GAZEWARP: GAZE-GUIDED ADAPTIVE WARPING FOR BANDWIDTH-EFFICIENT MULTIMODAL INTERACTION IN XR

Anonymous authors
Paper under double-blind review

ABSTRACT

Emerging wearable mixed reality (XR) systems increasingly adopt edge–cloud architectures, placing severe bandwidth constraints on the transmission of egocentric visual data to backend models. At the same time, natural XR interaction often involves ambiguous query expressions, with eye gaze serving as a practical cue to user intent. This intersection creates a critical challenge for gaze-guided multimodal interaction in XR: how to preserve the essential visual evidence needed to resolve a user's ambiguous query expression without exceeding a limited transmission budget. To study this problem, we introduce GazeRefer-GQA, a benchmark derived from GQA that combines user-view gaze cues, ambiguous query expressions, and structured reasoning annotations. We further propose GazeWarp, a gaze-guided visual enhancement framework that acts as a plug-and-play frontend before a frozen backend model. Using the user's gaze and question semantics, GazeWarp locates the intended target region and any structurally related regions, and drives a density-equalized warping process that spatially redistributes image resolution toward these regions before transmission. Across a range of transmission budgets and gaze input forms, GazeWarp generally outperforms resizing, cropping, and compression baselines under approximately matched payload budgets. Notably, despite using only a fraction of the bytes, it is comparable to the unconstrained full-resolution reference and numerically exceeds it in several settings. These gains hold across multiple backend vision-language models, and GazeWarp further generalizes strongly to additional gaze-guided datasets beyond our benchmark.

1 INTRODUCTION

Recent advances in wearable augmented and mixed reality (XR) systems are making always-available visual assistants increasingly practical (Wang et al., 2025e; Tang et al., 2025; Wang et al., 2025b; 2024b; Khelifi & Morris, 2024). To support such assistants in practice, recent XR systems adopt split or edge–cloud architectures, where the wearable device captures egocentric visual inputs while heavyweight multimodal inference is executed on a backend server (Srinidhi et al., 2024; Li et al., 2025b). While this design circumvents severe on-device computational, memory, and thermal constraints, it places visual uplink transmission directly on the critical path of interaction (Shinde et al., 2025; Sharshar et al., 2025; Kämäräinen & Siekkinen, 2023). Prior XR offloading studies show that transmitting high-fidelity camera streams without compression consumes substantial bandwidth and noticeably degrades user experience due to network latency (Gao et al., 2025; Jiang et al., 2025; Alhilal et al., 2024; Mohanto et al., 2022). A straightforward remedy is to shrink the payload by uniformly resizing the image before upload. However, treating all regions equally indiscriminately discards local detail; when an interaction hinges on fine-grained visual evidence, this often destroys the very cues the backend model needs to answer correctly. As illustrated in Fig. 1, the challenge is therefore not merely how to send less visual data, but how to preserve the most task-relevant evidence under a limited transmission budget (Chen & Qi, 2025).

Parallel to these system constraints, a second challenge arises at the user level, from how people naturally phrase their queries in wearable XR (Wei et al., 2023; Adhanom et al., 2023). Recent studies

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.

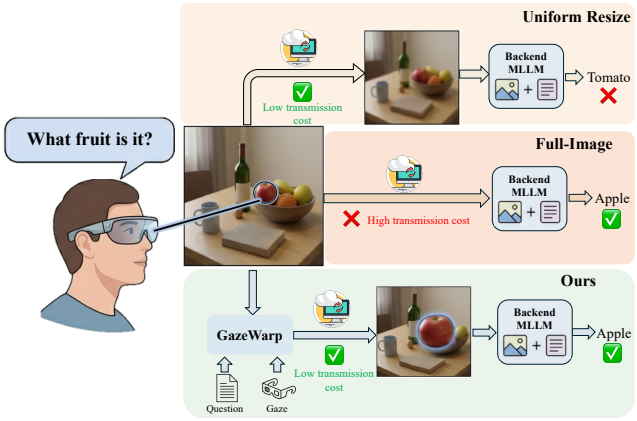

Figure 1: A user issues an ambiguous query while gazing at a target. The full image answers correctly but is costly to transmit; under the same compact budget, uniform resizing destroys the critical detail, whereas GazeWarp warps the image toward the query- and gaze-relevant evidence before transmission.

on multimodal querying suggest that when seeking information about their surroundings, users tend to refer to the object of current interest through context-dependent, ambiguous expressions—often pronouns—rather than naming it explicitly (Wang et al., 2024a; Plopski et al., 2023; Lee et al., 2025; Zhang & Rekimoto, 2025). For instance, a user may simply ask "What fruit is it?" while looking at an object, expecting the assistant to resolve the intended reference from the embodied visual context, as illustrated in Fig. 1. In such settings, eye gaze offers a practical proxy for user intent, allowing the system to disambiguate these expressions and ground the visual interaction (Lee et al., 2024; Wang et al., 2024a; Yan et al., 2024). These two factors intersect to form a concrete problem setting: gaze-guided multimodal interaction in XR under bandwidth constraints. Here, correctly answering a query hinges on resolving the user's intended reference and preserving its corresponding visual evidence for backend reasoning.

Studying this problem systematically requires a benchmark whose task formulation matches it, yet existing gaze resources are valuable but built for other purposes: they target video streams (Ilaslan et al., 2023; Rekanar et al., 2025), pair gaze with explicit rather than ambiguous questions and lack fine-grained structural annotations (Mondal et al., 2024; Sood et al., 2021; Inadumi et al., 2024; Sood et al., 2023), or evaluate open-ended generation from simulated mouse traces (Yan et al., 2024); we discuss this mismatch in detail in Sec. 2. To fill this gap, we introduce GazeRefer-GQA, built on GQA (Hudson & Manning, 2019): leveraging its structured object and relation annotations, we rewrite explicit questions into ambiguous query expressions and derive gaze cues from the referred object's annotation. It comprises two complementary subsets: Local Reasoning, where the answer depends primarily on the gaze-indicated target region, and Global Reasoning, where the model must additionally infer a related region conditioned on the gaze cue and question semantics.

