and shortcut-learning analyses study consistency and dataset-dependent behavior (Houlsby et al., 2019; Dettmers et al., 2023; Jie et al., 2024; Vasu et al., 2025; Penamakuri et al., 2025; Roy & Etemad, 2024; Dancette et al., 2021; Geirhos et al., 2020).

B Experimental comparisons and training configurations

The experiments comprise 72 backbone–dataset–objective–seed configurations: 30 in the five-backbone comparison, 16 additional PathVQA seed configurations, 18 alternative-loss configurations, and eight augmentation controls (Table 9). Ten Moondream2 configurations are evaluated from complete predictions; the other 62 are reported at summary level and marked accordingly where used.

Table 9: Experimental coverage. The 34 prediction-level evaluations comprise 10 of the 72 training configurations and 24 external checkpoint evaluations.

Experimental component	Count
Core: 5 backbones × 3 datasets × 2 objectives (seed 42)	30
Additional PathVQA seeds: 4 backbones × 2 objectives × 2 seeds	16
Alternative losses: 2 backbones × 3 losses × 3 seeds	18
Augmentation: 2 backbones × 2 datasets × 2 objectives	8
Distinct training configurations	**72**
External frozen checkpoints × 3 datasets	24
Primary Moondream2 runs with complete predictions	10
Complete prediction sets / answer rows	34 / 655,116

Table 10: Positive paired differences across the indicated, overlapping comparisons. P: rewrite accuracy; C: answer agreement; R: all-six correctness. Mean ΔR weights comparisons equally; the final row uses all-six retention on the common-original-correct subset.

Comparison	P	C	R	Mean ΔR
Moondream2, seed 42 × 3 datasets	2/3	3/3	3/3	+1.49 pp
Five backbones × 3 datasets	10/15	12/15	13/15	+0.98 pp
Moondream2, PathVQA × 3 seeds	3/3	3/3	2/3	+1.18 pp
Common-original-correct subsets	5/5	–	4/5	+1.96 pp

Table 11: Complete seed-42 comparison across five backbones and three datasets. Within each metric the two columns are CE and SemLoss. Moondream2 scores use complete predictions; the other backbones use saved seed-level summaries.

Dataset	Backbone	A		P		C		R	
PathVQA	BLIP-VQA	51.38	51.47	32.96	33.86	37.79	39.54	13.54	11.01
	InternVL2-1B	55.60	55.50	49.73	48.45	63.96	61.67	38.90	37.74
	Moondream2	57.21	57.24	48.92	49.82	60.40	63.84	37.80	39.65
	SmolVLM-256M	54.09	53.34	47.55	47.60	60.07	60.09	36.29	36.70
	Qwen2.5-VL-3B	58.30	58.24	55.01	55.65	73.85	74.98	45.59	47.12
SLAKE-English	BLIP-VQA	66.07	65.60	57.59	58.30	70.24	71.56	39.49	40.62
	InternVL2-1B	77.10	76.53	68.60	68.97	75.63	76.95	49.58	51.93
	Moondream2	78.51	78.70	70.78	71.23	78.32	78.40	55.14	57.30
	SmolVLM-256M	73.42	73.80	62.60	63.86	71.12	72.10	43.54	45.43
	Qwen2.5-VL-3B	77.38	78.13	69.84	70.88	77.74	78.38	56.83	59.00
VQA-RAD	BLIP-VQA	35.03	35.25	35.65	35.12	66.92	68.16	22.62	23.28
	InternVL2-1B	47.67	47.23	45.37	45.23	57.74	63.06	25.06	27.49
	Moondream2	45.01	44.79	45.01	45.01	67.23	67.89	30.16	30.60
	SmolVLM-256M	39.02	39.47	39.60	38.85	65.50	65.32	25.94	26.61
	Qwen2.5-VL-3B	55.65	55.65	52.99	53.30	69.98	69.89	41.02	41.69

The equal-weight mean of the 15 seed-42 paired all-six differences is +0.98 points. Table 10 reports this matrix alongside the three-seed Moondream2 study and common-original-correct subsets. Appendix D specifies backbone preprocessing, microbatch sizes, and state extraction.

