CoTBox-TTT: Grounding Medical VQA with Visual Chain-of-Thought Boxes During Test-time Training

Jiahe Qian[1*] Yuhao Shen[2*] Zhangtianyi Chen[2] Juexiao Zhou[2†] Peisong Wang[1]
[1]Institute of Automation, Chinese Academy of Sciences
[2]School of Data Science, The Chinese University of Hong Kong, Shenzhen

Abstract

Medical visual question answering could support clinical decision making, yet current systems often fail under domain shift and produce answers that are weakly grounded in image evidence. This reliability gap arises when models attend to spurious regions and when retraining or additional labels are impractical at deployment time. We address this setting with CoTBox-TTT, an evidence-first test-time training approach that adapts a vision–language model at inference while keeping all backbones frozen. The method updates only a small set of continuous soft prompts. It identifies question-relevant regions through a visual chain-of-thought signal and encourages answer consistency across the original image and a localized crop. The procedure is label free, and plug and play with diverse backbones. Experiments on medical VQA show that the approach is practical for real deployments. For instance, adding CoTBox-TTT to LLaVA increases closed-ended accuracy by 12.3% on pathVQA.

1. Introduction

Medical visual question answering plays a growing role in clinical decision support and in building trust through transparent interactions [2, 21, 32, 39, 44]. Real deployments face substantial domain shift across institutions, devices, acquisition protocols, and patient populations [23]. Tasks are open ended and answers are generated in natural language [10, 18, 26, 51]. Recent vision language models perform well in distribution but their answers often degrade under shift [15, 48, 49]. Models may attend to spurious regions and the generated text can drift or hallucinate [19, 24, 47, 52]. There is a strong need for an approach that adapts a model to each test case without extra labels and without costly retraining [35]. Our goal is to improve answer accuracy, stability, and interpretability at the moment of inference.

Existing strategies are limited. Conventional fine tuning requires supervision and multiple training rounds. Popular test time adaptation methods target discriminative classification and do not directly control the two prerequisites of reliable generation in medical VQA [4, 12, 14, 29, 34, 40, 41, 45, 46, 50]. A model must first attend to the correct visual evidence [6, 30, 33] and then produce an answer that remains consistent when the visual context changes slightly. Directly aligning answers without evidence constraints can reinforce self consistent yet incorrect outputs [1, 7]. These gaps motivate an evidence first design for test-time training that remains label free and low parameter [20, 22, 28, 53].

Our central intuition is that a strong answer should be grounded in explicit visual evidence and should remain consistent across complementary views of that evidence. We operationalize visual chain of thought as a bounding box that marks the region most helpful for answering the question. For each test case the model first predicts a box on the original image and then validates this evidence on a cropped view of the same region. We enforce self consistency of the two boxes. This anchors attention on the question relevant area and suppresses reliance on spurious cues. We then encourage answer consistency between the original view and the cropped view. The student distribution is guided by an exponential moving average teacher through sequence-level text alignment. The design is effective because it converts correct evidence selection into a direct supervisory signal on generation, which reduces hallucination and improves stability. The entire procedure uses only a small set of continuous soft prompts that are updated at test time.

Our approach is model agnostic and plug and play. The test time training head attaches to diverse vision language backbones without changing encoders or projection layers. The parameter and compute overhead are small which supports near line or online adaptation in clinical workflows. Across multiple medical VQA benchmarks and cross domain settings the method consistently improves accuracy and stability while producing an interpretable evidence trail

*Equal contribution.
†Corresponding author: juexiao.zhou@gmail.com.

that links each answer to an explicit region. In summary our work makes three contributions.

- An evidence driven test time training mechanism that enforces self consistency of visual chain of thought bounding boxes.
- A cross view answer consistency objective for generative vision language models guided by an exponential moving average teacher.
- A parameter efficient head that integrates with a wide range of models and requires no reinforcement learning.

2. Related Work

2.1. Medical VQA

Medical visual question answering has progressed from small single-domain settings toward open-ended and clinically oriented scenarios. PMC-VQA establishes large-scale visual instruction data for medical VQA and highlights brittleness in free-form generation under distribution shift [51]. Biomedical vision–language models such as LLaVA-Med show that instruction tuning can induce strong priors for clinical images, yet hallucination and evidence drift remain open challenges in deployment [21, 47]. Region-grounded pipelines bring explicit localization into the loop. R-LLaVA injects region-of-interest supervision to better connect textual rationales with image evidence [6]. VividMed broadens visual grounding operators for fine-grained medical dialogue and question answering [30]. Grounded evaluation protocols that require localizing before answering help disentangle evidence selection from language generation and diagnose evidence–answer mismatch [33]. Clinician-facing studies further suggest that carefully grounded and calibrated systems can assist expert-level medical question answering [39].