To address these intertwined challenges, we propose **GazeWarp**, a gaze-guided visual enhancement framework based on adaptive warping for bandwidth-efficient multimodal interaction in XR. As a plug-and-play frontend module, GazeWarp uses the user's gaze cue and question semantics to adaptively preserve task-relevant visual evidence before transmission. Specifically, the framework identifies a target region indicated by gaze and, for complex queries, an additional related region inferred from the question context. These regions of interest drive a density-equalized warping process that spatially redistributes the image's resolution, so that critical details are prioritized within a fixed transmission budget. This design allows GazeWarp to be integrated with various downstream vision-language models, providing a more bandwidth-efficient alternative to naive uniform resizing without requiring model-specific fine-tuning.

In summary, this work takes a step toward gaze-guided multimodal interaction in XR by studying how to preserve answer-relevant visual evidence under bandwidth constraints. Our contributions are as follows:

- We introduce **GazeRefer-GQA**, a GQA-derived benchmark for gaze-guided visual question answering in XR-style interaction, pairing user-view gaze cues with ambiguous query expressions under two reasoning modes (Local and Global) and evaluating models under simulated gaze noise rather than idealized object-center points.
- We propose **GazeWarp**, a plug-and-play frontend that requires no model-specific fine-tuning: it locates gaze- and question-relevant regions and spatially redistributes image resolution toward them, preserving task-relevant evidence under a limited transmission budget.
- Across a range of transmission budgets and gaze input forms, GazeWarp generally outperforms resizing, cropping, and compression baselines under approximately matched payloads, and is comparable to—in some settings numerically exceeding—the unconstrained full-resolution reference. These gains hold across multiple backend VLMs and transfer to additional gaze-guided datasets.

2 Related Work

2.1 XR Assistants under Edge–Cloud Constraints

Wearable XR increasingly adopts a split-processing paradigm, where lightweight headsets capture egocentric data while heavyweight multimodal inference is offloaded to a backend server, circumventing severe on-device computational, memory, and thermal constraints (Srinidhi et al., 2024; Li et al., 2025b; Shinde et al., 2025). However, offloading places visual uplink transmission on the critical path, making the user experience highly sensitive to network latency and bandwidth (Gao et al., 2025; Jiang et al., 2025; Wang et al., 2025a). While foveated streaming and perceptual optimization have been explored for media delivery (Illahi et al., 2021; Chen et al., 2022; Schwarz & Behnke, 2025), practical multimodal interaction further needs bandwidth-efficient transmission that prioritizes answer-relevant regions and preserves task-relevant details before backend reasoning.

2.2 Resolving Ambiguous Query Expressions with Gaze in XR

In XR assistants, natural interaction often involves ambiguous query expressions that are hard to resolve from language alone (Plopski et al., 2023; Zhang & Rekimoto, 2025; Lee et al., 2025). Recent studies show that users rely on context-dependent queries and coordinate eye gaze with speech during information seeking, as demonstrated by G-VOILA (Wang et al., 2024a), and that gaze combined with conversational context disambiguates underspecified intent (GazePointAR (Lee et al., 2024)). Existing data resources, however, are often misaligned with this setting: RefCOCO-Gaze (Mondal et al., 2024) and VQA-MHUG (Sood et al., 2021) analyze multimodal attention, GazeVQA (Ilaslan et al., 2023) targets video, and Gaze-grounded VQA (Inadumi et al., 2024) studies speaker-centric rather than wearer gaze. Voila-A (Yan et al., 2024) is closest but targets open-ended alignment with simulated mouse traces. In contrast, GazeRefer-GQA is a structured benchmark designed to evaluate how gaze resolves ambiguous query expressions under bandwidth-constrained VQA, while explicitly modeling gaze noise.

2.3 Auxiliary-Guided Image Processing

Recent work improves visual question answering by adapting image representations using auxiliary guidance (Wu & Xie, 2024; Wang et al., 2025c;d), such as query-guided visual prompting, question-conditioned augmentation, or token pruning (Yu et al., 2024; Liu et al., 2026; Huang et al., 2024; Li et al., 2025a). By contrast, cropping strategies transmit a high-fidelity local crop: gaze-centered methods such as GazeGPT (Konrad et al., 2024) and GazeLLM (Rekimoto, 2025) send a crop around the gaze-indicated object, while attention- or head-guided variants like ViCrop (Zhang et al., 2025) and head-aware cropping (Xie et al., 2026) localize crops from model or head-pose cues. Such crops, however, discard the broader scene context often needed for relational queries. GazeWarp instead uses gaze-guided adaptive warping to preserve both target details and global context in a single bandwidth-efficient input, retaining evidence for local and global reasoning.

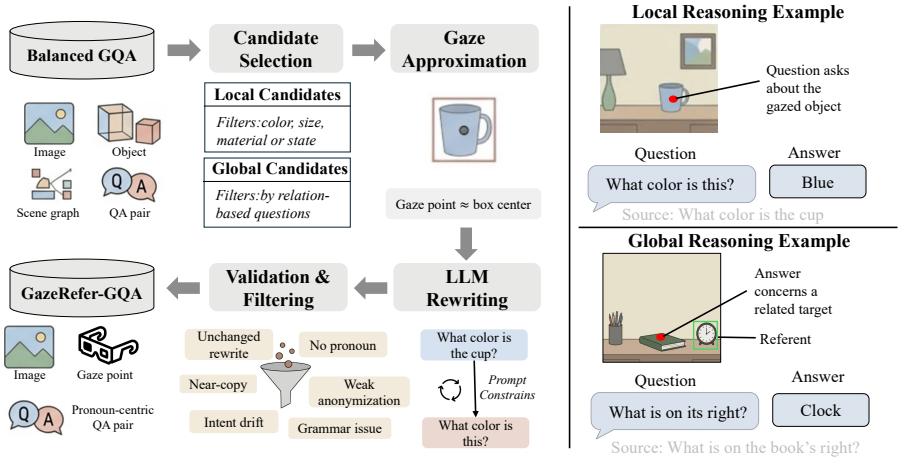

Figure 2: **Construction of GazeRefer-GQA.** Left: pipeline for deriving gaze-guided question-answer pairs from GQA, including candidate selection, gaze approximation, question rewriting, and quality filtering. Right: examples of the two reasoning modes.

