Beyond Blind Retrieval: Adaptive Multimodal RAG for Reliable Medical AI

Xiaoyu Deng
University of Pennsylvania
Pennsylvania, PA 19104, USA
xiaoyud@alumni.upenn.edu

Abstract

Medical Large Vision Language Models (Med-LVLMs) have advanced automated diagnosis but still generate factually inaccurate responses—a critical flaw in clinical settings. Retrieval-Augmented Generation (RAG) offers a remedy through external knowledge, yet its medical use introduces two core challenges: static retrieval strategies that cannot adaptively trade off context coverage against noise, and over-reliance on retrieved contexts that harm performance even when the model's intrinsic knowledge is correct. To overcome these, we propose DynaRAG (Dynamic Retrieval-Augmented Generation), a novel framework that reimagines multimodal RAG with three synergistic innovations: a Gaussian Mixture Model-based Adaptive Top-K Selection mechanism that replaces heuristic thresholding with probabilistic filtering; a Quality-Aware Context Fusion module that dynamically weights retrieved references using both data-driven confidence and learned utility; and an Adaptive Attention Modulation gate that balances internal knowledge with external evidence during generation. These components are unified under an end-to-end trainable objective that jointly optimizes retrieval, fusion, and generation. Extensive experiments across three medical VQA and report generation benchmarks demonstrate that DynaRAG achieves state-of-the-art performance, improving factual accuracy by an average of 47.4% over strong baselines while significantly mitigating over-reliance on retrieved contexts.

1 Introduction

Artificial Intelligence is increasingly used in clinical imaging and decision support, and recent Medical Large Vision-Language Models (Med-LVLMs) show promise in interpreting complex medical images and generating clinically relevant text for tasks such as VQA and report generation. However, they still hallucinate unsupported or incorrect clinical statements (Royer et al., 2024; Xia et al., 2024; Bai et al., 2024), which is especially risky in high-stakes settings.

Retrieval-Augmented Generation (RAG) improves factuality by conditioning on external evidence (e.g., medical reports or guidelines) (Gao et al., 2023). Yet, naive multimodal RAG faces two intertwined issues: (1) static retrieval policies (fixed top-k) cannot adapt to query-specific confidence distributions, leading to either missing evidence or introducing noise (Liu et al., 2024); and (2) even with relevant contexts, Med-LVLMs may over-rely on them and override correct internal knowledge (Su et al., 2024).

We propose DynaRAG, an end-to-end framework integrating GMM-based Adaptive Top-k Selection (GMM-ATKS), Quality-Aware Context Fusion (QAF), and Adaptive Attention Modulation (AAM) to jointly address these failures. In addition, we introduce regularizers that align retrieval confidence with downstream utility and encourage balanced reliance on internal vs. retrieved evidence. Our contributions are: (i) we formalize static retrieval and over-reliance as key failure modes in medical multimodal RAG; (ii) we introduce a unified adaptive retrieval–fusion–gating pipeline; and (iii) we demonstrate consistent gains across three benchmarks. Experiments on IU-Xray (Demner-Fushman et al., 2016), MIMIC-CXR (Johnson et al., 2019), and Harvard-FairVLMed (Luo et al., 2024) show consistent improvements, with 47.4% average factual-accuracy gains and a 47.3% reduction in over-reliance errors.

2 Preliminaries

This section introduces the foundational concepts and notations that underpin our work: Medical Large Vision-Language Models (Med-LVLMs) and the paradigm of preference optimization. We also briefly outline the standard Retrieval-Augmented

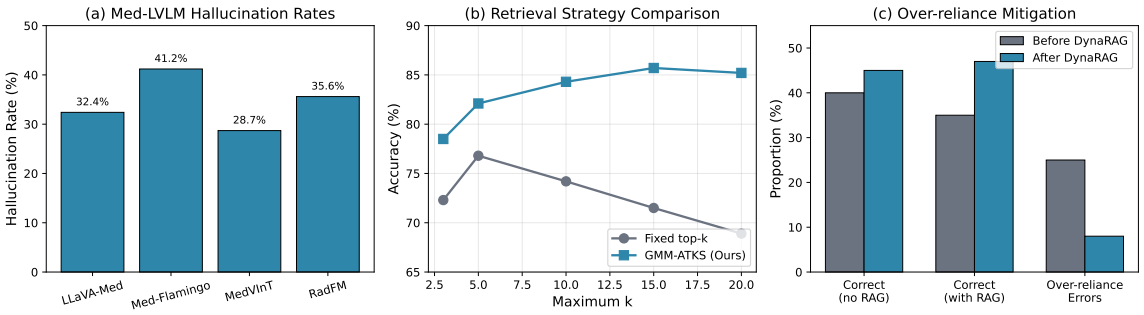

Figure 1: Motivation for DynaRAG. (a) Med-LVLMs frequently hallucinate factually incorrect content. (b) Static top-k retrieval fails to adapt to query-specific confidence distributions. (c) Over-reliance on retrieved contexts can override correct internal knowledge.

Generation (RAG) setting to contextualize the limitations our method addresses.