2.2. Test-Time Training

Test-time adaptation and training aim to mitigate distribution shift without labeled target data. [41] introduced a self-supervised auxiliary objective that is optimized on each test sample to improve robustness under shift. Entropy minimization then framed fully test-time adaptation as confidence-driven online optimization and demonstrated substantial gains on corrupted and shifted benchmarks [45]. Subsequent analyses clarified failure modes and proposed feature alignment strategies that stabilize adaptation across diverse shifts and batch regimes [29]. Recent studies have begun to explore test-time procedures for vision–language models with efficiency in mind, including dynamic adapters and prompt-tuning variants that preserve zero-shot competence while reducing computation [15, 53]. Together these developments provide a foundation for test-time procedures in multimodal systems and motivate designs that respect the requirements of generation under clinical distribution shift.

3. Method

This section is organized as follows. We first present an overview of CoTBox-TTT that defines the test-time training setting in Section 3.1. We then detail self consistency by evidence localization with the grounding model $\mathcal{G}$ and the validated box loss in Section 3.2. Next we describe cross-view answer consistency with the EMA teacher for $\mathcal{F}$ and the language modeling objective across views in Section 3.3. We formulate the overall objective and the two step optimization schedule in Section 3.4. Finally we provide implementation details including model choices, prompt sizes, initialization, learning rates, and the update schedule in Section 3.5.

3.1. Overview

CoTBox-TTT operates in a test-time training setting. Given an image-question pair (I, q), all backbone, encoder, and projection parameters remain frozen. Adaptation is restricted to two small sets of continuous soft prompts. The first set P_{vis} conditions a pre-trained multi-modality grounding model $\mathcal{G}$ that emits question-relevant bounding boxes under a fixed JSON schema. The second set P_{ans} conditions a vision-language model $\mathcal{F}$ that generates textual answers. An exponential moving average copy $\overline{P}_{\text{ans}}$ is maintained as a teacher and does not receive gradients.

As shown in Figure 1, the procedure contains two stages that share the same frozen backbones. The evidence stage runs $\mathcal{G}$ twice with shared prompts P_{vis}. It first predicts a box on the original image and then validates the evidence on the cropped view. A single language modeling loss on the validated box string updates P_{vis}. The answer stage runs $\mathcal{F}$ on the original image and on the crop. A single language modeling loss aligns student outputs to teacher sequences across the two views and updates P_{ans}. The teacher prompts $\overline{P}_{\text{ans}}$ are refreshed by an exponential moving average. Bounding boxes follow a strict schema and the crop is a deterministic function of the predicted coordinates. The design is fully label free.

3.2. Self-consistency by Evidence Localization

Given (I, q), the grounding model $\mathcal{G}$ is conditioned by the soft prompts P_{vis}. The two passes share the same P_{vis} and all other parameters are frozen. The model outputs a bounding box with absolute pixel coordinates under a fixed JSON schema. The schema is written as $\{$“bbox” : $[x_1, y_1, x_2, y_2]\}$. Let T_b denote the fixed sequence length under the shared template.

The first pass predicts on the original image

$$b_1 = \mathcal{G}(I, q; P_{\text{vis}}), \tag{1}$$

where $b_1 = [x_1, y_1, x_2, y_2]$ are integer pixel coordinates within image bounds. The crop for the second pass is com-