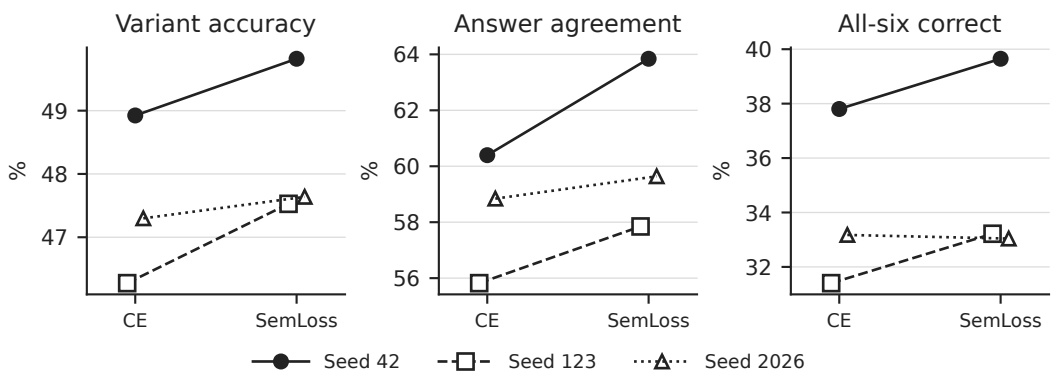

Figure 7: **Paired results for all three primary PathVQA seeds.** Rewrite accuracy and answer agreement improve in all three; all-six accuracy improves for seeds 42 and 123 and changes by −0.13 points for seed 2026. Marker offsets separate overlapping observations.

B.1 Objective and supervision controls

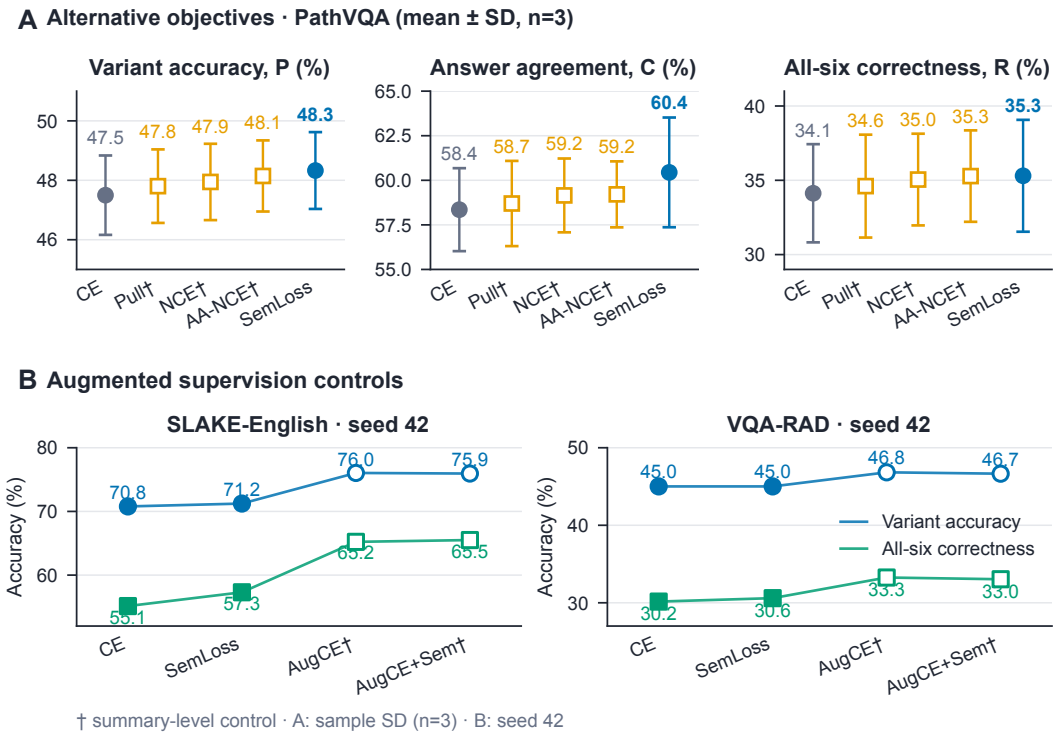

Figure 8: **Full objective comparison.** A: mean and sample SD over three PathVQA seeds. B: seed-42 augmentation controls. † marks summary-level results. Tables 6 and 5 give the scores.

Table 6 reports the aggregate objective comparison, and Table 7 gives the same-seed SemLoss–answer-aware-NCE differences. Direct positive-question supervision reaches 65.22% all-six correctness on SLAKE, and its combination with SemLoss reaches 65.50%. These controls compare representation alignment with the stronger supervision supplied by additional answer targets.

Table 12: All 72 training configurations (%). Each backbone–dataset–objective–seed tuple appears once. Ten primary Moondream2 runs have complete predictions; the other scores use seed-level summaries.