3 GAZEREFER-GQA

3.1 PROBLEM FORMULATION

We formulate gaze-guided visual question answering as a benchmark setting for gaze-guided multimodal interaction in XR. Each example is $x = (I, q, \mathcal{G})$, where q is a rewritten question and $\mathcal{G} = \{g_1, \ldots, g_K\}$ is a set of K gaze samples associated with the referred target; the goal is to predict the answer a. Modeling gaze as a set of samples, rather than a single point, is a deliberate design choice that reflects the imperfect, noisy nature of real eye tracking. Each of the K samples defines a separate evaluation instance (I, q, g_k) that is processed independently by the frontend and backend. The rewritten question is constructed so that its semantics cannot be reliably recovered from (I, q) alone, making gaze essential for resolving user intent.

Let r_t denote the target region induced by the gaze samples. In Local Reasoning, the answer depends primarily on r_t, i.e., $a \sim p(a \mid I, q, r_t)$. In Global Reasoning, the model must further infer a related region r_r conditioned on the gaze cue and question semantics, i.e., $r_r = G(I, q, r_t)$ and $a \sim p(a \mid I, q, r_t, r_r)$. The benchmark therefore tests both grounding the target region and, when needed, inferring an additional related region.

3.2 BENCHMARK CONSTRUCTION

We build GazeRefer-GQA from the balanced split of GQA (Hudson & Manning, 2019), retaining only question-answer pairs that can be rewritten into gaze-dependent questions while preserving the original answer and a unique spatial interpretation. Each retained source example is mapped to at most one rewritten example.

Following the Local/Global distinction of Sec. 3.1, we partition the retained set according to what each question asks about the gaze-indicated target region r_t. For the Local subset, we retain object-centric questions about r_t itself, covering its object category and intrinsic attributes such as color, size, material, and state, as well as its absolute location within the image. For the Global subset, we retain relation-based questions whose related region r_r can be uniquely determined from GQA annotations, using a restricted whitelist of relations including *on*, *in*, *under*, *behind*, *in front of*, *next to*, *left*, *right*, *holding*, *wearing*, and *near*; here r_r is taken as the bounding box of the object on the other end of the annotated relation, and directional relations such as *left* and *right* are kept only when this related object is unambiguous.

Because GazeRefer-GQA is derived from annotated VQA data rather than from an eye tracker, gaze is approximated from the referred object's bounding box. Rather than equating gaze with a single

box-center point, we model it as a set of $K = 5$ samples $\mathcal{G} = \{g_1, \ldots, g_5\}$ scattered around the target: a clean anchor fixation g_1 at the box center, plus four samples drawn from a *target-anchored Gaussian* whose per-axis spreads scale with the box geometry. This is a heuristic, target-relative jitter model rather than a physically calibrated angular-error model: it is qualitatively motivated by the fact that real gaze is imperfect—head-worn eye trackers exhibit non-negligible calibration error and human fixations disperse around the intended target due to microsaccades, ocular tremor, and slow drift (Adhanom et al., 2023; Plopski et al., 2023; Wei et al., 2023). Exact jitter parameters are given in App. A.2.

Finally, we rewrite the retained questions into gaze-dependent forms using a large language model, constrained to preserve the original answer and question intent, remove explicit object mentions, preserve relation direction, and require gaze for reliable disambiguation. A rule-based audit then discards low-quality rewrites—covering unchanged or near-copy outputs, missing deictic references, weak anonymization, intent or polarity drift, and grammatical anomalies—followed by deduplication; the prompt and the full audit criteria are given in App. A.3. In addition to this automatic filtering, we manually inspect a random sample of the curated examples to verify that the rewritten questions preserve the original answer and intent, genuinely rely on gaze for disambiguation, and remain grammatically fluent. The curated examples are then repartitioned into image-disjoint training and test splits. Figure 2 summarizes the full construction pipeline.

3.3 Benchmark Statistics and Evaluation Protocol

The resulting GazeRefer-GQA benchmark contains 26,328 training question-answer pairs over 20,210 images and 2,000 test question-answer pairs over 1,959 images, evenly split between the Local and Global subsets. Detailed per-split statistics (including survival rates) and the fine-grained question-type distribution are provided in the appendix.

We evaluate with exact-match accuracy against the preserved original answers. Letting $\mathrm{Acc_l}$ and $\mathrm{Acc_g}$ denote the accuracies on the Local and Global subsets, the overall accuracy is computed on the full test set, equivalently the sample-weighted mean $\mathrm{Acc_{all}} = (N_\mathrm{l}\mathrm{Acc_l} + N_\mathrm{g}\mathrm{Acc_g})/(N_\mathrm{l} + N_\mathrm{g})$. For every test example, the reported accuracy is further averaged over the five gaze samples defined in Sec. 3.2.

4 Method

4.1 Overview

Gaze-guided multimodal interaction in XR requires preserving answer-relevant visual evidence under limited transmission budgets. As illustrated in Fig. 3, GazeWarp addresses this challenge through a two-stage pipeline. Given an input image I, a question q, and a gaze cue g, GazeWarp first predicts, for the current query, a target box $\mathbf{b}_t$ and a related box $\mathbf{b}_r$ that estimate the target and related regions r_t, r_r of Sec. 3.1, together with a reasoning type $\hat{y} \in \{0, 1\}$. The predicted regions are then converted into a spatial importance signal that drives density-equalized warping, whose output is resized to a fixed transmission resolution to yield a bandwidth-efficient visual input I^{trans} for downstream inference.

4.2 Multimodal Region Prediction

Given a gaze-marked image I_g and a gaze-augmented question q_g, GazeWarp predicts a target region, a related region, and a reasoning type. The gaze cue is injected into both the visual and textual inputs: a visible gaze marker, rendered as a red circle, is overlaid on the image, while the textual input is augmented with gaze coordinates in the form of (x, y). During training, we further add a short instruction encouraging the model to use gaze when resolving the query.