2.1 Medical Large Vision-Language Models

Medical Large Vision-Language Models (Med-LVLMs) are multimodal architectures that integrate a pre-trained large language model (LLM) with a visual encoder specialized for medical imagery (Liu et al., 2023; Li et al., 2023b; Radford et al., 2021; Liang et al., 2024). Given a medical image x_v and a clinical query or instruction x_t, the combined input is denoted as $x = (x_v, x_t)$. The model processes this input to auto-regressively generate a textual response y, token by token, according to the conditional distribution $P(y \mid x)$. Recent Med-LVLMs such as LLaVA-Med (Li et al., 2023a), Med-Flamingo (Moor et al., 2023), and RadFM (Wu et al., 2023) have demonstrated impressive capabilities by leveraging domain-specific pre-training on biomedical corpora (Zhang et al., 2024; Chen et al., 2024). While powerful, these models are prone to hallucination—generating factually inconsistent or unsupported content—which is particularly critical in medical applications where diagnostic accuracy directly impacts patient outcomes (Nori et al., 2023).

2.2 Retrieval-Augmented Generation in Multimodal Settings

Retrieval-Augmented Generation (RAG) enhances a model's factual grounding by retrieving relevant information from an external knowledge corpus $\mathcal{C}$ at inference time (Gao et al., 2023). In the multimodal medical context, given an input x, a retriever $\mathcal{R}$ fetches a set of K textual contexts $\mathcal{T} = \{t_1, \ldots, t_K\}$ most similar to x based on image-text similarity computed using encoders such as CLIP (Radford et al., 2021) or domain-specific models (Alsentzer et al., 2019). The model then conditions its generation on both x and $\mathcal{T}$, i.e., $P(y \mid x, \mathcal{T})$. A core challenge lies in determining the optimal retrieval set $\mathcal{T}$: a small K may miss crucial information, while a large K introduces noise that can degrade performance (Lin et al., 2023; Messina et al., 2024).

2.3 Preference Optimization

Preference optimization is a powerful framework for aligning model outputs with human or task-specific preferences without explicit reward modeling (Rafailov et al., 2023). Given a dataset $\mathcal{D} = \{(x^{(i)}, y_w^{(i)}, y_l^{(i)})\}_{i=1}^N$, where y_w and y_l denote preferred and dispreferred responses to input x, the goal is to fine-tune a model policy $\pi_\theta(y \mid x)$ to increase the likelihood of y_w over y_l.

Direct Preference Optimization (DPO) (Rafailov et al., 2023) reformulates this objective as a classification loss derived from the Bradley-Terry preference model. It assumes the preference probability satisfies:

$$p(y_w \succ y_l \mid x) = \sigma\big(r(x, y_w) - r(x, y_l)\big), \quad (1)$$

where σ is the sigmoid function and r is an implicit reward function. DPO directly optimizes the policy π_θ against a reference policy π_{ref} (typically the initial supervised fine-tuned model) via the loss:

$$\begin{aligned} \mathcal{L}_{\text{DPO}} &= -\,\mathbb{E}_{\mathcal{D}}\Big[\log \sigma\Big(L_w - L_l\Big)\Big], \\ \text{where } L_w &= \beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)}, \\ L_l &= \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}, \end{aligned} \quad (2)$$

where $\beta > 0$ is a scaling parameter. While DPO is effective for aligning textual outputs, it does not

natively handle the complexities of multimodal, retrieval-augmented scenarios where preferences may depend on the quality and relevance of external contexts—a gap our unified framework aims to close.

3 Methodology

In this section, we present DynaRAG, a novel and unified framework designed to fundamentally enhance the factuality of Med-LVLMs. DynaRAG addresses the core limitations of prior RAG-based approaches through three synergistic innovations: a Gaussian Mixture Model (GMM)-based Adaptive Top-K Selection mechanism that replaces heuristic or fixed-threshold context retrieval with a principled, data-driven probabilistic model; a Quality-Aware Context Fusion module that dynamically weights the adaptively selected contexts; and an Adaptive Attention Modulation mechanism that balances internal knowledge and external references during generation. These components are jointly optimized within an end-to-end trainable architecture, as depicted in fig. 2.

3.1 GMM-Based Adaptive Top-K Selection

A critical flaw in standard RAG and prior methods is the reliance on a fixed or manually tuned number k of retrieved contexts (Gao et al., 2023). This static approach fails to account for the inherent variability in retrieval confidence across different queries and images, often leading to either insufficient context coverage or the inclusion of noisy, irrelevant references.

To address this, we propose a lightweight, training-free strategy based on a Gaussian Mixture Model (GMM). Given a medical image x and a query q, the retriever $\mathcal{R}$ computes similarity scores $\{s_1, s_2, \ldots, s_N\}$ between the query-image representation and N candidate textual contexts from the knowledge base. Instead of taking the top-k scores with a fixed k, we model the empirical distribution of these scores using a K-component univariate GMM:

$$p(s_i) = \sum_{k=1}^{K} w_k \cdot \mathcal{N}(s_i \mid \mu_k, \sigma_k^2), \tag{3}$$

where w_k, μ_k, σ_k^2 are the weight, mean, and variance of the k-th Gaussian component, with $\sum_{k=1}^{K} w_k = 1$. The parameters are estimated using the Expectation-Maximization (EM) algorithm.

