LOOKANYTHING: A GENERALIST FOUNDATION MODEL FOR GAZE TARGET DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Human gaze provides a rich window into attention and intent, making its understanding essential for applications ranging from social signal analysis to human-robot interaction. Existing methods typically employ highly customized architectures relying on auxiliary inputs like head crops or depth maps, causing cascading errors and inference overhead. Trained on single domains and confined to isolated gaze target detection with heatmap outputs, they lack flexibility and generalization for diverse interactive scenarios. To bridge this gap, we present **LookAnything**, a generalist foundation model unifying diverse gaze-centric tasks. Benefiting from large-scale multi-task cross-domain training, LookAnything flexibly adapts to various settings, generating diverse outputs (points, boxes, heatmaps, and language) and supporting broad gaze-centric localization, description, and reasoning. At its core, we introduce learnable gaze-prior queries to promote gaze-aware representation learning, eliminating the need for auxiliary external inputs. To facilitate this paradigm, we construct **HoliGaze**, a comprehensive corpus and benchmark comprising 311K question-answer pairs across 5 capability dimensions and 9 distinct tasks. Experiments demonstrate that LookAnything consistently outperforms 33 leading vision-language models on holistic gaze understanding, while remaining highly competitive with dedicated specialists across 5 conventional target detection benchmarks. Deployment on a Unitree G1 robot further validates its strong out-of-domain generalization and flexible adaptation in real-world scenarios. Our codes, model, and dataset will be released. The project website is available at `https://anonymous1541paper.github.io/iclr2027-1541`.

1 INTRODUCTION

Human gaze is a fundamental nonverbal cue that offers a rich window into human attention, cognitive state, and intent (Zhou et al., 2025b; Tafasca et al., 2024a). As a cognitive signal, gaze behavior reveals underlying preferences and social dynamics, which is essential in diverse domains such as social signal analysis and human–robot interaction (Zuo et al., 2026; Saran et al., 2018). Driven by these demands, the core visual task of *gaze target detection* has emerged to identify the exact spatial location a person is focusing on (Recasens et al., 2015; Chong et al., 2020). This capability is crucial for developing responsive embodied agents and enabling human-centric interactive applications.

Existing specialist methods typically employ highly customized architectures relying on auxiliary inputs like head crops or depth maps (Tonini et al., 2022; Gupta et al., 2022), which introduce cascading errors and inference overhead (Ryan et al., 2025; Tu et al., 2022). They are often trained on single domains and strictly confined to the isolated task of predicting gaze target heatmaps (Miao et al., 2023; Song et al., 2024). Such coarse outputs lack instance-level semantics, rendering them inflexible and difficult to generalize across diverse interactive scenarios and downstream reasoning tasks. While recent efforts have incorporated gaze to vision-language models (VLMs) (Wang et al., 2026a; Mathew et al., 2026), they remain limited in task scope and underexploit gaze-centric priors.

To bridge this gap, we present **LookAnything**, a generalist foundation model that unifies diverse gaze-centric tasks within a single multimodal framework. Benefiting from large-scale multi-task and cross-domain training, LookAnything flexibly adapts to various settings, generating diverse outputs—including points, bounding boxes, heatmaps, and natural language—while supporting broad gaze-centric localization, description, and reasoning. As shown in Figure 1, rather than relying on

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.

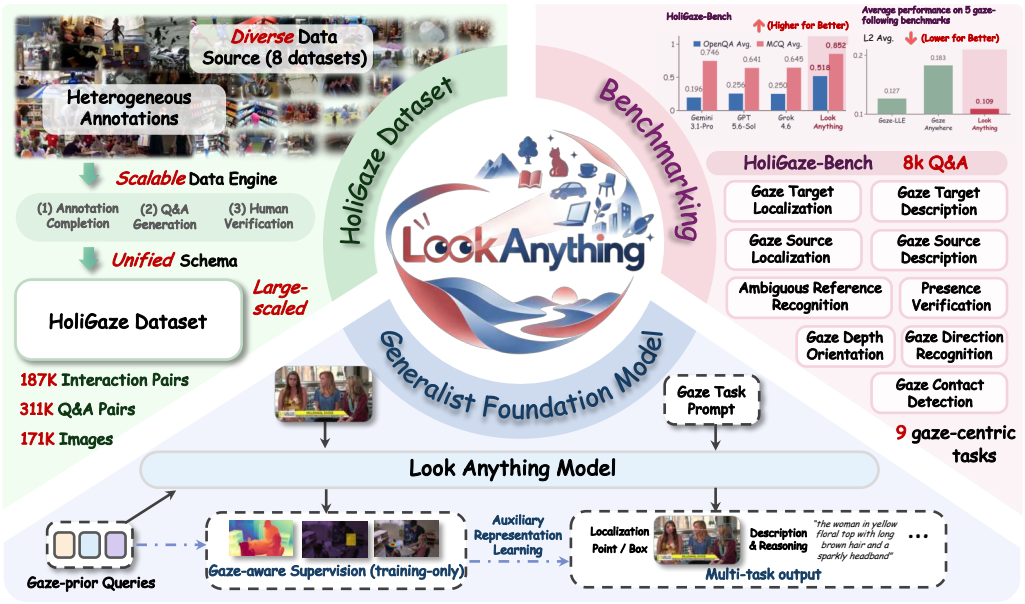


Figure 1: **LookAnything and HoliGaze Overview.** The proposed HoliGaze dataset comprises 311K Q&A samples across 9 distinct tasks, serving as comprehensive resource for holistic gaze understanding. LookAnything is a generalist foundation model that demonstrates strong performance across diverse gaze-centric tasks.

external auxiliary inputs, we introduce a set of learnable gaze-prior queries that internalize scene depth, source-head grounding, and target spatial geometry directly into the latent space, which promotes intrinsic gaze-aware representation learning.

Facilitating this unified paradigm demands comprehensive data and evaluation protocols. However, existing benchmarks remain fragmented and limited in scope (Recasens et al., 2015; Cao et al., 2026; Wang et al., 2026a). To address this, we construct **HoliGaze**, a comprehensive corpus and benchmark comprising 311K question-answer pairs across 5 capability dimensions and 9 gaze-centric tasks. Extensive experiments demonstrate that LookAnything substantially outperforms 33 leading VLMs on HoliGaze-Bench, while remaining highly competitive with dedicated specialists across 5 conventional benchmarks. Finally, deploying LookAnything on a real-world Unitree G1 robot validates its strong out-of-domain generalization and flexible adaptation to diverse physical scenarios.

Our key contributions are as follows:

- We introduce LookAnything, a generalist foundation model that internalizes gaze priors via large-scale multi-task and cross-domain learning, enabling flexible adaptation across various scenarios, task settings, and output formats without external auxiliary inputs.
- We construct HoliGaze, the first comprehensive corpus and benchmark for holistic gaze understanding. Standardizing 311K Q&A pairs across 5 capability dimensions and 9 distinct tasks, it addresses the fragmentation of existing datasets to facilitate large-scale multi-task learning and systematic evaluation.
- LookAnything substantially outperforms 33 leading VLMs on HoliGaze-Bench and remains highly competitive with specialists on 5 conventional gaze target detection benchmarks. Real-world deployment on a Unitree G1 robot further validates its strong out-of-domain generalization to diverse physical scenarios.

2 The HoliGaze Dataset

We propose **HoliGaze**, a unified and comprehensive dataset for holistic gaze understanding. It comprises **HoliGaze-Corpus** (303K training samples) and **HoliGaze-Bench** (8K human-verified evaluation samples), totaling 311K question–answering (Q&A) pairs.



Figure 2: **Task Taxonomy and Q&A Examples of HoliGaze.** Each panel represents a type of gaze-centric task, with panels of the same color belonging to the same capability dimension.

2.1 Task Taxonomy

As illustrated in Figure 2, HoliGaze defines 9 tasks grouped into 5 dimensions. Open-ended Q&A (openQA) covers generative localization and description, while multiple-choice Q&A (MCQ) supports stable automatic scoring.

1. Gaze Target Understanding. Given a textual or box reference of the gaze source, *Gaze Target Localization (GTL)* grounds the target as a 2D point (GTL-T2P) or bounding box (GTL-T2B). *Gaze Target Description (GTD)* requires a natural-language description of the attended entity. Training additionally includes a box-referred variant (GTL-B2P).

2. Gaze Source Understanding. Given a textual reference to the target, *Gaze Source Localization (GSL)* predicts the head boxes of all observers, and *Gaze Source Description (GSD)* describes their visual attributes.

3. Gaze Spatial Reasoning. *Gaze Depth Orientation (GDO)* asks whether the observer's head is closer to or farther from the camera than the target. *Gaze Direction Recognition (GDR)* classifies the gaze into four camera-relative directions.

4. Gaze Social Interaction. *Gaze Contact Detection (GCD)* is a binary MCQ asking whether a specified person is making eye contact with the camera.

5. Gaze Query Answerability. Models must refuse ill-posed queries. *Ambiguous Reference Recognition (ARR)* requires refusal when the referring expression matches multiple entities, and *Presence Verification (PV)* requires refusal when the target lies outside the frame.

2.2 Scalable Data Engine

We design a multi-stage data engine that converts heterogeneous gaze annotations from 8 source datasets into unified VQA pairs, as illustrated in Figure 3. First, we unify records into a shared schema and use off-the-shelf models (SAM3, Depth Anything v3, and Qwen3-VL) to complete missing fields such as head masks, scene depth, and referring expressions. Next, task-specific templates instantiate the schema into openQA and MCQ formats. We synthesize challenging MCQ distractors via entity expansion, attribute perturbation, and box permutation, while filtering ambiguous or out-of-frame targets for answerability tasks (ARR/PV). Finally, to prevent data leakage, we strictly follow each source's native split to curate the training corpus and evaluation benchmark. HoliGaze-Bench undergoes rigorous instance-level human verification, yielding a low modification rate (0.70%) that indicates strong agreement between the automated outputs and human judgment. We defer the comprehensive details of dataset construction to Appendix D.

Table 1: **Comparison of HoliGaze with Existing Gaze-Following Datasets.** Here, † denotes full-pool video frames; *Samples* counts instances in native formats (gaze annotations or Q&A pairs); *Gaze Pairs* counts source-target interactions.

Dataset	Publication	#Samples	#Images	#Gaze Pairs	#Tasks	Referring Types		Output Modalities			QA Formats	
						Text	Box	Box	Point	Text	MCQ	openQA
GazeFollow (2015)	NeurIPS 2015	130K	122K	130K	1	✗	✓	✗	✓	✗	✗	✗
VideoCoAtt (2018)	CVPR 2018	138K	492K†	138K	1	✗	✗	✓	✓	✗	✗	✗
VAT (2020)	CVPR 2020	165K	71K†	165K	2	✗	✓	✗	✓	✗	✗	✗
GOO (2021)	CVPRW 2021	172K	202K	172K	2	✗	✓	✓	✓	✗	✗	✗
ChildPlay (2023)	ICCV 2023	258K	121K†	258K	2	✗	✓	✗	✓	✗	✗	✗
GazeHOI (2024a)	NeurIPS 2024	56K	44K	56K	2	✗	✓	✗	✓	✓	✗	✗
VL4Gaze (2026a)	arXiv 2025	489K	124K	130K	4	✓	✗	✗	✓	✓	✗	✓
Gaze-Co (2026)	CVPR 2026	120K	120K	120K	1	✓	✗	✓	✓	✗	✗	✗
HoliGaze (Ours)		**311K**	**171K**	**187K**	**9**	✓	✓	✓	✓	✓	✓	✓

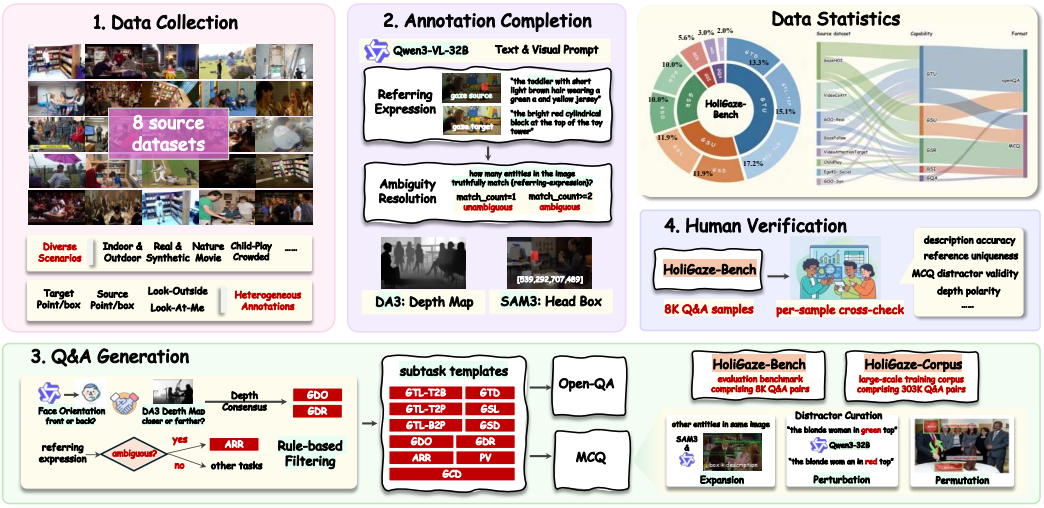


Figure 3: **Data Engine and Dataset Statistics of HoliGaze.** The pipeline converts eight source datasets into unified VQA samples through data collection, annotation completion, template-based Q&A generation, and human verification.

2.3 Dataset Statistics

HoliGaze-Bench comprises 8K Q&A pairs over 6,143 images, quota-sampled across 9 gaze-centric tasks: GTL-T2P (15.2%), GTL-T2B (15.2%), GTD (15.2%), GSL (11.9%), GSD (11.9%), GDO (10.0%), GDR (10.0%), GCD (5.6%), ARR (2.5%), and PV (2.5%), with a 1:1 openQA-to-MCQ ratio. **HoliGaze-Corpus** contributes 303K training samples. Jointly, they yield 311K Q&A pairs, 171K images, and 187K gaze interactions. As shown in Table 1, HoliGaze uniquely supports both referring types, all three output modalities, and both QA formats across 9 tasks, bridging the gap for holistic gaze understanding. Detailed dataset statistics and distributions are provided in Appendix C.

3 The LookAnything Model

3.1 Model Architecture

We formulate holistic gaze understanding as instruction-conditioned autoregressive generation. Given an image I and a language prompt P, LookAnything models the conditional distribution over an open-form answer Y: $p_\theta(Y \mid I, P) = \prod_{t=1}^{N_Y} p_\theta(y_t \mid I, P, y_{<t})$, where Y unifies localization, description, and reasoning into a single token sequence.

As illustrated in Figure 4, LookAnything builds upon the Qwen3-VL (Bai et al., 2025) architecture. Modality-specific encoders first map the image and prompt into latent tokens. To explicitly inject gaze-aware spatial priors, we introduce learnable continuous queries $\mathcal{Q} = [Q^d, Q^h, Q^g]$ designated

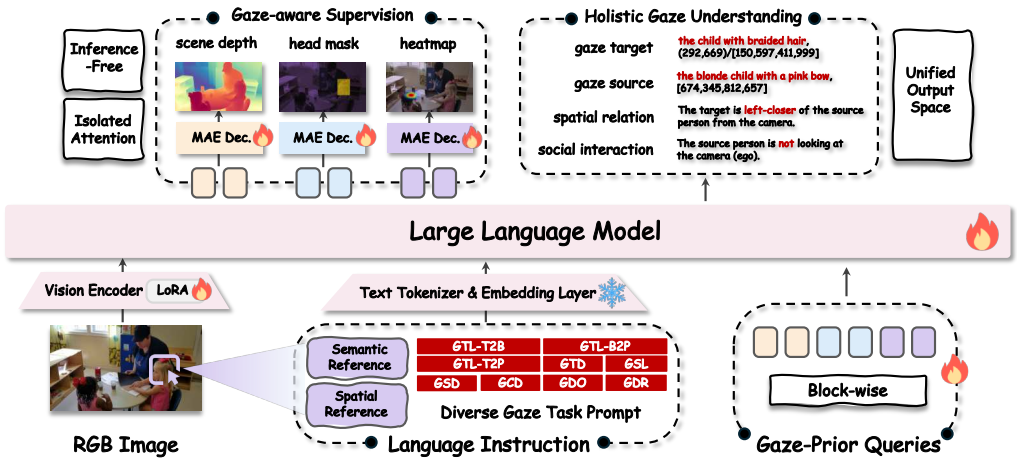


Figure 4: **Model Architecture Overview.** LookAnything appends learnable gaze-prior queries to multimodal tokens to learn gaze-aware representations via auxiliary geometric supervision (depth, head grounding, and target geometry). The language head then autoregressively decodes the answer in a unified output space. The auxiliary branches can be skipped during inference.

for scene depth, source head grounding, and target gaze geometry. These queries are appended to the multimodal embeddings.

The large language model backbone $\mathcal{M}$ integrates these inputs into compact latent gaze-context representations $H = \mathcal{M}(\mathcal{E}_v(I), \mathcal{E}_t(P) \mid \mathcal{Q})$. A set of lightweight auxiliary decoders $\mathcal{D}$ projects these query states into explicit geometric predictions, while the language head autoregressively generates the answer Y. During inference, the auxiliary decoders can be skipped to save computation. Detailed architectural configurations are provided in Appendix B.3.

3.2 Gaze Context-Aware Representation Learning

Instead of relying on explicit external inputs (e.g., depth maps or head crops) which introduce cascading errors and latency, LookAnything internalizes critical gaze priors into latent representations via auxiliary supervision. Detailed loss formulations are deferred to Appendix B.4.

Scene Depth Estimation. Scene depth provides essential 3D structural context, which is crucial for understanding the spatial relationship between the gaze source and the target. The depth query states H^d are projected by a lightweight MAE-style (He et al., 2022) decoder $\mathcal{D}_d$ to predict the relative monocular scene depth $\hat{D} = \mathcal{D}_d(H^d)$. This prediction is supervised using a scale-invariant logarithmic loss to emphasize relative structural depth.

Source Head Grounding. Accurate head localization is a prerequisite for gaze target estimation. The head query states H^h are decoded by an MAE-style decoder $\mathcal{D}_h$ to output a binary source head mask $\hat{M} = \mathcal{D}_h(H^h)$, which is optimized via a combination of Binary Cross-Entropy and soft Dice loss to mitigate foreground-background imbalance.

Target Spatial Prediction. The spatial distribution of the gaze target is a vital geometric cue. By interacting with the contextualized visual features F_v via cross-attention, the gaze queries H^g are decoded by an MAE-style decoder into a dense gaze heatmap $\hat{A} = \mathcal{D}_a(H^g, F_v)$, supervised by KL-divergence and Mean Squared Error. Concurrently, dedicated regression heads predict continuous geometric primitives, specifically a normalized target point $\hat{p} = \mathcal{D}_p(H^g)$ and a bounding box $\hat{b} = \mathcal{D}_b(H^g)$, optimized via Huber and GIoU losses.

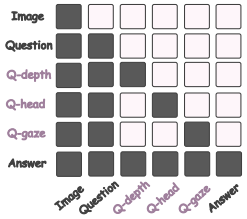


Figure 5: Structurally Disentangled Block-wise Attention.