Dataset	Backbone	Objective	Seed	A	P	C	R
Path	BLIP-VQA	AA-NCE mix	42	51.45	33.02	38.32	11.24
Path	BLIP-VQA	AA-NCE mix	123	51.11	31.85	37.64	4.02
Path	BLIP-VQA	AA-NCE mix	2026	51.14	31.66	36.87	4.35
Path	BLIP-VQA	CE	42	51.38	32.96	37.79	13.54
Path	BLIP-VQA	CE	123	51.23	33.82	38.83	8.51
Path	BLIP-VQA	CE	2026	51.29	33.01	37.42	11.77
Path	BLIP-VQA	NCE mix	42	51.50	32.98	38.28	10.25
Path	BLIP-VQA	NCE mix	123	51.08	31.72	37.46	4.09
Path	BLIP-VQA	NCE mix	2026	51.29	31.70	36.90	4.02
Path	BLIP-VQA	Pull	42	51.56	33.36	38.24	13.59
Path	BLIP-VQA	Pull	123	51.23	33.36	38.78	5.89
Path	BLIP-VQA	Pull	2026	51.30	32.61	37.61	10.46
Path	BLIP-VQA	SemLoss	42	51.47	33.86	39.54	11.01
Path	BLIP-VQA	SemLoss	123	51.14	31.87	37.53	3.99
Path	BLIP-VQA	SemLoss	2026	51.24	31.67	36.81	4.35
Path	InternVL2-1B	CE	42	55.60	49.73	63.96	38.90
Path	InternVL2-1B	CE	123	55.26	48.03	59.10	36.49
Path	InternVL2-1B	CE	2026	56.06	50.40	65.51	38.96
Path	InternVL2-1B	SemLoss	42	55.50	48.45	61.67	37.74
Path	InternVL2-1B	SemLoss	123	55.57	48.51	60.09	37.04
Path	InternVL2-1B	SemLoss	2026	55.75	50.40	65.34	38.19
Path	Moondream2	AA-NCE mix	42	57.27	49.50	61.14	38.74
Path	Moondream2	AA-NCE mix	123	57.43	47.22	57.45	32.85
Path	Moondream2	AA-NCE mix	2026	57.14	47.72	59.05	34.26
Path	Moondream2	CE	42	57.21	48.92	60.40	37.80
Path	Moondream2	CE	123	57.60	46.27	55.82	31.40
Path	Moondream2	CE	2026	57.24	47.30	58.84	33.17
Path	Moondream2	NCE mix	42	57.05	49.40	61.21	38.46
Path	Moondream2	NCE mix	123	57.37	46.96	57.07	32.45
Path	Moondream2	NCE mix	2026	57.24	47.48	59.17	34.25
Path	Moondream2	Pull	42	57.39	49.19	61.03	38.37
Path	Moondream2	Pull	123	57.64	46.81	56.25	31.55
Path	Moondream2	Pull	2026	57.14	47.41	58.82	33.90
Path	Moondream2	SemLoss	42	57.24	49.82	63.84	39.65
Path	Moondream2	SemLoss	123	57.57	47.53	57.85	33.22
Path	Moondream2	SemLoss	2026	57.00	47.64	59.64	33.04
Path	SmolVLM-256M	CE	42	54.09	47.55	60.07	36.29
Path	SmolVLM-256M	CE	123	54.89	48.58	60.47	36.58
Path	SmolVLM-256M	CE	2026	54.23	47.58	61.73	35.09
Path	SmolVLM-256M	SemLoss	42	53.34	47.60	60.09	36.70
Path	SmolVLM-256M	SemLoss	123	54.49	47.81	60.95	35.97
Path	SmolVLM-256M	SemLoss	2026	54.26	48.66	61.79	36.17
Path	Qwen2.5-VL-3B	CE	42	58.30	55.01	73.85	45.59
Path	Qwen2.5-VL-3B	SemLoss	42	58.24	55.65	74.98	47.12
SLAKE	BLIP-VQA	AugCE	42	66.26	63.86	83.64	52.40
SLAKE	BLIP-VQA	AugCE+Sem	42	65.88	63.79	83.32	52.69
SLAKE	BLIP-VQA	CE	42	66.07	57.59	70.24	39.49
SLAKE	BLIP-VQA	SemLoss	42	65.60	58.30	71.56	40.62
SLAKE	InternVL2-1B	CE	42	77.10	68.60	75.63	49.58
SLAKE	InternVL2-1B	SemLoss	42	76.53	68.97	76.95	51.93
SLAKE	Moondream2	AugCE	42	78.51	76.04	88.03	65.22
SLAKE	Moondream2	AugCE+Sem	42	79.26	75.95	88.12	65.50
SLAKE	Moondream2	CE	42	78.51	70.78	78.32	55.14