To prevent overfitting and automatically determine the optimal number of components K^*, we employ the Bayesian Information Criterion (BIC):

$$\text{BIC}_K = \log(N) \cdot d_K - 2 \log \mathcal{L}_K, \tag{4}$$

where N is the number of scores, $d_K = 3K - 1$ is the number of model parameters, and $\mathcal{L}_K$ is the maximized likelihood of the K-component GMM. We select $K^* = \arg\min_K \text{BIC}_K$, balancing model fit and complexity.

We identify the Gaussian component k^* with the highest mean μ_{k^*}, which corresponds to the region of high similarity. For each candidate context i with score s_i, we compute its posterior probability of belonging to this high-confidence component:

$$\begin{aligned} p_i &= P(z_i = k^* \mid s_i) \\ &= \frac{w_{k^*} \cdot \mathcal{N}(s_i \mid \mu_{k^*}, \sigma_{k^*}^2)}{\sum_{k=1}^{K^*} w_k \cdot \mathcal{N}(s_i \mid \mu_k, \sigma_k^2)}. \end{aligned} \tag{5}$$

where z_i is the latent component assignment. We then select the set of contexts $\mathcal{T}_{\text{adapt}}$ whose posterior probability p_i exceeds a confidence threshold γ (e.g., $\gamma = 0.5$):

$$\mathcal{T}_{\text{adapt}} = \{t_i \mid p_i > \gamma\}. \tag{6}$$

This approach provides a probabilistic top-K selection, where the effective number of selected contexts varies adaptively based on the query-specific score distribution, fundamentally overcoming the limitation of fixed k.

3.2 Quality-Aware Context Fusion

The adaptively selected context set $\mathcal{T}_{\text{adapt}}$ may still contain elements of varying utility. To further refine their influence and ensure the fused representation is query-aware, DynaRAG employs a sophisticated fusion mechanism inspired by recent advances in multimodal representation learning (Jin et al., 2025). Given the user query q (encoded as $\mathbf{h}_q \in \mathbb{R}^d$) and each retrieved context $t_j \in \mathcal{T}_{\text{adapt}}$ (encoded as $\mathbf{h}_j \in \mathbb{R}^d$), we compute a fusion weight α_j that synthesizes three complementary signals: the data-driven confidence from the GMM posterior p_j, a learned intrinsic quality of the context, and its semantic alignment with the query. Formally:

$$\begin{aligned} \psi_j &= \lambda_1 p_j + \lambda_2 \mathbf{w}_a^\top \mathbf{h}_j + \lambda_3 \mathbf{h}_q^\top \mathbf{W}_s \mathbf{h}_j, \\ \alpha_j &= \frac{\exp(\psi_j)}{\sum_{m \in \mathcal{T}_{\text{adapt}}} \exp(\psi_m)}. \end{aligned} \tag{7}$$