Table 2: **Results on HoliGaze-Bench.** Group-wise best/second-best are in color / color ; overall best/second-best are **bolded**/underlined.

Methods	Avg.		GTL		GTD			GSL	GSD			GDO	GDR	GCD	ARR	PV
	openQA	MCQ	T2P	T2B	MCQ	openQA		openQA	MCQ	openQA		MCQ	MCQ	MCQ	openQA	openQA
			$L_2\downarrow$	mIoU↑	Acc↑	R-L↑	B-4↑	mIoU↑	F1↑	R-L↑	B-4↑	Acc↑	Acc↑	Acc↑	F1↑	F1↑
Closed-source MLLMs																
Gemini-2.5-Flash (2025)	0.111	0.537	0.237	0.032	0.432	0.152	0.065	0.074	0.493	0.267	0.109	0.557	0.385	0.816	0.000	0.111
Gemini-2.5-Pro (2025)	0.191	0.644	0.189	0.085	0.532	0.159	0.065	0.197	0.588	0.227	0.109	0.792	0.575	0.731	0.049	0.529
Gemini-3.7-Flash (2025)	0.298	0.770	0.131	0.038	0.609	0.237	0.149	0.047	0.717	0.276	0.124	0.901	0.785	0.840	0.621	0.630
Gemini-3.1-Pro (2025)	0.196	0.746	0.194	0.029	0.581	0.137	0.102	0.040	0.742	0.204	0.107	0.875	0.734	0.800	0.140	0.642
DeepSeek-V4-Flash (2026)	0.181	0.394	0.341	0.023	0.258	0.125	0.115	0.026	0.314	0.137	0.067	0.615	0.292	0.493	0.000	0.797
GLM-5.3-Flash (2026)	0.170	0.639	0.254	0.090	0.515	0.175	0.097	0.247	0.606	0.277	0.117	0.770	0.666	0.640	0.095	0.180
Grok-4.6 (2026)	0.250	0.645	0.241	0.076	0.505	0.236	0.153	0.128	0.586	0.205	0.092	0.785	0.642	0.707	0.404	0.533
Qwen3-Max (2025)	0.192	0.405	0.342	0.019	0.271	0.161	0.096	0.014	0.306	0.189	0.086	0.580	0.274	0.593	0.000	0.800
GPT-5.6-Sol (2025)	0.256	0.641	0.176	0.116	0.510	0.257	0.179	0.248	0.513	0.267	0.122	0.785	0.573	0.824	0.048	0.672
Open-source MLLMs																
Qwen3-VL-2B (2025)	0.087	0.465	0.387	0.040	0.394	0.210	0.116	0.081	0.483	0.128	0.061	0.550	0.237	0.660	0.000	0.011
Qwen3.5-2B (2025)	0.142	0.421	0.372	0.070	0.377	0.241	0.153	0.104	0.340	0.140	0.045	0.461	0.209	0.720	0.182	0.126
InternVL3-2B (2025)	0.100	0.453	0.610	0.020	0.370	0.212	0.119	0.045	0.444	0.235	0.092	0.532	0.249	0.671	0.000	0.000
InternVL3.5-2B (2025)	0.113	0.463	0.536	0.005	0.434	0.195	0.123	0.177	0.495	0.218	0.081	0.590	0.196	0.602	0.000	0.000
RynnBrain-2B (2026)	0.146	0.557	0.369	0.000	0.494	0.153	0.115	0.000	0.540	0.043	0.053	0.711	0.334	0.704	0.000	0.658
Embodied-R1-3B (2026a)	0.194	0.469	0.765	0.004	0.389	0.136	0.080	0.022	0.438	0.160	0.069	0.546	0.292	0.682	0.115	0.777
Qwen3-VL-4B (2025)	0.147	0.550	0.329	0.058	0.496	0.223	0.127	0.121	0.568	0.223	0.102	0.545	0.352	0.789	0.000	0.233
Qwen2.5-VL-7B (2025)	0.143	0.506	0.591	0.008	0.403	0.142	0.066	0.038	0.445	0.174	0.074	0.522	0.385	0.773	0.261	0.244
MiMo-Embodied-7B (2025)	0.114	0.210	0.859	0.004	0.120	0.101	0.042	0.005	0.001	0.121	0.031	0.268	0.044	0.616	0.000	0.496
Pelican-VL-7B (2025)	0.088	0.532	0.732	0.000	0.379	0.189	0.107	0.000	0.450	0.128	0.064	0.721	0.352	0.758	0.118	0.012
RoboBrain-2.0-7B (2025)	0.081	0.561	0.511	0.004	0.391	0.184	0.095	0.002	0.391	0.180	0.082	0.757	0.535	0.731	0.025	0.000
VeBrain-7B (2025)	0.226	0.496	0.476	0.026	0.399	0.285	0.148	0.058	0.508	0.121	0.052	0.449	0.362	0.760	0.157	0.760
InternVL3-8B (2025)	0.237	0.548	0.326	0.044	0.468	0.200	0.109	0.179	0.490	0.264	0.107	0.551	0.449	0.780	0.025	0.777
InternVL3.5-8B (2025)	0.207	0.467	0.227	0.088	0.452	0.159	0.108	0.177	0.304	0.183	0.088	0.613	0.241	0.724	0.000	0.734
Embodied-R1.5-8B (2026b)	0.203	0.617	0.246	0.068	0.508	0.215	0.128	0.241	0.611	0.164	0.091	0.725	0.506	0.736	0.000	0.583
RoboBrain-2.5-8B (2026)	0.188	0.569	0.319	0.049	0.488	0.132	0.053	0.237	0.566	0.210	0.081	0.589	0.411	0.793	0.000	0.603
RynnBrain-9B (2026)	0.165	0.322	0.687	0.007	0.154	0.091	0.036	0.068	0.149	0.131	0.049	0.527	0.198	0.580	0.000	0.779
Qwen3.5-27B (2025)	0.227	0.498	0.173	0.137	0.339	0.262	0.167	0.340	0.060	0.210	0.095	0.724	0.576	0.791	0.049	0.467
Qwen3.6-27B (2025)	0.239	0.487	0.229	0.112	0.340	0.270	0.181	0.395	0.048	0.211	0.095	0.724	0.534	0.789	0.025	0.498
Qwen3.8-27B (2025)	0.232	0.401	0.223	0.152	0.159	0.282	0.179	0.294	0.032	0.221	0.085	0.640	0.365	0.811	0.025	0.539
Hy-Embodied-30B (2026b)	0.062	0.603	0.262	0.000	0.434	0.079	0.100	0.002	0.542	0.171	0.079	0.748	0.500	0.791	0.000	0.000
Qwen3-VL-32B (2025)	0.228	0.641	0.255	0.108	0.569	0.221	0.135	0.366	0.667	0.239	0.103	0.693	0.482	0.796	0.000	0.529
InternVL3-38B (2025)	0.163	0.655	0.336	0.092	0.552	0.185	0.112	0.253	0.592	0.259	0.113	0.755	0.557	0.818	0.025	0.197
InternVL3.5-38B (2025)	0.178	0.577	0.694	0.092	0.462	0.205	0.098	0.290	0.478	0.210	0.088	0.728	0.429	0.787	0.000	0.358
LookAnything (2B, Ours)	**0.518**	**0.852**	**0.094**	**0.333**	**0.764**	**0.459**	**0.283**	**0.610**	**0.865**	**0.466**	**0.260**	**0.924**	**0.863**	**0.847**	**0.805**	**0.932**

Structurally Disentangled Block-wise Attention. If heterogeneous query groups freely attended to one another, dense depth gradients or low-level head contours might interfere with gaze target representations, creating shortcut dependencies that degrade semantic generalization. To strictly isolate functional specialization, we introduce a block-wise structured attention mask (Figure 5). Each query group independently attends to the shared multimodal prompt but remains isolated from other query groups. Crucially, the downstream autoregressive language tokens jointly attend to all contextualized gaze representations, elegantly integrating them for holistic reasoning.

4 Experiments

4.1 Experimental Settings

Evaluation Metrics. For HoliGaze-Bench, point and box localization are evaluated by L_2 distance and mIoU. OpenQA is measured by ROUGE-L and BLEU-4, while MCQs use accuracy (set-F1 for multi-select). For native benchmarks, we follow established protocols and report L_2 and AUC, together with Min L_2 on GazeFollow and GazeAcc on GazeHOI (Recasens et al., 2015; Tafasca et al., 2024a). We compute the AUC metric using the query-based heatmap output.

Implementation Details. LookAnything is built upon Qwen3-VL-2B (Bai et al., 2025). We adapt the vision encoder via LoRA (Hu et al., 2021) and jointly train the language backbone, gaze-prior queries, and auxiliary decoders. We optimize the model using AdamW (Loshchilov & Hutter, 2017) with cosine decay and a peak learning rate of $2e-5$ for 2 epochs on 8 NVIDIA H20 GPUs. Training uses the 303K HoliGaze-Corpus samples. Detailed hyperparameters are provided in Appendix B.

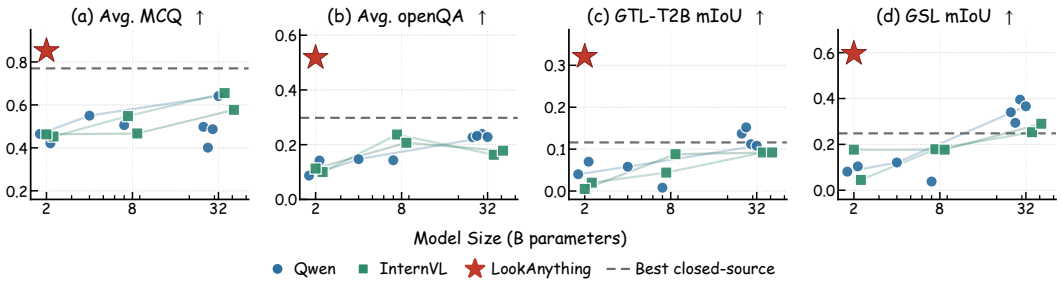


Figure 6: **HoliGaze-Bench performance vs. model size.** Our 2B model (red star) achieves superior performance despite its compact scale.

Table 3: **Results on 5 gaze-following benchmarks.** Best and second-best are **bold** and underlined.

	Avg.		VAT		GazeFollow			ChildPlay		GazeHOI			GOO-Real	
Method	L2↓	AUC↑	L2↓	AUC↑	L2↓	Min L2↓	AUC↑	L2↓	AUC↑	L2↓	AUC↑	GazeAcc↑	L2↓	AUC↑
Recasens et al. (2015)	0.253	0.809	0.242	0.789	0.227	0.155	0.857	0.292	0.754	0.259	0.817	0.371	0.245	0.829
Lian et al. (2018)	0.226	0.835	0.242	0.789	0.189	0.119	0.885	0.292	0.754	0.200	0.877	0.561	0.205	0.868
Chong et al. (2020)	0.172	0.857	0.126	0.864	0.138	0.077	0.923	0.128	0.858	0.197	0.869	0.564	0.272	0.770
HGTTR (2022)	0.224	0.825	0.266	0.816	0.169	0.109	0.847	0.204	0.859	0.216	0.824	0.528	0.265	0.780
MADGTD (2022)	0.172	0.857	0.128	0.860	0.138	0.077	0.925	0.133	0.866	0.195	0.870	0.578	0.265	0.762
Gupta et al. (2022)	0.181	0.867	0.128	0.898	0.156	0.091	0.914	0.144	0.893	0.198	0.867	0.556	0.279	0.765
Miao et al. (2023)	0.170	0.877	0.117	0.906	0.132	0.072	0.928	0.122	0.919	0.194	0.864	0.584	0.287	0.768
Tafasca et al. (2023)	0.121	0.921	0.123	0.904	0.128	0.065	0.935	0.112	0.924	–	–	–	–	–
ViTGaze (2024)	0.136	0.900	0.103	0.908	0.105	0.047	0.949	0.112	0.899	0.163	0.880	0.714	0.196	0.862
Sharingan (2024b)	0.150	0.900	0.104	0.920	0.112	0.055	0.944	0.110	0.930	0.177	0.886	0.633	0.248	0.819
Gaze-LLE (ViT-B) (2025)	0.132	0.930	0.105	0.932	0.104	0.045	0.956	0.115	0.945	0.162	0.906	0.705	0.175	0.909
Gaze-LLE (ViT-L) (2025)	0.127	0.929	**0.100**	0.937	0.098	**0.041**	**0.958**	0.102	0.950	0.160	0.907	0.710	0.173	0.915
GazeAnywhere (2026)	0.183	0.897	0.236	0.845	0.179	0.115	0.923	0.139	0.916	0.196	0.885	0.625	0.165	0.918
LookAnything (Ours)	**0.109**	**0.942**	**0.098**	**0.941**	**0.091**	0.048	0.945	**0.094**	**0.956**	**0.130**	**0.940**	**0.771**	**0.132**	**0.927**

4.2 Results on HoliGaze-Bench

As depicted in Table 2, we compare LookAnything with leading closed-source and open-source MLLMs (e.g., Gemini-3.1-Pro (Comanici et al., 2025), GPT-5.6-Sol (Singh et al., 2025), etc.), including generalist and embodied models. We highlight three observations:

(1) ***Geometric grounding is the primary gap for generalist MLLMs.*** While closed-source models like Gemini-3.7-Flash (Comanici et al., 2025) excel on descriptive MCQs, most of them struggle on localization. The strongest open-source localizers (Qwen3.8-27B, Qwen3.6-27B (Bai et al., 2025)) lag LookAnything by large margins (e.g., 0.152 vs 0.321 GTL-T2B mIoU). Broad language and reasoning capability therefore does not intrinsically yield precise gaze-conditioned grounding.

(2) ***Embodied pretraining does not automatically transfer.*** Embodied models (e.g., Embodied-R1 (Yuan et al., 2026a), RoboBrain (Team et al., 2025; Tan et al., 2026)) perform adequately on GCD/GDO but remain weak on point and box localization, indicating that generic spatial or action-centric supervision cannot substitute for gaze-specific priors.

(3) ***LookAnything delivers broad gains at a compact scale.*** Our 2B model outperforms all baselines across all metrics. Compared to its Qwen3-VL-2B backbone, it reduces GTL-T2P L_2 from 0.387 to 0.094, validating the effectiveness of large-scale multi-task training and gaze-aware representation learning on HoliGaze-Corpus. Figure 6 further shows that these gains are not the result of model scale: LookAnything lies well above the open-source scaling trend, and even 27B–38B generalists fall far short of its localization averages.

4.3 Results on Native Gaze Target Detection

To verify that task unification does not compromise conventional gaze target detection, we evaluate LookAnything on the official test splits of GazeFollow (Recasens et al., 2015), VAT (Chong et al., 2020), ChildPlay (Tafasca et al., 2023), GazeHOI (Tafasca et al., 2024a), and GOO-Real (Tomas et al., 2021), against representative specialist models (See Appendix B.1 for detailed introductions).

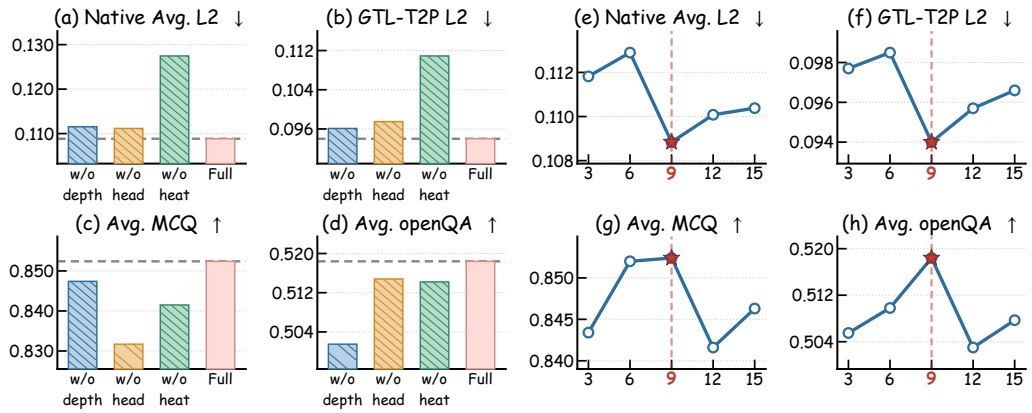


Figure 7: **Ablation of Gaze-Prior Query Groups and Query-Token Budget.** *Left (a–d)*: Ablation of the three gaze-prior query groups. *Right (e–h)*: Sweeping the query-token budget $k \in \{3, 6, 9, 12, 15\}$. The red star and vertical line mark the adopted setting $k = 9$.

As shown in Table 3, LookAnything achieves the best overall performance with an average L_2 of 0.109 and AUC of 0.942 across all 5 benchmarks. Specifically, it achieves the best L_2 and AUC on 4 benchmarks (VAT, ChildPlay, GazeHOI, and GOO-Real). On GazeFollow, it obtains the best average L_2 despite being trained jointly across diverse tasks. Overall, LookAnything matches or exceeds dedicated specialists on core localization.

4.4 Ablation Study

In this section, we conduct extensive ablation studies to investigate the following questions.

Q1: What is the contribution of each gaze-prior query? To understand the individual impact of the injected priors, we remove one type of auxiliary supervision at a time. As illustrated in Figure 7 (left), omitting the target heatmap supervision causes the most severe degradation in precise localization, as it directly promotes the learning of target spatial representations. Conversely, discarding head mask supervision primarily hampers multiple-choice reasoning, underscoring that reliably grounding the source head is vital for downstream inference. Furthermore, removing the depth prior noticeably impairs open-ended semantic generation, as 3D structural context is crucial for resolving spatial depth ambiguities. Together, these results validate their complementary contributions.

Q2: How does the number of query tokens affect performance? We allocate a fixed budget of k learnable query tokens for each geometric prior group. Sweeping $k \in \{3, 6, 9, 12, 15\}$ (Figure 7, right) shows that a restricted budget ($k \leq 6$) forces queries to compress too much spatial information, leading to suboptimal performance. The model achieves its peak performance at $k = 9$, striking an ideal balance of representational bandwidth without overwhelming the language backbone. Expanding the capacity to $k \geq 12$ yields no additional benefits and causes slight regressions, likely due to optimization difficulties and attention dilution.

Q3: Is block-wise structured attention necessary? To validate our block-wise structured attention (Figure 5), we replace it with a standard causal mask. Under a causal mask, queries from later groups can attend to earlier ones, allowing dense depth gradients or low-level head contours to inadvertently interfere with gaze target representations. This shortcut dependency introduces gradient noise and degrades semantic generalization, causing a noticeable performance drop (Table 4). Instead, our structured design ensures each domain independently extracts relevant features from the shared context, preserving representational purity and improving overall accuracy.

Q4: Are specialized queries necessary for different geometric priors? An alternative choice is employing a unified group of queries to simultaneously predict scene depth, head masks, and target heatmaps. As reported in Table 5, this shared-query design consistently underperforms. Forcing a single set of tokens to capture disparate physical and geometric properties inevitably causes feature

Table 4: Performance comparison between causal and block-wise structured attention.

Setting	HoliGaze-Bench			Native
	Avg. MCQ ↑	Avg. openQA ↑	GTL-T2P L_2 ↓	Avg. L_2 ↓
Causal	0.848	0.498	0.097	0.111
Structured	**0.852**	**0.518**	**0.094**	**0.109**

Table 5: Performance comparison between shared and separated query groups.