Dataset	Backbone	Objective	Seed	A	P	C	R
SLAKE	Moondream2	SemLoss	42	78.70	71.23	78.40	57.30
SLAKE	SmolVLM-256M	CE	42	73.42	62.60	71.12	43.54
SLAKE	SmolVLM-256M	SemLoss	42	73.80	63.86	72.10	45.43
SLAKE	Qwen2.5-VL-3B	CE	42	77.38	69.84	77.74	56.83
SLAKE	Qwen2.5-VL-3B	SemLoss	42	78.13	70.88	78.38	59.00
RAD	BLIP-VQA	AugCE	42	35.48	35.79	69.49	22.84
RAD	BLIP-VQA	AugCE+Sem	42	34.81	36.10	69.67	23.73
RAD	BLIP-VQA	CE	42	35.03	35.65	66.92	22.62
RAD	BLIP-VQA	SemLoss	42	35.25	35.12	68.16	23.28
RAD	InternVL2-1B	CE	42	47.67	45.37	57.74	25.06
RAD	InternVL2-1B	SemLoss	42	47.23	45.23	63.06	27.49
RAD	Moondream2	AugCE	42	47.45	46.83	70.55	33.26
RAD	Moondream2	AugCE+Sem	42	47.45	46.65	70.16	33.04
RAD	Moondream2	CE	42	45.01	45.01	67.23	30.16
RAD	Moondream2	SemLoss	42	44.79	45.01	67.89	30.60
RAD	SmolVLM-256M	CE	42	39.02	39.60	65.50	25.94
RAD	SmolVLM-256M	SemLoss	42	39.47	38.85	65.32	26.61
RAD	Qwen2.5-VL-3B	CE	42	55.65	52.99	69.98	41.02
RAD	Qwen2.5-VL-3B	SemLoss	42	55.65	53.30	69.89	41.69

C Data and question-variant construction

Dataset provenance. PathVQA originates from the PathVQA dataset report; the experiments use `flaviagiammarino/path-vqa` (revision `1685832`). SLAKE-English derives from SLAKE and uses `mdwiratathya/SLAKE-vqa-english` (`8d18b4d`). VQA-RAD originates from the radiology VQA dataset release and uses `flaviagiammarino/vqa-rad` (`bcf91e7`). These revisions define the evaluated dataset packages.

Splits and image identity. Each dataset is trained and evaluated separately using the counts in Table 13. PathVQA uses a fixed valid-positive index set of 17,654 questions from the 19,654-item source training split, with identical indices for CE and SemLoss. Each retained item has one to four positive questions after normalized-original and within-group duplicate removal; the index-file hash is recorded in the data manifest. SLAKE uses its English train/validation/test splits. VQA-RAD has no source validation split, so 10% of its training images are held out once with split seed 42. The three datasets contain 39,909 split-specific question–answer records and 5,045 distinct RGB images; each backbone applies its own preprocessing to the lossless PNG exports.

Table 13: Data inventory. Counts before parentheses are questions; parentheses give distinct images within that split. Positive instances count training question variants, not additional images or answer labels.

Dataset	Train	Validation	Test	Positive instances
PathVQA	17,654 (2,400)	6,259 (832)	6,719 (858)	56,761
SLAKE-English	4,919 (450)	1,053 (96)	1,061 (96)	24,595
VQA-RAD	1,618 (282)	175 (31)	451 (203)	8,090

RGB content hashes show one PathVQA image shared between training and validation and none shared with test; SLAKE has no pairwise overlap. VQA-RAD contains the source benchmark's image reuse: 182 images overlap between training and test and 20 between validation and test, while training and validation remain image-disjoint. Its results therefore follow the benchmark's question-level partition.

PathVQA training positives. The PathVQA training file contains 17,654 original questions and five candidate rows per question from GPT-5-nano, Grok-4.1-fast-non-reasoning, Gemini-3-flash, DeepSeek-chat, and Kimi-K2-chat. Removing normalized copies of the original and within-group duplicates leaves 56,761 positives: five questions have one, 541 have two, 12,758 have three, and

4,350 have four. Training samples one positive uniformly at each access after matching the original question and answer to its source index.

PathVQA test variants. Each of the 6,719 test questions is preserved verbatim inside five fixed image-context instructions: "In the displayed pathology image, [question]", "Based on the displayed pathology image, [question]", "With reference to the pathology image, please answer: [question]", "Examine the pathology image and determine the following: [question]", and "[question] Answer with respect to the displayed pathology image." These answer-blind context wrappers preserve the original question text.