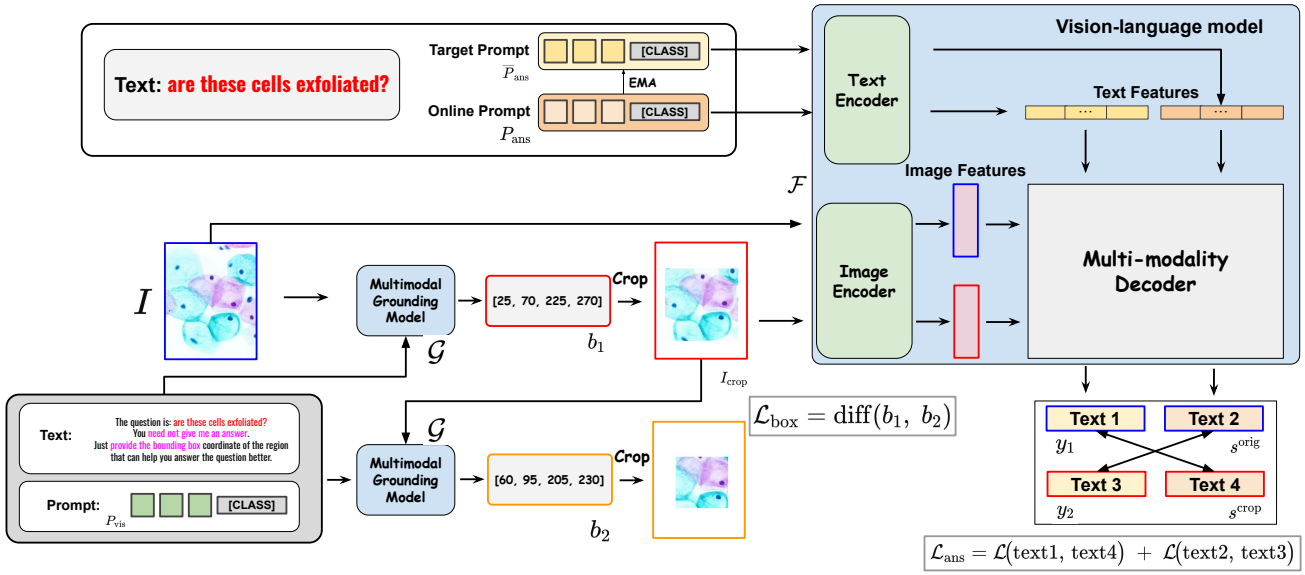

Figure 1. Overview of CoTBox-TTT. (a) Evidence localization: a grounding model conditioned on visual prompts predicts a bounding box on the original image, crops a localized view, and performs a second pass to validate the evidence. (b) Answer consistency: a vision–language model generates student answers on the original and cropped views while an EMA teacher provides targets on the same two views, and the student is aligned to the teacher across views. (c) Test-time adaptation updates only small soft prompts and keeps encoders and decoders frozen, yielding an interpretable evidence trail and consistent gains across backbones.

puted by a deterministic operator

$$I_{\text{crop}} = \text{Crop}(I, b_1), \tag{2}$$

where $\text{Crop}(\cdot)$ clips coordinates to the valid pixel range of I, extracts the rectangular region, pads the result back to the original image size using white pixels, and resizes the padded image to the input resolution of $\mathcal{G}$. The second pass validates the evidence on the crop

$$b_2 = \mathcal{G}(I_{\text{crop}}, q; P_{\text{vis}}), \tag{3}$$

where $b_2 = [x_1', y_1', x_2', y_2']$ follows the same schema. Let $P_{\mathcal{G}}(\cdot \mid \cdot)$ be the next token conditional probability defined by $\mathcal{G}$ under P_{vis} with teacher forcing on the prefix of the string b_2. CoTBox-TTT uses a single language modeling loss

$$\mathcal{L}_{\text{box}} = -\frac{1}{T_b} \sum_{t=1}^{T_b} \log P_{\mathcal{G}}(b_{2,t} \mid b_{2,<t}, I, q; P_{\text{vis}}), \tag{4}$$

where $b_{2,t}$ is the target token at position t in the string b_2 and $b_{2,<t}$ is its prefix.

3.3. Cross-view Consistency with EMA Teacher

Given (I, q) and the crop I_{crop}, the answer model $\mathcal{F}$ is conditioned by the soft prompts P_{ans}. The network weights are shared across branches and remain frozen. Let $\overline{P}_{\text{ans}}$ be an exponential moving average copy of P_{ans} that serves as a teacher. At the beginning of an adaptation episode the teacher is initialized to the current student prompts.

Student outputs are generated on both views

$$\begin{aligned} y_1 &= \mathcal{F}(I, q; P_{\text{ans}}), \\ y_2 &= \mathcal{F}(I_{\text{crop}}, q; P_{\text{ans}}), \end{aligned} \tag{5}$$

where y_1 and y_2 are token sequences produced by teacher forced decoding on the corresponding inputs. Teacher sequences are produced with the EMA prompts

$$\begin{aligned} s^{\text{orig}} &= \mathcal{F}(I, q; \overline{P}_{\text{ans}}), \\ s^{\text{crop}} &= \mathcal{F}(I_{\text{crop}}, q; \overline{P}_{\text{ans}}), \end{aligned} \tag{6}$$

where the lengths are T_{orig} and T_{crop}. Let $P_{\mathcal{F}}(\cdot \mid \cdot)$ denote the next–token conditional probability defined by $\mathcal{F}$ under the given prompts. We compute the cross–view objective as the mean of two view–specific language modeling losses. The first loss supervises the student with the teacher sequence on the original view

$$\mathcal{L}_{\text{ans}}^{\text{orig}} = -\frac{1}{T_{\text{orig}}} \sum_{t=1}^{T_{\text{orig}}} \log P_{\mathcal{F}}\left(s_t^{\text{orig}} \mid s_{<t}^{\text{orig}}, I, q; P_{\text{ans}}\right), \tag{7}$$

where s_t^{orig} is the teacher token at position t and $s_{<t}^{\text{orig}}$ is its prefix. The second loss supervises the student with the

teacher sequence on the cropped view

$$\mathcal{L}_{\text{ans}}^{\text{crop}} = -\frac{1}{T_{\text{crop}}} \sum_{t=1}^{T_{\text{crop}}} \log P_{\mathcal{F}}(s_t^{\text{crop}} \mid s_{<t}^{\text{crop}},\, I_{\text{crop}},\, q;\, P_{\text{ans}}), \tag{8}$$

where s_t^{crop} is the teacher token at position t and $s_{<t}^{\text{crop}}$ is its prefix.