Setting	HoliGaze-Bench			Native
	Avg. MCQ ↑	Avg. openQA ↑	GTL-T2P L_2 ↓	Avg. L_2 ↓
Shared	0.845	0.493	0.098	0.112
Separated	**0.852**	**0.518**	**0.094**	**0.109**

entanglement, feeding noisy, conflicting signals to the language head. By explicitly disentangling the queries into specialized groups, LookAnything maintains clean, domain-specific representations, yielding clear improvements across all benchmarks.

Q5: Is fine-tuning the vision encoder necessary? While modern MLLMs possess strong general visual representations, fine-grained gaze understanding requires capturing subtle cues like eye orientation and head pose. Table 6 contrasts the performance with and without applying LoRA to the vision encoder. Freezing the vision tower results in a severe degradation across all metrics, particularly increasing the average L_2 on native benchmarks from 0.1118 to 0.1410. This confirms that task-specific adaptation of the visual backbone is indispensable for extracting the nuanced geometric features necessary for accurate gaze target detection.

Table 6: Effect of LoRA to the vision tower.

Setting	HoliGaze-Bench			Native
	Avg. MCQ ↑	Avg. openQA ↑	GTL-T2P L_2 ↓	Avg. L_2 ↓
w/o LoRA	0.774	0.467	0.130	0.138
w LoRA	**0.852**	**0.518**	**0.094**	**0.109**

4.5 Real-world Experiments



Figure 8: Real-world Experiments.

To evaluate the out-of-domain (OOD) generalization of LookAnything, we deploy it on a Unitree G1 robot in real-world physical environments. Figure 8 illustrates the experimental setup and qualitative visualization results. From the robot's egocentric perspective, the 3 rows depict 3 different subjects in various shopping scenarios, with the model predicting gaze targets in the form of bounding boxes, heatmaps, and points, respectively. The blue overlays indicate the predicted head masks of the source persons, while the top-left corner displays the gaze contact detection results (i.e., whether the source person is looking at the ego-camera). As observed, LookAnything accurately grounds the source person in multi-person scenes, reliably recognizes social gaze signals, and precisely tracks gaze targets under both static and moving perspectives. These results demonstrate its strong adaptability and robustness for real-world embodied applications.

5 Conclusion

In this paper, we introduced LookAnything, a generalist foundation model that unifies diverse gaze-centric tasks. By internalizing gaze priors via large-scale cross-domain learning, our framework bypasses explicit auxiliary inputs, enabling flexible adaptation across various scenarios, task settings, and output formats. To facilitate this paradigm, we construct HoliGaze, a comprehensive corpus and benchmark standardizing 311K Q&A samples across 9 distinct tasks. Empirical results confirm that LookAnything surpasses leading VLM in holistic gaze understanding while remaining highly competitive with specialists in conventional gaze target detection. Real-world experiments on a Unitree G1 robot further underscores its robust generalization in complex and diverse physical environments. We believe LookAnything and HoliGaze will inspire future advances in human-centric interactive applications and embodied AI, envisioning them as a foundational step toward socially intelligent embodied agents capable of seamless non-verbal cognitive alignment with humans.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing, LaTeX and table formatting. We have not used generative AI tools for research idea generation and experimental design decisions. Additionally, we use large vision-language models (e.g., Qwen3-VL-32B) as part of our automated data engine to construct the HoliGaze dataset, which is explicitly described in Section 2.2 and Appendix D. We have reviewed all AI-assisted work. Generated or revised code was verified and tested for correctness; and all reported results, interpretations, and final manuscript content were independently reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work introduces LookAnything, a generalist foundation model for gaze target detection, and HoliGaze, a comprehensive dataset for holistic gaze understanding. The construction of the HoliGaze dataset relies entirely on publicly available academic datasets (e.g., GazeFollow, VideoAttentionTarget, ChildPlay, etc.) strictly according to their intended research purposes and usage licenses. The experiments do not involve private user data, personally identifiable information, or any sensitive data. Real-world deployment experiments on the Unitree G1 robot were conducted in controlled laboratory environments without capturing or compromising the privacy of non-consenting individuals. We anticipate that our unified model will advance human-centric interactive applications and embodied AI, while we also recognize the importance of deploying such gaze-understanding technologies responsibly to prevent potential misuse in unauthorized surveillance.

REPRODUCIBILITY STATEMENT

To ensure the full reproducibility of our findings, we detail all experimental settings, data construction processes, and implementation specifics throughout the paper and appendix. Section 4.1 describes the evaluation metrics and training configurations. Appendix B provides comprehensive implementation details, including the LookAnything model architecture hyperparameters (Table 7), training objectives, loss formulations, hardware specifications, and optimization details (Table 8). Furthermore, Appendix D details the automated data engine pipeline, providing the exact prompts used for the vision-language models (Figures 16 to 23) and the canonical Q&A generation templates (Table 10). The source datasets and their split compositions are detailed in Appendix C. To facilitate future research, the complete training code, evaluation scripts, data generation pipeline, pretrained LookAnything models, and the HoliGaze dataset will be publicly released.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Xu Cao, Houze Yang, Vipin Gunda, Zhongyi Zhou, Tianyu Xu, Adarsh Kowdle, Inki Kim, and James M Rehg. Gaze target estimation anywhere with concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31304–31315, 2026.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5395–5405. IEEE, 2020.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the

frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Zhuangzhuang Dai, Zhongxi Lu, Vincent G Zakka, Luis J Manso, Jose M Alcaraz Calero, and Chen Li. Gazemoe: Perception of gaze target with mixture-of-experts. *arXiv preprint arXiv:2603.06256*, 2026.

Ronghao Dang, Jiayan Guo, Bohan Hou, Sicong Leng, Kehan Li, Xin Li, Jiangpin Liu, Yunxuan Mao, Zhikai Wang, Yuqian Yuan, et al. Rynnbrain: Open embodied foundation models. *arXiv preprint arXiv:2602.14979*, 2026.

Lifeng Fan, Yixin Chen, Ping Wei, Wenguan Wang, and Song-Chun Zhu. Inferring shared attention in social scene videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6460–6468. IEEE, 2018.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Anshul Gupta, Samy Tafasca, and Jean-Marc Odobez. A modular multimodal architecture for gaze target prediction: Application to privacy-sensitive settings. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 5037–5046. IEEE, 2022.

Xiaoshuai Hao, Lei Zhou, Zhijian Huang, Zhiwen Hou, Yingbo Tang, Lingfeng Zhang, Guang Li, Zheng Lu, Shuhuai Ren, Xianhui Meng, et al. Mimo-embodied: X-embodied foundation model technical report. *arXiv preprint arXiv:2511.16518*, 2025.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

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

Isaac Kasahara, Simon Stent, and Hyun Soo Park. Look both ways: Self-supervising driver gaze estimation and road scene saliency. In *European Conference on Computer Vision*, pp. 126–142. Springer, 2022.

Dongze Lian, Zehao Yu, and Shenghua Gao. Believe it or not, we know what you are looking at! In *Asian Conference on Computer Vision*, pp. 35–50. Springer, 2018.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025a.

Zhi-Yi Lin, Jouh Yeong Chew, Jan van Gemert, and Xucong Zhang. Gazehta: End-to-end gaze target detection with head-target association. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9447–9454. IEEE, 2025b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Gen Luo, Ganlin Yang, Ziyang Gong, Guanzhou Chen, Haonan Duan, Erfei Cui, Ronglei Tong, Zhi Hou, Tianyi Zhang, Zhe Chen, et al. Visual embodied brain: Let multimodal large language models see, think, and control in spaces. *arXiv preprint arXiv:2506.00123*, 2025.

Athul M Mathew, Haithem Hermassi, Thariq Kadavil, and Arshad Ali Khan. Gazevlm: A vision-language model for multi-task gaze understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9241–9250, 2026.

Jiajie Mi, Xinyu Liu, Mengke Song, and Chenglizhao Chen. Multi-scale object-aware gaze estimation via geometric reasoning. *arXiv preprint arXiv:2606.29334*, 2026.

Qiaomu Miao, Minh Hoai, and Dimitris Samaras. Patch-level gaze distribution prediction for gaze following. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 880–889. IEEE, 2023.

Qiaomu Miao, Haoyu Wu, Jingyi Xu, Minh Hoai, and Dimitris Samaras. Omnigf: A dual-branch vision-language framework for unified gaze following. *arXiv preprint arXiv:2605.26399*, 2026.

Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? *Advances in neural information processing systems*, 28, 2015.

Fiona Ryan, Ajay Bati, Sangmin Lee, Daniel Bolya, Judy Hoffman, and James M Rehg. Gazelle: Gaze target estimation via large-scale learned encoders. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28874–28884. IEEE, 2025.

Akanksha Saran, Srinjoy Majumdar, Elaine Schaertl Short, Andrea Thomaz, and Scott Niekum. Human gaze following for human-robot interaction. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8615–8621. IEEE, 2018.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Yuehao Song, Xinggang Wang, Jingfeng Yao, Wenyu Liu, Jinglin Zhang, and Xiangmin Xu. Vitgaze: gaze following with interaction features in vision transformers. *Visual Intelligence*, 2 (1):31, 2024.

Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Childplay: A new benchmark for understanding children's gaze behaviour. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20878–20889. IEEE, 2023.

Samy Tafasca, Anshul Gupta, Victor Bros, and Jean-Marc Odobez. Toward semantic gaze target detection. *Advances in neural information processing systems*, 37:121422–121448, 2024a.

Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Sharingan: A transformer architecture for multi-person gaze following. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2008–2017. IEEE, 2024b.

Huajie Tan, Enshen Zhou, Zhiyu Li, Yijie Xu, Yuheng Ji, Xiansheng Chen, Cheng Chi, Pengwei Wang, Huizhu Jia, Yulong Ao, et al. Robobrain 2.5: Depth in sight, time in mind. *arXiv preprint arXiv:2601.14352*, 2026.

BAAI RoboBrain Team, Mingyu Cao, Huajie Tan, Yuheng Ji, Xiansheng Chen, Minglan Lin, Zhiyu Li, Zhou Cao, Pengwei Wang, Enshen Zhou, et al. Robobrain 2.0 technical report. *arXiv preprint arXiv:2507.02029*, 2025.

Henri Tomas, Marcus Reyes, Raimarc Dionido, Mark Ty, Jonric Mirando, Joel Casimiro, Rowel Atienza, and Richard Guinto. Goo: A dataset for gaze object prediction in retail environments. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3119–3127. IEEE, 2021.

Francesco Tonini, Cigdem Beyan, and Elisa Ricci. Multimodal across domains gaze target detection. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pp. 420–431, 2022.

Danyang Tu, Xiongkuo Min, Huiyu Duan, Guodong Guo, Guangtao Zhai, and Wei Shen. End-to-end human-gaze-target detection with transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2192–2200. IEEE, 2022.

Shijing Wang, Yaping Huang, Chaoqun Cui, David Wong, Yihua Cheng, Alexandros Neophytou, and Hyung Jin Chang. Enhancing gaze reasoning in vision foundation models for gaze following. *arXiv preprint arXiv:2605.22607*, 2026a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Ziyi Wang, Xumin Yu, Yongming Rao, Yonggen Ling, Yunheng Li, Oran Wang, Mingqi Gao, Yuchen Zhou, Yves Liang, Zuyan Liu, et al. Hy-embodied-vlm-1.0: Efficient physical-world agents. *arXiv preprint arXiv:2607.12894*, 2026b.

xAI. Model card: Grok 4.6. Model card, August 2026. URL `https://media.x.ai/v1/website/card-4p6-4cd2dc57.pdf`. Revision 2026-08-17.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Zhoutong Ye, Chengwen Zhang, Zhaibin Cui, Mingze Sun, Jiaqi Liu, Xiangwu Li, Qingyang Wan, Chang Liu, Xutong Wang, Huan-ang Gao, et al. Page: Towards practical human-level gaze target estimation. *arXiv preprint arXiv:2607.04860*, 2026.

Yifu Yuan, Haiqin Cui, Yaoting Huang, Yibin Chen, Fei Ni, Zibin Dong, Pengyi Li, Yan Zheng, Hongyao Tang, and Jianye Hao. Embodied-r1: Reinforced embodied reasoning for general robotic manipulation. In *International Conference on Learning Representations*, volume 2026, pp. 8984–9009, 2026a.

Yifu Yuan, Yaoting Huang, Xianze Yao, Yutong Li, Shuoheng Zhang, Linqi Han, Pengyi Li, Jiangeng Sun, Wenting Jia, Zhao Zhang, et al. Embodied-r1. 5: Evolving physical intelligence via embodied foundation models. *arXiv preprint arXiv:2606.11324*, 2026b.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Yi Zhang, Che Liu, Xiancong Ren, Hanchu Ni, Shuai Zhang, Zeyuan Ding, Jiayu Hu, Hanzhe Shan, Zhenwei Niu, Zhaoyang Liu, et al. Pelican-vl 1.0: A foundation brain model for embodied intelligence. *arXiv preprint arXiv:2511.00108*, 2025.

Yuchen Zhou, Linkai Liu, and Chao Gou. Learning from observer gaze: Zero-shot attention prediction oriented by human-object interaction recognition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28390–28400. IEEE, 2024.

Yuchen Zhou, Chao Gou, Zipeng Guo, Yihua Cheng, and Hyung Jin Chang. Behavior-aware knowledge-embedded model for driver attention prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(10):10199–10212, 2025a.

Yuchen Zhou, Jiayu Tang, Xiaoyan Xiao, Yueyao Lin, Linkai Liu, Zipeng Guo, Hao Fei, Xiaobo Xia, and Chao Gou. Where, what, why: Towards explainable driver attention prediction. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2675–2685. IEEE, 2025b.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Kuangji Zuo, Gen Li, Bofan Lyu, Yanshuo Lu, Boyu Ma, Shijia Han, Xinyu Zhou, Xichen Yuan, Chuhao Zhou, Jiaqi Bai, et al. Gaze2act: Gaze-conditioned vision-language-action policies for interactive robot manipulation. *arXiv preprint arXiv:2605.30282*, 2026.

A Related Work

Human Gaze Target Detection. Gaze target detection aims to infer where people look in visual scenes, providing vital cues for attention estimation, intent understanding, and human–robot interaction (Zhou et al., 2025b; 2024; 2025a; Kasahara et al., 2022). Pioneered by Recasens et al. (2015),

classical frameworks formulate this task as predicting a 2D spatial heatmap by fusing cropped head representations with full-scene context in dual-stream architectures (Lian et al., 2018; Chong et al., 2020). Subsequent studies advance this line of research by incorporating scene depth priors (Gupta et al., 2022; Tonini et al., 2022), patch-level distributions (Miao et al., 2023), and transformer-based head–target associations (Tu et al., 2022; Lin et al., 2025b; Song et al., 2024; Mi et al., 2026). More recently, methods such as Gaze-LLE (Ryan et al., 2025), PAGE (Ye et al., 2026), and GazeMoE (Dai et al., 2026) scale up visual representations using large pretrained vision backbones or mixture-of-experts designs. Despite steady progress, these specialists rely on highly customized pipelines with pre-extracted crops or depth maps that induce cascading errors. Moreover, their coarse heatmap outputs lack instance-level semantics, hindering seamless integration with downstream interactive and embodied tasks (Saran et al., 2018; Zuo et al., 2026). In contrast, LookAnything reformulates gaze localization and holistic understanding within a generalist vision-language framework, supporting both instance-level grounding and semantic description.

Multi-modal Large Language Models (MLLMs). MLLMs have demonstrated remarkable capabilities in open-vocabulary recognition, visual reasoning, and conversational interaction (Bai et al., 2025; Comanici et al., 2025; Singh et al., 2025; Xu et al., 2026). Motivated by these advances, recent studies adapt MLLMs to human gaze tasks (Wang et al., 2026a; Mathew et al., 2026; Miao et al., 2026). Data-driven fine-tuning approaches like GazeVLM (Mathew et al., 2026) and VL4Gaze (Wang et al., 2026a) formulate gaze following as standard visual question answering without incorporating architectural gaze prior contexts. Conversely, architectural adaptations like OmniGF (Miao et al., 2026) incorporate specialized localization branches but remain restricted to isolated gaze target estimation. In contrast, LookAnything injects gaze-relevant priors (e.g., head masks and scene depth) into latent representations via learnable context queries and, scaled with diverse gaze corpora, establishes the first generalist foundation model for holistic gaze understanding.

Benchmarks for Gaze Following and Understanding. Foundational gaze following benchmarks, such as GazeFollow (Recasens et al., 2015), VideoAttentionTarget (Chong et al., 2020), ChildPlay (Tafasca et al., 2023), and GOO (Tomas et al., 2021), predominantly focus on gaze target detection across diverse contexts, spanning static images, dynamic videos, and retail environments. Subsequently, Ego4D-Social (Grauman et al., 2022) benchmarks egocentric gaze contact in social interactions, while GazeHOI (Tafasca et al., 2024a) bridges gaze following and human-object interactions to enable simultaneous target localization and semantic category recognition. More recently, to advance beyond pure gaze target localization, Gaze-Co (Cao et al., 2026) incorporates semantic object descriptions, and VL4Gaze (Wang et al., 2026a) introduces vision-language question answering for gaze understanding. Despite these efforts, current benchmarks remain fragmented, each isolating specific gaze facets within narrow task definitions. To bridge this gap, we propose HoliGaze, a unified benchmark encompassing 5 capability dimensions across 9 distinct tasks, providing the first holistic and multi-faceted evaluation suite for gaze understanding.

B Additional Implementation Details

B.1 Baseline Models for Conventional Gaze Target Detection

We include foundational dual-stream CNN methods by Recasens et al. (2015), Lian et al. (2018), and Chong et al. (2020); transformer-based specialists HGTTR (Tu et al., 2022), ViTGaze (Song et al., 2024), and Sharingan (Tafasca et al., 2024b); depth- or multimodal-enhanced approaches MADGTD (Tonini et al., 2022) and Gupta et al. (2022); and patch-level prediction by Miao et al. (2023). The most recent baselines include Gaze-LLE, which adapts large-scale pretrained visual encoders (Ryan et al., 2025), and GazeAnywhere, which extends gaze estimation with concept prompts (Cao et al., 2026). Together, these methods cover the principal architectural families in specialist gaze target detection.

B.2 Baseline Models on HoliGaze-Bench

To comprehensively benchmark holistic gaze understanding capabilities, we curate a diverse evaluation pool comprising 33 distinct models. These models are systematically categorized into three primary families: proprietary multimodal APIs, general-purpose open-weight models, and special-

ized embodied foundation models. All evaluations are conducted in a zero-shot setting, utilizing either the official provider APIs or the canonical checkpoints hosted on Hugging Face.