We encode the image–text pair with a small pretrained VLM backbone, producing multimodal tokens $H = \mathrm{Backbone}(I_g, q_g)$. These tokens are projected into a task-specific space by $M = \phi(H)$ and refined by a lightweight fusion encoder as $\tilde{M} = \mathrm{Enc}(M)$. Three learned queries, q_t, q_r, and q_c, are then decoded from $\tilde{M}$ to obtain query features z_t, z_r, and z_c, corresponding to the target region, related region, and reasoning type, respectively. We further compute a global context fea-

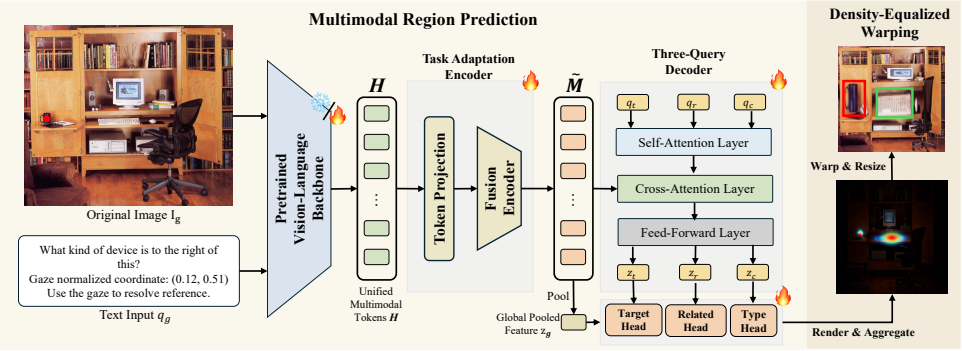

Figure 3: Overview of GazeWarp. A small pretrained VLM backbone encodes the gaze-marked image and gaze-augmented question; a lightweight fusion encoder and three learned queries decode the target region, related region, and reasoning type. These regions form a spatial importance map driving density-equalized warping. Example image from GQA (Hudson & Manning, 2019).

ture $z_g = \frac{1}{L}\sum_{i=1}^{L} \tilde{M}_i$ by mean-pooling all multimodal tokens. The prediction heads then output $\mathbf{b}_t = f_t([z_t; z_g])$, $\mathbf{b}_r = f_r([z_r; z_g])$, and $p = \sigma(f_c([z_c; z_g]))$, where $\mathbf{b}_t$ and $\mathbf{b}_r$ are normalized boxes (c_x, c_y, w, h) and p is the probability of Global Reasoning. At inference time, we use hard gating $\hat{y} = \mathbb{I}(p > \tau)$ to decide between Local ($\hat{y} = 0$) and Global ($\hat{y} = 1$) Reasoning: the model always predicts both boxes, but the related box is used only when the query is predicted as Global.

For training, let $\mathbf{b}_t^*$, $\mathbf{b}_r^*$, and $y^* \in \{0, 1\}$ denote the ground-truth target box, related box, and reasoning type, respectively. Both box branches use the same regression objective (Rezatofighi et al., 2019),

$$\mathcal{L}_{\text{box}}(\mathbf{b}, \mathbf{b}^*) := \lambda_{\ell_1} \|\mathbf{b} - \mathbf{b}^*\|_1 + \lambda_{\text{giou}} \mathcal{L}_{\text{giou}}(\mathbf{b}, \mathbf{b}^*). \tag{1}$$

The three training losses are computed as $\mathcal{L}_{\text{t}} = \mathcal{L}_{\text{box}}(\mathbf{b}_t, \mathbf{b}_t^*)$, $\mathcal{L}_{\text{r}} = y^* \, \mathcal{L}_{\text{box}}(\mathbf{b}_r, \mathbf{b}_r^*)$, and $\mathcal{L}_{\text{c}} = \text{BCE}(p, y^*)$, and combined into the final objective $\mathcal{L} = \lambda_t \mathcal{L}_{\text{t}} + \lambda_r \mathcal{L}_{\text{r}} + \lambda_c \mathcal{L}_{\text{c}}$. In practice, we keep the pretrained backbone mostly frozen and perform lightweight adaptation on top of it, while fully training the task-specific projection, fusion, decoder, and prediction heads. To make the predictor robust to gaze noise, we train it on the five-sample expansion of GazeRefer-GQA (Sec. 3.2), where each of the five per-example gaze samples yields a distinct single-gaze training instance sharing the same ground-truth boxes.

4.3 Density-Equalized Warping

Given the predicted target region, related region, and reasoning type, we next construct a bandwidth-efficient visual input through adaptive warping. We first render the selected boxes into adaptive box-shaped Gaussian maps and aggregate them into a spatial importance map,

$$A(x, y) = G(\mathbf{b}_t)(x, y) + \hat{y}\, G(\mathbf{b}_r)(x, y), \tag{2}$$

where $G(\mathbf{b})$ renders box $\mathbf{b} = (c_x, c_y, w, h)$ as an adaptive box-shaped Gaussian centered at (c_x, c_y), with horizontal and vertical spreads proportional to its width and height, and $\hat{y} \in \{0, 1\}$ determines whether the related region is included. This yields a smooth spatial importance map that adapts to both region location and box geometry. Thus, Local Reasoning preserves only the target region, while Global Reasoning preserves both the target and related regions.

We then reduce A to axis-wise semantic densities $s_x(x) = \int_0^1 A(x, y)\, dy$ and $s_y(y) = \int_0^1 A(x, y)\, dx$, and convert them into positive density fields

$$\kappa_x(x) = \alpha + s_x(x), \qquad \kappa_y(y) = \alpha + s_y(y), \tag{3}$$

where $\alpha > 0$ provides a minimum background allocation level and controls the overall warping strength, so that higher semantic density induces denser sampling.

Rather than interpreting these profiles as probability distributions, we construct the warp by requiring each output sampling interval to carry approximately equal semantic mass. Let $u, v \in [0, 1]$

Method	Scale	Ratio	point			cross			text		
			Local	Global	Overall	Local	Global	Overall	Local	Global	Overall
Full-resolution reference	–	100%	46.6	37.6	42.1	53.5	36.4	45.0	58.0	37.5	47.8
JPEG compression	–	4.7%	35.3	28.4	31.9	37.1	**27.2**	32.2	40.9	26.5	33.7
	–	16.3%	40.6	34.0	37.3	44.8	33.6	39.2	44.3	30.6	37.5
	–	33.9%	43.0	35.4	39.2	46.9	34.6	40.7	45.3	31.6	38.5
Uniform resize	112	6.4%	32.8	25.5	29.1	31.7	24.0	27.9	30.2	23.8	27.0
	224	19.8%	39.7	31.6	35.7	44.9	30.9	37.9	44.2	30.2	37.2
	336	38.3%	42.9	34.5	38.7	47.7	33.9	40.8	47.4	33.4	40.4
GazeGPT	112	5.2%	34.8	18.9	26.8	31.3	16.9	24.1	28.3	18.2	23.2
	224	17.1%	37.7	28.3	33.0	36.3	27.4	31.9	32.9	27.9	30.4
	336	32.2%	36.0	33.4	34.7	36.4	32.7	34.5	31.6	32.2	31.9
ViCrop	71	5.2%	41.1	28.0	34.6	**44.5**	26.9	**35.7**	**45.0**	27.6	36.3
	146	17.2%	44.9	34.4	39.7	47.6	**34.0**	40.8	47.0	32.8	39.9
	226	35.9%	46.4	36.6	41.5	53.0	**36.7**	44.9	46.8	35.1	41.0
GazeWarp (ours)	112	6.2%	**43.1**	**30.4**	**36.7**	43.3	27.1	35.2	44.8	**29.6**	**37.2**
	224	18.7%	**47.5**	**35.1**	**41.3**	**52.5**	33.7	**43.1**	**57.3**	**35.6**	**46.5**
	336	34.9%	**47.2**	**37.2**	**42.2**	**56.4**	36.3	**46.3**	**58.6**	**36.1**	**47.3**

