CLINKD: CROSS-MODAL DISTILLATION FOR FINE-GRAINED MEDICAL VQA

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-grained medical visual question answering requires a model to identify findings, ground them spatially, and generate clinically relevant text. We investigate two potential weaknesses in adapting a general-purpose multimodal model to this setting: contiguous position indices do not distinguish image–text boundaries, and uniform distillation can overemphasize uncertain teacher targets. ClinKD addresses these issues with Modality-aware Cross-modal Gap Rotary Position Embedding (MCG-RoPE) and confidence-margin curriculum distillation. MCG-RoPE inserts explicit boundary gaps while preserving the native spatial grid, whereas the distillation objective weights teacher-forced token distributions by confidence and top-two probability margin and gradually relaxes its admission threshold. On Med-GRIT-Test30k, the full pipeline obtains 67.51, 82.35, 70.56, and 65.69 on visual grounding, referring object classification, referring captioning, and medical image analysis, respectively, averaging 71.53. This is 14.87 points above the reported BiRD baseline, which uses a different backbone and is therefore not architecture-matched; on LLaVA-Med-qa0.2k, mBMR increases from 21.06 to 23.54. Single-component experiments show gains from MCG-RoPE and distillation separately, although the archived results do not isolate every interaction in the full pipeline.

1 INTRODUCTION

Medical visual question answering (Med-VQA) extends image understanding beyond recognition by requiring models to connect visual evidence with domain-specific language (Lin et al., 2023; Abacha et al., 2019; Hasan et al., 2018). This requirement becomes more demanding in fine-grained settings that combine visual grounding (VG), referring object classification (ROC), referring captioning (RC), and medical image analysis (MIA). A successful model must localize a region, identify the corresponding anatomy or finding, and describe the image without losing the relation between the textual query and spatial evidence. Multimodal large language models (MLLMs) provide a flexible interface for these tasks, but adapting a general-purpose model to specialized medical imagery remains difficult (Caffagni et al., 2024; Nam et al., 2025; Maity & Saikia, 2025).

LLaVA-Med adapts visual instruction tuning to biomedical data and demonstrates the value of paired medical images and language (Li et al., 2023). BiRD subsequently introduces Med-GRIT, a refer-and-ground instruction corpus spanning four tasks and eight imaging modalities, and equips a vision-language model with fine-grained medical grounding capabilities (Huang et al., 2024). These systems establish strong foundations but leave two design questions open. Should a position sequence explicitly distinguish a within-image step from an image–text transition? Can teacher uncertainty regulate noisy domain predictions more effectively than uniform distillation? We study these as representational and optimization hypotheses rather than established causes of model error, and we do not treat downstream scores as direct measurements of latent alignment or medical knowledge.

We propose ClinKD, a framework for cross-modal distillation in fine-grained Med-VQA. Its first component, Modality-aware Cross-modal Gap Rotary Position Embedding (MCG-RoPE), inserts an explicit index gap at modality boundaries while retaining the native temporal, height, and width indices of the visual grid. The gap changes the relative rotation between image and text tokens without introducing a new encoder or additional attention parameters. Its second component, Adaptive Confidence-Margin Curriculum Pseudo-KD, weights teacher supervision using both the maximum

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.

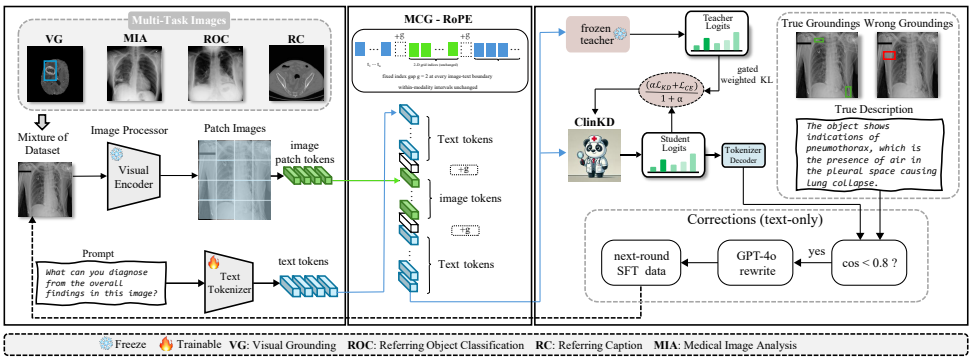


Figure 1: ClinKD training and inference pipeline. MCG-RoPE marks modality transitions while preserving the visual grid. Confidence-margin curriculum distillation precedes supervised fine-tuning; low-similarity textual responses can then be corrected and reused. Semantic reranking is an optional inference-time stage.

probability and the margin between the two most likely tokens. A high threshold initially admits only reliable targets, and a linear schedule gradually expands the set of distilled examples. Reflective text correction and semantic reranking are auxiliary stages; because neither stage changes ground-truth boxes, we do not present them as mechanisms for correcting localization errors.

Our contributions are threefold:

- We formulate MCG-RoPE, a parameter-free modification of multimodal position indices that makes modality transitions explicit while preserving two-dimensional visual structure.
- We introduce confidence-margin curriculum distillation for autoregressive medical responses, including a token-masked objective and a schedule that limits unreliable teacher supervision early in training.
- We evaluate ClinKD on four Med-GRIT tasks drawn from eight imaging modalities and on an external Med-VQA set. Under the reported protocol, and relative to a BiRD baseline that is not architecture-matched, the full pipeline improves the Med-GRIT aggregate from 56.66 to 71.53 and external-set mBMR from 21.06 to 23.54.

2 Related Work

2.1 Medical Vision-Language Models

Biomedical MLLMs combine a visual encoder with an instruction-following language model for image interpretation (Nazi & Peng, 2024; Reale-Nosei et al., 2024; Ryu et al., 2025). LLaVA-Med constructs biomedical instruction data (Li et al., 2023), BioMedGPT studies a general biomedical foundation model (Zhang et al., 2024b), and Med-Flamingo focuses on few-shot reasoning (Moor et al., 2023); more recent work augments medical VQA with caption-based reasoning and parameter-efficient tuning (Li et al., 2025). BiRD is the closest comparison because it targets fine-grained medical referring and grounding (Huang et al., 2024); related general-domain systems include Ferret, Ferret-v2, and GroundingGPT (You et al., 2023; Zhang et al., 2024a; Li et al., 2024), and the broader landscape of referring and grounding is surveyed in (Xiao et al., 2025). Parameter-efficient medical grounding and uncertainty-aware phrase grounding have also been studied independently (He et al., 2025; Zou et al., 2025). ClinKD instead examines modality-aware positions and uncertainty-aware distillation in BiRD's multi-task output space.