Proprietary Multimodal APIs. This category encompasses state-of-the-art closed-source models accessed via commercial endpoints, which typically leverage massive scale and proprietary training recipes to set the upper bound for general multimodal reasoning. Within the Google Gemini lineage (Comanici et al., 2025), we evaluate multiple generations: Gemini-3.7-Flash serves as the highly efficient, reasoning-optimized endpoint supporting a massive 1M-token context window and interleaved multimodal inputs, while Gemini-3.1-Pro represents the higher-compute tier designed for complex, multi-step spatial and logical reasoning. We additionally include the previous-generation Gemini-2.5-Flash and Gemini-2.5-Pro to track cross-generational capability improvements. From OpenAI, we evaluate GPT-5.6-Sol (Singh et al., 2025), the flagship vision-language model exhibiting exceptional zero-shot spatial grounding and instruction-following capabilities. To provide a comprehensive view of the commercial landscape, we also incorporate DeepSeek-V4-Flash (Xu et al., 2026), a highly optimized Mixture-of-Experts (MoE) model; GLM-5.3-Flash (Zeng et al., 2026); Grok-4.6 (xAI, 2026); and Qwen3-Max (Bai et al., 2025). These models provide robust alternatives with strong visual grounding, spatial relationship understanding, and state-of-the-art performance on broad vision-language benchmarks.

General-Purpose Open-Weight Models. This group consists of publicly accessible vision-language models that allow for local inference and transparent architectural analysis, varying significantly in parameter scale and architectural design. We comprehensively evaluate the Qwen ecosystem (Bai et al., 2025), including the latest Qwen3-VL series (2B, 4B, 32B), which features dynamic image resolution and temporal-aware positional encoding, making them highly adept at fine-grained spatial localization. To analyze the scaling laws of gaze understanding within a unified architectural family, we also assess various intermediate checkpoints: the dense Qwen3.5-2B and Qwen3.5-27B, the MoE-based Qwen3.6-27B, and the advanced Qwen3.8-27B. The earlier Qwen2.5-VL-7B is included as a baseline for instruction-tuned dynamic-resolution VLMs. Furthermore, we evaluate the InternVL family (Zhu et al., 2025; Wang et al., 2025), specifically InternVL3 (2B, 8B, 38B) and the subsequently aligned InternVL3.5 (2B, 8B, 38B). Known for their high-resolution patch processing and strong visual encoders, these models excel in capturing intricate visual details, which is theoretically advantageous for detecting subtle gaze cues and small target objects.

Embodied Foundation Models. Given the strong connection between gaze understanding and human-robot interaction, we evaluate a suite of VLMs specifically fine-tuned on embodied, egocentric, and robotic manipulation datasets. This allows us to investigate whether action-centric spatial supervision naturally transfers to human gaze target detection. We evaluate vision-language-action (VLA) models such as Embodied-R1-3B (Yuan et al., 2026a) and Embodied-R1.5-8B (Yuan et al., 2026b), which are optimized for embodied control and spatial grounding. We also include the RoboBrain series (RoboBrain-2.0-7B (Team et al., 2025) and RoboBrain-2.5-8B (Tan et al., 2026)), tailored for 3D spatial reasoning and complex robotic manipulation tasks. To assess lightweight, edge-deployable models focusing on efficient visual-motor policies, we incorporate RynnBrain-2B and RynnBrain-9B (Dang et al., 2026). Finally, specialized robotic VLMs including MiMo-Embodied-7B (Hao et al., 2025), Pelican-VL-7B (Zhang et al., 2025), VeBrain-7B (Luo et al., 2025), and Hy-Embodied-30B (Wang et al., 2026b) are evaluated to provide a comprehensive view of how current embodied foundation models, trained on diverse large-scale robotic datasets, handle human-centric cognitive signals like gaze.

B.3 LookAnything Model Architecture

LookAnything is built upon the generalist multimodal large language model, Qwen3-VL (Bai et al., 2025). The architecture integrates a vision encoder, a language model backbone, learnable gaze-prior queries, and specialized geometric decoders. Specifically, the vision encoder (ViT) processes the input image into a sequence of visual patch tokens, which are then concatenated with textual tokens from the prompt. To inject gaze-aware spatial inductive biases, we introduce three groups of learnable continuous queries: depth queries Q^d, head queries Q^h, and gaze target queries Q^g. Each group consists of $k = 9$ tokens. These queries are appended to the multimodal embeddings and processed by the language backbone. To ensure strict functional specialization, we implement a

block-wise structured attention mask as an additive 4D causal mask in PyTorch Scaled Dot-Product Attention (SDPA). This mask allows each query group to independently attend to the shared multimodal context (image and text tokens) and itself, but strictly masks out attention to other query groups, preventing shortcut dependencies such as dense depth gradients interfering with gaze target representations.

The contextualized query states are then projected into explicit geometric predictions by lightweight auxiliary decoders. The depth decoder $\mathcal{D}_d$ and head mask decoder $\mathcal{D}_h$ are instantiated as MAE-style Transformer decoders (He et al., 2022). They take the 9 query tokens and 256 (16×16) learnable mask tokens with 2D sin-cos positional embeddings, process them through 2 self-attention layers (hidden dimension 512, 8 heads), and project the output to reconstruct 64×64 resolution maps. The target spatial distribution decoder also uses this MAE-style architecture but is augmented with 2 layers of cross-attention, where the spatial mask tokens attend to the visual patch tokens from the vision encoder to output a 64×64 dense gaze heatmap.

Simultaneously, the target coordinate regressor directly predicts continuous coordinates. The point head applies 1 layer of cross-attention to the visual tokens, mean-pools the 9 gaze queries, and processes them through a 2-layer MLP (Linear $\rightarrow$ GELU $\rightarrow$ Linear) to regress the normalized continuous point (x, y). The bounding box head similarly mean-pools the queries and uses a 2-layer MLP to regress the box coordinates $[x_1, y_1, x_2, y_2]$. During inference, these auxiliary decoders can be entirely skipped to reduce computational overhead, allowing the language head to autoregressively generate the final answers based on the compact gaze-context representations. The detailed model architecture hyperparameters are summarized in Table 7.

Table 7: **Model Architecture Hyperparameters of LookAnything.**

Hyperparameter	**Value**
LLM Backbone	Qwen3-VL-2B
Vision Encoder	ViT
Visual Patch Size	14
Number of Depth Queries (Q^d)	9
Number of Head Queries (Q^h)	9
Number of Gaze Target Queries (Q^g)	9
MAE Decoder Mask Tokens	$16 \times 16 = 256$
MAE Decoder Hidden Dimension	512
MAE Decoder Attention Heads	8
Depth Decoder Layers ($\mathcal{D}_d$)	2 Self-Attention
Head Mask Decoder Layers ($\mathcal{D}_h$)	2 Self-Attention
Target Heatmap Decoder	2 Self-Attention + 2 Cross-Attention
Target Point Regressor	1 Cross-Attention + Mean-Pool + 2-layer MLP
Target Box Regressor	Mean-Pool + 2-layer MLP
Output Spatial Resolution	64×64

B.4 Training Objectives and Loss Formulations

We train LookAnything end-to-end on the HoliGaze-Corpus, which comprises 303K training samples. During training, the language backbone, the newly introduced gaze-prior queries, and the geometric decoders are jointly optimized, while the vision encoder is adapted using Low-Rank Adaptation (LoRA) (Hu et al., 2021) to capture fine-grained gaze and geometric features. The LoRA rank is set to $r = 16$ and the scaling factor is $\alpha = 32$.

Loss Formulations. LookAnything is trained end-to-end with a composite objective combining the gaze context-aware geometric losses and an autoregressive language loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{AR}} + \lambda_d \mathcal{L}_{\text{depth}} + \lambda_h \mathcal{L}_{\text{head}} + \lambda_t \mathcal{L}_{\text{target}}, \tag{1}$$

where $\mathcal{L}_{\text{AR}}$ is the token-level cross-entropy over the open-form answer Y:

$$\mathcal{L}_{\text{AR}} = -\frac{1}{N_Y}\sum_{t=1}^{N_Y}\log p_\theta\big(y_t \mid I, P, y_{<t}\big). \tag{2}$$

The auxiliary geometric supervision utilizes specific loss formulations for each prior. The depth representation is optimized using a scale-invariant logarithmic (SiLog) loss to emphasize relative structural depth while remaining robust to global scale ambiguity:

$$\mathcal{L}_{\text{depth}} = \sqrt{\frac{1}{|\Omega|}\sum_{i\in\Omega} e_i^2 - \lambda\Big(\frac{1}{|\Omega|}\sum_{i\in\Omega} e_i\Big)^2}, \tag{3}$$

where $e_i = \log \hat{D}_i - \log D_i^*$ is the logarithmic error for valid pixels Ω, with D^* being the ground-truth depth map.

The head mask representation is supervised via a combination of Binary Cross-Entropy (BCE) and soft Dice loss to mitigate foreground-background pixel imbalance:

$$\mathcal{L}_{\text{head}} = \mathcal{L}_{\text{BCE}}(\hat{M}, M^*) + \mathcal{L}_{\text{Dice}}(\hat{M}, M^*). \tag{4}$$

The target spatial distribution is supervised using a combination of spatial KL-divergence and Mean Squared Error (MSE) against a Gaussian ground-truth distribution ($\sigma = 3.0$), while the continuous coordinates are optimized via Huber loss ($\delta = 0.05$) for points and a joint L1 and Generalized IoU (GIoU) loss for bounding boxes:

$$\mathcal{L}_{\text{target}} = \mathcal{L}_{\text{KL_MSE}}(\hat{A}, A^*) + \mathcal{L}_{\text{Huber}}(\hat{p}, p^*) + \mathcal{L}_{\text{box}}(\hat{b}, b^*). \tag{5}$$

Following the main text, the loss weights for the auxiliary geometric supervision are empirically set to $\lambda_d = 0.1$ for scene depth, $\lambda_h = 1.0$ for source head grounding, and $\lambda_t = 1.0$ for target spatial distribution.

Optimization Details. We optimize the model using the AdamW optimizer (Loshchilov & Hutter, 2017) with a weight decay of 0.01 and a maximum gradient norm of 1.0. The learning rate follows a cosine decay schedule with a linear warmup ratio of 0.03. The language backbone and vision LoRA use a peak learning rate of 1×10^{-5}, while the newly initialized geometric heads use a higher learning rate of 1×10^{-4}. The training is conducted on 8 NVIDIA H20 GPUs with a per-device batch size of 2 and gradient accumulation steps of 2, resulting in an effective global batch size of 32. The model is trained for 2 epochs. The detailed training hyperparameters are provided in Table 8.

C Additional Analysis of HoliGaze

In this section, we detail the statistics information of HoliGaze.

Semantic Diversity of Gaze Sources and Targets. Figure 9 presents the word clouds of the semantic descriptions for gaze sources and targets. The source descriptions are predominantly centered around human subjects (e.g., "man", "woman", "boy", "girl"), reflecting the dataset's focus on human gaze behaviors. In contrast, the target descriptions exhibit a much broader vocabulary, encompassing various objects (e.g., "ball", "phone", "book"), scene elements (e.g., "ground", "table"), and other people. This semantic richness ensures that models trained on HoliGaze can generalize to unconstrained, real-world gaze targets.

Distributions of Source-Target Spatial Relationship. To evaluate the model's understanding of source-target spatial relationship, we analyze the answer distributions for the Gaze Contact Detection (GCD), Gaze Direction Recognition (GDR), and Gaze Depth Orientation (GDO) tasks, as shown in Figure 10. For the GCD task, the majority of subjects are not looking at the camera, which aligns with naturalistic third-person observations. The GDR task exhibits a balanced distribution across the four directional quadrants (left-closer, left-farther, right-closer, right-farther), ensuring diverse horizontal and depth orientations. Similarly, the GDO task shows a balanced distribution between targets that are closer to or farther from the camera compared to the gaze source, preventing the model from exploiting trivial depth biases.

Table 8: **Training Hyperparameters of LookAnything.**

Hyperparameter	Value
Training Data	HoliGaze-Corpus (303K)
Optimizer	AdamW
Optimizer Hyperparameters	$\beta_1 = 0.9, \beta_2 = 0.95, \epsilon = 10^{-8}$
Weight Decay	0.01
Gradient Clipping	1.0
Learning Rate Schedule	Cosine Decay
Backbone Peak Learning Rate	1×10^{-5}
Geometric Heads Peak Learning Rate	1×10^{-4}
Warmup Ratio	0.03
Global Batch Size	32
Training Epochs	2
Vision Encoder LoRA Rank (r)	16
Vision Encoder LoRA Alpha (α)	32
Depth Loss Weight (λ_d)	0.1
Head Mask Loss Weight (λ_h)	1.0
Target Loss Weight (λ_t)	1.0
Target Heatmap Gaussian σ	3.0
Target Point Huber δ	0.05
Hardware	8 $\times$ NVIDIA H20 GPUs



Figure 9: Word clouds of semantic descriptions of gaze sources and targets.

Spatial Distribution and Shared Attention. Figure 11(a) visualizes the spatial distribution of gaze targets in HoliGaze-Bench. The targets are widely distributed across the image plane, with a noticeable concentration towards the center and lower-middle regions, which is typical for human-object interactions and social scenarios. Furthermore, Figure 11(b) illustrates the distribution of correct option counts for the Gaze Source Detection (GSD) task. While most targets are attended by a single person, a significant portion of targets are looked at by two or more people, highlighting the presence of shared attention and complex social dynamics in the dataset.

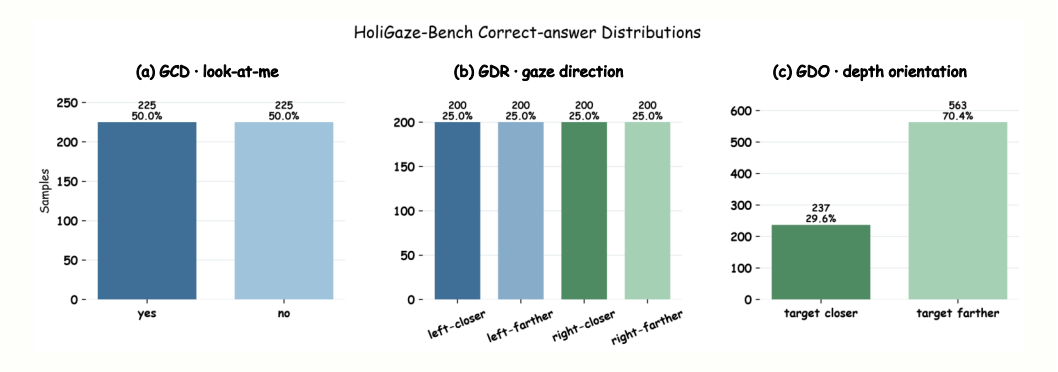


Figure 10: Distributions of correct answers across GCD, GDR, and GDO on HoliGaze-Bench.

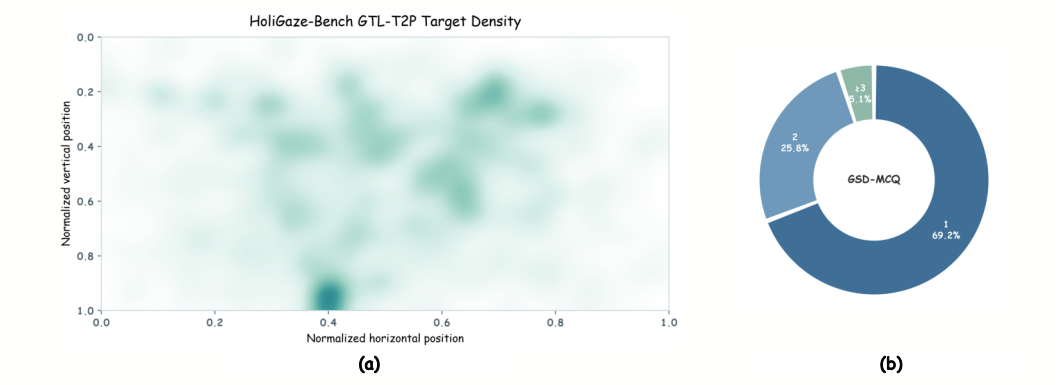


Figure 11: (a) Spatial distribution of gaze targets on HoliGaze-Bench. (b) Distribution of correct option counts on HoliGaze-Bench.

Source–Target Joint Geometry. Beyond the marginal target distribution in Figure 11, we analyze the relative geometry between each gaze source and its target. Figure 12 translates the source-head center to the origin and expresses the target displacement in normalized image coordinates for all 1,210 answerable GTL-T2P items. The horizontal distribution is nearly symmetric, with 48.5% of targets lying to the left of the source, whereas 78.6% lie below the source head. This downward tendency reflects the common composition of gaze-following scenes, in which faces appear above attended hands, objects, or interaction regions. The overall median source–target distance is 0.381 in normalized image coordinates. The distributions remain broad across all three contributing datasets, with median distances of 0.422, 0.368, and 0.331 for GazeFollow, VideoAttentionTarget, and ChildPlay, respectively, demonstrating substantial variation in both gaze direction and spatial extent.

Details of HoliGaze-Corpus and Source-Dataset Composition. Figure 13 provides a comprehensive overview of the statistics for HoliGaze-Corpus, detailing the distribution of different task types and answer lengths. The corpus contains a balanced mix of localization and reasoning tasks, providing a robust foundation for training holistic gaze understanding models. Furthermore, Table 9 summarizes the composition of HoliGaze from various source datasets. To construct a diverse and comprehensive benchmark, we aggregate data from multiple domains, including everyday scenes (GazeFollow, VideoAttentionTarget, VideoCoAtt), human-object interactions (GazeHOI), child-centric environments (ChildPlay), social interactions (Ego4D-Social), and retail environments (GOO). Moreover, Figures 38 to 51 further present diverse data samples across the 9 task formats in the HoliGaze-Corpus. These examples illustrate the high-quality annotations, varied geometric distributions, and rich semantic context provided by our scalable data engine, which collectively empower the holistic capabilities of LookAnything.

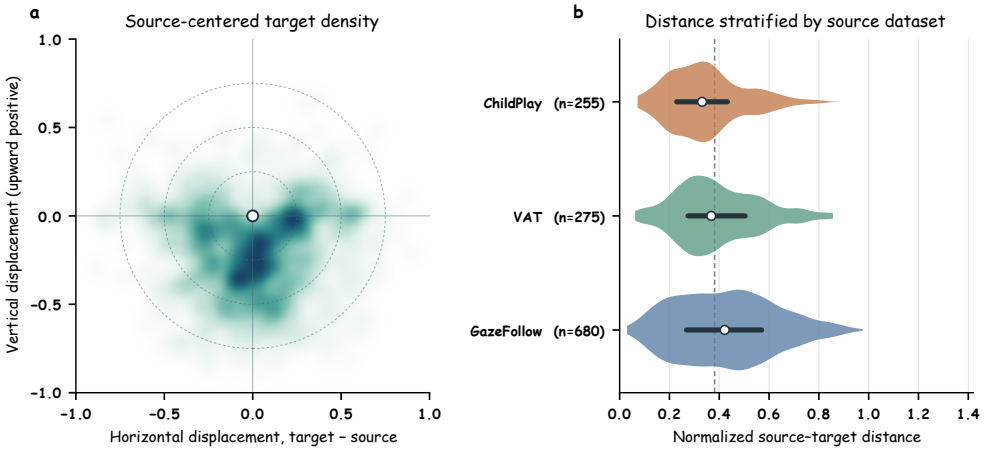


Figure 12: **Source–target joint geometry on HoliGaze-Bench.** **(a)** Log-scaled density of gaze-target displacement relative to the source-head center for the 1,210 answerable GTL-T2P items. Coordinates are normalized by image width and height, and positive vertical displacement denotes a target above the source. **(b)** Source–target distance distributions stratified by source dataset (VAT denotes VideoAttentionTarget). Violins show the full distributions, white points denote medians, thick segments denote interquartile ranges, and the dashed line marks the overall median.