Table 1: Main results on GazeRefer-GQA (exact-match accuracy, %). "Ratio" is the mean per-image transmitted bytes relative to the original (App. C.4). Best constrained result per payload tier, input form, and metric block in bold; the full-resolution row is an unconstrained reference.

denote normalized coordinates in the warped image with source coordinates $x(u), y(v)$. Imposing equal mass per interval and taking the continuous limit yields the monotone coordinate mappings

$$\kappa_x(x(u))\frac{dx(u)}{du} = M_x, \qquad \kappa_y(y(v))\frac{dy(v)}{dv} = M_y, \tag{4}$$

where $M_x = \int_0^1 \kappa_x$ and $M_y = \int_0^1 \kappa_y$ are the total masses. Intuitively, a larger κ shrinks dx/du, so denser regions occupy more pixels after warping. Integrating gives strictly monotone cumulative allocation functions C_x, C_y, whose inverses yield the source coordinates $x^*(u) = C_x^{-1}(u)$ and $y^*(v) = C_y^{-1}(v)$. We use them to resample the source by bilinear interpolation (Jaderberg et al., 2015), and additionally resize the warped result to a fixed transmission resolution,

$$I^{\text{trans}} := \text{Resize}(\text{BilinearSample}(I,\, x^*(u),\, y^*(v)))\,. \tag{5}$$

Full derivations and numerical solution details are provided in App. B.1. This design reallocates more spatial resolution to task-relevant content while preserving sufficient global context for downstream gaze-guided visual question answering (Avidan & Shamir, 2023; Wang et al., 2008).

5 Experiments

5.1 Experimental Setup

Implementation details. Unless otherwise stated, we use Qwen2.5-VL-7B-Instruct (Qwen Team, 2025a) as the default backend VLM, and additionally evaluate cross-model generalization with Qwen3-VL-2B/8B/32B-Instruct (Qwen Team, 2025b) and LLaVA-1.5-7B (Liu et al., 2023). We compare GazeWarp against five transmission baselines under approximately matched payload budgets: (i) uniform resize, (ii) GazeGPT (Konrad et al., 2024), which transmits a high-fidelity square crop centered on the gaze-indicated location, (iii) JPEG compression at the original resolution, (iv) ViCrop (Zhang et al., 2025), a two-image crop-based approach, and (v) a full-resolution reference that ignores the transmission budget. All methods operate as frontends before the same frozen backend VLM, so accuracy differences reflect how the transmitted visual content is allocated under a comparable payload rather than to backend fine-tuning. We consider three transmission scales ($112/224/336$), whose per-image transmitted bytes relative to the original image are reported in the "Ratio" column of Table 1 and average roughly $5\%/18\%/35\%$ across methods, and three downstream input forms that differ only in how the same gaze position is conveyed to the backend—an overlaid dot (point), a crosshair (cross), or textual coordinates appended to the question with no marker drawn (text); see App. C.3.

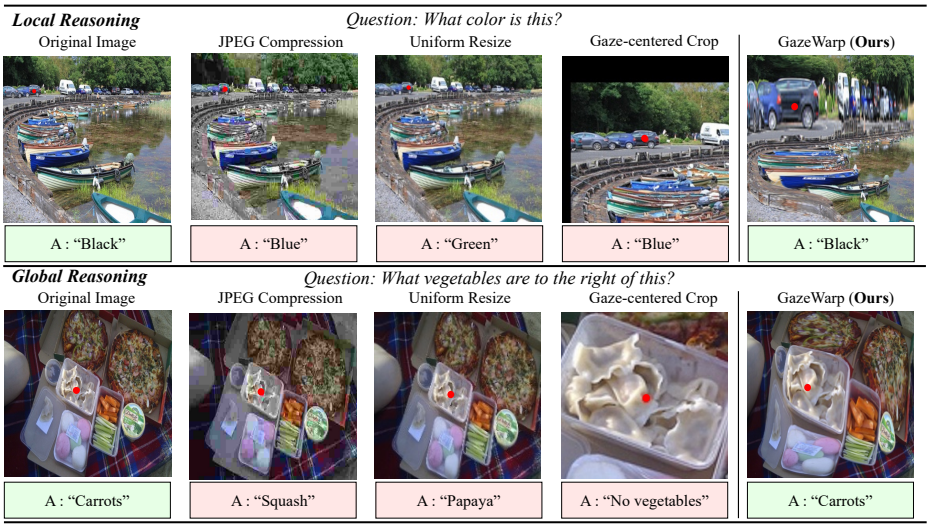

Figure 4: Qualitative comparison under the main setting (Qwen2.5-VL-7B-Instruct, scale $= 224$). At a comparable payload, GazeWarp retains more task-relevant detail than JPEG, uniform resize, and GazeGPT. Example image from GQA (Hudson & Manning, 2019).

Evaluation protocol. Following the gaze model in Sec. 3, every test query is associated with $K = 5$ gaze samples (the object-center anchor and four Gaussian-perturbed samples around the referred object), and all reported accuracies are averaged over these five samples, so that each number reflects expected performance under simulated gaze noise. We report strict exact-match accuracy (%) after light answer normalization; details are in App. C.