2.2 Position Encoding for Multimodal Inputs

Rotary Position Embedding (RoPE) represents relative token offsets through rotations of query and key features (Su et al., 2024), and its behavior under long-context extrapolation continues to be analyzed systematically (Zhao et al., 2024). Extensions to vision introduce axis-specific or learnable

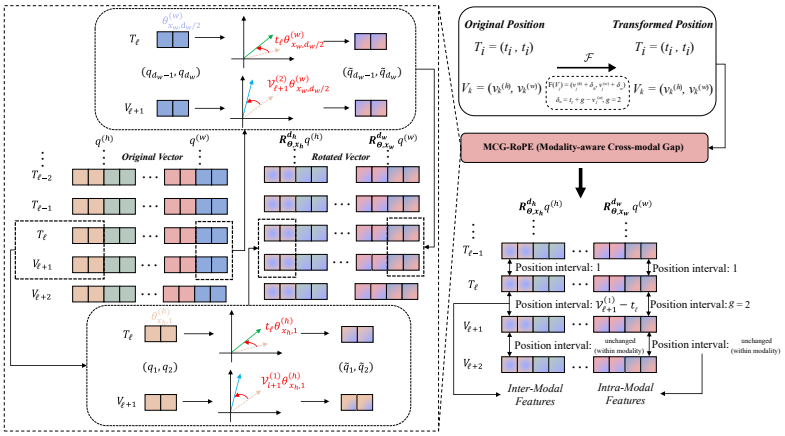


Figure 2: MCG-RoPE preserves the native visual grid but inserts an explicit gap at each image–text boundary. The discontinuity changes relative query–key rotation across modalities while retaining within-modality order.

frequencies for two-dimensional structure (Heo et al., 2024), VideoRoPE extends the design to temporal inputs (Wei et al., 2025), and sequence-level schemes learn position representations directly (Li et al., 2026a). These approaches primarily refine geometry within a modality. MCG-RoPE instead preserves the underlying grid and modifies only the index discontinuity at a modality transition; the present experiments use one visual backbone and do not claim encoder-independent effectiveness.

2.3 Knowledge Distillation with Noisy Targets

Knowledge distillation transfers teacher representations or distributions to a student (Hinton et al., 2015; Tung & Mori, 2019; Miles et al., 2024; Gou et al., 2021), with recent surveys dedicated to medical imaging (Li et al., 2026b) and to distilling large language models (Fang et al., 2025). Pseudo-labeling extends supervision to model-generated targets, but uncertainty can propagate errors, particularly in medical prediction (Xue et al., 2022). Curriculum distillation has therefore incorporated prediction and label uncertainty (Jiménez-Sánchez et al., 2022). ClinKD adapts this principle to autoregressive targets through confidence, top-two margin, and a curriculum threshold.

3 Method

Figure 1 summarizes ClinKD. Given a medical image and an instruction, the visual encoder produces a two-dimensional token grid and the language model autoregressively predicts a response. ClinKD modifies the pipeline at two points. MCG-RoPE marks image–text transitions in the position sequence, and Adaptive Confidence-Margin Curriculum Pseudo-KD controls which teacher predictions influence the student. Reflective correction and semantic-aware selective generation are auxiliary training- and inference-time stages.

3.1 Modality-Aware Rotary Position Encoding

Standard RoPE rotates query and key vectors as a deterministic function of their position indices (Su et al., 2024). Multimodal implementations extend the index to temporal, height, and width dimensions, but a contiguous assignment can make a modality transition numerically indistinguishable from a neighboring within-modality step. Our modality-aware cross-modal gap scheme (MCG-RoPE) introduces a fixed boundary gap to distinguish these relations without modifying the attention mechanism or adding learned parameters.

Let a packed text–image–text sequence contain text positions $T_i = (t_i, t_i)$ and visual positions $V_j = (v_j^{(h)}, v_j^{(w)})$. Qwen2-VL also maintains a temporal stream; for a still image, its temporal index is constant within the visual segment, so we write only the spatial streams. We transform visual

indices as

$$\mathcal{F}(V_j) = \big(v_j^{(h)} + \delta_h,\ v_j^{(w)} + \delta_w\big), \tag{1}$$

where $\delta_h = t_\ell + g - v_1^{(h)}$ and $\delta_w = t_\ell + g - v_1^{(w)}$ place the first visual token g units after the last preceding text token T_ℓ. The implementation uses $g = 2$, whereas adjacent text positions retain unit spacing. If $m = \max_j \max\{\mathcal{F}(V_j)^{(h)}, \mathcal{F}(V_j)^{(w)}\}$, trailing text begins at $m + g$ and resumes unit spacing. Thus MCG-RoPE changes cross-modal relative offsets without invoking a separate encoder or rescaling the internal visual grid.

For a query or key vector $q = [q^{(h)}; q^{(w)}]$, let x_h, x_w be transformed indices and $\omega_i^{(a)} = 10{,}000^{-2i/d_a}$ for $a \in \{h, w\}$. Applying the standard block rotation $R(\theta)$ with $\theta_{x_a,i}^{(a)} = x_a \omega_i^{(a)}$ gives

$$\begin{aligned} \widetilde{q} &= \left[\mathbf{R}^{d_h}_{\mathbf{\Theta},x_h} q^{(h)}; \mathbf{R}^{d_w}_{\mathbf{\Theta},x_w} q^{(w)}\right], \\ R(\theta) &= \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \end{aligned} \tag{2}$$

Because attention depends on relative rotations, the boundary gap distinguishes cross-modal interactions from local movement inside the image. We evaluate its effect through downstream tasks rather than treating those scores as a direct alignment measurement.

3.1.1 Effect on Relative Attention

For one rotary frequency, the dot product between a query at index a and a key at index b can be written as $q_a^\top R(b-a)k_b$. A conventional packed sequence assigns the last prompt token and the first visual token consecutive indices, so their relative phase is the same as that of two neighboring text tokens. MCG-RoPE replaces this boundary difference by g while keeping differences inside each segment unchanged. It therefore changes attention terms whose token pair either crosses a modality boundary or spans the inserted gap; within-segment pairs—text tokens on the same side of the image, and local image–image pairs—retain their original offsets, whereas a pre-image and a post-image text token are separated by the additional gap. This locality is useful because the modification does not require relearning the visual grid, but it also limits the claim: MCG-RoPE does not directly enforce semantic correspondence and cannot guarantee that an attended region is clinically relevant.

The fixed gap introduces no trainable parameter and does not change asymptotic attention complexity. Position construction is linear in sequence length, while attention remains quadratic in the packed token count. Nevertheless, the optimal gap may depend on sequence length, token ordering, and the frequency spectrum of the backbone. A large gap can move high-frequency channels through rapidly varying phases, whereas a gap close to one becomes nearly indistinguishable from ordinary packing. The archived experiment evaluates $g = 2$ only, so sensitivity to g and transfer to multi-image or video sequences remain open questions.