SLAKE and VQA-RAD generation and screening. SLAKE and VQA-RAD provide five screened variants for each of 9,277 questions, totaling 46,385 variants. Candidate generation uses Qwen2.5-7B-Instruct (revision `a09a35458c702b33eeacc393d103063234e8bc28`). The candidate pool combines two generation stages, with all candidates reviewed by Qwen2.5-32B-Instruct (revision `5ede1c97bbab6ce5cda5812749b4c0bdf79b18dd`).

The prompts preserve clinical entities, modality, anatomy, laterality, numbers, negation, uncertainty, comparisons, spatial relations, and answer space while changing vocabulary or syntax. Images and reference answers are withheld. Automatic screening removes empty, duplicate, unchanged, overlong, multi-line, and non-question outputs; it also verifies numeric, laterality, negation, modality, uncertainty, and comparison signatures before requiring a positive model review. Every variant used in the experiments passes this pipeline; none receives expert review. VQA-RAD additionally receives a targeted repair pass.

D Training implementation and alternative objectives

Shared protocol and backbone adaptations. Pretrained parameters are frozen; joint training updates LoRA and the auxiliary projection head. The primary Moondream2 backbone uses the last prompt-token state and official 378-pixel dynamic crops. The other causal backbones also use an answer-prefix state; the BLIP-VQA comparator uses its fused text-encoder CLS state. Comparison preprocessing uses 384 pixels for BLIP-VQA, a single 448-pixel tile for InternVL2, and processor image splitting for SmolVLM. Each backbone uses the same specified representation and image preprocessing for CE and SemLoss.

The effective optimization batch is eight. PathVQA micro-batches are eight for BLIP, four for SmolVLM, and two for InternVL2, Moondream2, and Qwen2.5-VL, yielding seven, three, and one in-batch negatives per anchor, respectively. Gradient accumulation completes the optimization batch while semantic comparisons remain within each micro-batch. CE and SemLoss share the same backbone-specific configuration.

Moondream2 implementation. The primary training backbone is `vikhyatk/moondream2`, revision `2025-06-21`. Each language block's fused QKV projection receives low-rank updates to its query and value components, with no low-rank update to the key component. The frozen visual encoder runs once per training image, and its embeddings are reused for the original and positive question. The text forward pass uses the model's causal attention with its configured prefix-attention region and does not write an inference KV cache. Question and answer tokens are truncated to 64 and 24 tokens, respectively, and an EOS token is appended to the answer. Only answer tokens and EOS receive CE supervision; other target positions are masked. CE is averaged over valid target tokens in the batch.

Section 3 defines the shared pre-answer projection head and objective. Table 14 gives the complete Moondream2 settings.

Alternative losses. The alternative-loss experiments use the same normalized representations as the margin objective. Write $d_i^+ = 1 - z_i^\top z_i^+$, $S_{ij} = z_i^\top z_j^+ / \tau$, and $\tau = 0.07$. The pull-only objective

Table 14: Moondream2 training and selection settings. The auxiliary weight is zero for CE and 0.2 for the main SemLoss setting.

Setting	Value	Setting	Value
Epochs	3	Optimizer	AdamW
Learning rate	2×10^{-4}	Weight decay	0.01
Warmup ratio	0.06	Schedule	Linear warmup/decay
Micro-batch	2	Accumulation steps	4
Gradient clipping	1.0	Evaluation batch	2
LoRA rank / alpha	16 / 32	LoRA dropout	0.1
Projection output	256	Margin	0.2
Pull mixture weight	0.5	Auxiliary weight	0 or 0.2
Question/answer cap	64 / 24 tokens	Generation cap	24 tokens
Training precision	BF16 autocast	Training loader	Drop incomplete batch

is $B^{-1} \sum_i d_i^+$. The NCE mix combines pull with a temperature-scaled symmetric contrastive term:

$$\mathcal{C}(S) = -\frac{1}{2B} \sum_{i=1}^{B} \left[\log \frac{\exp S_{ii}}{\sum_j \exp S_{ij}} + \log \frac{\exp S_{ii}}{\sum_j \exp S_{ji}} \right], \tag{6}$$

$$\mathcal{L}_{\text{NCE-mix}} = \gamma \frac{1}{B} \sum_i d_i^+ + (1-\gamma)\tau \mathcal{C}(S), \qquad \gamma = 0.5. \tag{7}$$

Both NCE alternatives include the pull and contrastive terms. The answer-aware variant excludes an off-diagonal entry S_{ij} from the denominator when normalized training answers a_i and a_j are equal, while the diagonal remains the positive pair. Answer labels affect this training mask only. For a micro-batch with fewer than two samples, the contrastive variants reduce to the pull term; the main margin objective uses its unmasked instance negatives.