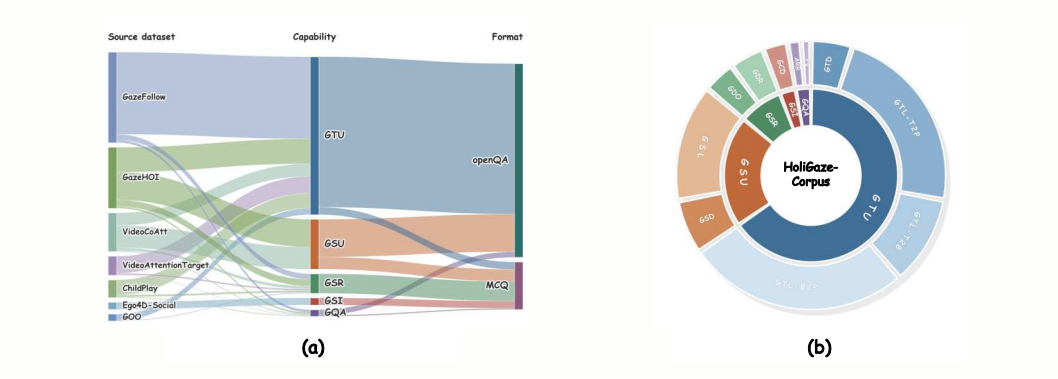


Figure 13: Statistics of HoliGaze-Corpus.

D Additional Dataset Construction Details

In this section, we provide further details on the construction process of the HoliGaze dataset.

D.1 Data Collection

We collect static images and video keyframes from eight source datasets, grouped by complementary coverage:

- **In-the-wild gaze following:** GazeFollow (Recasens et al., 2015) and VideoAttentionTarget (VAT) (Chong et al., 2020) provide large-scale annotations of people looking at targets in diverse, unconstrained everyday scenes. ChildPlay (Tafasca et al., 2023) specifically focuses on child-centric environments, capturing the unique gaze behaviors of children during play.
- **Retail environments:** GOO-Real and GOO-Syn (Tomas et al., 2021) capture shoppers' gaze behaviors in real and synthetic grocery store aisles, offering dense object-level targets.

Table 9: **Source-dataset composition.**

Source	Corpus	%	Bench	%	Overall	%
GazeFollow	170,299	56.20	940	11.75	171,239	55.06
GazeHOI	42,497	14.03	2,404	30.05	44,901	14.44
VideoAttentionTarget	30,094	9.93	540	6.75	30,634	9.85
VideoCoAtt	26,960	8.90	1,696	21.20	28,656	9.21
ChildPlay	21,765	7.18	460	5.75	22,225	7.15
GOO	6,385	2.11	1,510	18.88	7,895	2.54
Ego4D-Social	5,000	1.65	450	5.62	5,450	1.75
Total	**303,000**	100	**8,000**	100	**311,000**	100

- **Human–object interactions:** GazeHOI (Tafasca et al., 2024a) provides rich annotations linking human gaze to specific objects they are interacting with, facilitating joint reasoning of attention and action.
- **Multi-person co-gaze:** VideoCoAtt (Fan et al., 2018) emphasizes shared attention, where multiple subjects in the scene simultaneously focus on the same target.
- **Egocentric gaze contact:** Ego4D-Social (Grauman et al., 2022) captures first-person social interactions, providing critical annotations for mutual gaze and camera contact.

Each source contributes only the fields it natively supports.

Unified Schema. We unify every record into a shared schema comprising the RGB image, the source head region, the target point or box, a monocular depth map, textual referring expressions for the source and the target, and flags for referential ambiguity and out-of-frame targets.

D.2 Annotation Completion

Off-the-shelf models complete missing fields in the unified schema.

Head Masks. Datasets that already provide a head or face box (e.g., GazeFollow, VideoAttentionTarget) do not require external segmentation. The head mask is the axis-aligned rectangle of that box, which is converted to a binary mask. For datasets (GOO) providing only a head point, SAM3 (Carion et al., 2026) recovers the source head boxes and masks. Specifically, SAM3 is prompted with the text "head". Among the returned masks, we select the highest-scoring mask that contains the annotated point. If no mask contains the point, we fall back to the mask whose tight bounding box contains the point, or the box nearest to the point within a strict distance limit. The final head mask is the rasterized rectangle of this selected box.

Scene Depth. Depth Anything v3 (DA3) (Lin et al., 2025a) is employed to estimate a monocular scene depth map for each image.

Referring Expressions. Qwen3-VL-32B (Bai et al., 2025) is utilized to generate referring expressions for both the source person and the gaze target from visual box prompts. Crucially, the image shown to the model is modified to carry a single red bounding box with a thin black outline to explicitly highlight the referred entity. The model writes one long expression and one short expression for each valid entity. The prompt templates for generating long and short referring descriptions of gaze sources and targets are depicted in Figure 16, Figure 17, Figure 18, and Figure 19, respectively.

Ambiguity Check. To construct the ambiguity labels, Qwen3-VL-32B is queried to count how many entities each referring expression matches in the image, using the prompts depicted in Figure 20 and Figure 21. Unique matches (count of 1) proceed as regular samples, while non-unique matches (count ≥ 2) are reserved for Ambiguous Reference Recognition (ARR) tasks.

Head-Orientation Judgment. For gaze depth and direction questions, Qwen3-VL-32B predicts the head orientation relative to the camera, following the prompt template shown in Figure 22. The predictions are restricted to `front`, `back`, or `uncertain`.

Table 10: **Canonical HoliGaze question and answer templates.** {person} and {target} are filled by referring expressions. GTL-B2P replaces {person} with a head-box token such as `head box [x1,y1,x2,y2]`.

Task	Question templates	Answer grammar
GTD	What is {person} looking at? / What is the gaze target of {person}?	{target}.
GTL-T2P	Where is {person} looking at? Output only its normalized point as (x,y), using integer coordinates from 0 to 1000.	(x,y)
GTL-T2B	Where is {person} looking at? Output only its normalized bounding box as [x1,y1,x2,y2], using integer coordinates from 0 to 1000.	[x1,y1,x2,y2]
GTL-B2P	Locate the gaze target of the person with head box [x1,y1,x2,y2]. Output only its normalized point as (x,y), ...	(x,y)
GSD	Who is looking at {target}? Provide all gaze sources that satisfy the condition.	{persons}.
GSL	Locate the person who is looking at {target}. / Find the gaze source for {target}. Output only the person's normalized head bounding box as [x1,y1,x2,y2], ... If multiple people satisfy the condition, provide all of their head bounding boxes.	[x1,y1,x2,y2]; ...
GDO	Is the head of {person} farther from / closer to the camera than what the person is looking at?	yes / no
GDR	From the camera's perspective, relative to the head of {person}, is the gaze target left or right in the image, and closer to or farther from the camera?	left-closer / left-farther / right-closer / right-farther
GCD	Is {person} looking toward the camera lens? / Is {person} looking at the camera?	yes / no
ARR	Host GTD or GTL question with a non-unique referring expression.	I cannot answer because the referred entity in the question is not unique.
PV	Host GTL-T2P question whose target lies outside the frame.	I cannot answer because the gaze target is not visible in the image.

Relative-Contrast Depth Filter. Using the depth map from DA3, we compute the relative contrast between the head point and target point. Let d_s and d_t be the median depth values in a 5×5 window around the source and target points, respectively, and let f be the 5th percentile of positive depths in the image (with a floor of 10^{-3}). The relative contrast is defined as $r = |d_t - d_s| / \max(\min(|d_t|, |d_s|), f)$. A sample pair is removed if $r \leq 0.10$. Otherwise, the pair is labeled by the depth difference polarity (i.e., the sign of $d_t - d_s$).

Depth Consensus Filtering. To ensure high-quality supervision for the Gaze Depth Ordering (GDO) and Gaze Direction Recognition (GDR) tasks, we apply a strict consensus filter. A sample is retained only if the facial orientation polarity predicted by Qwen3-VL-32B logically agrees with the DA3 depth polarity. For instance, if the head orientation is judged as "back" (facing away from the camera), the DA3 depth map must show that the target is farther from the camera than the source head ($d_t > d_s$) and pass the relative contrast threshold. This consensus prevents relative closer/farther labels from being inverted by monocular scale ambiguity.

D.3 Q&A Construction

Task-specific templates instantiate the schema into openQA and MCQ formats.

Canonical Templates. To unify diverse gaze understanding tasks into a standard visual question answering (VQA) format, we design canonical templates for each task, as detailed in Table 10. The templates utilize referring expressions to ground the gaze sources and targets, ensuring natural language interaction. For tasks requiring precise localization, such as GTL-B2P, the templates incorporate explicit bounding box coordinates to guide the model's spatial reasoning.

OpenQA Generation. The openQA format encompasses GTL, GTD, GSL, GSD, as well as the refusal tasks ARR and PV. For the standard localization and description tasks, the canonical templates directly prompt the model to generate the corresponding spatial coordinates (points or boxes) or semantic descriptions.

Ambiguity and Presence Verification (Refusal). To evaluate the model's ability to refuse ill-posed queries in the openQA format, ambiguous expressions identified during the Ambiguity Check are kept as ARR positives. Similarly, samples where the gaze target is known to be out-of-frame are utilized to construct Presence Verification (PV) questions. For both ARR and PV, the model is expected to output standardized refusal templates (see Table 10) rather than guessing a location or description.

MCQ Distractors. MCQ covers all 9 main tasks. GTD-MCQ provides 5 single-choice options, each pairing a description with a bounding box; an option is correct only if both match. GSD-MCQ provides 6 options and allows multi-select (1–3) to accommodate co-gaze. Distractors for both tasks are drawn from four distinct sources to ensure robustness: (i) **Expansion**: Other in-image entities are proposed by SAM3 and subsequently re-described by Qwen3-VL-32B (Bai et al., 2025)[1]. (ii) **Attribute Perturbation**: Qwen3-32B rewrites the gold description into hard negatives by altering 1–4 specific attributes (e.g., changing "red shirt" to "blue shirt") while preserving the rest of the sentence, utilizing the template illustrated in Figure 23. (iii) **Permutation**: We swap bounding boxes across options so that a correct description can be paired with a wrong location (e.g., a translated copy of the ground-truth box or another entity's box), and vice versa. (iv) **Refusal Templates**: Standardized ARR refusal templates are included as distractors to test if the model incorrectly identifies a valid query as ambiguous.

Data Splits. To prevent leakage, we follow each source's native split to curate **HoliGaze-Corpus** and **HoliGaze-Bench**, with the former sampled only from training sets, and the latter from test sets (except Ego4D-Social, whose official validation split is used). Sampling quotas over the 9 gaze-centric tasks and QA formats are reported in §2.3.

D.4 Human Verification

To ensure a high-quality evaluation standard, HoliGaze-Bench undergoes rigorous instance-level human verification. We engage 5 trained annotators to cross-check every sample in the benchmark. The annotators meticulously verify multiple dimensions of correctness: (1) the accuracy and visual faithfulness of the referring descriptions; (2) the strict uniqueness of the references within the image; (3) the logical validity and mutual exclusivity of the MCQ distractors; (4) the spatial consistency and tightness of the predicted points and bounding boxes; (5) the correctness of the depth polarity for GDO/GDR tasks; and (6) the accurate labeling of out-of-frame targets for PV tasks. Any sample failing these checks is either manually corrected or discarded. Ultimately, the annotators modified only 56 of the 8,000 instances (a modification rate of 0.70%). This exceptionally low rate indicates that the automated data engine already enforces strong quality control, and the subsequent human review further guarantees a highly reliable evaluation set.

E Additional Experimental Results

E.1 Scaling Analysis

As a foundation model for holistic gaze understanding, LookAnything is expected to benefit from larger training data volumes. To examine this, we perform a scaling analysis by evaluating the model checkpoints at different pretraining stages, which correspond to the amount of training data seen by the model (ranging from 0.25 to 2.0 epochs).

As shown in Figure 14, model performance improves consistently with increased training data across both the holistic capabilities and conventional tasks. On the HoliGaze-Bench (top row), the multiple-choice accuracy (Avg. MCQ), the generation quality for open-ended queries (Avg. openQA), and

[1]Note: For GOO, distractor objects are taken from ground-truth product boxes rather than being proposed by SAM3. However, following the main text, we broadly characterize the expansion as being proposed by SAM3.

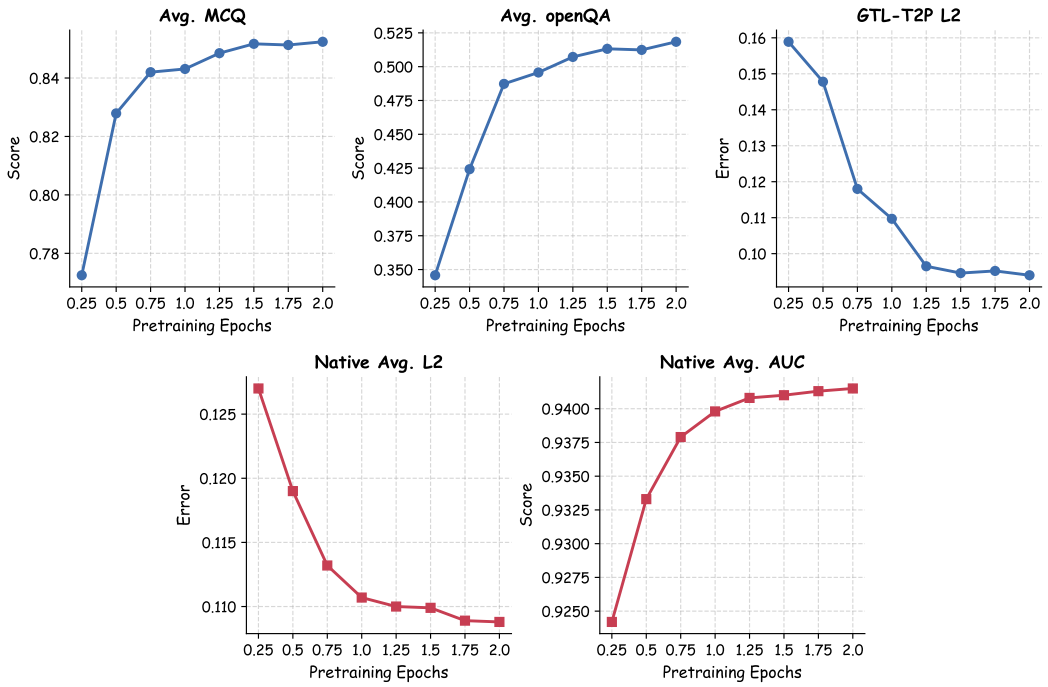


Figure 14: Effect of training data volume (pretraining epochs) on the LookAnything performance. The top row illustrates the holistic gaze understanding metrics on HoliGaze-Bench, while the bottom row demonstrates the average performance across the 5 native benchmarks.

the precise spatial localization error (GTL-T2P L_2) all show steady gains as the data size increases. A similar trend is observed on the conventional native benchmarks (bottom row), where the average L_2 error continuously decreases and the Area Under Curve (Avg. AUC) progressively rises.

Notably, the performance gains are especially prominent in the early stages of training (from 0.25 to 1.0 epochs). As the data volume scales further, the improvement begins to plateau, demonstrating the classic diminishing returns of neural scaling laws. These results strongly support the core motivation of building a foundation model for gaze: scaling up the diverse and heterogeneous gaze data fundamentally enhances the model's representational power and downstream generalizability.

E.2 Visualization of Auxiliary Predictions

As shown in Figure 15, we visualize the auxiliary predictions generated by the structurally disentangled queries in LookAnything, including the scene depth maps, the source head masks, and the gaze target spatial heatmaps. Although the auxiliary decoders adopt a lightweight MAE-style architecture with patch-level reconstruction, LookAnything is able to reconstruct semantically meaningful geometric and physical representations of the entire scene.

The effectiveness of these predictions highlights the value of our auxiliary supervision design. Empowered by the large-scale multi-task pretraining on the extensive HoliGaze-Corpus, the gaze-prior queries are driven to implicitly integrate global visual context, such as background structures, spatial layout, and object appearances, with localized gaze behaviors. As a result, the model yields coherent and accurate predictions for the critical physical priors, successfully internalizing them without relying on explicit external inputs like cropped head images or pre-computed depth maps.

Although the predicted depth maps are relatively coarse due to the inherent patch-level resolution of the auxiliary decoders (He et al., 2022), the source head masks remain remarkably accurate and well-localized. Together, they provide essential geometric guidance for downstream tasks. In particular, the language backbone fundamentally benefits from anticipating the scene depth and grounding the

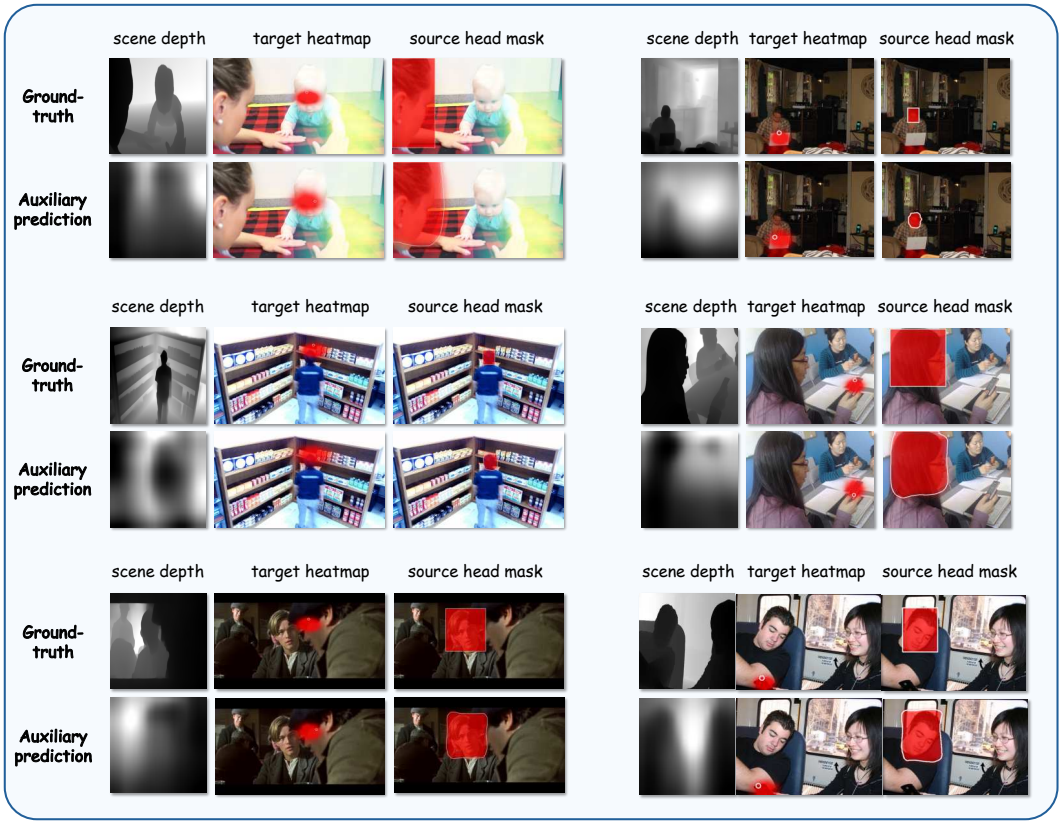


Figure 15: Visualizations of auxiliary predictions generated by the structurally disentangled queries.

human head, which helps resolve spatial ambiguities, refine precise target localization, and improve overall holistic gaze understanding.