The overall cross–view objective is the sum of the two view–specific losses

$$\mathcal{L}_{\text{ans}} = \mathcal{L}_{\text{ans}}^{\text{orig}} + \mathcal{L}_{\text{ans}}^{\text{crop}}. \tag{9}$$

3.4. Overall Objective and Algorithmic Scheme

This subsection specifies the joint objective used by CoTBox-TTT and how it is applied during test-time adaptation. The procedure is shown in Algorithm 1

The total objective is written as

$$\mathcal{L}_{\text{total}} = \alpha\, \mathcal{L}_{\text{box}} + \beta\, \mathcal{L}_{\text{ans}}, \tag{10}$$

where $\alpha > 0$ and $\beta > 0$ are scalar weights. In practice CoTBox-TTT applies a two-step schedule that decouples the updates.

During the evidence step the prompts P_{vis} are updated by minimizing the evidence loss while all other parameters are frozen

$$\min_{P_{\text{vis}}} \mathcal{L}_{\text{box}}(I, q; \mathcal{G}, P_{\text{vis}}). \tag{11}$$

After the update the crop $I_{\text{crop}} = \text{Crop}(I, b_1)$ is fixed for the next step

$$(\alpha, \beta) = (1, 0). \tag{12}$$

The gradient of $\mathcal{L}_{\text{box}}$ with respect to P_{ans} is zero by construction

$$\frac{\partial \mathcal{L}_{\text{box}}}{\partial P_{\text{ans}}} = 0. \tag{13}$$

During the answer step the prompts P_{ans} are updated by minimizing the cross-view loss while all other parameters are frozen

$$\min_{P_{\text{ans}}} \mathcal{L}_{\text{ans}}(I, I_{\text{crop}}, q; \mathcal{F}, P_{\text{ans}}, \overline{P}_{\text{ans}}). \tag{14}$$

The schedule sets

$$(\alpha, \beta) = (0, 1). \tag{15}$$

No gradient flows from the answer loss to the evidence prompts

$$\frac{\partial \mathcal{L}_{\text{ans}}}{\partial P_{\text{vis}}} = 0. \tag{16}$$

For completeness the per-step updates follow standard first-order optimization on the prompts. The evidence update is

$$P_{\text{vis}} \leftarrow P_{\text{vis}} - \eta_{\text{vis}}\, \nabla_{P_{\text{vis}}} \mathcal{L}_{\text{box}}, \tag{17}$$

Algorithm 1 CoTBox-TTT procedure

Require: Image–question (I, q), grounding model $\mathcal{G}$, answer model $\mathcal{F}$, prompts $P_{\text{vis}}, P_{\text{ans}}$, EMA decay β, mini epochs E

$\overline{P}_{\text{ans}} \leftarrow P_{\text{ans}}$
for $e = 1$ to E **do**
  $b_1 \leftarrow \mathcal{G}(I, q; P_{\text{vis}})$
  $I_{\text{crop}} \leftarrow \text{Crop}(I, b_1)$
  $b_2 \leftarrow \mathcal{G}(I_{\text{crop}}, q; P_{\text{vis}})$
  $P_{\text{vis}} \leftarrow \text{SGD}(P_{\text{vis}}, \nabla \mathcal{L}_{\text{box}}(b_2 \mid I, q))$
  $s^{\text{orig}} \leftarrow \mathcal{F}(I, q; \overline{P}_{\text{ans}})$
  $s^{\text{crop}} \leftarrow \mathcal{F}(I_{\text{crop}}, q; \overline{P}_{\text{ans}})$
  $P_{\text{ans}} \leftarrow \text{SGD}\big(P_{\text{ans}}, \nabla \mathcal{L}_{\text{ans}}(s^{\text{orig}}, s^{\text{crop}} \mid I, I_{\text{crop}}, q)\big)$
  $\overline{P}_{\text{ans}} \leftarrow \beta\, \overline{P}_{\text{ans}} + (1 - \beta)\, P_{\text{ans}}$
return final answer $\mathcal{F}(I, q; P_{\text{ans}})$ and box b_1

where $\eta_{\text{vis}} > 0$ is the learning rate. The answer update is

$$P_{\text{ans}} \leftarrow P_{\text{ans}} - \eta_{\text{ans}}\, \nabla_{P_{\text{ans}}} \mathcal{L}_{\text{ans}}, \tag{18}$$

where $\eta_{\text{ans}} > 0$ is the learning rate. The teacher prompts are refreshed by an exponential moving average

$$\overline{P}_{\text{ans}} \leftarrow \beta\, \overline{P}_{\text{ans}} + (1 - \beta)\, P_{\text{ans}}, \tag{19}$$

where $\beta \in (0, 1)$ is the decay factor. This schedule yields a total objective that is simple to state yet applied in two sequential minimizations with disjoint gradient flow.