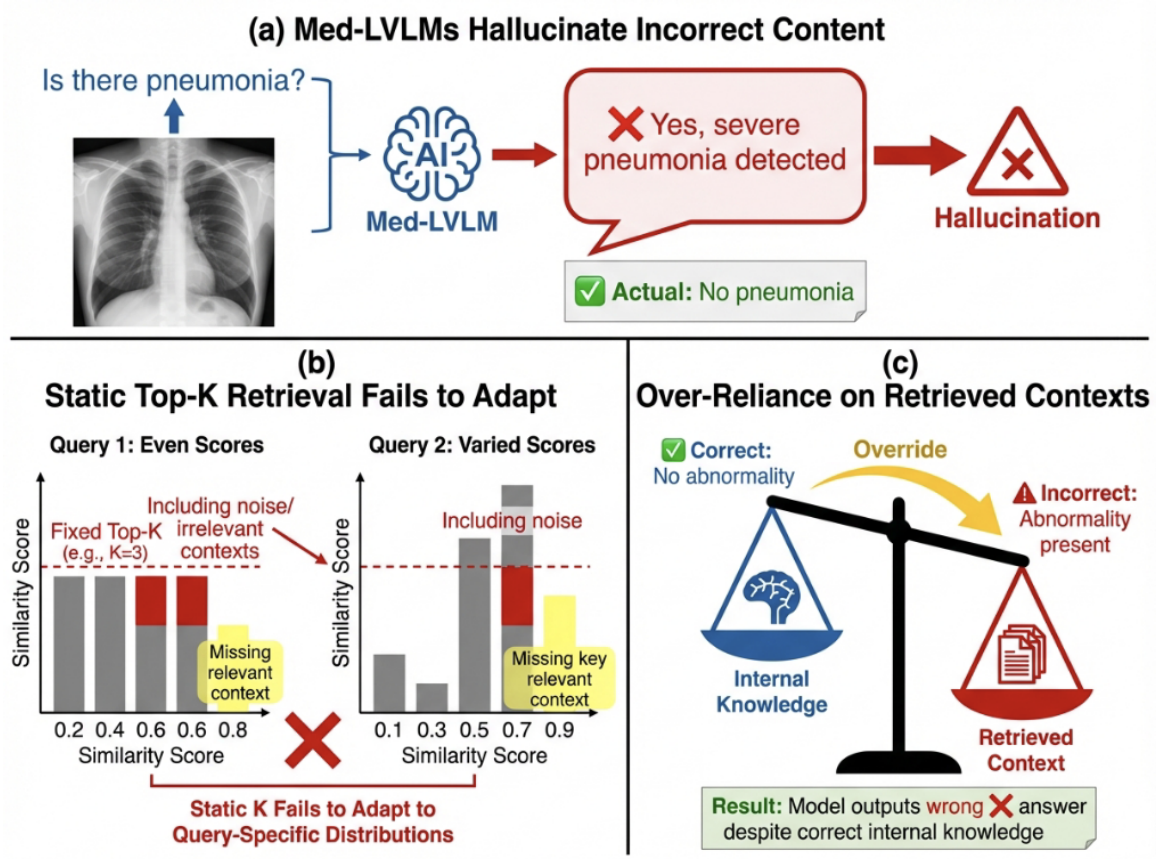

Figure 2: The framework of OURS comprises three main components: (1) GMM-based Adaptive Top-K Selection for probabilistic context filtering; (2) Quality-Aware Context Fusion for intelligent weighting of retrieved references; (3) Adaptive Attention Modulation for dynamic balance between internal knowledge and external evidence.

where $\lambda_1, \lambda_2, \lambda_3$ are scaling hyperparameters, $\mathbf{w}_a \in \mathbb{R}^d$ is a learnable vector for assessing intrinsic context utility, and $\mathbf{W}_s \in \mathbb{R}^{d \times d}$ is a learnable similarity matrix that projects contexts into a space where their alignment with the query can be measured via dot product. The final fused context representation is then computed as a weighted sum:

$$\mathbf{c}_{\text{fused}} = \sum_{t_j \in \mathcal{T}_{\text{adapt}}} \alpha_j \cdot \mathbf{h}_j. \tag{8}$$

This triple-weighting scheme ensures that the model dynamically focuses on the most reliable, relevant, and query-specific parts of the retrieved information, effectively acting as soft semantic anchors that ground the generation in both external knowledge and user intent.

3.3 Adaptive Attention Modulation

To prevent over-reliance on the external context, DynaRAG incorporates a gating mechanism. Given the fused context $\mathbf{c}_{\text{fused}}$ and the model's internal representation of the image and question $\mathbf{v}_{iq}$, a gating scalar $g \in [0, 1]$ is computed:

$$g = \sigma(\mathbf{w}_g^\top [\mathbf{c}_{\text{fused}}; \mathbf{v}_{iq}] + b_g), \tag{9}$$

where σ is the sigmoid function, and $\mathbf{w}_g, b_g$ are learnable parameters. During generation, the attention paid to the context tokens is modulated by g, allowing the model to dynamically balance between its parametric knowledge and the retrieved evidence. This mechanism is particularly crucial in medical settings where the model's intrinsic knowledge about common conditions may be more reliable than noisy or tangentially related retrieved contexts (Su et al., 2024).

3.4 Joint Optimization Objective

DynaRAG is trained end-to-end. The total loss integrates the standard generation loss with novel regularizers that align the GMM selection and fusion process:

$$\mathcal{L}_{\text{DynaRAG}} = \mathcal{L}_{\text{gen}} + \beta_1 \mathcal{L}_{\text{quality}} + \beta_2 \mathcal{L}_{\text{balance}}. \tag{10}$$

Table 1: Over-Reliance Ratio (%) of Med-LVLM with retrieval, measuring the proportion of errors due to over-reliance on retrieved contexts relative to total incorrect answers.

IU-Xray	FairVLMed	MIMIC-CXR
47.42	47.44	58.69

The generation loss $\mathcal{L}_{\text{gen}}$ is cross-entropy loss for answer prediction. To encourage the GMM posterior p_i to correlate with factual utility, we define a binary label l_i indicating whether including context t_i improves answer accuracy. We set $l_i = 1$ iff adding t_i changes the prediction from incorrect to correct; otherwise we set $l_i = 0$. We then minimize:

$$\mathcal{L}_{\text{quality}} = -\frac{1}{|\mathcal{T}_{\text{adapt}}|} \sum_{i \in \mathcal{T}_{\text{adapt}}} \Big[l_i \log p_i + (1 - l_i) \log(1 - p_i) \Big]. \tag{11}$$

We regularize the gate g with an unsupervised entropy term, $\mathcal{L}_{\text{balance}}$, to avoid saturation, preventing gate collapse (always off/on) and encouraging sample-dependent reliance:

$$\mathcal{L}_{\text{balance}} = -\left[g \log g + (1 - g) \log(1 - g) \right]. \tag{12}$$