E.3 Qualitative Visualization Results

In this section, we present additional qualitative visualization results on HoliGaze-Bench across diverse gaze-centric tasks. Figures 24 to 37 showcase the model's capabilities across the 9 defined tasks.

Specifically, Figures 24 to 28 demonstrate the precise grounding capabilities in Gaze Target Localization (GTL) across both point and box formats, handling complex spatial relationships and distant targets. Figure 29 visualizes Gaze Target Description (GTD), where the model accurately attributes semantic meaning to the localized targets. Conversely, Figures 30 to 33 illustrate Gaze Source Localization (GSL) and Description (GSD), highlighting the model's ability to trace back from objects to the attending human sources, even in multi-person shared attention scenarios.

Furthermore, Figure 34 demonstrates strong geometric reasoning in the Gaze Depth Orientation (GDO) and Gaze Direction Recognition (GDR) tasks, where the model successfully discerns fine-grained 3D spatial arrangements and camera-relative directions. Figure 35 shows accurate Gaze Contact Detection (GCD), which is crucial for identifying social intent. Finally, Figures 36 and 37 highlight the model's robustness in answering ill-posed queries via Ambiguous Reference Recognition (ARR) and Presence Verification (PV), appropriately issuing refusals instead of hallucinating answers.

Overall, as observed in these diverse examples, LookAnything effectively and robustly processes complex multimodal queries, grounding visual entities precisely while providing accurate geometric reasoning and semantic interpretations.

Long Referring Description Prompt · Source Person

[System Prompt]

[Task]

You are an expert multimodal annotator for gaze and attention VQA datasets. The image contains exactly one clearly drawn RED bounding box with a thin BLACK outline. Describe only the entity enclosed by that visual box so a reader can distinguish it from similar entities in the same image.

[Rules]

1. Use only clearly visible evidence. Never invent identity, brand, occupation, relationships, emotion causes, or off-screen content.

2. Select exactly 3 to 5 concise, discriminative attributes. Prefer appearance, clothing, pose, position, color, material, shape, or visibility as appropriate.

3. Produce one concise English noun phrase that uniquely refers to the entity.

4. The red and black border pixels belong only to the annotation. Never use the border colors as evidence for the entity's color or appearance.

5. Do not mention the box, coordinates, annotations, gaze, or these rules.

6. Output strict JSON with exactly two top-level keys: "attributes" and "referring_expression". Output no markdown and no commentary.

[User Prompt]

[Task]

Describe only the person enclosed by the red visual bounding box with a black outline. The marked region is the person's head / face (the body may be only partly visible).

[Rules]

Choose exactly 3 to 5 visible attributes that best distinguish this person from other people in the image, then write one overall referring phrase.

[Output]

Return only this JSON shape:

```
{
  "attributes": [
    {"name": "attribute_name", "value": "short visible attribute"},
    {"name": "attribute_name", "value": "short visible attribute"},
    {"name": "attribute_name", "value": "short visible attribute"}
  ],
  "referring_expression": "one concise, discriminative English noun phrase"
}
```

Figure 16: Prompt template for long referring description generation of source person.

Long Referring Description Prompt · Gaze Target

[System Prompt]

[Task]

You are an expert multimodal annotator for gaze and attention VQA datasets. The image contains exactly one clearly drawn RED bounding box with a thin BLACK outline. Describe only the entity enclosed by that visual box so a reader can distinguish it from similar entities in the same image.

[Rules]

1. Use only clearly visible evidence. Never invent identity, brand, occupation, relationships, emotion causes, or off-screen content.

2. Select exactly 3 to 5 concise, discriminative attributes. Prefer appearance, clothing, pose, position, color, material, shape, or visibility as appropriate.

3. Produce one concise English noun phrase that uniquely refers to the entity.

4. The red and black border pixels belong only to the annotation. Never use the border colors as evidence for the entity's color or appearance.

5. Do not mention the box, coordinates, annotations, gaze, or these rules.

6. Output strict JSON with exactly two top-level keys: "attributes" and "referring_expression". Output no markdown and no commentary.

[User Prompt]

[Task]

Describe only the entity enclosed by the red visual bounding box with a black outline. It may be an object, a region, or a person.

[Rules]

Choose exactly 3 to 5 visible attributes that best distinguish it from similar entities in the image, then write one overall referring phrase.

[Output]

Return only this JSON shape:

```
{
  "attributes": [
    {"name": "attribute_name", "value": "short visible attribute"},
    {"name": "attribute_name", "value": "short visible attribute"},
    {"name": "attribute_name", "value": "short visible attribute"}
  ],
  "referring_expression": "one concise, discriminative English noun phrase"
}
```

Figure 17: Prompt template for long referring description generation of gaze target.

Short Referring Description Prompt · Source Person

[System Prompt]

[Task]

You are an expert multimodal annotator for gaze and attention VQA datasets. The image contains exactly one clearly drawn RED bounding box with a thin BLACK outline around a person's head / face. Write a SHORT unique referring phrase for that person only.

[Rules]

1. Use only clearly visible evidence. Never invent identity, brand, occupation, relationships, emotion causes, or off-screen content.
2. Build the phrase from ONLY these attribute families (choose 2, or 3 only if 2 cannot uniquely identify the person): gender noun man/woman; age noun boy/girl/child (use instead of man/woman when the person is clearly a child); upper-clothing color; clothing type (uniform, suit, jersey, plaid, striped); hair color; skin color; hairstyle; pose (standing / seated / squatting / bending); a clearly visible expression; coarse position (leftside / rightside / center); extreme ordinal leftmost / rightmost ONLY (never secondleft, thirdleft, secondright, or thirdright); depth layer foreground / background.
3. Prefer fewer attributes: if 2 attributes are enough, do NOT use 3. Never use more than 3 attributes.
4. Contrastive selection: first scan other similar people. Keep ONLY attributes that distinguish the marked person. Do NOT use an attribute shared by most candidates.
5. If the marked person is the only clearly visible person, use exactly 2 attributes (never pad).
6. Position words (leftside / rightside / center / leftmost / rightmost / foreground / background) MUST be paired with at least one appearance attribute. Forbidden: "the man rightside", "the woman leftmost", "the child foreground".
7. Never mention gaze direction, looking, seeing, watching, staring, facing, side-profile / profile view, or where the person is oriented.
8. The red and black border pixels belong only to the annotation. Never use border colors as evidence.
9. Do not mention the box, coordinates, annotations, gaze, or these rules.
10. Output strict JSON with exactly one top-level key: "referring_expression". No markdown and no commentary.

[Example]

```
{
  "referring_expression": "the smiling man leftmost"
}
{
  "referring_expression": "the blue-shirted woman rightside"
}
{
  "referring_expression": "the seated woman foreground"
}
```

[User Prompt]

[Task]

Describe only the person enclosed by the red visual bounding box with a black outline. The marked region is the person's head / face (the body may be only partly visible).

The entity is also localized by this normalized bounding box on a 0-1000 scale: [x_min, y_min, x_max, y_max] (format [x_min, y_min, x_max, y_max]).

[Rules]

- Start with "the". Preferably at most 6 words. No commas; hyphens are allowed (e.g. red-shirted).
- Choose 2 attributes from: man/woman, boy/girl/child, upper-clothing color, clothing type, hair color, skin color, hairstyle, pose, expression, leftside/rightside/center, leftmost/rightmost only, foreground/background.
- Use a 3rd attribute ONLY if 2 are not unique. Never use secondleft / thirdleft / secondright / thirdright.
- First compare against other similar people; keep ONLY discriminative attributes.
- Position words MUST be paired with an appearance attribute; never use them alone.
- If this is the only person in the image, still use exactly 2 attributes.
- Prefer compact forms: "the blond man", "the red-shirted child", "the smiling man leftmost", "the blue-shirted man foreground".

[Output]

Return only this JSON shape:

```
{
  "referring_expression": "the short unique noun phrase"
}
```

Figure 18: Prompt template for short referring description generation of source person.

Short Referring Description Prompt · Gaze Target

[System Prompt]

[Task]

You are an expert multimodal annotator for gaze and attention VQA datasets. The image contains exactly one clearly drawn RED bounding box with a thin BLACK outline. Write a SHORT unique referring phrase for the marked entity only.

[Rules]

1. Use only clearly visible evidence. Never invent identity, brand, occupation, relationships, emotion causes, or off-screen content.
2. The phrase MUST start with lowercase "the", should preferably use at most 6 words, and must NOT contain commas. Hyphens are allowed inside compound words. No other punctuation and no digits.
3. Prefer fewer cues: use 2 cues when enough; use a 3rd cue only if needed for uniqueness.
4. Contrastive selection: first scan other similar entities. Keep ONLY cues that distinguish the marked entity.
5. If the marked entity is a person, use ONLY: gender (man/woman), age noun boy/girl/child, upper-clothing color, clothing type, hair color, skin color, hairstyle, pose, expression, coarse position (leftside/rightside/center), foreground/background, and leftmost/rightmost ONLY.
6. If the marked entity is an object or region, use ONLY: object/region type, main color, coarse position, foreground/background, and leftmost/rightmost among similar objects.
7. Position words MUST be paired with at least one appearance or type cue. Forbidden: "the man rightside", "the cup leftmost", "the box foreground".
8. Never mention gaze direction, looking, seeing, watching, staring, facing, side-profile, or orientation.
9. Never use the red or black border colors as evidence.
10. Do not mention the box, coordinates, annotations, gaze, or these rules.
11. Output strict JSON with exactly one top-level key: "referring_expression".

[Example]

```
{
  "referring_expression": "the red cup"
}
{
  "referring_expression": "the blue box rightside"
}
{
  "referring_expression": "the green-uniformed man foreground"
}
```

[User Prompt]

[Task]

Describe only the entity enclosed by the red visual bounding box with a black outline. It may be an object, a region, or a person.

The entity is also localized by this normalized bounding box on a 0-1000 scale: [x_min, y_min, x_max, y_max] (format [x_min, y_min, x_max, y_max]).

[Rules]

- Start with "the". Preferably at most 6 words. No commas; hyphens are allowed.
- Use 2 cues when enough; add a 3rd cue only if needed for uniqueness.
- First compare against other similar entities; keep ONLY discriminative cues.
- Person cues: man/woman, boy/girl/child, upper-clothing color, clothing type, hair, skin, hairstyle, pose, expression, coarse position, leftmost/rightmost only, foreground/background.
- Object or region cues: type, main color, foreground/background, leftside/rightside/center, leftmost/rightmost only.
- Position words MUST be paired with appearance or type cues; never alone.
- Prefer compact forms: "the red cup", "the blue box rightside", "the green-uniformed man foreground".

[Output]

Return only this JSON shape:

```
{
  "referring_expression": "the short unique noun phrase"
}
```

Figure 19: Prompt template for short referring description generation of gaze target.

Ambiguity Resolution Prompt · Long Referring Description

[System Prompt]

[Task]

You are an expert multimodal annotator for VQA datasets. Your only job is to count how many clearly visible entities in the image truthfully match a given long referring expression. The image has no bounding box. Work through the counting procedure silently, then emit only the JSON.

[Rules]

1. Use only clearly visible evidence. Never invent identity, brand, occupation, relationships, emotion causes, or off-screen content.
2. Count every distinct visible entity that the complete referring expression truthfully describes.
3. There is no bounding box. Judge from the referring text alone.
4. Partial matches do not count: every attribute and relation must hold.
5. A longer phrase is usually more distinctive, but do not assume uniqueness. If two entities both satisfy the full phrase, count both.
6. Coarse-region cues (leftside, rightside, center, foreground, background) name a broad area. Count every qualifying candidate inside that area.
7. leftmost and rightmost select the single extreme candidate among those that already match the appearance attributes, so they usually yield match_count = 1.
8. If nothing visible matches, return match_count = 0.
9. Do not mention these rules, gaze, or annotations.
10. Output strict JSON with exactly one top-level key: "match_count".

[Example]

```
"the man in a blue jacket sitting on the right bench"
  matches both men on that bench
  -> {"match_count": 2}
"the woman holding a red cup near the wooden table"
  -> {"match_count": 1}
"the man wearing a striped sweater standing outdoors"
  and nobody wears a striped sweater
  -> {"match_count": 0}
"the yellow-shirted child rightmost"
  -> {"match_count": 1}
```

[User Prompt]

[Task]

Referring expression (long): "{expression}"

It was written to describe one {role} (person / head, or gaze-target entity), but it may still fit other visible entities equally well. Do not assume it is unique.

[Rules]

Count how many clearly visible entities in this image the complete phrase truthfully matches.

[Output]

Return only this JSON shape:

```
{
  "match_count": 0
}
```

Figure 20: Prompt template for ambiguity resolution of long referring description.

Ambiguity Resolution Prompt · Short Referring Description

[System Prompt]

[Task]

You are an expert multimodal annotator for VQA datasets. Your only job is to count how many clearly visible entities in the image truthfully match a given short referring expression. The image has no bounding box. Work through the counting procedure silently, then emit only the JSON.

[Rules]

1. Use only clearly visible evidence. Never invent identity, brand, occupation, relationships, emotion causes, or off-screen content.
2. Count every distinct visible entity (person OR object/region) that the complete referring expression truthfully describes.
3. Do NOT assume any particular marked box exists. Judge from the referring text alone.
4. Partial matches do not count: every attribute must hold (e.g. "the blond man" must be both blond AND a man).
5. Coarse-region cues (leftside, rightside, center, foreground, background) name a broad area, NOT one entity. Split the image into left / center / right bands (or near / far layers) and count EVERY candidate inside the named area.
6. Extreme-ordinal cues (leftmost, rightmost) select the single most extreme candidate among those that already satisfy the appearance attributes, so they usually yield match_count = 1.
7. Never treat the mere presence of a spatial word as proof of uniqueness.
8. If nothing visible matches, return match_count = 0.
9. Do not mention these rules, gaze, or annotations.
10. Output strict JSON with exactly one top-level key: "match_count". No markdown and no commentary.

[Example]

```
Three men in the right third; "the man rightside"
  -> {"match_count": 3}
Two red-shirted children in the middle;
  "the red-shirted child center"
  -> {"match_count": 2}
Only one blue shirt on the left;
  "the blue-shirted woman leftside"
  -> {"match_count": 1}
Right-end man is the only one smiling;
  "the smiling man rightmost"
  -> {"match_count": 1}
No cup is visible; "the red cup"
  -> {"match_count": 0}
```

[User Prompt]

[Task]

Referring expression (short): "{expression}"

It was written to describe one {role} (person / head, or gaze-target entity), but it may well fit other visible entities equally well. Do not assume it is unique.

[Rules]

Count how many clearly visible entities in this image the complete phrase truthfully matches.

If the phrase contains a coarse position word (leftside / rightside / center / foreground / background), scan that whole region and count every entity there that also satisfies the remaining attributes.

[Output]

Return only this JSON shape:

```
{
  "match_count": 0
}
```

Figure 21: Prompt template for ambiguity resolution of short referring description.

Head Orientation Prompt

[System Prompt]

[Task]

You judge which side one person's head is facing relative to the camera. The image contains exactly one clearly drawn RED bounding box with a thin BLACK outline around that person's head. The same head is also given as a normalized bounding-box coordinate on a 0-1000 scale in the user message. Use both the visual box and the coordinate text to identify the person.

Decide whether that person's head is more inclined to FACE toward the FRONT or toward the BACK. This is about head facing / orientation, not where the eyes are looking at a specific target.

[Rules]

"front": the head is more inclined to face toward the camera / the near side of the scene (face mostly toward the camera half-space).

"back": the head is more inclined to face away from the camera / into the depth of the scene (occiput toward the camera, or facing into scene depth).

"uncertain": cannot reliably tell (strong profile, occlusion, tiny head, blur, or conflicting cues).

Also report confidence as "high", "medium", or "low".

Do not invent off-screen content. Do not output left/right or any other fields.

[User Prompt]

[Task]

The red box with a black outline marks one person's head.

The person's head is also localized by this normalized bounding box on a 0-1000 scale: [x_min, y_min, x_max, y_max] (format [x_min, y_min, x_max, y_max]).

[Rules]

Is this person's head more inclined to face toward "front", "back", or "uncertain"?

Also give confidence as "high", "medium", or "low".

Optional person referring expression: {expression}

[Output]

Return only this JSON shape:

```
{
  "gaze_direction": "front" | "back" | "uncertain",
  "confidence": "high" | "medium" | "low"
}
```

Figure 22: Prompt template for face orientation prediction.

Attribute Perturbation Prompt

[System Prompt]

[Task]

You create minimally perturbed hard-negative referring expressions for multiple-choice gaze-following questions. Every replacement must be a genuinely mutually exclusive attribute value for the same entity slot.

[User Prompt]

[Task]

Create ONE perturbed referring expression for use as an INCORRECT MCQ option.

Role of the original phrase: {role} (gaze source, gaze target, or in-image candidate).

Length tier: {tier} (short = a compact phrase; long = a richer description).

Required number of attribute replacements: {1-2 if short, 2-4 if long}.

Original referring expression:
{expression}

[Rules]

1. Make a MINIMAL perturbation: change exactly that many attribute values. Each list item is exactly ONE atomic attribute slot. Never bundle multiple changes (for example color + pattern) into one replacement.
2. Each old/new pair MUST be mutually exclusive for the SAME slot and the same referent. Good pairs: male/female, red shirt/blue shirt, short hair/long hair, striped/solid, left/right, wooden/metal, cat/dog. Bad pairs: hat/glasses, red shirt/blue trousers, standing/wearing a coat, or any two features that could coexist.
3. Prefer replacing explicit attribute words already present in the original. Preserve the entity category and every non-selected attribute. Do not add or remove unrelated details.
4. Keep fluent English and roughly the same length and style. This must remain a light edit, not a newly invented description.
5. The result must clearly be an INCORRECT option, not a paraphrase, synonym swap, weaker or stronger wording, or an article- or punctuation-only edit.
6. Do not invent gaze verbs or mention looking/gazing. Output a noun phrase only.
7. Each original_value must occur verbatim in the original expression, each replacement_value must occur verbatim in the perturbed expression, and the old value must no longer occur in the perturbed expression.
8. The perturbed expression MUST equal the original expression after applying only the listed substitutions. Every other word, attribute, article, and phrase must remain unchanged.

[Output]

```
{
  "perturbed_expression": "<string>",
  "attribute_replacements": [
    {
      "attribute": "<specific attribute slot, e.g. shirt_color>",
      "original_value": "<exact words from original>",
      "replacement_value": "<exact mutually exclusive words>",
      "exclusivity_reason": "<why both values cannot hold>"
    }
  ],
  "rationale": "<short explanation of the minimal edit>"
}
```

Figure 23: Prompt template for attribute perturbation of referring description.