3.5. Implementation Details

The grounding model is VisCoT [38] with weights frozen. The answer model, such as LLaVA-Med, keeps its visual encoder, projection layers, and language backbone frozen. Adaptation uses two continuous soft prompts inserted as prefix embeddings and initialized to zero, with 24 tokens for evidence and 32 tokens for answers. The tokenizer and vocabulary are shared across original and cropped views. The maximum answer length is 128 tokens and the tokenized evidence string uses a fixed padded length of 32 tokens for stable loss scaling. Teacher decoding is deterministic with unit temperature and greedy selection. Learning rates are 1e-3 for evidence prompts and 5e-4 for answer prompts. The exponential moving average teacher uses a decay of 0.9 and is synchronized to the student at the start of each adaptation episode. Training follows a two step schedule with 20 epochs per image, where each mini epoch first updates evidence prompts with the evidence loss and then updates answer prompts with the cross view loss. Computation uses eight RTX 4090 GPUs. Other hyperparameters follow backbone defaults and remain unchanged during test time training.

4. Experiments

4.1. Settings

Baselines

- *LLaVA [27].* A general-purpose vision-language assistant trained by visual instruction tuning that connects a CLIP-based image encoder with a large language model. The baseline follows the official inference recipe including image resolution, decoding temperature, maximum answer length, and tokenization.
- *LLaVA-Med [21].* A biomedical adaptation of LLaVA trained with a two-stage curriculum. Stage one aligns biomedical concepts using large-scale PubMed Central figure–caption pairs. Stage two performs instruction tuning on GPT-4 generated multimodal conversations. The public release includes checkpoints and evaluation scripts for three medical VQA benchmarks.
- *Hulu-Med [13].* A transparent generalist medical vision-language model that unifies text, 2D images, 3D volumes, and videos through a patch-based vision encoder and an LLM decoder. The official resources provide inference pipelines across 30 medical benchmarks. For comparability here only the 2D VQA setting is considered.

Datasets and Tasks

- *VQA-RAD [18].* A clinician-authored radiology VQA benchmark with 315 images and 3,515 question–answer pairs. Questions cover close-ended types such as yes–no and open-ended free-form answers. The standard split and task protocol are used to ensure consistency with prior medical VQA practice.
- *SLAKE [26].* A bilingual medical VQA dataset with physician-verified annotations, comprehensive semantic labels, and an associated medical knowledge base. Images span multiple radiology modalities and body regions. The English subset is typically adopted for head-to-head comparisons and both open-ended and close-ended settings are reported.
- *PathVQA [10].* A pathology VQA dataset with 4,998 images and 32,799 questions collected from textbooks and a public digital library. The dataset mixes open-ended questions and close-ended yes–no items. The commonly used evaluation split is followed and questions are grouped under the open-ended and close-ended protocols.

Evaluation Metrics

- *Close-ended.* Accuracy is computed by exact match against the reference answer set for yes–no and other close-ended types.
- *Open-ended.* Recall is computed for free-form generation by checking whether gold keywords appear in the generated answer. Main tables are reported in recall to match established protocols. F1 is additionally provided in the appendix when applicable.

For each baseline model results are reported under two conditions on all datasets and task types. The first condition is the native inference without CoTBox-TTT. The second condition augments inference with CoTBox-TTT while keeping preprocessing and decoding identical. This design isolates the contribution of CoTBox-TTT under the open-ended and close-ended metrics on VQA-RAD, SLAKE and PathVQA. The data statistics are provided in Table 1.

Dataset	VQA-RAD		SLAKE			PathVQA		
	Train	Test	Train	Val	Test	Train	Val	Test
# Images	313	203	450	96	96	2,599	858	858
# QA Pairs	1,797	451	4,919	1,053	1,061	19,755	6,279	6,761
# Open	770	179	2,976	631	645	9,949	3,144	3,370
# Closed	1,027	272	1,943	422	416	9,806	3,135	3,391

Table 1. Dataset statistics for medical VQA datasets.

4.2. Main Results

We compare each backbone in two conditions that differ only by the presence of CoTBox–TTT. Decoding temperature, maximum answer length, preprocessing, and scoring are matched across conditions. Performance is reported on VQA–RAD, SLAKE, and PathVQA under open–ended recall and close–ended accuracy.