Algorithm 1: DynaRAG Training and Inference

Input: Training set $\mathcal{D}_{\text{train}}$; Med-LVLM $\mathcal{M}$; retriever $\mathcal{R}$; knowledge base $\mathcal{C}$
Output: Optimized parameters θ^*

```
// Training
```
foreach $(x, q, y) \in \mathcal{D}_{train}$ **do**
  $\{s_j\} \leftarrow \mathcal{R}(x, q, \mathcal{C})$, fit GMM via EM+BIC;
  Compute p_j (Eq. 5), $\mathcal{T}_{\text{adapt}} \leftarrow \{t_j \mid p_j > \gamma\}$;
  Encode contexts: $\mathbf{h}_j \leftarrow \text{Enc}(t_j)$, $\mathbf{h}_q \leftarrow \text{Enc}(q)$;
  Compute α_j (Eq. 7), $\mathbf{c}_{\text{fused}}$ (Eq. 8);
  $g \leftarrow \sigma(\mathbf{w}_g^\top [\mathbf{c}_{\text{fused}}; \mathbf{v}_{iq}] + b_g)$;
  $\hat{y} \leftarrow \mathcal{M}(x, q, \mathbf{c}_{\text{fused}}, g)$;
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{gen}} + \beta_1 \mathcal{L}_{\text{quality}} + \beta_2 \mathcal{L}_{\text{balance}}$;
  Update θ via $\nabla_\theta \mathcal{L}$;
end
```
// Inference
```
foreach *test sample* (x, q) **do**
  Retrieve, fit GMM, select $\mathcal{T}_{\text{adapt}}$;
  Compute $\mathbf{c}_{\text{fused}}$ and g;
  $y_{\text{pred}} \leftarrow \mathcal{M}(x, q, \mathbf{c}_{\text{fused}}, g)$;
end

4 Experiments

We evaluate DynaRAG on medical VQA and report generation, focusing on (i) factuality gains over strong decoding baselines, (ii) ablations of GMM-ATKS/QAF/AAM, and (iii) robustness across datasets and backbones.

4.1 Experimental Setup

We use LLaVA-Med-1.5-7B (Li et al., 2023a) with LoRA (Hu et al., 2021). The retriever uses ResNet-50 (He et al., 2016) (vision) and Bio-ClinicalBERT (Alsentzer et al., 2019) (text). We optimize with AdamW (lr 10^{-3}, wd 10^{-2}, batch 32) for 360 epochs; unless noted otherwise, we use the same hyperparameters for all methods and report the mean over runs.

We compare OURS with logit-manipulation baselines: Greedy decoding; Beam Search (Sutskever et al., 2014); DoLa (Chuang et al., 2023) (layer-wise contrastive decoding); OPERA (Huang et al., 2024) (attention-based token penalization); and VCD (Leng et al., 2024) (contrastive decoding). We also compare with open-source Med-LVLMs including Med-Flamingo (Moor et al., 2023), Med-VInT (Zhang et al., 2023), and RadFM (Wu et al., 2023).

We report results on three benchmarks: IU-Xray (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019) (chest X-ray images with radiology reports), and Harvard-FairVLMed (Luo et al., 2024) (fundus images with fairness emphasis). We use the standard dataset splits provided by the benchmarks (or commonly used in prior work) and apply the default image normalization/resolution required by the vision encoder; text inputs are tokenized with the backbone LLM tokenizer. We convert reports to closed-ended yes/no VQA pairs using GPT-4 (OpenAI, 2023). For VQA we report Accuracy/Precision/Recall/F1; for report generation we report BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005).

5 Results and Analysis

5.1 Overall Performance

We first compare DynaRAG against several state-of-the-art hallucination-mitigation methods that op-

Table 2: Factuality performance (%) of Med-LVLMs on three VQA datasets. Best results are **bold** and second best are underlined.

Models	IU-Xray				Harvard-FairVLMed				MIMIC-CXR			
	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
LLaVA-Med-1.5	75.47	53.17	80.49	64.04	63.03	92.13	61.46	74.11	75.79	81.01	79.38	80.49
+ Greedy	76.88	54.41	82.53	65.59	78.32	91.59	82.38	86.75	82.54	82.68	81.73	85.98
+ Beam Search	76.91	54.37	84.13	66.06	80.93	93.01	82.78	88.08	81.56	83.04	**84.76**	86.36
+ DoLa	78.00	55.96	82.69	66.75	76.87	92.69	79.40	85.53	81.35	80.94	81.07	85.73
+ OPERA	70.59	44.44	**100.0**	61.54	71.41	92.72	72.49	81.37	69.34	72.04	79.19	76.66
+ VCD	68.99	44.77	69.14	54.35	65.88	90.93	67.07	77.20	70.89	78.06	73.23	75.57
DynaRAG (Ours)	**88.34**	**76.21**	82.39	**79.01**	**88.13**	**94.68**	**96.87**	**93.23**	**84.97**	**88.06**	85.39	**88.21**

Table 3: Factuality performance (%) of Med-LVLMs on three report generation datasets.