3.2 Confidence-Margin Curriculum Distillation

Pseudo-label distillation can amplify teacher mistakes when uncertain predictions receive the same weight as confident ones. ClinKD combines confidence, decisiveness, and training progress in a token-masked autoregressive objective. For example x and answer token r, the frozen teacher and student consume the same ground-truth prefix through teacher forcing. The teacher probabilities are

$$p_{t,r,j}(x) = \frac{\exp(z_{t,r,j}/T)}{\sum_k \exp(z_{t,r,k}/T)}, \tag{3}$$

where T is temperature and j indexes the shared vocabulary. We compute confidence $C_r(x) = \max_j p_{t,r,j}(x)$ and probability margin $\Delta_r(x) = p_{t,r,(1)}(x) - p_{t,r,(2)}(x)$. Padding and instruction tokens are excluded by answer mask m_r.

At normalized progress $u \in [0, 1]$, the admission threshold decreases linearly:

$$\tau(u) = \tau_{\min} + (\tau_0 - \tau_{\min})(1 - u), \tag{4}$$

and the token weight is

$$w_r(x) = m_r\, \mathbb{I}[C_r(x) \geq \tau(u)]\, C_r(x)^\gamma \Delta_r(x)^\beta. \tag{5}$$

Algorithm 1 Confidence-Margin Curriculum Distillation

Require: Training batches $\mathcal{B}$, frozen teacher f_t, student f_s
Require: $T, \alpha, \beta, \gamma, \tau_0, \tau_{\min}$
for training step $s = 1, \ldots, S$ **do**
  Set $u \leftarrow (s-1)/(S-1)$ and compute $\tau(u)$
  Run f_t and f_s with the same teacher-forced answer prefix
  for each answer token r in $\mathcal{B}$ **do**
    Compute C_r, Δ_r, and w_r using Equation 5
    Accumulate weighted forward KL on the shared vocabulary
  end for
  Normalize accepted-token mass using Equation 6
  Update f_s with cross-entropy and weighted KD
end for

Early optimization therefore emphasizes confident, separated teacher predictions; later optimization admits harder targets while downweighting ambiguous tokens. Let $p_{s,r}$ be the student's temperature-scaled distribution. Normalizing by accepted token mass yields

$$\mathcal{L}_{\mathrm{KD}} = \frac{T^2 \sum_{i,r} w_r(x_i) \mathbb{D}_{\mathrm{KL}}(p_{t,r}(x_i) \| p_{s,r}(x_i))}{\sum_{i,r} w_r(x_i) + \epsilon}. \tag{6}$$

The final objective combines answer-token cross-entropy and distillation:

$$\mathcal{L} = \frac{\mathcal{L}_{\mathrm{CE}} + \alpha \mathcal{L}_{\mathrm{KD}}}{1 + \alpha}. \tag{7}$$

3.2.1 Curriculum Behavior

Equation 5 separates three decisions. The indicator determines whether a token enters the distillation set, confidence controls the magnitude of admitted supervision, and the top-two margin discounts distributions whose most likely alternatives remain difficult to distinguish. At $u = 0$, only tokens above τ_0 contribute; at $u = 1$, the criterion becomes $C_r(x) \geq \tau_{\min}$. This schedule changes the accepted set rather than increasing the global KD coefficient. Consequently, α retains the same interpretation throughout training while the content of $\mathcal{L}_{\mathrm{KD}}$ becomes progressively more difficult.

The normalization in Equation 6 prevents examples with many admitted answer tokens from dominating solely because of sequence length. If a mini-batch contains no accepted token, the numerator is zero and the ϵ term keeps the loss finite; the update then reduces to answer-token cross-entropy. This fallback avoids undefined gradients but also means that poorly calibrated teacher confidence can effectively disable distillation early in training. Temperature scaling affects both KL targets and the confidence used for admission, so T should not be interpreted independently of τ_0 and $\tau_{\min}$. The current results report one joint setting and do not establish robustness to teacher miscalibration.

3.3 Auxiliary Correction and Reranking

After distillation, supervised fine-tuning records predicted text $\hat{t}$ and reference text t. Their embeddings are compared using $s(\hat{t}, t) = \langle e(\hat{t}), e(t) \rangle / (\|e(\hat{t})\| \|e(t)\|)$. If $s(\hat{t}, t) < 0.8$, GPT-4o receives both texts and produces a paraphrase constrained to the reference intent, which is reused in the next fine-tuning round. This diversifies labels rather than recovering information absent from the reference. Coordinates remain unchanged, so the stage does not correct predicted boxes. At inference, Semantic-Aware Selective Generation (SASG) optionally generates several candidates, scores image–text similarity with a CLIP-style medical encoder (Zhao et al., 2025; Khattak et al., 2024; Lian et al., 2025), and returns the highest-scoring response. Its contribution is isolated in Table 3 because it increases test-time computation.

Model	VG	ROC	RC	MIA	Avg.
LLaVA-Med	0.00	2.75	8.18	11.20	5.53
BiRD	53.92	65.33	55.23	52.17	56.66
ClinKD	**67.51**	**82.35**	**70.56**	**65.69**	**71.53**

Table 1: Med-GRIT-Test30k results against LLaVA-Med (Li et al., 2023) and BiRD (Huang et al., 2024). Avg. is the mean of four task metrics.

Model	MIA (mBMR)
LLaVA-Med (Li et al., 2023)	20.04
BiRD (Huang et al., 2024)	21.06
ClinKD	**23.54**

Table 2: Results on LLaVA-Med-qa0.2k.

4 Experiments

4.1 Experimental Setup

Datasets. We train on the 240k training portion of Med-GRIT and reserve Med-GRIT-Test30k for evaluation (Huang et al., 2024). The benchmark combines VG, ROC, RC, and MIA over CT, MR, X-ray, PET, endoscopy, dermoscopy, fundus, and ultrasound images. Pseudo-label construction, correction, and parameter selection use training examples only. We additionally evaluate MIA on the 193-question LLaVA-Med-qa0.2k set (Li et al., 2023). Few-shot transfer is reported for VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), and Path-VQA (He et al., 2020); because the archive omits the shot-sampling protocol and repeated draws, these results are secondary evidence.

No hyperparameter was tuned against any evaluation set: the optimizer and loss settings follow the base fine-tuning recipe, and the boundary gap was fixed at $g = 2$ before evaluation. **Metrics.** Following BiRD, we use Recall@0.5 for VG, recall for ROC, SPICE for RC (Anderson et al., 2016), and mBMR for MIA (Huang et al., 2024). The Med-GRIT aggregate is the unweighted mean of these heterogeneous metrics and serves only as a compact summary; task-level values remain primary. Transfer sets use BLEU-1 and F1 where available. All headline values are point estimates, so we report numerical rather than statistically significant differences.