CoTBox–TTT improves every metric entry in Table 2. All forty eight backbone–dataset–metric cells show positive shifts. This holds for generic instruction models and for medically adapted models and for generalist medical models. The gains are largest where open–domain generation is weak at baseline and remain consistently positive on strong backbones. For example LLaVA shows a mean open recall improvement of 11.05 points and a mean closed accuracy improvement of 9.21 points across the three datasets. LLaVA–Med with a CLIP vision encoder improves open recall by 4.75 on VQA–RAD, 4.41 on SLAKE, and 2.68 on PathVQA, with closed accuracy gains of 2.94, 4.32, and 1.50. Hulu–Med–14B improves open recall by 4.73 on VQA–RAD, 3.31 on SLAKE, and 4.71 on PathVQA, and improves closed accuracy by 5.89, 3.12, and 2.59.

Closed accuracy increases are consistent on all datasets and backbones, which indicates that cross–view alignment stabilizes classification–style questions. On VQA–RAD LLaVA closed accuracy rises from 65.07 to 73.16 and Hulu–Med–14B rises from 82.35 to 88.24. On SLAKE LLaVA closed accuracy rises from 63.22 to 70.43 and LLaVA–Med–Vicuna rises from 83.17 to 87.26. On PathVQA LLaVA closed accuracy rises from 63.20 to 75.52 and LLaVA–Med–BioMedCLIP rises from 91.09 to 93.02.

Open–ended recall also improves across the board, which shows that the evidence–driven procedure strengthens free–form generation. The largest shift appears on PathVQA for LLaVA where open recall rises

from 7.74 to 32.60. Strong models benefit as well. Hulu–Med–32B on SLAKE rises from 85.06 to 90.26 and LLaVA–Med–BioMedCLIP on SLAKE rises from 87.11 to 90.24. On VQA–RAD the open recall gains are 5.44 for LLaVA and 3.41 to 4.75 across LLaVA–Med variants and 2.26 to 5.73 across Hulu–Med variants.

Taken together these results indicate that CoTBox–TTT acts as a model–agnostic adapter that consistently improves answer reliability. It improves closed accuracy by stabilizing cross–view predictions and improves open recall by grounding free–form generation on explicit visual evidence.

Figure 2 shows some examples. In the first example the baseline answers are vague and drift toward nonspecific mucosal changes, whereas CoTBox-TTT guides the model to the dysplastic epithelial strip along the mucosal edge and produces the correct diagnosis of oral epithelial dysplasia. The bounding box constrains attention to the atypical epithelium with nuclear crowding and loss of maturation and the crop suppresses distracting keratin debris in the surrounding field. In the second example the PAS stained renal biopsy contains a thickened arteriole with a glassy eosinophilic wall. The baseline mixes interstitial findings with vascular changes and fails to name the vascular lesion, while CoTBox-TTT places the box on the arteriole, attends to the concentric hyaline deposition, and outputs severe hyaline arteriolosclerosis. Across both cases the evidence-guided crop and the cross-view alignment reduce spurious focus and convert ambiguous descriptions into specific clinicopathologic terms that match the ground truth.

4.3. Ablation Study

We study four settings on VQA–RAD for each backbone. Row one enables both components and is the full CoTBox–TTT. Row two enables evidence consistency and disables the EMA teacher where the teacher parameters are tied to the student. Row three enables the EMA teacher and replaces the two pass localization with a single pass grounding so there is no second pass to enforce evidence consistency. Row four disables both components and is the native model without test time training.

Both modules contribute and they do so in complementary ways. On LLaVA the full method improves from 50.00 to 55.44 for open and from 65.07 to 73.16 for closed. The evidence only setting reaches 51.77 and 68.01 which confirms that consistent evidence improves both recall and accuracy. The EMA only setting reaches 53.20 and 70.59 which shows that cross view guidance improves both metrics even when localization uses a single pass. On LLaVA–Med–CLIP the full method reaches 66.27 and 87.13 while evidence only reaches 64.30 and 86.03 and EMA only reaches 64.46 and 85.66. On Hulu–Med–7B the full method reaches 69.46 and 90.07 while evidence only reaches 65.52 and 88.24 and EMA only reaches 65.85 and 88.97.

Taken together these comparisons indicate that the evidence module strengthens grounding which translates into higher recall and accuracy and that the EMA teacher stabilizes answer alignment which further improves both metrics. The full configuration outperforms either single component on LLaVA by 2.24 open and 2.57 closed over EMA only and on Hulu–Med–7B by 3.61 open and 1.10 closed over EMA only. Similar margins appear on LLaVA–Med–CLIP. These patterns support the view that evidence consistency and EMA guidance address complementary error modes and that combining them yields additive gains.

5. Discussion

CoTBox-TTT is a test-time training framework for medical-VQA that is model-agnostic and plug-and-play. The method freezes all backbones and adapts only small sets of soft prompts. It turns explicit visual evidence and aligns answers across the original image and a localized crop with a teacher based on exponential moving average. Across three datasets and multiple backbones it improves accuracy while providing an interpretable evidence trail that links each answer to a concrete region.