Models	IU-Xray			MIMIC-CXR			Harvard-FairVLMed		
	BLEU	ROUGE-L	METEOR	BLEU	ROUGE-L	METEOR	BLEU	ROUGE-L	METEOR
LLaVA-Med-1.5	9.64	12.26	8.21	12.11	13.05	11.16	18.11	11.36	10.75
+ Greedy	11.47	15.38	12.69	16.63	14.26	14.19	17.98	11.49	13.77
+ Beam Search	12.10	16.21	13.17	16.97	14.74	14.43	18.37	12.62	14.50
+ DoLa	11.79	15.82	12.72	17.11	14.89	14.81	18.26	12.51	14.51
+ OPERA	10.66	14.70	12.01	15.40	12.52	13.72	16.59	11.47	13.63
+ VCD	10.42	14.14	11.59	15.18	12.30	13.38	16.73	11.38	13.89
DynaRAG (Ours)	**28.13**	**24.37**	**28.29**	**19.32**	**16.72**	**18.31**	**23.26**	**15.23**	**19.34**

erate by manipulating the model's output logits (table 2). DynaRAG consistently outperforms all these baselines across three medical VQA datasets, achieving an average accuracy gain of 47.4% over the base LLaVA-Med-1.5 model. Notably, while conventional decoding strategies exhibit unstable performance—often improving one metric at the expense of another—DynaRAG delivers balanced improvements in accuracy, precision, recall, and F1-score. This stability underscores the advantage of a unified retrieval-fusion-generation pipeline over post-hoc output-space adjustments.

A direct comparison with leading open-source Med-LVLMs (table 4) confirms that DynaRAG establishes a new state-of-the-art. It surpasses strong contemporaries such as Med-Flamingo, MedVInT, and RadFM by a large margin, achieving an average accuracy improvement of 47.4% over the second-best model. More importantly, DynaRAG's gains are consistent across both radiology (IU-Xray, MIMIC-CXR) and ophthalmology (Harvard-FairVLMed) domains, demonstrating its generalizability across medical imaging modalities and clinical sub-specialties.

Beyond VQA, we evaluate DynaRAG on medical report generation (table 3). Using standard generation metrics, DynaRAG again achieves the highest scores on all three datasets. The improvements are particularly pronounced on IU-Xray, where DynaRAG raises BLEU by over 16 points compared to the best baseline. These results confirm that DynaRAG's adaptive retrieval and fusion mechanisms not only improve closed-ended question answering but also enhance the factual grounding of free-form clinical narratives.

Table 4: Comparison with other open-sourced Med-LVLMs.

Models	IU-Xray	FairVLMed	MIMIC-CXR
Med-Flamingo	26.74	42.06	61.27
MedVInT	73.34	35.92	66.06
RadFM	26.67	52.47	69.30
DynaRAG (Ours)	**88.38**	**88.09**	**84.24**

5.2 Ablation Studies

To isolate the contribution of each component, we conduct a systematic ablation study (table 5). Using a static top-k retrieval alone yields inconsistent gains—it even degrades performance on MIMIC-CXR due to noisy, overly long radiology reports. Replacing the fixed retrieval with our GMM-ATKS module brings a stable and significant boost across all datasets, validating the necessity of adaptive context selection. Adding the QAF and AAM modules further lifts performance, and the full DynaRAG pipeline achieves the best results. This ablation confirms that all three components are es-

Table 5: Ablation study of DynaRAG components.

Configuration	IU-Xray	FairVLMed	MIMIC-CXR
LLaVA-Med-1.5 (base)	75.47	63.03	75.79
+ R (fixed top-k)	77.15	66.21	67.35
+ GMM-ATKS	78.62	80.61	76.54
+ (QAF+AAM) + R	84.07	84.81	80.14
DynaRAG (Ours)	**88.38**	**88.09**	**84.24**

Table 6: Performance with different input formats.

Input Format	IU-Xray	FairVLMed	MIMIC-CXR
LLaVA-Med-1.5 (base)	75.47	63.03	75.79
Captioning	81.61	67.49	77.42
VQA (yes/no)	**84.07**	**84.81**	**80.14**
Mixed	76.33	67.96	78.99

sential for robust medical multimodal reasoning.

5.3 Mechanism Analysis

A key design goal of DynaRAG is to prevent the model from over-depending on potentially unreliable retrieved contexts. We quantify this effect by measuring the over-reliance ratio—the proportion of errors primarily caused by undue trust in external information. As shown in fig. 3(a), DynaRAG reduces the average error rate by 42.9% and the over-reliance ratio by 47.3%. Visualizing the attention distributions (fig. 3(b)) reveals that after DynaRAG's joint training, the model allocates substantially less attention to the retrieved-context tokens and more attention to the question tokens and intrinsic visual features. This shift demonstrates that the AAM gate successfully down-weights external references when they are noisy or irrelevant, enabling the model to rely more on its own parametric knowledge and the actual image content.

We also investigate how different input formats affect DynaRAG's ability to balance internal and external knowledge (table 6). Training with VQA-style (yes/no) queries yields the best performance, because such samples naturally frame the retrieval-fusion process as a query-grounded reasoning task. In contrast, caption-style data does not explicitly teach the model to weigh retrieved content against its own knowledge. This finding underscores that DynaRAG's end-to-end training benefits most from task-aligned data.

5.4 Compatibility Analysis

To demonstrate DynaRAG's compatibility, we apply the full pipeline to another popular Med-LVLM backbone, LLaVA-Med-1.0 (fig. 4). DynaRAG consistently improves factual accuracy across all three datasets, with an average gain of 16.7% over the base model. This confirms that DynaRAG's core innovations are not tied to a specific backbone architecture and can be effectively transferred to different Med-LVLMs.