5.2 Main Results on GazeRefer-GQA

Table 1 reports the main results. GazeWarp delivers the strongest overall performance across nearly all transmission scales and input forms, with especially clear gains at medium and high scales. The improvements hold on both Local and Global subsets, showing that it preserves detail not only around the gazed target but also around the related region required for Global Reasoning. Under approximately matched payload budgets it also beats JPEG compression and ViCrop in most settings, indicating that task-adaptive reallocation of resolution is more effective than uniform compression or simple target-centric cropping. At the highest scale GazeWarp approaches, and in several cases exceeds, the full-resolution reference, suggesting that adaptive warping can yield a more task-focused input than the full image. All results in this section use Qwen2.5-VL-7B-Instruct as the backend; results for the other backend VLMs exhibit the same trend and are reported in App. D.5. Figure 4 shows representative qualitative examples; more are in App. E.

5.3 Ablation Study

We ablate the core design choices of the region predictor under a representative setting (Qwen2.5-VL-7B-Instruct, scale $= 224$, point input), so that all variants differ only in the component under study: how the spatial importance map is constructed, and how the gaze cue is injected into the predictor.

Spatial importance map construction. Table 2a compares choices for building the importance map. Our default first predicts target and related boxes and renders them into adaptive box-shaped Gaussian maps, which outperforms a hard binary box (abrupt warping boundaries), a fixed-point Gaussian (size/aspect mismatch), and a directly predicted map (weaker than explicit box prediction). The cross-attention-derived map (Dalal et al., 2026) performs worst, showing that for ambiguous gaze-guided questions downstream attention is a less reliable spatial signal than explicit region prediction.

Variant	Local	Global	Overall
Binary box fill	39.0	31.9	35.5
Fixed-point Gaussian	39.1	32.3	35.7
Cross-attn.-derived map	37.4	29.1	33.3
Direct predicted map	46.7	33.3	40.0
Adaptive box Gauss. (ours)	**47.5**	**35.1**	**41.3**

(a) Spatial importance map construction.

Variant	Vis.	Txt.	Tok.	Overall
Learnable token			✓	41.1
Visual only	✓			39.7
Text only		✓		40.2
Ours	✓	✓		**41.3**

(b) Gaze input design.

Table 2: Ablations on GazeRefer-GQA (Qwen2.5-VL-7B-Instruct, scale $= 224$, point input); all values are exact-match accuracy (%). (a) How the spatial importance map is built. (b) How the gaze cue is injected into the predictor ("Vis." = visual mark, "Txt." = text coordinate, "Tok." = learnable token); per-subset Local/Global numbers are in App. D.1.

Baseline	Scale	Win	Tie	Lose
Full-res. ref.	–	42.4	28.7	28.9
Uniform resize	224	38.3	28.4	33.4
JPEG compression	–	41.8	26.9	31.3
GazeGPT	224	39.6	28.5	31.8
ViCrop	146	38.3	29.8	31.9

Table 3: GPT-RANK results on VOILA-COCO (Qwen2.5-VL-7B-Instruct) under approximately matched payload budgets, using the *point* gaze input form. "Scale" denotes the transmission scale (target resolution) of each baseline where applicable. Each row reports GazeWarp's Win / Tie / Lose rate (%) against the given baseline.

Gaze input design. Table 2b ablates how the gaze cue is injected into the region predictor, comparing a visual mark overlaid on the image, text coordinates, and a dedicated learnable token. Our final design combines the visual mark and text coordinate and achieves the best accuracy (41.3%, ahead of a learnable token at 41.1%, text only at 40.2%, and visual only at 39.7%): marking the gazed location while retaining its coordinate in language may provide a more informative grounding signal than either alone.

5.4 Generalization to Other Gaze-Guided Datasets

To test whether GazeWarp generalizes beyond GazeRefer-GQA, we further evaluate it on VOILA-COCO (Yan et al., 2024), a gaze-grounded visual question answering set that pairs COCO images with ambiguous queries and simulated gaze traces. As its answers are open-ended and ill-suited to exact match, we adopt the pairwise GPT-RANK protocol of Voila-A (Yan et al., 2024) and report GazeWarp's Win/Tie/Lose rate against each baseline (Table 3). Under the matched budget and the *point* gaze input, GazeWarp is preferred over the full-resolution reference and all transmission baselines (Win exceeding Lose in every comparison), indicating that its benefits transfer to a different gaze-guided dataset.

6 Conclusion

We studied bandwidth-constrained image transmission for gaze-guided multimodal interaction in wearable XR. We introduced GazeRefer-GQA, a benchmark pairing ambiguous deictic questions with gaze cues, and GazeWarp, a plug-and-play frontend that reallocates resolution to gaze-indicated target and query-dependent related regions. Under approximately matched payload budgets and simulated gaze noise, GazeWarp generally outperforms competing methods across heterogeneous backend VLMs, supporting task-adaptive resolution reallocation for bandwidth-efficient XR interaction.

AI USE STATEMENT

Generative AI tools were used exclusively for benchmark construction and language editing. During benchmark construction, a large language model was used to rewrite source questions into gaze-dependent deictic forms. All generated rewrites were subjected to rule-based automatic checks, and a randomly sampled subset was additionally inspected manually, as described in App. A.3. Generative AI was also used to improve the clarity and readability of the manuscript. It was not used to formulate the research questions, develop the proposed method, design or conduct experiments, analyze or interpret the results, or generate experimental findings. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Our benchmark construction pipeline, including question rewriting, automatic filtering rules, and gaze simulation, is described in App. A. The architecture, training objective, full derivations, and hyperparameters of GazeWarp are specified in Apps. B and C, with the training and inference pipelines summarized in Algorithms 1 and 2. Evaluation protocols, including payload-tier matching, downstream gaze input forms, and answer normalization, are detailed in App. C.

REFERENCES

Isayas Berhe Adhanom, Paul MacNeilage, and Eelke Folmer. Eye tracking in virtual reality: a broad review of applications and challenges. *Virtual Reality*, 27(2):1481–1505, 2023. doi: 10.1007/s10055-022-00738-z.

Ahmad Alhilal, Ze Wu, Yuk Hang Tsui, and Pan Hui. Fovoptix: Human vision-compatible video encoding and adaptive streaming in vr cloud gaming. In *Proceedings of the 15th ACM Multimedia Systems Conference*, pp. 67–77, 2024. doi: 10.1145/3625468.3647612.

Shai Avidan and Ariel Shamir. Seam carving for content-aware image resizing. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 609–617. Association for Computing Machinery, 2023. doi: 10.1145/3596711.3596776.