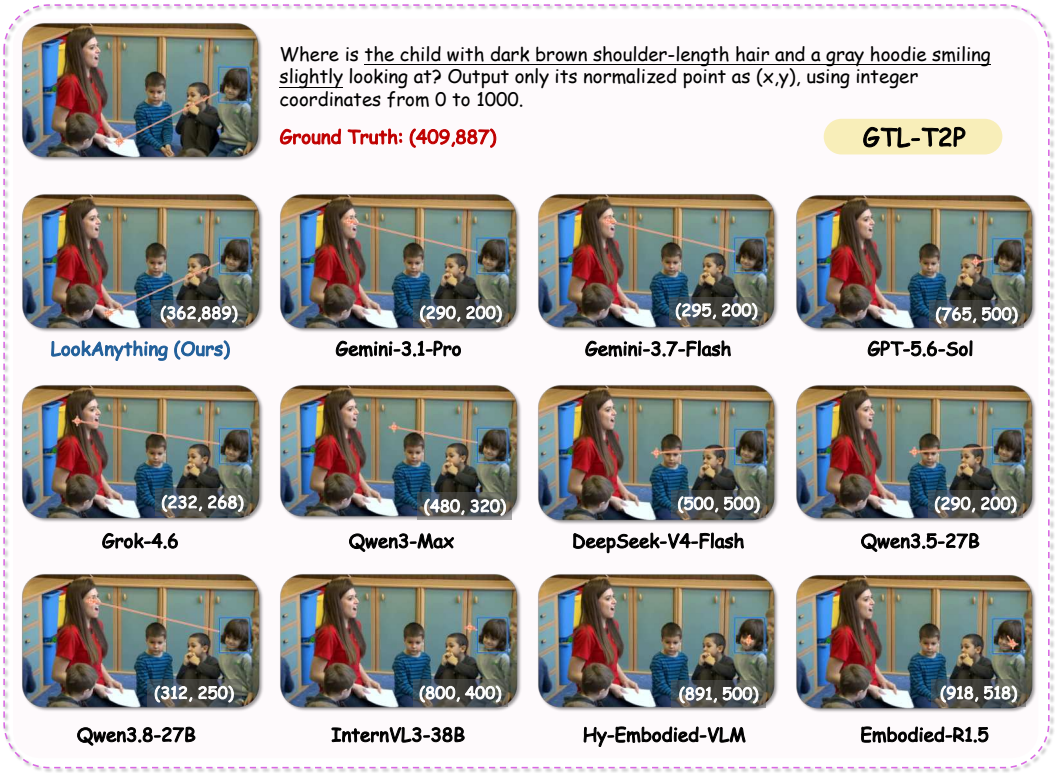


Figure 24: Qualitative Cases of GTL-T2P.

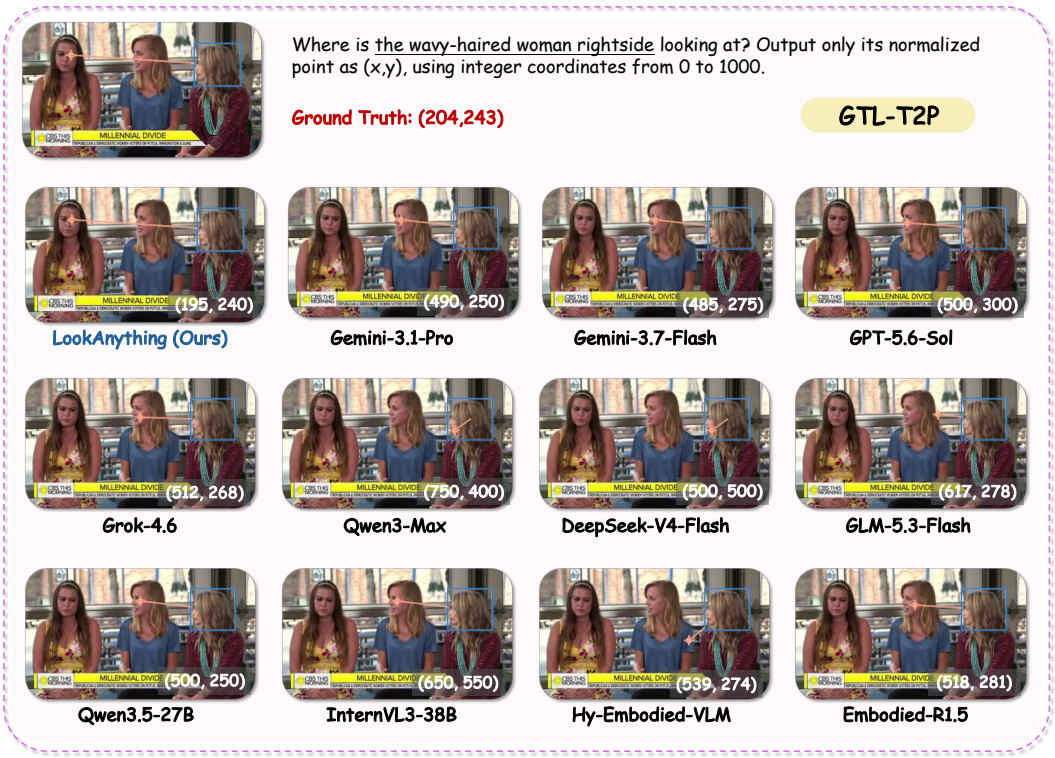


Figure 25: Qualitative Cases of GTL-T2P.

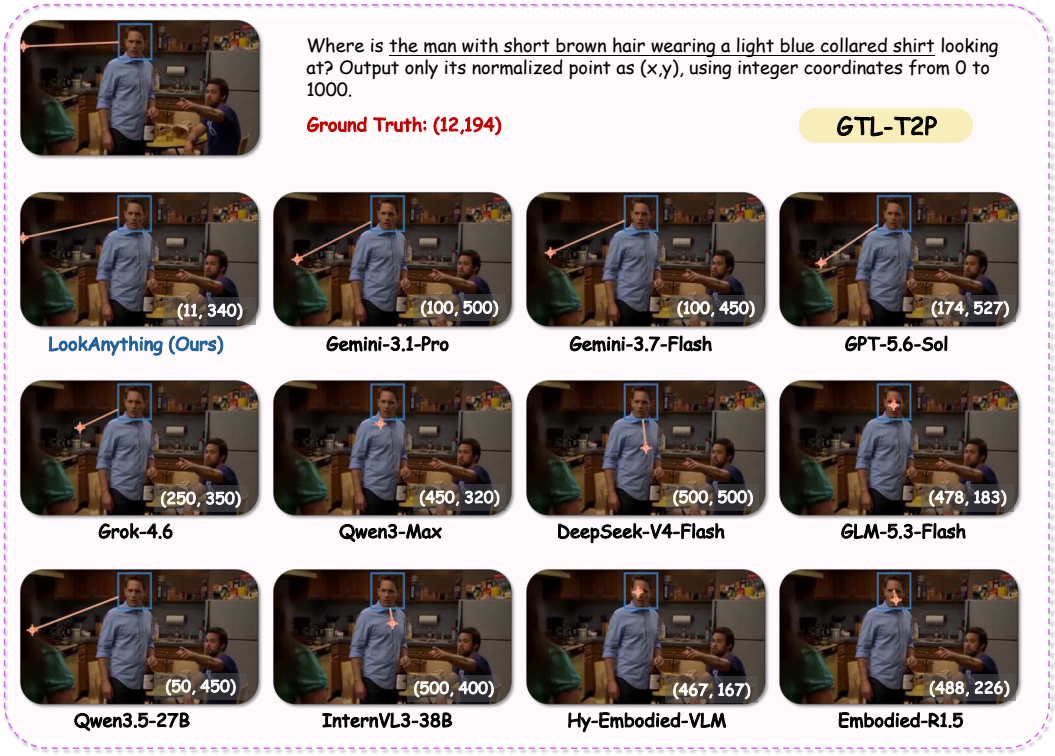


Figure 26: Qualitative Cases of GTL-T2P.

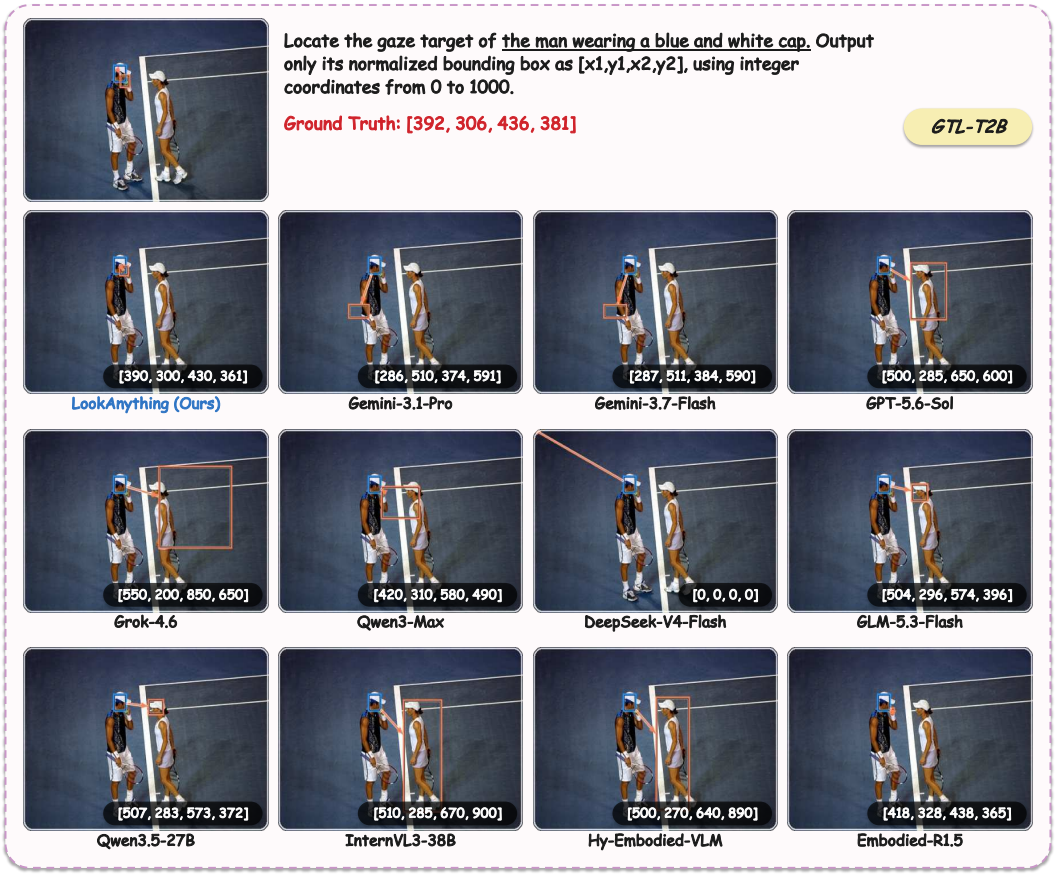


Figure 27: Qualitative Cases of GTL-T2B.

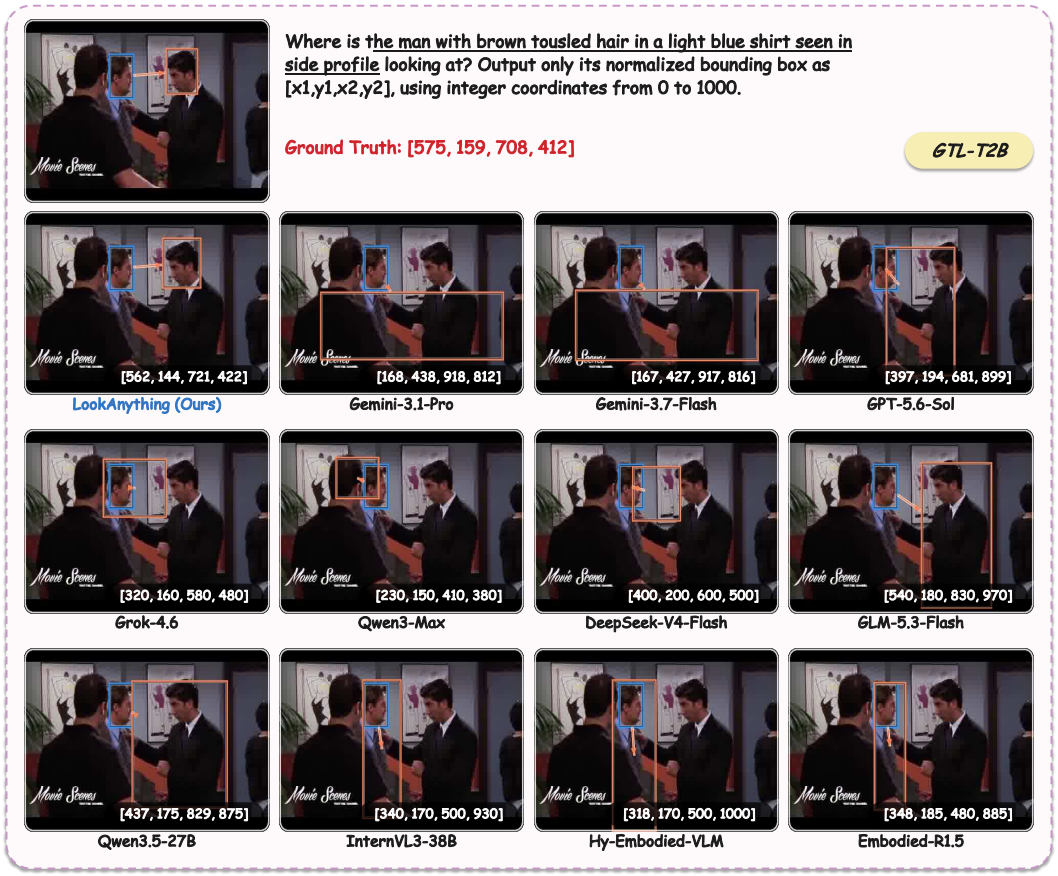


Figure 28: Qualitative Cases of GTL-T2B.

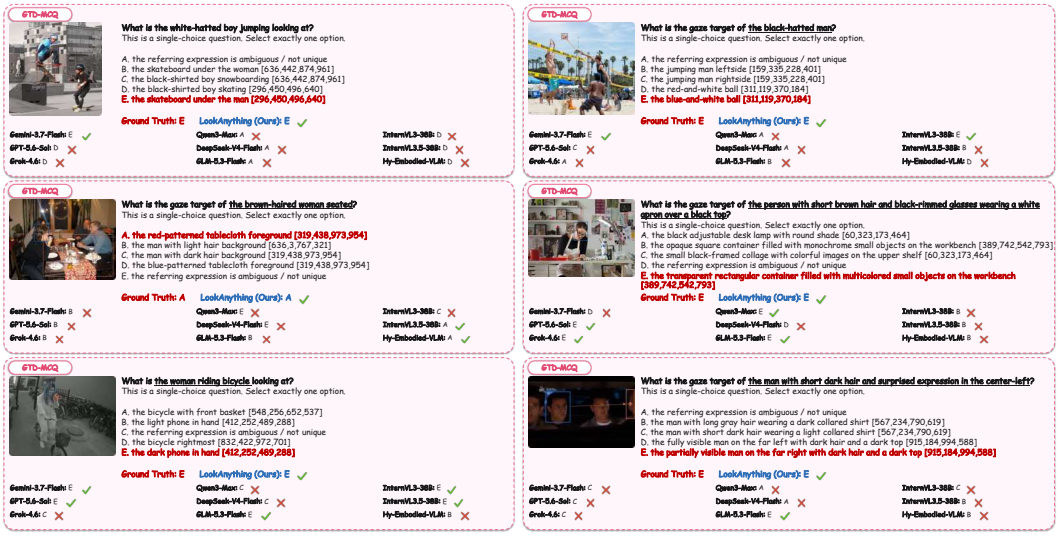


Figure 29: Qualitative Cases of GTD.

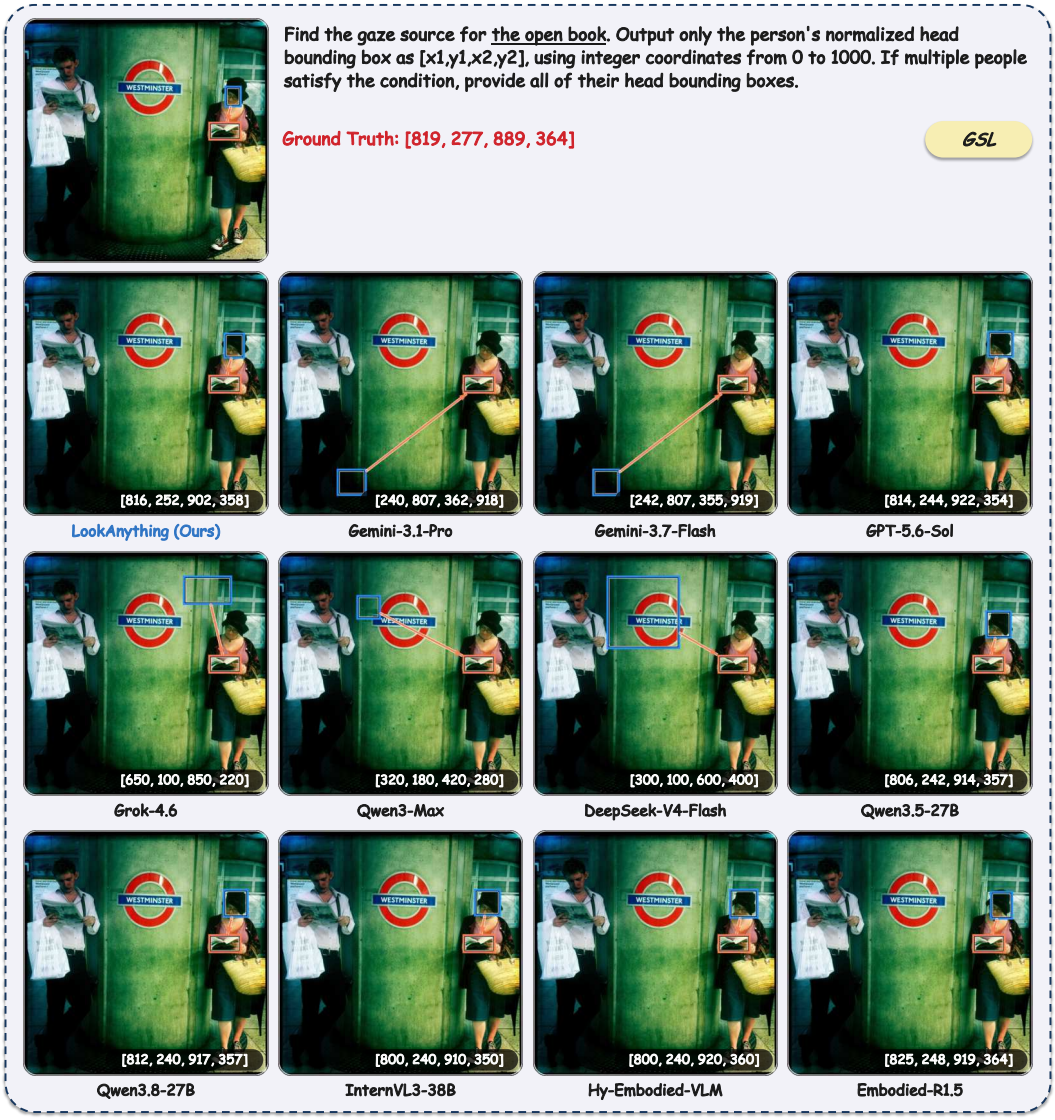


Figure 30: Qualitative Cases of GSL.



Figure 31: Qualitative Cases of GSL.