Additional answer supervision. The primary SemAlign-VQA objective applies answer CE only to the original question. The augmentation controls additionally supervise the positive question with the same answer and use

$$\mathcal{L}_{\text{augCE}} = \tfrac{1}{2} \left[\mathcal{L}_{\text{CE}}(a \mid x, q) + \mathcal{L}_{\text{CE}}(a \mid x, q^+) \right]. \tag{8}$$

The combined control adds $0.2\mathcal{L}_{\text{sem}}$. Averaging the two CE terms maintains their scale relative to the auxiliary term.

E Checkpoint selection, inference, and reproducibility

Selection and seeds. Original-question validation accuracy selects the checkpoint after each epoch, with the earliest epoch winning ties. The four-backbone PathVQA seed study uses 42, 123, and 2026; the five-backbone, three-dataset matrix uses seed 42. Table 15 lists the primary Moondream2 scores.

Table 15: PathVQA per-training-seed results (%). Scores are computed from saved predictions.

Seed	Objective	A	P	C	PC	R
42	CE	57.21	48.92	60.40	59.38	37.80
42	SemLoss	57.24	49.82	63.84	63.88	39.65
123	CE	57.60	46.27	55.82	54.70	31.40
123	SemLoss	57.57	47.53	57.85	57.67	33.22
2026	CE	57.24	47.30	58.84	56.97	33.17
2026	SemLoss	57.00	47.64	59.64	57.79	33.04

Parameter accounting and inference. The full Moondream2 training adapter contains 1,935,107,824 parameters. SemLoss updates 7,870,720 of them: 3,145,728 LoRA parameters and

4,724,992 projection-head parameters. The CE baseline's zero auxiliary weight updates only LoRA. Generation bypasses the head, giving both objectives the same 1,930,382,832-parameter inference path (Table 16).

Table 16: **Parameter updates on Moondream2.** The projector is bypassed during generation in both settings. Exact inference counts appear in Appendix E.

Updated component	CE	SemLoss
Q/V LoRA	3,145,728	3,145,728
Projection head	0	4,724,992
Total updated	**3,145,728**	**7,870,720**

Moondream2 answers each image–question pair through its official query interface with temperature zero, top-p zero, and a 24-token cap. The evaluator decodes each of the six variants independently through this common inference path and scores the six outputs directly.

Scoring protocol. Prediction files store the original text. Normalization lowercases, replaces non-alphanumeric punctuation with spaces, removes the articles *a*, *an*, and *the*, and collapses whitespace. Six-view all-correct accuracy requires every normalized output to match the reference; agreement separately records whether outputs match one another. The scorer validates question and variant identities, row counts, and reference answers before computing metrics.

Image-cluster intervals. Let image hash g identify its question groups, u_g their summed metric contributions, and n_g their count. Each replicate samples G image hashes with replacement and applies the same multiplicities w_g to CE and SemLoss, producing $\sum_g w_g u_g / \sum_g w_g n_g$. The reported pointwise percentile 95% intervals use 2,000 replicates and bootstrap seed 2026, preserving question weighting and within-image dependence. They quantify test-image variation for fixed checkpoints; the three-seed sample SD separately quantifies training variation.

Scoring definitions. Original-to-variant agreement is $C = (5N)^{-1} \sum_{i,v>0} \mathbf{1}[n(p_{iv}) = n(p_{i0})]$. Pairwise agreement averages the same indicator over all 15 unordered view pairs per item. Accuracy drop is $A - P$. ROUGE-1 uses multiset token-overlap F1 and ROUGE-L uses longest-common-subsequence F1, both after the same normalization and averaged over all $6N$ responses. Empty outputs remain in the denominator, with overlap F1 equal to one only when both token sequences are empty. Exact-match scoring uses normalized reference strings without medical-synonym mapping.

Experiments use Python 3.10, PyTorch 2.5.1+cu121, Transformers 4.57.6, PEFT 0.18.1, and Datasets 4.6.0. Experimental records include configurations, validation logs, checkpoint-selection metadata, primary predictions, fixed model identifiers, and SHA-256 manifests.

F Outcome composition and transition counts

Figure 9 shows the seed-42 outcome proportions; Table 17 lists the corresponding counts and the additional PathVQA seeds.