Qinyu Chen and Jiawen Qi. Eye gaze tells you where to compute: Gaze-driven efficient vlms, 2025. URL https://arxiv.org/abs/2509.16476.

Shaoyu Chen, Budmonde Duinkharjav, Xin Sun, Li-Yi Wei, Stefano Petrangeli, Jose Echevarria, Claudio Silva, and Qi Sun. Instant reality: Gaze-contingent perceptual optimization for 3d virtual reality streaming. *IEEE Transactions on Visualization and Computer Graphics*, 28(5):2157–2167, 2022. doi: 10.1109/tvcg.2022.3150522.

Dwip Dalal, Gautam Vashishtha, Utkarsh Mishra, Jeonghwan Kim, Madhav Kanda, Hyeonjeong Ha, Svetlana Lazebnik, Heng Ji, and Unnat Jain. Constructive distortion: Improving MLLMs with attention-guided image warping. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2510.09741.

Steven Gao, Jeffrey Liu, Qinjun Jiang, Finn Sinclair, William Sentosa, Brighten Godfrey, and Sarita Adve. Xrgo: Design and evaluation of rendering offload for low-power extended reality devices. In *Proceedings of the 16th ACM Multimedia Systems Conference*, pp. 124–135, 2025. doi: 10.1145/3712676.3714444.

Kai Huang, Hao Zou, Ye Xi, BoChen Wang, Zhen Xie, and Liang Yu. Ivtp: Instruction-guided visual token pruning for large vision-language models. In *European Conference on Computer Vision*, pp. 214–230. Springer, 2024. doi: 10.1007/978-3-031-72643-9_13.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019. doi: 10.1109/cvpr.2019.00686.

Muhammet Ilaslan, Chenan Song, Joya Chen, Difei Gao, Weixian Lei, Qianli Xu, Joo Lim, and Mike Shou. Gazevqa: A video question answering dataset for multiview eye-gaze task-oriented collaborations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10462–10479, 2023. doi: 10.18653/v1/2023.emnlp-main.648.

Gazi Karam Illahi, Matti Siekkinen, Teemu Kämäräinen, and Antti Ylä-Jääski. Foveated streaming of real-time graphics. In *Proceedings of the 12th ACM Multimedia Systems Conference*, pp. 214–226, 2021. doi: 10.1145/3458305.3463383.

Shun Inadumi, Seiya Kawano, Akishige Yuguchi, Yasutomo Kawanishi, and Koichiro Yoshino. A gaze-grounded visual question answering dataset for clarifying ambiguous japanese questions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 558–571, 2024. doi: 10.63317/42hxdk7mehhr.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2017–2025, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/33ceb07bf4eeb3da587e268d663aba1a-Abstract.html.

Qinjun Jiang, Yihan Pang, William Sentosa, Steven Gao, Muhammad Huzaifa, Jeffrey Zhang, Javier Perez-Ramirez, Dibakar Das, David Gonzalez-Aguirre, Brighten Godfrey, et al. Remotevio: Offloading head tracking in an end-to-end xr system. In *Proceedings of the 16th ACM Multimedia Systems Conference*, pp. 101–112, 2025. doi: 10.1145/3712676.3714442.

Teemu Kämäräinen and Matti Siekkinen. Foveated spatial compression for remote rendered virtual reality. In *Proceedings of the First Workshop on Metaverse Systems and Applications*, pp. 7–13, 2023. doi: 10.1145/3597063.3597359.

Syrine Khelifi and Alexis Morris. Mixed reality iot smart environments with large language model agents. In *2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS)*, pp. 1–7. IEEE, 2024. doi: 10.1109/ichms59971.2024.10555610.

Robert Konrad, Nitish Padmanaban, J. Gabriel Buckmaster, Kevin C. Boyle, and Gordon Wetzstein. Gazegpt: Augmenting human capabilities using gaze-contingent contextual ai for smart eyewear, 2024. URL https://arxiv.org/abs/2401.17217.

Jaewook Lee, Jun Wang, Elizabeth Brown, Liam Chu, Sebastian S Rodriguez, and Jon E Froehlich. Gazepointar: A context-aware multimodal voice assistant for pronoun disambiguation in wearable augmented reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–20, 2024. doi: 10.1145/3613904.3642230.

Jaewook Lee, Tianyi Wang, Jacqui Fashimpaur, Naveen Sendhilnathan, and Tanya R Jonker. Walkie-talkie: Exploring longitudinal natural gaze, llms, and vlms for query disambiguation in xr. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–9, 2025. doi: 10.1145/3706599.3720236.

Shuai Li, Jian Xu, Xiao-Hui Li, Chao Deng, and Lin-Lin Huang. Qg-vtc: Question-guided visual token compression in mllms for efficient vqa. *arXiv preprint arXiv:2504.00654*, 2025a. doi: 10.48550/ARXIV.2504.00654. URL https://arxiv.org/abs/2504.00654.

Yuyang Li, Devika Gumaste, Mehmet Kerem Turkcan, Javad Ghaderi, Gil Zussman, and Zoran Kostic. Distributed vlms: Efficient vision-language processing through cloud-edge collaboration. In *2025 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pp. 280–286. IEEE Computer Society, 2025b. doi: 10.1109/percomworkshops65533.2025.00078.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Jing Liu, Lizong Zhang, Chong Mu, Guangxi Lu, Ben Zhang, and Junsong Li. Question-guided multigranular visual augmentation for knowledge-based visual question answering. *Computer Vision and Image Understanding*, 263:104569, 2026. doi: 10.1016/j.cviu.2025.104569.

Bipul Mohanto, ABM Tariqul Islam, Enrico Gobbetti, and Oliver Staadt. An integrative view of foveated rendering. *Computers & Graphics*, 102:474–501, 2022. doi: 10.1016/j.cag.2021.10.010.

Sounak Mondal, Seoyoung Ahn, Zhibo Yang, Niranjan Balasubramanian, Dimitris Samaras, Gregory Zelinsky, and Minh Hoai. Look hear: Gaze prediction for speech-directed human attention, 2024. URL https://arxiv.org/abs/2407.19605.

Alexander Plopski, Teresa Hirzle, Nahal Norouzi, Long Qian, Gerd Bruder, and Tobias Langlotz. The eye in extended reality: A survey on gaze interaction and eye tracking in head-worn extended reality. *ACM Computing Surveys (CSUR)*, 55(3):1–39, 2023. doi: 10.1145/3491207.