Building on these gains, we see three natural directions that follow the same principles of evidence grounding, cross-view agreement, and parameter-efficient adaptation. The first direction is to replace box consistency with mask consistency. Pixel level masks capture shape and boundary and can represent multi focal or diffuse patterns that a rectangle cannot, as shown by strong medical segmentation baselines [11, 37]. Consistency can be defined on masks predicted in the original view and in the localized view and mapped by the same crop and padding operator, while promptable segmenters provide a practical mechanism for conditioning the evidence on the question [17, 31]. Weakly supervised consistency objectives can be transferred to the mask domain to reduce annotation demands and to stabilize updates with very few adaptation steps [5]. These ingredients can be combined within the current two stage schedule so that the update remains focused on soft prompts.

A second direction is to introduce adaptive momentum for the teacher. A fixed decay can be too rigid when the student changes at different speeds across cases. Mean teacher style targets have proven effective for stabilizing semi supervised training and suggest that teacher dynamics matter for generalization [43]. Momentum encoders in self supervised learning also highlight the importance of well tuned temporal smoothing and temperature schedules [3, 9]. Recent analyses of diffusion training further show that revisiting or scheduling EMA can improve stability without increasing annotation cost [16]. Adapting these insights to test time training suggests scheduling momentum across

Model	CoTBox–TTT	VQA–RAD		SLAKE		PathVQA	
		Open ↑	Closed ↑	Open ↑	Closed ↑	Open ↑	Closed ↑
LLaVA		50.00	65.07	78.18	63.22	7.74	63.20
	✓	**55.44**	**73.16**	**81.04**	**70.43**	**32.60**	**75.52**
LLaVA–Med–CLIP		61.52	84.19	83.08	85.34	37.95	91.21
	✓	**66.27**	**87.13**	**87.49**	**89.66**	**40.63**	**92.71**
LLaVA–Med–Vicuna		64.39	81.98	84.71	83.17	38.87	91.65
	✓	**68.25**	**86.40**	**88.12**	**87.26**	**42.90**	**93.03**
LLaVA–Med–BioMedCLIP		64.75	83.09	87.11	86.78	39.60	91.09
	✓	**68.16**	**87.13**	**90.24**	**90.39**	**42.73**	**93.02**
Hulu–Med–7B		63.73	87.50	84.27	90.87	38.41	92.63
	✓	**69.46**	**90.07**	**88.33**	**92.07**	**41.64**	**93.34**
Hulu–Med–14B		66.92	82.35	86.20	87.02	39.34	89.37
	✓	**71.65**	**88.24**	**89.51**	**90.14**	**44.05**	**91.96**
Hulu–Med–32B		71.11	88.24	85.06	86.54	44.37	90.08
	✓	**73.37**	**90.07**	**90.26**	**89.42**	**46.04**	**92.22**

Table 2. Main results on VQA–RAD, SLAKE, and PathVQA. Open denotes open–ended Recall. Closed denotes close–ended Accuracy. The CoTBox–TTT column indicates whether the model uses CoTBox–TTT at inference.

Model	Components		Subset	
	Evidence Consistency	EMA Teacher	Open↑	Closed↑
LLaVA	✓	✓	**55.44**	**73.16**
	✓		51.77	68.01
		✓	53.20	70.59
			50.00	65.07
LLaVA–Med–CLIP	✓	✓	**66.27**	**87.13**
	✓		64.30	86.03
		✓	64.46	85.66
			61.52	84.19
Hulu–Med–7B	✓	✓	**69.46**	**90.07**
	✓		65.52	88.24
		✓	65.85	88.97
			63.73	87.50

Table 3. Ablation on VQA–RAD. Each backbone is evaluated with controlled toggles for the two CoTBox–TTT components. The first row per block is the full method (both components enabled), the second removes the validation pass, the third removes the EMA teacher, and the fourth is the native model without test–time training. Metrics are open–ended recall and close–ended accuracy.

mini epochs and modulating it by uncertainty or student teacher disagreement.

A third direction is to extend CoTBox-TTT to three dimensional and to other domains. For CT or MRI volumes the evidence and answer can be aligned across neighboring slices with architectures that already support volumetric features [8, 42]. For ultrasound or endoscopy videos the same idea applies across neighboring frames where temporal coherence is essential for reliable assessment [36]. Practical measures such as 2.5D sampling, feature caching, and lightweight prompt updates can keep the adaptation efficient. The same principles can transfer to remote sensing and industrial inspection where robust grounding and cross view agreement are also important and where emerging video language models and generalist medical models provide compatible backbones [25].

Figure 2. Qualitative examples of CoTBox-TTT. Each case shows the original image, the crop guided by the grounding model with the predicted bounding box, the model responses from baseline model with and without CoTBox-TTT, and the ground truth label. In each case the baseline model without test-time adaptation produces an answer that is either incomplete or inconsistent with the image content, while CoTBox-TTT first localizes the clinically relevant region and then aligns the answers across the original and cropped views in order to convert ambiguous or incorrect predictions into medically specific and image supported statements.