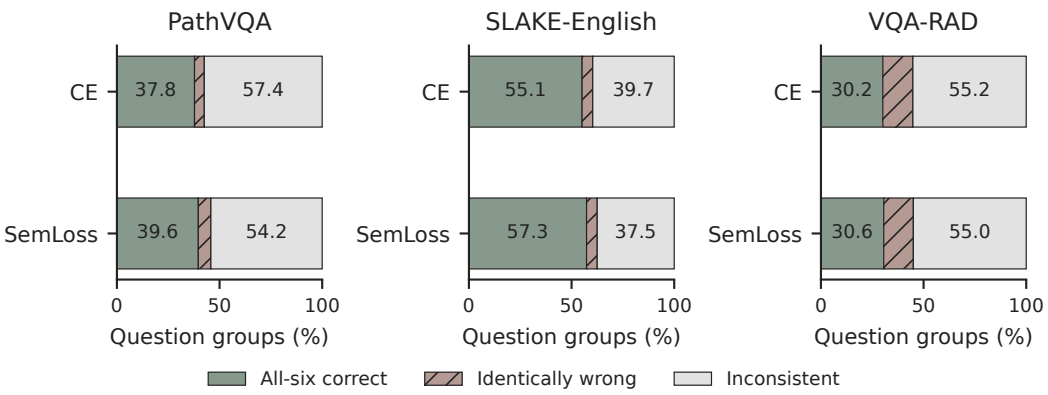

Figure 9: **Six-view outcome composition for Moondream2 at seed 42.** All-six correct, identically wrong, and inconsistent groups partition each dataset. Hatching identifies stable wrong answers independently of color.

Table 17: Counts of mutually exclusive answer-group outcomes from raw predictions.

Dataset	Seed	Objective	Correct	Wrong	Inconsistent
PathVQA	42	CE	2540	319	3860
PathVQA	42	SemLoss	2664	412	3643
PathVQA	123	CE	2110	240	4369
PathVQA	123	SemLoss	2232	284	4203
PathVQA	2026	CE	2229	324	4166
PathVQA	2026	SemLoss	2220	328	4171
SLAKE-English	42	CE	585	55	421
SLAKE-English	42	SemLoss	608	55	398
VQA-RAD	42	CE	136	66	249
VQA-RAD	42	SemLoss	138	65	248

G External checkpoint evaluation

All external checkpoints use FP16 inference, model-native image processing, greedy generation, and a 24-token budget on a shared Quadro RTX 6000. Their prompts include a common short-answer suffix; Moondream2 uses the original question text. Training exposure also varies: Lingshu and MedMO report benchmark-family data in their training mixtures, while exact exclusion of the evaluated questions is unverified. Tables 19–21 report the resulting fixed-query comparison.

The eight external checkpoints contribute 24 evaluations and 395,088 answers. On PathVQA, MedMO records 77.96% original accuracy and 40.14% all-six correctness, while Lingshu records 65.47% and 60.72%. Their ordering therefore reverses across the two metrics: 12.49 points lower original accuracy accompanies 20.58 points higher all-six correctness. These checkpoint comparisons are descriptive because their training data and prompts differ; the in-house comparisons hold the backbone fixed while changing the objective.

VGRefine uses HuatuoGPT-Vision with its published attention-refinement settings: 20 sensitive heads, layer 16, normalization enabled, Q disabled, and percentile 80. Table 18 records all model repositories and versions.

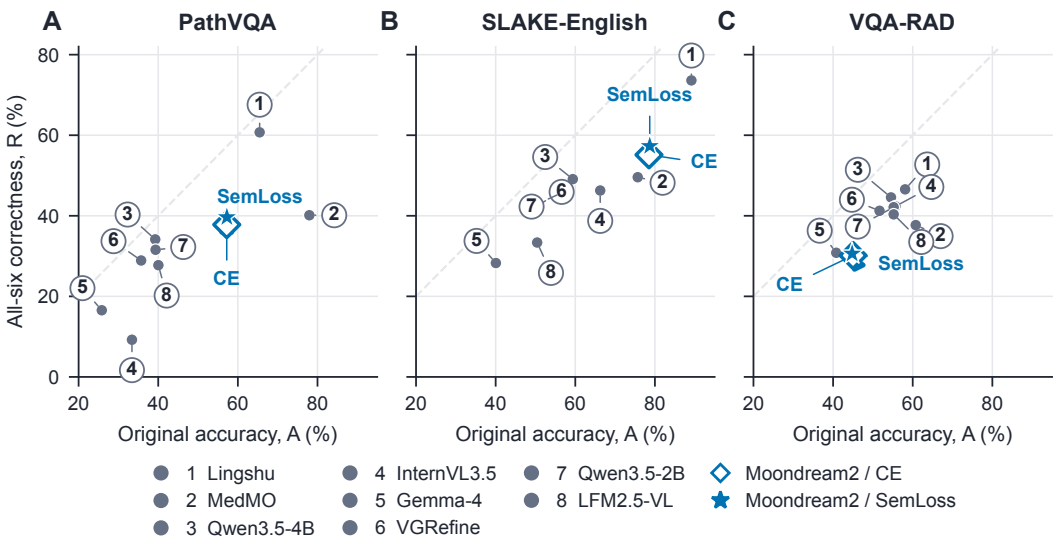

Figure 10: **Original accuracy and all-six correctness provide different model rankings.** Eight external checkpoints and Moondream2 CE/SemLoss use the same six-view queries. The connected Moondream2 points show the paired objective change; external checkpoints differ in training exposure and prompts. Full scores are in Tables 19–21.