Qwen Team. Qwen2.5-vl. https://qwenlm.github.io/blog/qwen2.5-vl/, 2025a.

Qwen Team. Qwen3-VL technical report, 2025b. URL https://arxiv.org/abs/2511.21631.

Kaavya Rekanar, John M Joyce, Martin Hayes, and Ciarán Eising. Drivqa: A gaze-based dataset for visual question answering in driving scenarios. *Data in Brief*, 59:111367, 2025. doi: 10.1016/j.dib.2025.111367.

Jun Rekimoto. Gazellm: Multimodal llms incorporating human visual attention, 2025. URL https://arxiv.org/abs/2504.00221.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 658–666, 2019. doi: 10.1109/cvpr.2019.00075.

Max Schwarz and Sven Behnke. Foveated compression for immersive telepresence visualization, 2025. URL https://arxiv.org/abs/2510.19848.

Ahmed Sharshar, Latif U Khan, Waseem Ullah, and Mohsen Guizani. Vision-language models for edge networks: A comprehensive survey. *IEEE Internet of Things Journal*, 12(16):32701–32724, 2025. doi: 10.1109/JIOT.2025.3579032.

Gaurav Shinde, Anuradha Ravi, Emon Dey, Shadman Sakib, Milind Rampure, and Nirmalya Roy. A survey on efficient vision-language models. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 15(3):e70036, 2025. doi: 10.1002/widm.70036.

Ekta Sood, Fabian Kögel, Florian Strohm, Prajit Dhar, and Andreas Bulling. Vqa-mhug: A gaze dataset to study multimodal neural attention in visual question answering. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 27–43, 2021. doi: 10.18653/v1/2021.conll-1.3.

Ekta Sood, Fabian Kögel, Philipp Müller, Dominike Thomas, Mihai Bâce, and Andreas Bulling. Multimodal integration of human-like attention in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2648–2658, 2023. doi: 10.1109/cvprw59228.2023.00265.

Sruti Srinidhi, Edward Lu, and Anthony Rowe. Xair: An xr platform that integrates large language models with the physical world. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 759–767. IEEE, 2024. doi: 10.1109/ismar62088.2024.00091.

Yiliu Tang, Jason Situ, Andrea Yaoyun Cui, Mengke Wu, and Yun Huang. Llm integration in extended reality: A comprehensive review of current trends, challenges, and future perspectives. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–24, 2025. doi: 10.1145/3706598.3714224.

Haopeng Wang, Haiwei Dong, and Abdulmotaleb El Saddik. Immersive multimedia communication: state-of-the-art on extended reality streaming. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(7):1–33, 2025a. doi: 10.1145/3721292.

Jingyan Wang, Yang Zhao, Haotian Mao, and Xubo Yang. Survey of large language models in extended reality: Technical paradigms and application frontiers. *arXiv preprint arXiv:2508.03014*, 2025b. doi: 10.48550/ARXIV.2508.03014. URL https://arxiv.org/abs/2508.03014.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025c. doi: 10.1609/aaai.v39i8.32852.

Wenbin Wang, Yongcheng Jing, Liang Ding, Yingjie Wang, Li Shen, Yong Luo, Bo Du, and Dacheng Tao. Retrieval-augmented perception: High-resolution image perception meets visual rag. *arXiv preprint arXiv:2503.01222*, 2025d. doi: 10.48550/ARXIV.2503.01222. URL https://arxiv.org/abs/2503.01222.

Yu-Shuen Wang, Chiew-Lan Tai, Olga Sorkine, and Tong-Yee Lee. Optimized scale-and-stretch for image resizing. In *ACM SIGGRAPH Asia 2008 papers*, pp. 1–8, 2008. doi: 10.1145/1457515.1409071.

Zeyu Wang, Yuanchun Shi, Yuntao Wang, Yuchen Yao, Kun Yan, Yuhan Wang, Lei Ji, Xuhai Xu, and Chun Yu. G-voila: gaze-facilitated information querying in daily scenarios. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(2):1–33, 2024a. doi: 10.1145/3659623.

Zhan Wang, Lin-Ping Yuan, Liangwei Wang, Bingchuan Jiang, and Wei Zeng. Virtuwander: Enhancing multi-modal interaction for virtual tour guidance through large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–20, 2024b. doi: 10.1145/3613904.3642235.

Zhi-Min Wang, Mao-Hang Rao, Shang-Hua Ye, Wei-Tao Song, and Feng Lu. Towards spatial computing: recent advances in multimodal natural interaction for extended reality headsets. *Frontiers of Computer Science*, 19(12):1912708, 2025e. doi: 10.1007/s11704-025-41123-8.

Yushi Wei, Rongkai Shi, Difeng Yu, Yihong Wang, Yue Li, Lingyun Yu, and Hai-Ning Liang. Predicting gaze-based target selection in augmented reality headsets based on eye and head endpoint distributions. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pp. 1–14, 2023. doi: 10.1145/3544548.3581042.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024. doi: 10.1109/cvpr52733.2024.01243.

Junfei Xie, Peng Pan, and Xulong Zhang. Head-aware visual cropping: Enhancing fine-grained vqa with attention-guided subimage. In *2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026. doi: 10.1109/ICASSP55912.2026.11463583. URL https://arxiv.org/abs/2601.22483.

Kun Yan, Zeyu Wang, Lei Ji, Yuntao Wang, Nan Duan, and Shuai Ma. Voila-a: Aligning vision-language models with user's gaze attention. *Advances in neural information processing systems*, 37:1890–1918, 2024. doi: 10.52202/079017-0060.

Runpeng Yu, Weihao Yu, and Xinchao Wang. Attention prompting on image for large vision-language models. In *European Conference on Computer Vision*, pp. 251–268. Springer, 2024. doi: 10.1007/978-3-031-73404-5_15.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Mllms know where to look: Training-free perception of small visual details with multimodal llms. *arXiv preprint arXiv:2502.17422*, 2025. doi: 10.48550/ARXIV.2502.17422. URL https://arxiv.org/abs/2502.17422.

Qing Zhang and Jun Rekimoto. Look and talk: Seamless ai assistant interaction with gaze-triggered activation. In *Proceedings of the Augmented Humans International Conference 2025*, pp. 418–421, 2025. doi: 10.1145/3745900.3746107.