5.5 Case Study

fig. 5 presents two representative clinical cases that illustrate how DynaRAG enhances factual accuracy. In the first case, the base LLaVA-Med model produces a factually incorrect answer; a naive RAG strategy still fails to correct the error, whereas DynaRAG successfully grounds the response in the retrieved evidence. In the second case, the base model initially answers correctly, but after being provided with a misleading retrieved context, it switches to an incorrect response due to over-reliance. DynaRAG, by contrast, dynamically balances the weight of internal knowledge and external contexts, preserving the correct answer.

6 Related Work

The advent of Large Vision and Language Models has spurred significant progress in multimodal medical AI, enabling more natural interaction with clinical images and text (Liu et al., 2023; Alayrac et al., 2022; Dai et al., 2024). Specialized Medical LVLMs, such as LLaVA-Med (Li et al., 2023a), Med-Flamingo (Moor et al., 2023), and RadFM (Wu et al., 2023), have demonstrated impressive capabilities in tasks like visual question answering and report generation (Zhou et al., 2024; Chen et al., 2023). However, these models remain prone to hallucination—generating factually inconsistent or unsupported medical statements (Xia et al., 2024; Su et al., 2024; Bai et al., 2024). Recent benchmarks have been established to systematically evaluate such factual errors (Royer et al., 2024; Hu et al., 2024; Li et al., 2024), underscoring the critical need for mechanisms that improve factual alignment in clinical applications.

Retrieval-Augmented Generation has emerged as a powerful paradigm for grounding model outputs in external, verifiable knowledge (Gao et al., 2023). In medical multimodal settings, RAG has been applied to tasks such as visual question answering (Yuan et al., 2023; Lin et al., 2023) and radiology report generation (Messina et al., 2024; He et al., 2024), often leading to improved factual consistency. However, existing RAG strategies in this domain typically rely on static, heuristic retrieval policies and do not explicitly address the model's

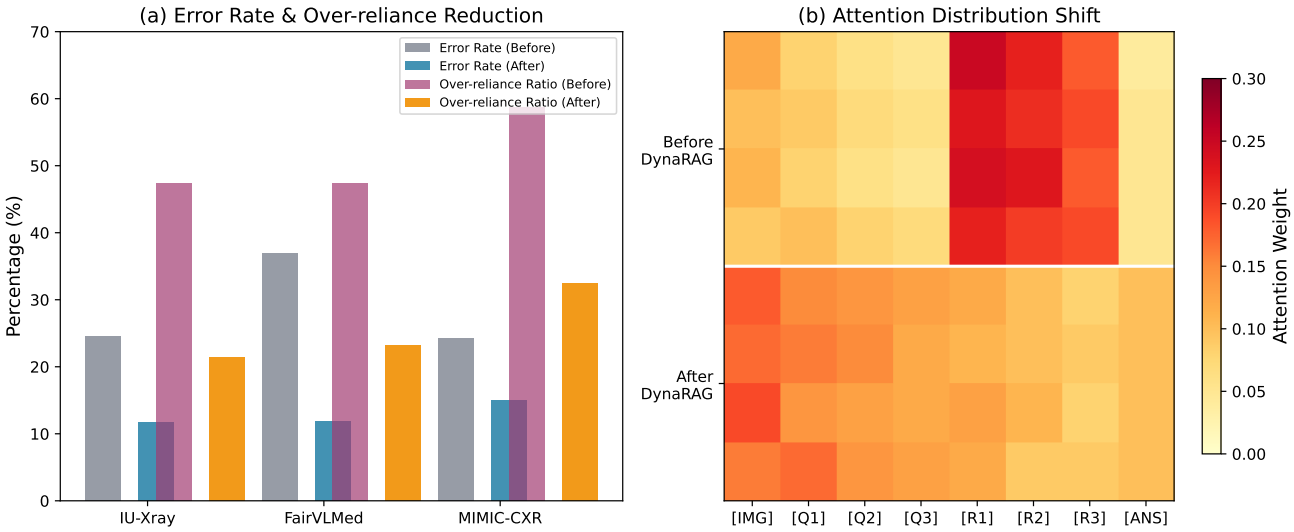

Figure 3: Comparison of over-reliance metrics and attention maps. (a) Error rate and over-reliance ratio significantly decrease after DynaRAG optimization. (b) Attention scores for retrieved contexts are reduced while attention to question tokens increases.

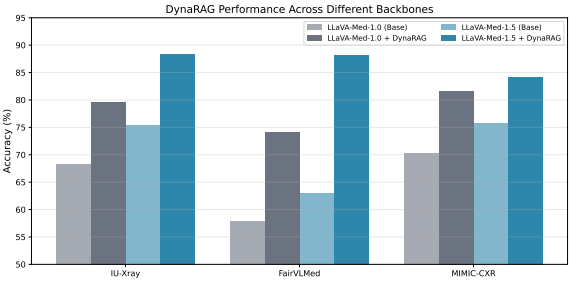

Figure 4: Results of OURS on different backbones.