Table 18: Pinned external checkpoints; links resolve to full revision IDs.

Checkpoint	Revision prefix
`Lingshu-I-8B`	`b004bfc0554d`
`MedMO-4B-Next`	`33dae044fdf6`
`Qwen3.5-4B`	`851bf6e806ef`
`InternVL3_5-4B-HF`	`6bd448740211`
`gemma-4-E4B-it`	`ee0ef6023621`
`HuatuoGPT-Vision-7B-hf`	`a092ca6536fe`
`Qwen3.5-2B`	`15852e8c1636`
`LFM2.5-VL-1.6B`	`919fde3d022e`

Table 19: PathVQA checkpoint scores (%). R1/RL: six-view token-overlap F1.

Checkpoint	A	P	C	PC	R	R1	RL
Moondream2 / CE	57.21	48.92	60.40	59.38	37.80	51.79	51.79
Moondream2 / SemLoss	57.24	49.82	63.84	63.88	39.65	52.49	52.48
Lingshu-I-8B	65.47	65.29	87.62	87.08	60.72	67.52	67.49
MedMO-4B-Next	77.96	58.88	63.39	64.64	40.14	66.26	66.19
Qwen3.5-4B	39.25	38.49	73.40	75.77	34.14	40.25	40.23
InternVL3.5-4B	33.40	26.89	37.73	37.95	9.23	32.08	32.05
Gemma-4-E4B-it	25.82	24.85	53.72	62.44	16.55	26.41	26.40
VGRefine / Huatuo	35.70	34.19	73.41	80.98	28.90	35.46	35.46
Qwen3.5-2B	39.34	37.43	76.09	79.40	31.57	38.94	38.94
LFM2.5-VL-1.6B	40.04	36.51	59.47	62.74	27.74	38.15	38.15

Table 20: SLAKE-English checkpoint scores (%). R1/RL: six-view token-overlap F1.

Checkpoint	A	P	C	PC	R	R1	RL
Moondream2 / CE	78.51	70.78	78.32	77.31	55.14	75.77	75.50
Moondream2 / SemLoss	78.70	71.23	78.40	78.40	57.30	75.99	75.76
Lingshu-I-8B	89.16	83.37	88.43	88.06	73.61	88.35	88.25
MedMO-4B-Next	75.68	67.37	73.67	73.14	49.58	72.98	72.93
Qwen3.5-4B	59.38	57.27	83.07	82.10	49.10	61.75	61.49
InternVL3.5-4B	66.26	60.87	73.21	71.74	46.28	66.94	66.65
Gemma-4-E4B-it	40.06	38.02	72.06	71.30	28.28	43.10	42.98
VGRefine / Huatuo	50.33	50.56	83.17	81.97	42.51	53.40	53.35
Qwen3.5-2B	56.64	55.55	79.36	78.61	45.71	60.25	59.86
LFM2.5-VL-1.6B	50.42	47.13	75.66	73.84	33.36	51.66	51.38

Table 21: VQA-RAD checkpoint scores (%). R1/RL: six-view token-overlap F1.

Checkpoint	A	P	C	PC	R	R1	RL
Moondream2 / CE	45.01	45.01	67.23	68.54	30.16	48.72	48.65
Moondream2 / SemLoss	44.79	45.01	67.89	69.16	30.60	48.69	48.61
Lingshu-I-8B	58.09	58.00	77.43	78.27	46.56	63.30	63.22
MedMO-4B-Next	60.75	55.83	65.85	70.04	37.69	61.66	61.64
Qwen3.5-4B	54.55	53.75	79.47	80.53	44.57	59.50	59.48
InternVL3.5-4B	55.65	54.06	70.91	74.46	42.35	59.08	58.93
Gemma-4-E4B-it	40.80	41.95	70.73	71.83	30.82	47.20	47.18
VGRefine / Huatuo	51.66	51.66	77.78	77.86	41.24	55.90	55.90
Qwen3.5-2B	55.21	54.10	77.34	78.95	42.13	58.88	58.85
LFM2.5-VL-1.6B	55.21	52.90	68.29	71.72	40.35	57.56	57.46