Implementation. The reported experiments initialize the student from Qwen2-VL-7B-Instruct (Wang et al., 2024) and preserve a frozen visual tower during language-side adaptation. Images are resampled to a square grid. Distillation uses $\alpha = 0.5$, $\beta = 0.6$, $\gamma = 0.4$, $\tau_0 = 0.9$, $\tau_{\min} = 0.4$, and $T = 0.5$, followed by two supervised fine-tuning epochs. Optimization uses AdamW, a cosine schedule, learning rate 2×10^{-5}, and six NVIDIA A100 40GB GPUs; non-reranked inference uses beam size three. "BiRD" denotes the published checkpoint and evaluation protocol, while "BiRD + component" denotes the corresponding archived run; these are not architecture-matched Qwen2-VL retrainings. The archive does not preserve the teacher checkpoint, random seeds, batch size, image resolution, GPT-4o prompt version, or complete SASG configuration, preventing exact reproduction.

4.2 Main Results

Table 1 shows that ClinKD reaches 71.53 on Med-GRIT-Test30k, 14.87 points above BiRD. Gains cover VG (13.59), ROC (17.02), RC (15.33), and MIA (13.52). LLaVA-Med was not trained for refer-and-ground output, making BiRD the more informative comparison. On LLaVA-Med-qa0.2k, ClinKD improves mBMR by 2.48 points over BiRD and 3.50 over LLaVA-Med. The smaller external gain indicates that in-domain improvement does not eliminate distribution and response-style shifts.

4.3 Component Analysis

Table 3 separates reported component effects. The Qwen2-VL SFT run exceeds BiRD by 4.19 points, with most of the difference in MIA; because architectures differ, this row is contextual rather than controlled. Adding MCG-RoPE to BiRD raises the aggregate from 56.66 to 64.43, with all four tasks improving. Pseudo-KD produces the largest isolated gain, reaching 69.34. The full pipeline reaches 71.53, but the archive lacks a row combining MCG-RoPE and Pseudo-KD without auxiliary stages, so their interaction is not isolated.

Configuration	MCG	KD	SASG	VG	ROC	RC	MIA	Avg.
Qwen2-VL SFT	–	–	–	55.43	68.33	57.39	62.23	60.85
BiRD	–	–	–	53.92	65.33	55.23	52.17	56.66
BiRD + MCG-RoPE	✓	–	–	64.03	72.74	61.32	59.63	64.43
BiRD + Pseudo-KD	–	✓	–	66.33	79.52	67.21	64.32	69.34
BiRD + SASG	–	–	✓	55.89	67.84	54.32	51.27	57.33
ClinKD	✓	✓	✓	**67.51**	**82.35**	**70.56**	**65.69**	**71.53**

Table 3: Component analysis on Med-GRIT-Test30k. Reflective correction is not independently ablated in the archived results.

Task	Model	CT	MR	X-ray	PET	Endo.	Derm.	Fundus	US
VG	BiRD	44.47	29.26	41.73	56.46	53.60	75.63	84.15	46.04
VG	ClinKD	**65.52**	**52.23**	**48.56**	**69.25**	**60.37**	**94.22**	**86.59**	**64.47**
ROC	BiRD	34.76	61.79	53.74	–	60.40	96.61	–	84.65
ROC	ClinKD	**61.42**	**71.30**	**97.00**	–	**68.84**	**100.00**	–	**95.54**
RC	BiRD	41.88	51.69	37.39	47.95	**54.07**	77.44	48.73	**82.65**
RC	ClinKD	**53.94**	**58.41**	**71.47**	**79.91**	52.10	**83.92**	**84.51**	80.19
MIA	BiRD	47.01	49.35	37.17	57.15	39.91	72.13	48.87	65.78
MIA	ClinKD	**66.85**	**56.82**	**54.01**	**67.08**	**50.58**	**81.00**	**70.02**	**76.05**

Table 4: Per-modality Med-GRIT-Test30k scores. "–" indicates a modality without an archived ROC value. Endo., Derm., and US denote endoscopy, dermoscopy, and ultrasound.

SASG increases the aggregate by only 0.67 points, improving VG and ROC but reducing RC and MIA. CLIP selection can favor short responses with strong lexical-image correspondence over richer descriptions, and additional candidates create unmatched test-time compute. We therefore report SASG as optional. Reflective correction is present in the full pipeline but lacks a controlled ablation; its independent contribution remains unresolved.

4.4 Alternative Components and Transfer

RoPE-Mixed (Heo et al., 2024) obtains 59.08, 69.80, 55.54, and 57.03 on VG, ROC, RC, and MIA when added to BiRD. MCG-RoPE reaches 64.03, 72.74, 61.32, and 59.63, gains of 4.95, 2.94, 5.78, and 2.60. For distillation, V_kD (Miles et al., 2024) obtains 63.93, 76.00, 64.15, and 61.42, whereas Pseudo-KD reaches 66.33, 79.52, 67.21, and 64.32. These limited comparisons do not support a broad state-of-the-art claim.

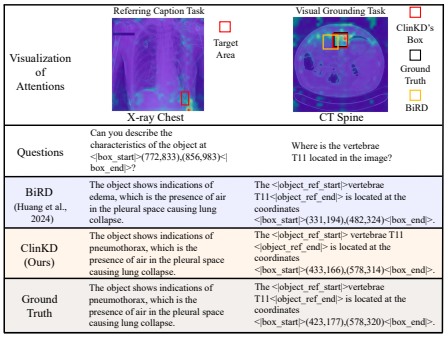
Table 5 records available transfer results. ClinKD is numerically best on all metrics, with gains over BiRD from 8.86 to 13.09 points. They are compatible with improved transfer but not definitive few-shot estimates because the archive omits the number of shots, sampled demonstrations, and repeated-draw variability.

4.5 Modality-Level Behavior

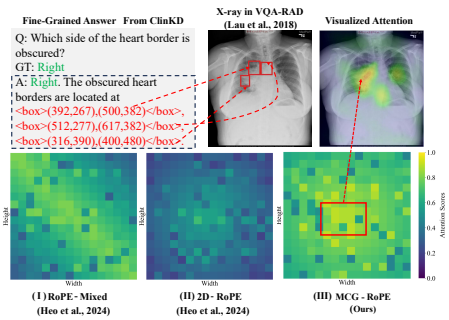
The modality breakdown in Table 4 tests whether the aggregate gain is concentrated in one image family. ClinKD improves over BiRD for every reported modality on VG and MIA, and on six of the eight RC modalities. On VG, the largest absolute increases occur for MR (22.97 points), CT (21.05), dermoscopy (18.59), and ultrasound (18.43), while fundus changes by 2.44 points because BiRD already scores 84.15. RC exhibits broad gains as well, including 31.96 points on PET and 35.78 on fundus, but endoscopy and ultrasound decrease by 1.97 and 2.46 points. ROC has missing PET and fundus annotations in the archived table; among evaluated modalities, ClinKD improves on every column, most strongly on X-ray ($+43.26$) and CT ($+26.66$), while ClinKD saturates dermoscopy at 100.00 (BiRD: 96.61). These results show broad numerical coverage rather than uniform superiority, and the unreported number of cases per modality prevents interpreting the macro pattern as a patient-level robustness result.