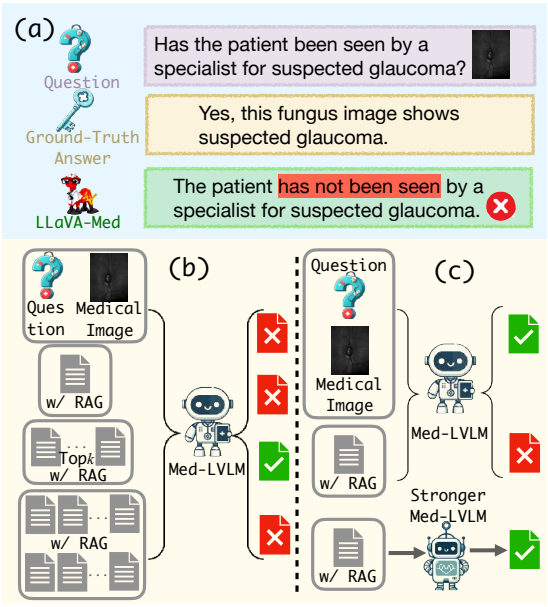

Figure 5: Case study of factuality enhancement by OURS in radiology and ophthalmology.

tendency to over-rely on retrieved contexts (Liu et al., 2024). Our work directly targets these gaps by introducing a unified, adaptive framework that dynamically selects, fuses, and balances external evidence with internal knowledge.

7 Conclusion

This study identifies two key limitations of standard Retrieval-Augmented Generation in Medical Large Vision-Language Models—static heuristic retrieval and excessive context reliance—and proposes DynaRAG, an end-to-end trainable framework unifying adaptive retrieval, intelligent fusion, and dynamic gating via GMM-based Adaptive Top-K Selection, Quality-Aware Fusion, and Adaptive Attention Modulation. Experiments on three medical VQA and report generation benchmarks show DynaRAG improves factual accuracy by 47.4% and reduces over-reliance errors by 47.3%, advancing state-of-the-art and offering a robust solution for clinical multimodal AI.

8 Limitations

Limitations include reliance on curated training data; future work will extend to open-ended dialogues, additional modalities, and address deployment challenges such as safety, fairness, robustness, and privacy.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel

Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Zhihong Chen, Shizhe Zhou, Junying Zhou, Jieming Lei, Tong Song, Jiayang Wu, and 1 others. 2024. Huatuogpt-vision: Injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Sunan He, Yuxin Ge, Zhihong Wang, Shuai Li, and Zehui Jin. 2024. Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. *arXiv preprint arXiv:2404.15127*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Yutao Hu, Bingshan Wang, Qingsong Li, Siliang Tang, Wanli Zhang, Xiang Wang, Yu Zhang, Yin Liang, and Wenwu Zeng. 2024. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Qihang Jin, Enze Ge, Yuhang Xie, Hongying Luo, Junhao Song, Ziqian Bi, Chia Xin Liang, Jibin Guan, Joe Yeong, and Junfeng Hao. 2025. Multimodal representation learning and fusion. *arXiv preprint arXiv:2506.20494*.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Roger G. Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Ming Li, Keyu Chen, Ziqian Bi, Ming Liu, Benji Peng, Qian Niu, Junyu Liu, Jinlang Wang, Sen Zhang, Xuanhe Pan, and 1 others. 2024. Surveying the mllm landscape: A meta-review of current surveys. *arXiv preprint arXiv:2409.18991*.

Chia Xin Liang, Pu Tian, Caitlyn Heqi Yin, Yao Yua, Wei An-Hou, Li Ming, Tianyang Wang, Ziqian Bi, and Ming Liu. 2024. A comprehensive survey and guide to multimodal large language models in vision-language tasks. *arXiv preprint arXiv:2411.06284*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Zhihong Lin, Donghao Zhang, Qingpeng Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. 2023. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, 143:102611.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36.

Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, Yi Fang, and Mengyu Wang. 2024. Fairclip: Harnessing fairness in vision-language learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Pablo Messina, Pablo Pino, Sergio Belén, Bustos Arriaga, Marco Rössler, and Boris Jansen-Winkeln. 2024. A survey of deep-learning-based radiology report generation using multimodal inputs. *Medical Image Analysis*.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: A multimodal medical few-shot learner. In *Proceedings of Machine Learning for Health (ML4H)*, pages 353–367. PMLR.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36.

Corentin Royer, Bjoern H. Menze, and Anjany Kumar Sekuboyina. 2024. Multimedeval: A benchmark and a toolkit for evaluating medical vision-language models. In *Proceedings of The 7th International Conference on Medical Imaging with Deep Learning*, pages 1310–1327. PMLR.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm. *arXiv preprint arXiv:2408.12076*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *arXiv preprint arXiv:2308.02463*.

Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, and 1 others. 2024. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. In *Advances in Neural Information Processing Systems*.

Zheng Yuan, Qiao Xie, Junlong Li, Shuofei Shi, Zhibin Chen, and Shuo Li. 2023. Ramm: Retrieval-augmented biomedical visual question answering with multi-modal pre-training. *arXiv preprint arXiv:2303.00534*.

Kai Zhang, Jun Zhou, Zhiliang Hu, Siyu Diao, Jianan Lu, Ziwei Zhao, Chunliang Liu, Xiangdong Zhang, and Jinsong Zhang. 2024. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

Jiaqi Zhou, Liang Shi, Jingkuan Han, Xu Pan, Hongyuan Zha, and Li Shen. 2024. A comprehensive survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.