6. Conclusion

This work introduced CoTBox-TTT, a model agnostic and plug and play test time training framework for medical visual question answering. The approach freezes backbone networks and adapts only small sets of soft prompts, grounds reasoning on explicit visual evidence with aligned bounding boxes, and aligns answers across the original image and a localized crop with an exponential moving average teacher. The framework requires no additional labels and leaves encoders and decoders unchanged, and it delivers consistent gains in open-ended recall and close-ended accuracy on VQA-RAD, SLAKE, and PathVQA across diverse backbones. These results indicate that lightweight test time adaptation can make medical vision language systems more reliable and interpretable without architectural changes or extra supervision.

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4971–4980, 2018. 1

[2] Yakoub Bazi, Mohamad Mahmoud Al Rahhal, Laila Bashmal, and Mansour Zuair. Vision–language model for visual question answering in medical imagery. *Bioengineering*, 10 (3):380, 2023. 1

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 6

[4] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 295–305, 2022. 1

[5] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2613–2622, 2021. 6

[6] Xupeng Chen, Zhixin Lai, Kangrui Ruan, Shichu Chen, Jiaxiang Liu, and Zuozhu Liu. R-llava: Improving med-vqa understanding through visual region of interest. *arXiv preprint arXiv:2410.20327*, 2024. 1, 2

[7] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1

[8] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022. 7

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 6

[10] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. 1, 5

[11] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 6

[12] Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. *Advances in Neural Information Processing Systems*, 34:2427–2440, 2021. 1

[13] Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025. 5

[14] KANG JUWON and KIM NAYEONG. Leveraging proxy of training data for test-time adaptation. ML Research Press, 2023. 1

[15] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14162–14171, 2024. 1, 2

[16] Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24174–24184, 2024. 6

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 6

[18] 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):1–10, 2018. 1, 5

[19] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024. 1

[20] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 1

[21] 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. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023. 1, 2, 5

[22] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 1

[23] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018. 1

[24] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 1

[25] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 7

[26] 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)*, pages 1650–1654. IEEE, 2021. 1, 5

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 5

[28] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021. 1

[29] Yuejiang Liu, Parth Kothari, Bastien Van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? *Advances in Neural Information Processing Systems*, 34: 21808–21820, 2021. 1, 2

[30] Lingxiao Luo, Bingda Tang, Xuanzhong Chen, Rong Han, and Ting Chen. Vividmed: Vision language model with versatile visual grounding for medicine. *arXiv preprint arXiv:2410.12694*, 2024. 1, 2

[31] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 6

[32] 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 *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023. 1

[33] Dung Nguyen, Minh Khoi Ho, Huy Ta, Thanh Tam Nguyen, Qi Chen, Kumar Rav, Quy Duong Dang, Satwik Ramchandre, Son Lam Phung, Zhibin Liao, et al. Localizing before answering: A hallucination evaluation benchmark for grounded medical multimodal llms. *arXiv preprint arXiv:2505.00744*, 2025. 1, 2

[34] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pages 16888–16905. PMLR, 2022. 1

[35] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400*, 2023. 1

[36] David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P Langlotz, Paul A Heidenreich, Robert A Harrington, David H Liang, Euan A Ashley, et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020. 7

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 6

[38] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024. 4

[39] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025. 1, 2

[40] Yongyi Su, Xun Xu, and Kui Jia. Revisiting realistic test-time training: Sequential inference and adaptation by anchored clustering. *Advances in Neural Information Processing Systems*, 35:17543–17555, 2022. 1

[41] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020. 1, 2

[42] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20730–20740, 2022. 7

[43] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 6

[44] Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*, 2023. 1

[45] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 1, 2

[46] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022. 1

[47] Jinge Wu, Yunsoo Kim, and Honghan Wu. Hallucination benchmark in medical visual question answering. *arXiv preprint arXiv:2401.05827*, 2024. 1, 2

[48] Maxime Zanella, Clément Fuchs, Christophe De Vleeschouwer, and Ismail Ben Ayed. Realistic test-time adaptation of vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25103–25112, 2025. 1

[49] Jian Zhang, Lei Qi, Yinghuan Shi, and Yang Gao. Domainadaptor: A novel approach to test-time adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18971–18981, 2023. 1

[50] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems*, 35:38629–38642, 2022. 1

[51] 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. 1, 2

[52] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023. 1

[53] Yuhan Zhu, Guozhen Zhang, Chen Xu, Haocheng Shen, Xiaoxin Chen, Gangshan Wu, and Limin Wang. Efficient test-time prompt tuning for vision-language models. *arXiv preprint arXiv:2408.05775*, 2024. 1, 2