Dataset	Metric	L-Med	BiRD	M-Fla.	ClinKD
VQA-RAD	F1	42.91	46.12	41.79	**59.21**
SLAKE	B1	43.35	48.23	43.27	**57.09**
SLAKE	F1	44.28	49.25	45.86	**58.64**
Path-VQA	B1	43.35	48.23	41.98	**58.63**
Path-VQA	F1	44.28	49.25	42.56	**59.21**

Table 5: Archived few-shot results; sampling uncertainty is unavailable.



(a) Med-GRIT case.



(b) VQA-RAD case.

Figure 3: Qualitative grounding and attention comparison. The Med-GRIT case (left) contrasts a ClinKD grounding box with the reference annotation; the VQA-RAD case (right) compares attention maps for RoPE-Mixed and the axial 2D-RoPE variant, both from Heo et al. (2024), and our modality-aware scheme.

4.6 Qualitative Analysis

Figure 3 illustrates one grounding case and one attention case. ClinKD produces a box closer to the Med-GRIT annotation, while MCG-RoPE concentrates more attention around the annotated VQA-RAD finding than the RoPE-Mixed and axial 2D-RoPE variants of Heo et al. (2024). These examples are illustrative; they neither establish causality nor replace quantitative evaluation.

4.7 Error Analysis and Computational Cost

The component results reveal two recurring failure modes. First, methods optimized for image–text similarity do not necessarily improve long-form medical responses: SASG raises VG and ROC but reduces RC and MIA, consistent with a selector preferring concise lexical overlap over complete descriptions. Second, modality gains are uneven. Endoscopy RC and ultrasound RC are lower for ClinKD than for BiRD despite improvements on the other tasks, which suggests that a fixed position gap and one global curriculum may not match every visual distribution. A practical extension would calibrate the admission threshold by task or modality, but the current evidence does not establish that such conditioning would improve performance.

The method also introduces costs beyond ordinary supervised adaptation. Distillation requires teacher and student forward distributions over the shared vocabulary, while reflective correction invokes an external model for low-similarity responses. SASG multiplies decoding and CLIP-scoring cost by the number of candidates. The archived setup records six A100 40GB GPUs but not wall-clock time, accepted-token rate, number of corrected responses, SASG candidate count, or inference latency. Consequently, we cannot provide a reliable accuracy–compute frontier. Reporting these quantities is necessary before the method can be compared fairly with parameter-efficient or single-pass medical MLLMs.

4.8 Threats to Validity

The principal internal-validity threat is confounding between backbone, training recipe, and proposed components. Qwen2-VL SFT and BiRD are not matched initializations, while the full ClinKD row combines MCG-RoPE, Pseudo-KD, reflective correction, and SASG. The single-component rows establish numerical associations but do not identify interaction effects. A complete factorial design would include the MCG-RoPE plus Pseudo-KD combination with correction and reranking disabled, followed by separate additions of each auxiliary stage under identical decoding budgets.

Construct validity is limited by the metrics. Recall@0.5 measures overlap at one threshold, SPICE evaluates scene-graph propositions, and mBMR aggregates text-similarity measures; none directly measures clinical correctness or harm. Their unweighted average is convenient for model selection but has no natural clinical unit. External validity is likewise restricted because Med-GRIT and LLaVA-Med do not capture the full range of institutions, scanners, prevalence, or reporting conventions. The per-modality table broadens descriptive coverage, but without patient-level grouping and site metadata it cannot establish robustness to deployment shift.

Finally, conclusion validity is constrained by absent repeated runs. Several component differences are large, but magnitude alone does not reveal variability from initialization, data order, or demonstration sampling. A confirmatory study should predefine seeds, report paired per-example bootstrap intervals, and use matched compute for every baseline to distinguish stable method effects from one-run rankings.

5 Clinical Interpretation

ClinKD improves most on ROC, followed by RC and VG, while the external-set MIA gain is smaller. This ordering is consistent with its emphasis on localized image–language relations, but it does not prove clinically meaningful evidence: a better-overlap box can still select the wrong structure, and a caption can match reference propositions while omitting a decisive finding. Specialist review is needed to distinguish metric gains from medically consequential gains. Medical VQA is also heterogeneous: dermoscopy and fundus images often contain compact two-dimensional findings, CT and MR examples may be isolated slices from volumetric studies, and endoscopy differs in viewpoint and illumination. A fixed rotary gap does not encode these acquisition differences or adjacent-slice context.

6 Practical Scope and Evaluation Priorities

The four tasks expose different error costs. VG can point to the wrong structure, ROC can assign the wrong category, and RC or MIA can omit or invent findings despite high surface similarity. Deployment-oriented evaluation should therefore preserve task decomposition; vary grounding thresholds and lesion size; distinguish unsupported, omitted, and misattributed findings; and stratify by modality, institution, protocol, and patient subgroup. Auxiliary stages and uncertainty also require controls: corrected text should be audited against its reference, while SASG should be compared under fixed decoding budgets with random-selection and multi-sample controls. Because teacher confidence governs admission, future work should report calibration, selective risk, and coverage under abstention. The aggregate remains a research summary, not a clinical-utility measure.

7 Discussion and Conclusion

ClinKD combines modality-aware position indices with confidence-margin curriculum distillation for fine-grained medical VQA, and the full archived pipeline reaches 71.53 on Med-GRIT-Test30k. The single-component runs are compatible with the intended roles of boundary-aware positions and selective teacher supervision, but they do not mechanistically verify improved alignment or knowledge transfer. The evidence also has important limits: baselines are narrow and not fully architecture-matched, results are single-run point estimates, and the archive lacks a factorial ablation separating the four components. ClinKD should therefore be viewed as a research prototype, not a clinical decision aid.

AI USE STATEMENT

AI tools were used for language polishing.

ETHICS STATEMENT

Fine-grained grounding can make medical VLM outputs easier to inspect because a textual answer is paired with a localized region. This does not make the system clinically reliable: incorrect boxes and fluent descriptions can jointly create a false impression of evidence, and hallucination in medical vision-language models is a documented and safety-critical failure mode (Gu et al., 2025; Sadanandan & Behzadan, 2026). The training data span heterogeneous public imaging sources, and the current evaluation does not characterize demographic representation, acquisition-site shift, rare diseases, or calibration under prevalence changes. ClinKD should therefore not be used for diagnosis, triage, or treatment decisions without prospective validation, clinician oversight, and task-specific risk controls. Reflective correction also sends generated medical text to a proprietary model; any deployment using patient data would require privacy review and a compliant data-governance process.

REPRODUCIBILITY STATEMENT

The released artifact and manuscript are insufficient for exact replication. The student initialization, principal loss hyperparameters, optimizer, learning rate, epoch count, and GPU type are recorded, but the teacher checkpoint, random seeds, effective batch size, exact image resolution, correction prompt, reranking configuration, and few-shot demonstration sampling are absent. These omissions do not change the reported point estimates, but they limit verification. We commit to releasing the training and evaluation code, the model configurations, prompts, and per-example predictions under a permissive open-source license upon publication, with repeated runs reporting mean, standard deviation, confidence intervals, and compute cost for all matched baselines, following emerging reporting guidelines for language-model studies (Gallifant et al., 2025).

REFERENCES

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*, volume 2380 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. *SPICE: Semantic Propositional Image Caption Evaluation*, pp. 382–398. Springer International Publishing, 2016. doi: 10.1007/978-3-319-46454-1_24.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13590–13618. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.807.

Luyang Fang, Xiaowei Yu, Jiazhang Cai, Yongkai Chen, Shushan Wu, Zhengliang Liu, Zhenyuan Yang, Haoran Lu, Xilin Gong, Yufang Liu, Terry Ma, Wei Ruan, Ali Abbasi, Jing Zhang, Tao Wang, Ehsan Latif, Wei Liu, Wei Zhang, Soheil Kolouri, Xiaoming Zhai, Dajiang Zhu, Wenxuan Zhong, Tianming Liu, and Ping Ma. Knowledge distillation and dataset distillation of large language models: emerging trends, challenges, and future directions. *Artificial Intelligence Review*, 59(1), November 2025. ISSN 1573-7462. doi: 10.1007/s10462-025-11423-3.

Jack Gallifant, Majid Afshar, Saleem Ameen, Yindalon Aphinyanaphongs, Shan Chen, Giovanni Cacciamani, Dina Demner-Fushman, Dmitriy Dligach, Roxana Daneshjou, Chrystinne Fernandes, Lasse Hyldig Hansen, Adam Landman, Lisa Lehmann, Liam G. McCoy, Timothy Miller, Amy Moreno, Nikolaj Munch, David Restrepo, Guergana Savova, Renato Umeton, Judy Wawira Gichoya, Gary S. Collins, Karel G. M. Moons, Leo A. Celi, and Danielle S. Bitterman. The

tripod-llm reporting guideline for studies using large language models. *Nature Medicine*, 31(1): 60–69, January 2025. ISSN 1546-170X. doi: 10.1038/s41591-024-03425-5.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, March 2021. ISSN 1573-1405. doi: 10.1007/s11263-021-01453-z. URL http://dx.doi.org/10.1007/s11263-021-01453-z.

Zishan Gu, Jiayuan Chen, Fenglin Liu, Changchang Yin, and Ping Zhang. Medvh: Toward systematic evaluation of hallucination for large vision language models in the medical context. *Advanced Intelligent Systems*, 8(1), July 2025. ISSN 2640-4567. doi: 10.1002/aisy.202500255.

Sadid A. Hasan, Yuan Ling, Oladimeji Farri, Joey Liu, Henning Müller, and Matthew P. Lungren. Overview of imageclef 2018 medical domain visual question answering task. In *Working Notes of CLEF 2018*, volume 2125 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018.

Jinlong He, Pengfei Li, Gang Liu, and Shenjun Zhong. Parameter-efficient fine-tuning medical multimodal large language models for medical visual grounding, 2025. URL https://arxiv.org/abs/2410.23822.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering, 2020.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. *Rotary Position Embedding for Vision Transformer*, pp. 289–305. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72684-2_17.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

Xiaoshuang Huang, Haifeng Huang, Lingdong Shen, Yehui Yang, Fangxin Shang, Junwei Liu, and Jia Liu. *A Refer-and-Ground Multimodal Large Language Model for Biomedicine*, pp. 399–409. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72390-2_38.

Amelia Jiménez-Sánchez, Diana Mateus, Sonja Kirchhoff, Chlodwig Kirchhoff, Peter Biberthaler, Nassir Navab, Miguel A. González Ballester, and Gemma Piella. Curriculum learning for improved femur fracture classification: Scheduling data with prior knowledge and uncertainty. *Medical Image Analysis*, 75:102273, January 2022. ISSN 1361-8415. doi: 10.1016/j.media.2021.102273. URL http://dx.doi.org/10.1016/j.media.2021.102273.

Muhammad Uzair Khattak, Shahina Kunhimon, Muzammal Naseer, Salman Khan, and Fahad Shahbaz Khan. Unimed-clip: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities, 2024. URL https://arxiv.org/abs/2412.10372.

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), 2018. doi: 10.1038/sdata.2018.251.

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, 2023.

Huayang Li, Yahui Liu, Hongyu Sun, Deng Cai, Leyang Cui, Wei Bi, Peilin Zhao, and Taro Watanabe. Seqpe: Transformer with sequential position encoding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–14, 2026a. ISSN 1939-3539. doi: 10.1109/TPAMI.2026.3695094.

Xiang Li, Like Li, Minglei Li, Pengfei Yan, Ting Feng, Hao Luo, Yong Zhao, and Shen Yin. Knowledge distillation and teacher–student learning in medical imaging: Comprehensive overview, pivotal role, and future directions. *Medical Image Analysis*, 107:103819, January 2026b. ISSN 1361-8415. doi: 10.1016/j.media.2025.103819.

Yong Li, Jianping Man, Yi Zhou, and Likeng Liang. Caption-augmented reasoning model with hierarchical rank lora finetuing for medical visual question answering. *Journal of Biomedical Informatics*, 172:104964, December 2025. ISSN 1532-0464. doi: 10.1016/j.jbi.2025.104964.

Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, YiQing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, Zhida Huang, and Tao Wang. Groundinggpt: Language enhanced multi-modal grounding model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6657–6678. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.360. URL http://dx.doi.org/10.18653/v1/2024.acl-long.360.

Chenyu Lian, Hong-Yu Zhou, Dongyun Liang, Jing Qin, and Liansheng Wang. Efficient medical vision-language alignment through adapting masked vision models. *IEEE Transactions on Medical Imaging*, 44(11):4499–4510, November 2025. ISSN 1558-254X. doi: 10.1109/TMI.2025.3575853.

Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, 143:102611, 2023. ISSN 0933-3657. doi: 10.1016/j.artmed.2023.102611.

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*, pp. 1650–1654. IEEE, 2021. doi: 10.1109/isbi48211.2021.9434010.

Subhankar Maity and Manob Jyoti Saikia. Large language models in healthcare and medical applications: A review. *Bioengineering*, 12(6):631, June 2025. ISSN 2306-5354. doi: 10.3390/bioengineering12060631.

Roy Miles, Ismail Elezi, and Jiankang Deng. Vkd: Improving knowledge distillation using orthogonal projections. *CoRR*, abs/2403.06213, 2024. doi: 10.48550/ARXIV.2403.06213.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-flamingo: a multimodal medical few-shot learner, 2023.

Yoojin Nam, Dong Yeong Kim, Sunggu Kyung, Jinyoung Seo, Jeong Min Song, Jimin Kwon, Jihyun Kim, Wooyoung Jo, Hyungbin Park, Jimin Sung, Sangah Park, Heeyeon Kwon, Taehee Kwon, Kanghyun Kim, and Namkug Kim. Multimodal large language models in medical imaging: Current state and future directions. *Korean Journal of Radiology*, 26(10):900, 2025. ISSN 2005-8330. doi: 10.3348/kjr.2025.0599.

Zabir Al Nazi and Wei Peng. Large language models in healthcare and medical domain: A review. *Informatics*, 11(3):57, August 2024. ISSN 2227-9709. doi: 10.3390/informatics11030057. URL http://dx.doi.org/10.3390/informatics11030057.

Gabriel Reale-Nosei, Elvira Amador-Domínguez, and Emilio Serrano. From vision to text: A comprehensive review of natural image captioning in medical diagnosis and radiology report generation. *Medical Image Analysis*, 97:103264, October 2024. ISSN 1361-8415. doi: 10.1016/j.media.2024.103264. URL http://dx.doi.org/10.1016/j.media.2024.103264.

Ji Seung Ryu, Hyunyoung Kang, Yuseong Chu, and Sejung Yang. Vision-language foundation models for medical imaging: a review of current practices and innovations. *Biomedical Engineering Letters*, 15(5):809–830, June 2025. ISSN 2093-985X. doi: 10.1007/s13534-025-00484-6.

Binesh Sadanandan and Vahid Behzadan. Consistent but dangerous: Per-sample safety classification reveals false reliability in medical vision-language models, 2026. URL https://arxiv.org/abs/2603.20985.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Fred Tung and Greg Mori. Similarity-preserving knowledge distillation. In *2019 IEEE/CVF International Conference on Computer Vision*, pp. 1365–1374. IEEE, 2019. doi: 10.1109/iccv.2019.00145.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Jian Tong, Haodong Duan, Qipeng Guo, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. Videorope: What makes for good video rotary position embedding?, 2025.

Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. Towards visual grounding: A survey, 2025. URL https://arxiv.org/abs/2412.20206.

Cheng Xue, Lequan Yu, Pengfei Chen, Qi Dou, and Pheng-Ann Heng. Robust medical image classification from noisy labeled data with global and local representation guided co-training. *IEEE Transactions on Medical Imaging*, 41(6):1371–1382, June 2022. ISSN 1558-254X. doi: 10.1109/TMI.2021.3140140. URL http://dx.doi.org/10.1109/TMI.2021.3140140.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity, 2023. URL https://arxiv.org/abs/2310.07704.

Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, and Yinfei Yang. Ferret-v2: An improved baseline for referring and grounding with large language models, 2024a. URL https://arxiv.org/abs/2404.07973.

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, 2024b. doi: 10.1038/s41591-024-03185-2.

Liang Zhao, Xiachong Feng, Xiaocheng Feng, Weihong Zhong, Dongliang Xu, Qing Yang, Hongtao Liu, Bing Qin, and Ting Liu. Length extrapolation of transformers: A survey from the perspective of positional encoding, 2024. URL https://arxiv.org/abs/2312.17044.

Zihao Zhao, Yuxiao Liu, Han Wu, Mei Wang, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Zhiming Cui, Qian Wang, and Dinggang Shen. Clip in medical imaging: A survey, 2025. URL https://arxiv.org/abs/2312.07353.

Ke Zou, Yang Bai, Bo Liu, Yidi Chen, Zhihao Chen, Yang Zhou, Xuedong Yuan, Meng Wang, Xiaojing Shen, Xiaochun Cao, Yih Chung Tham, and Huazhu Fu. Uncertainty-aware medical diagnostic phrase identification and grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12):11315–11329, December 2025. ISSN 1939-3539. doi: 10.1109/TPAMI.2025.3596878.

A Data Interfaces and Implementation Settings

A.1 Datasets and task interfaces

Med-GRIT supplies the four-task benchmark used in this study (Huang et al., 2024). Adaptation uses its 240k training portion; evaluation uses Med-GRIT-Test30k. The eight modalities are CT, MR, X-ray, PET, endoscopy, dermoscopy, fundus, and ultrasound. Pseudo-label construction, reflective correction, and parameter selection use training examples only. External MIA evaluation uses the 193 questions associated with 50 images in LLaVA-Med-qa0.2k (Li et al., 2023).

Table A1: Task interfaces. Every query is accompanied by the image.

Task	Query information	Expected response
VG	Textual reference to a finding	Bounding box for the referred finding
ROC	Specified image region	Category of the finding in that region
RC	Specified image region	Textual description of that region
MIA	Question about the medical image	Textual answer to the question

A.2 Conversation representation

The data adapter uses ShareGPT records with an `images` list and a `messages` sequence. Each message contains `role` and `content`; dataset registration identifies the roles as `user` and `assistant`. A source image identifier is converted to its `.png` path. The box-delimiter replacement preserves the enclosed coordinate text while converting the opening delimiter to `<|box_start|>` and the closing delimiter to `<|box_end|>`. These fields specify the stored interface. Task instructions and targets remain in message content.

A.3 Settings by stage

Table A2 collects the settings reported in the main paper. The student initialization is Qwen2-VL-7B-Instruct (Wang et al., 2024). The visual tower remains frozen during language-side adaptation.

Table A2: Reported configuration. Unrecorded run parameters are not inferred from generic training defaults.

Stage	Setting
Images and positions	Square image grid; MCG-RoPE boundary gap $g = 2$; no rescaling of the internal visual grid.
Distillation	$\alpha = 0.5$, $\beta = 0.6$, $\gamma = 0.4$, $T = 0.5$; linear admission schedule from $\tau_0 = 0.9$ to $\tau_{\min} = 0.4$.
Supervised adaptation	Two SFT epochs after distillation.
Optimization and hardware	AdamW; cosine schedule; learning rate 2×10^{-5}; six NVIDIA A100 40GB GPUs.
Reflective correction	Text similarity threshold 0.8; reference-constrained GPT-4o text correction; coordinates unchanged.
Decoding	Beam size three without reranking; SASG is optional and has a separate candidate-selection stage.

The experiment archive does not preserve the teacher checkpoint, random seeds, effective batch size, exact image resolution, correction prompt version, or complete SASG configuration. The table therefore summarizes recorded settings without claiming exact reproducibility.

B Position Indices and Auxiliary Stages

B.1 Boundary-gap construction

Let t_ℓ be the last text position before an image and let $v_j^{(h)}, v_j^{(w)}$ be the native height and width indices of visual token j. For each spatial axis $a \in \{h, w\}$, the translated index is

$$v_j'^{(a)} = v_j^{(a)} + \delta_a, \qquad \delta_a = t_\ell + g - v_1^{(a)}, \qquad g = 2. \tag{A1}$$

The first visual position is consequently $t_\ell + g$ on both axes. After the visual segment, text resumes at $m + g$, where $m = \max_j \max\{v_j'^{(h)}, v_j'^{(w)}\}$, and subsequent text positions increase by one. The construction translates the visual grid; it does not scale its internal coordinates or add learned position parameters. The spatial notation follows the main text; the backbone also maintains a temporal stream.

Relative rotations within and across segments. For a single rotary frequency ω, let $R(\omega x)$ denote the rotation at position x. The rotary contribution to an attention score satisfies

$$(R(\omega a)q)^\top (R(\omega b)k) = q^\top R\big(\omega(b-a)\big)k. \tag{A2}$$

If two positions belong to the same translated segment, $a' = a + c$ and $b' = b + c$, so $b' - a' = b - a$. Their relative rotation is preserved. For positions in different segments, with offsets c_a and c_b,

$$b' - a' = (b - a) + (c_b - c_a). \tag{A3}$$

This includes text tokens on opposite sides of an image. Thus the invariant applies within a segment, not to every text–text pair in a packed sequence. The result describes the position construction and does not establish semantic alignment or clinical correctness.

B.2 Reflective correction and selective generation

Reflective correction applies when the cosine similarity between a generated text and its reference is below 0.8. The correction model receives both texts and revises the generated wording subject to the reference intent. Bounding-box coordinates remain unchanged. This stage therefore augments textual targets; it is not a procedure for repairing localization annotations.

At inference, SASG selects a response from a candidate set using image–text similarity. Algorithm A1 states the selection operation for an arbitrary candidate budget K and fixed generation configuration d. Neither is inferred from the non-reranked beam size. The archived results do not preserve K, d, or the exact scoring checkpoint, so the algorithm is a description of the operation rather than a complete runtime configuration.

Algorithm A1 Semantic-Aware Selective Generation

Require: Image I, query Q, model M, candidate budget $K \geq 1$, generation configuration d, image–text scorer s

1: Generate candidates $y_1, \ldots, y_K$ using $M(I, Q)$ and d
2: **for** $i = 1, \ldots, K$ **do**
3: Compute $z_i \leftarrow s(I, y_i)$
4: **end for**
5: Choose $i^* \in \arg\max_{i \in \{1,\ldots,K\}} z_i$
6: **return** y_{i^*}

When candidates tie, a fixed tie rule is required for reproducible execution; the historical tie rule is not recorded. The component comparison in the main paper reports the available SASG results; this specification adds no new result.

C Extended Results by Imaging Modality

Tables A3–A6 provide the archived modality-level point estimates for the full model and intermediate component configurations. Entries shared with the main-text modality table agree. These tables add no repeated-run uncertainty estimates. The highest score in each column is bold, including ties.

A row beginning with + denotes the indicated component added to BiRD; ClinKD denotes the full pipeline. Endo., Derm., and US denote endoscopy, dermoscopy, and ultrasound. A dash indicates that no ROC score was reported for that modality.

Method	CT	MR	X-ray	PET	Endo.	Derm.	Fundus	US
BiRD	44.47	29.26	41.73	56.46	53.60	75.63	84.15	46.04
+ RoPE-Mixed	46.90	**53.11**	42.03	60.84	53.49	78.25	85.16	52.83
+ MCG-RoPE	51.22	49.40	45.92	68.60	**60.37**	88.28	**86.59**	61.88
+ $V_k D$	58.77	50.95	45.69	65.89	57.63	93.32	79.27	59.90
+ Pseudo-KD	62.59	52.86	47.19	66.28	59.84	93.68	85.37	62.87
ClinKD	**65.52**	52.23	**48.56**	**69.25**	**60.37**	**94.22**	**86.59**	**64.47**

Table A3: Visual grounding (VG): Recall@0.5 on Med-GRIT-Test30k, by imaging modality.

Method	CT	MR	X-ray	PET	Endo.	Derm.	Fundus	US
BiRD	34.76	61.79	53.74	–	60.40	96.61	–	84.65
+ RoPE-Mixed	39.78	62.13	74.27	–	61.33	96.61	–	84.65
+ MCG-RoPE	44.47	64.20	83.75	–	62.19	96.61	–	85.22
+ $V_k D$	53.19	69.10	90.84	–	61.71	**100.00**	–	81.15
+ Pseudo-KD	58.47	69.92	95.83	–	66.32	**100.00**	–	86.58
ClinKD	**61.42**	**71.30**	**97.00**	–	**68.84**	**100.00**	–	**95.54**

Table A4: Referring object classification (ROC): Recall on Med-GRIT-Test30k, by imaging modality.

Method	CT	MR	X-ray	PET	Endo.	Derm.	Fundus	US
BiRD	41.88	51.69	37.39	47.95	54.07	77.44	48.73	**82.65**
+ RoPE-Mixed	43.51	51.88	54.61	52.83	54.07	68.15	53.99	65.28
+ MCG-RoPE	50.73	53.11	60.33	70.16	54.07	70.18	61.52	70.46
+ $V_k D$	51.33	54.39	62.83	74.31	54.07	76.23	67.89	72.15
+ Pseudo-KD	52.78	**58.41**	70.39	75.85	**55.22**	81.34	71.54	72.15
ClinKD	**53.94**	**58.41**	**71.47**	**79.91**	52.10	**83.92**	**84.51**	80.19

Table A5: Referring captioning (RC): SPICE on Med-GRIT-Test30k, by imaging modality.

Method	CT	MR	X-ray	PET	Endo.	Derm.	Fundus	US
BiRD	47.01	49.35	37.17	57.15	39.91	72.13	48.87	65.78
+ RoPE-Mixed	57.10	50.89	42.55	**69.28**	41.63	73.32	59.63	61.80
+ MCG-RoPE	60.53	53.77	**54.10**	62.01	43.88	75.33	62.89	64.53
+ $V_k D$	62.33	53.77	52.27	62.01	47.29	75.33	68.04	70.31
+ Pseudo-KD	**66.85**	55.73	52.27	64.74	**50.58**	78.21	**70.02**	**76.16**
ClinKD	**66.85**	**56.82**	54.01	67.08	**50.58**	**81.00**	**70.02**	76.05

Table A6: Medical image analysis (MIA): mBMR on Med-GRIT-Test30k, by imaging modality.