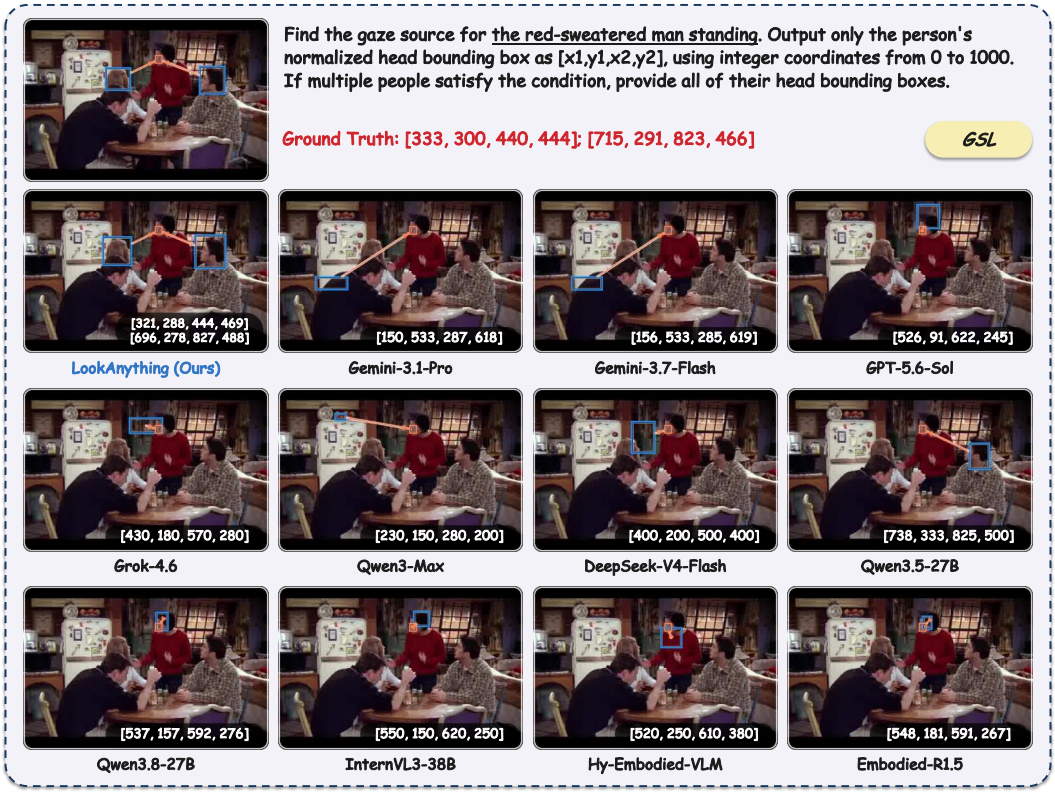


Figure 32: Qualitative Cases of GSL.

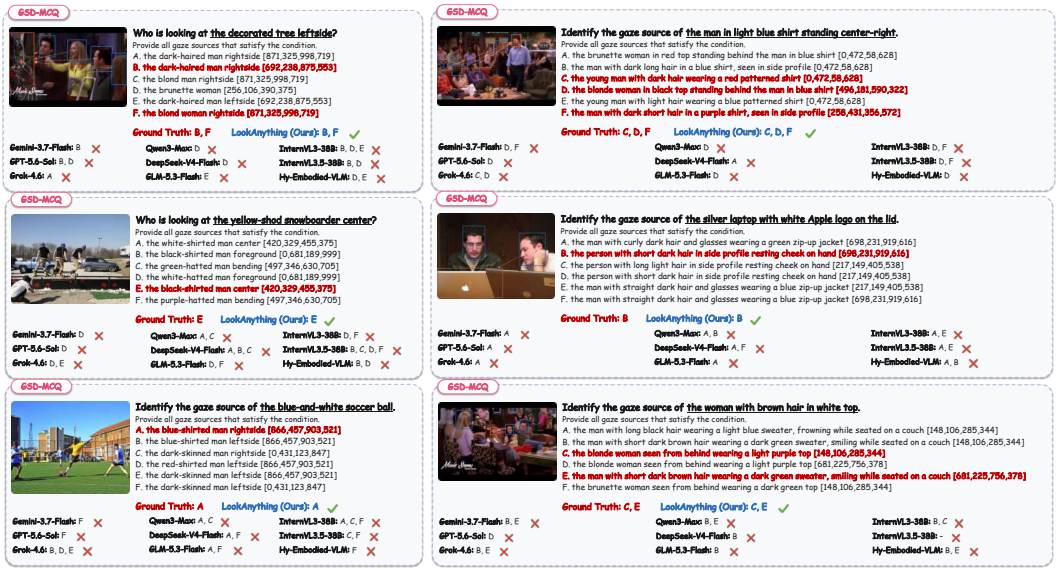


Figure 33: Qualitative Cases of GSD.

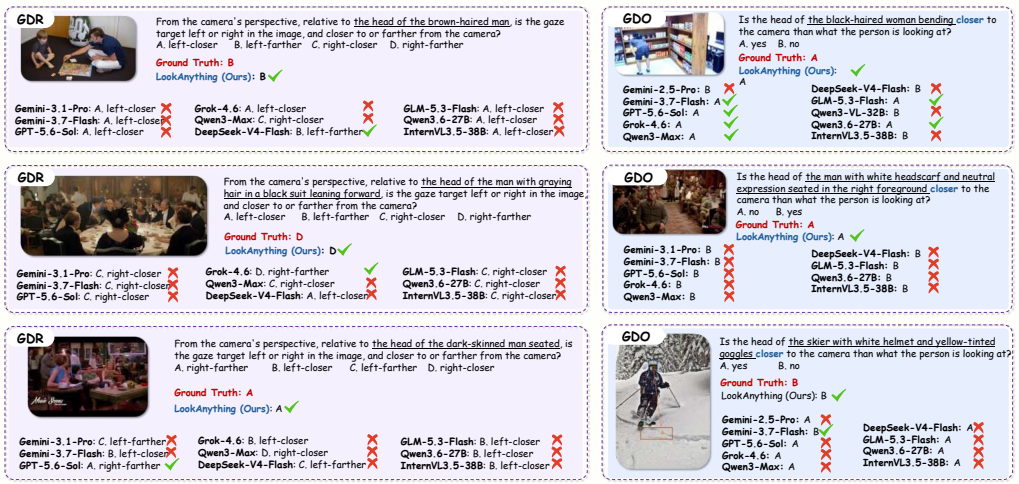


Figure 34: Qualitative Cases of GDO/GDR.



Figure 35: Qualitative Cases of GCD.

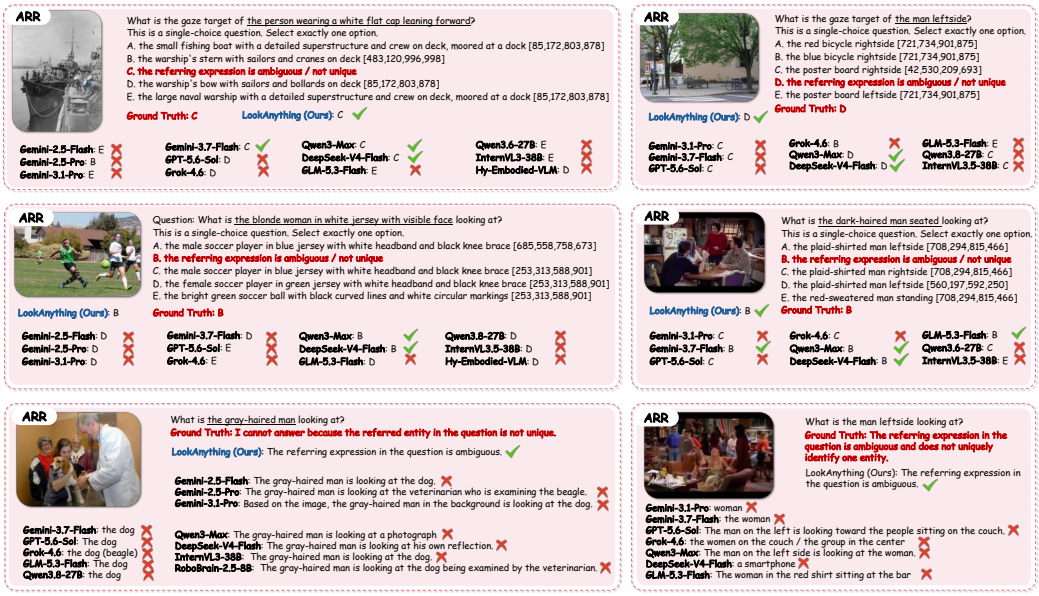


Figure 36: Qualitative Cases of ARR.

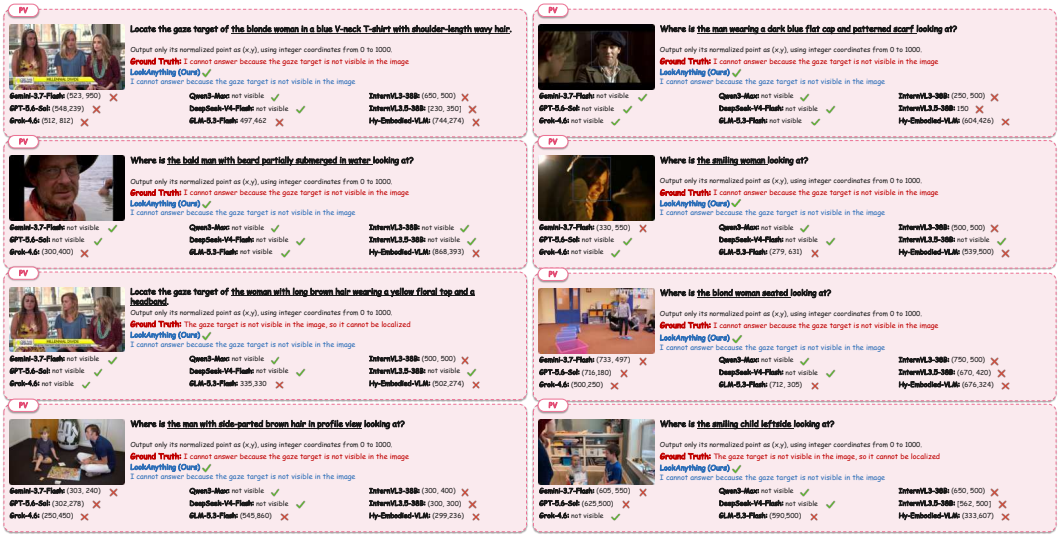


Figure 37: Qualitative Cases of PV.

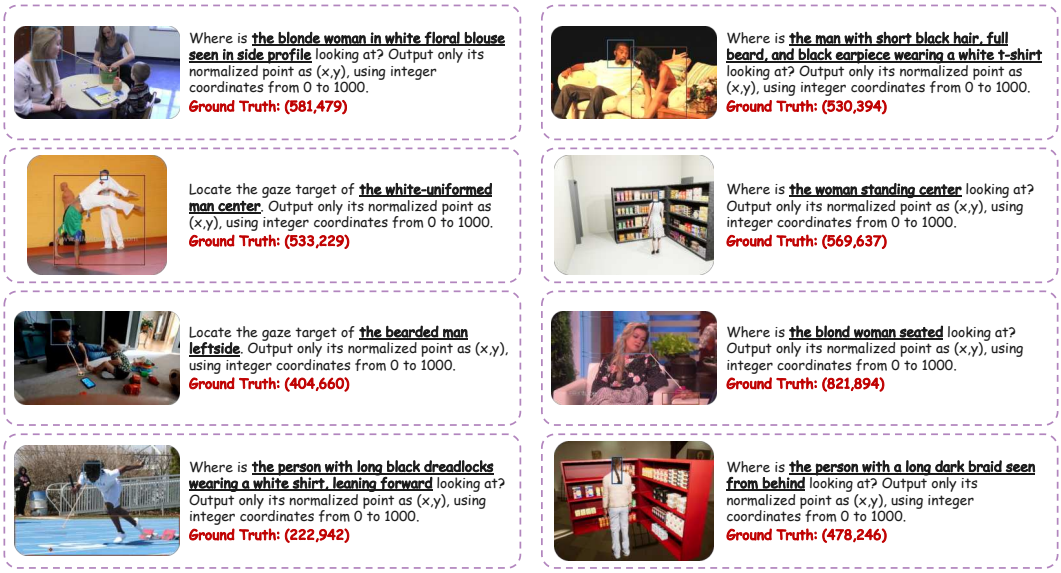


Figure 38: Sample cases of GTL-T2P in HoliGaze-Corpus.

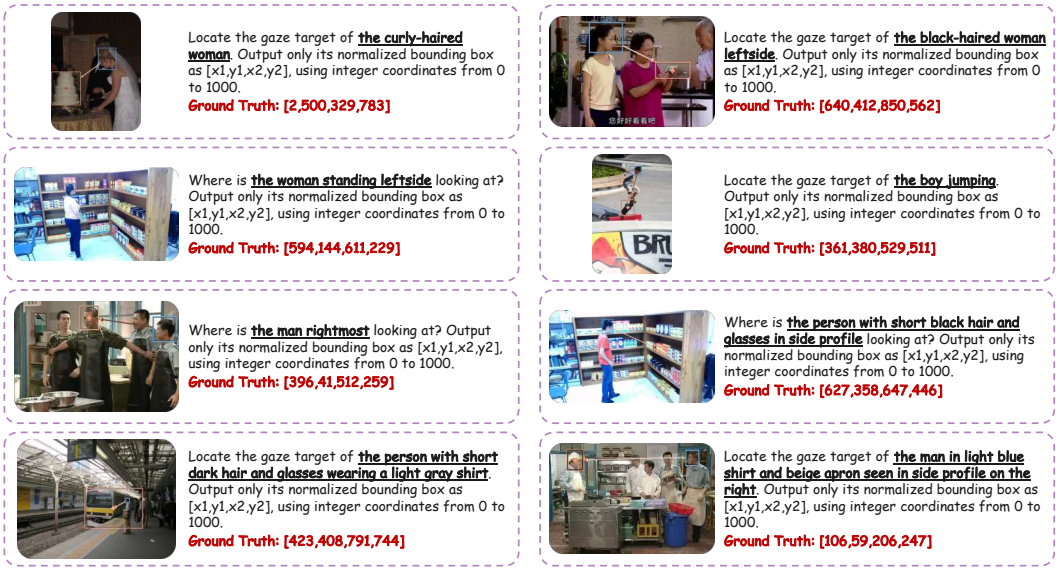


Figure 39: Sample cases of GTL-T2B in HoliGaze-Corpus.

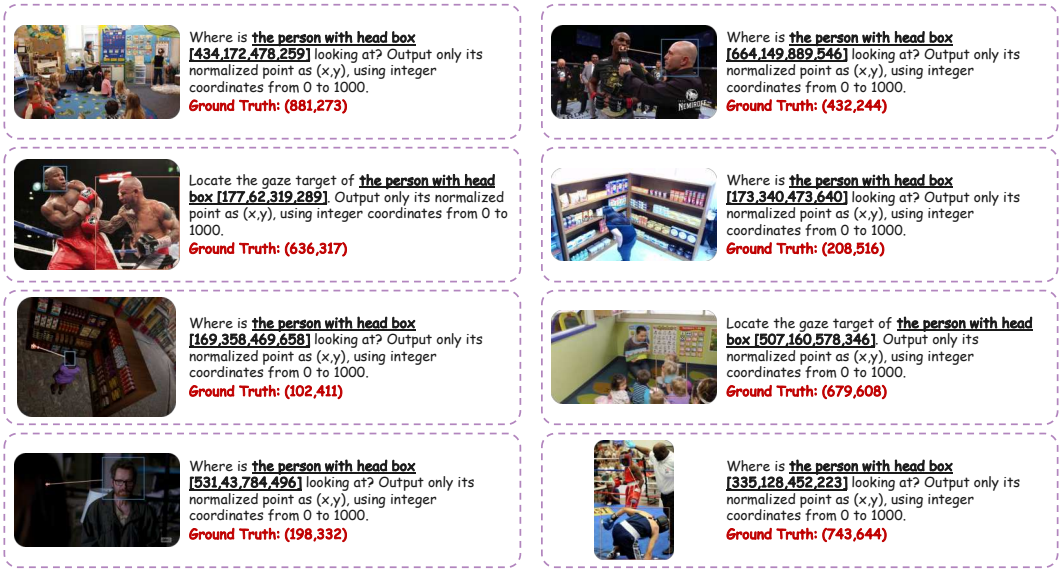


Figure 40: Sample cases of GTL-B2P in HoliGaze-Corpus.



Figure 41: Sample cases of GTD-MCQ in HoliGaze-Corpus.



Figure 42: Sample cases of GTD-OpenQA in HoliGaze-Corpus.



Figure 43: Sample cases of GSL in HoliGaze-Corpus.

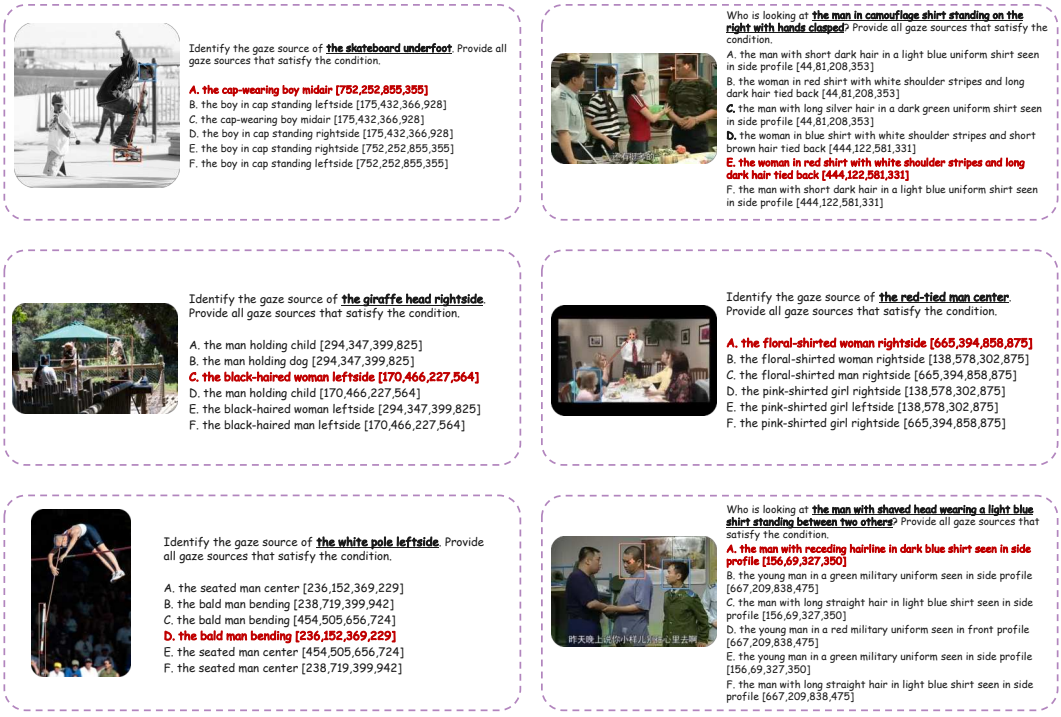


Figure 44: Sample cases of GSD in HoliGaze-Corpus.



Figure 45: Sample cases of GSD-OpenQA in HoliGaze-Corpus.

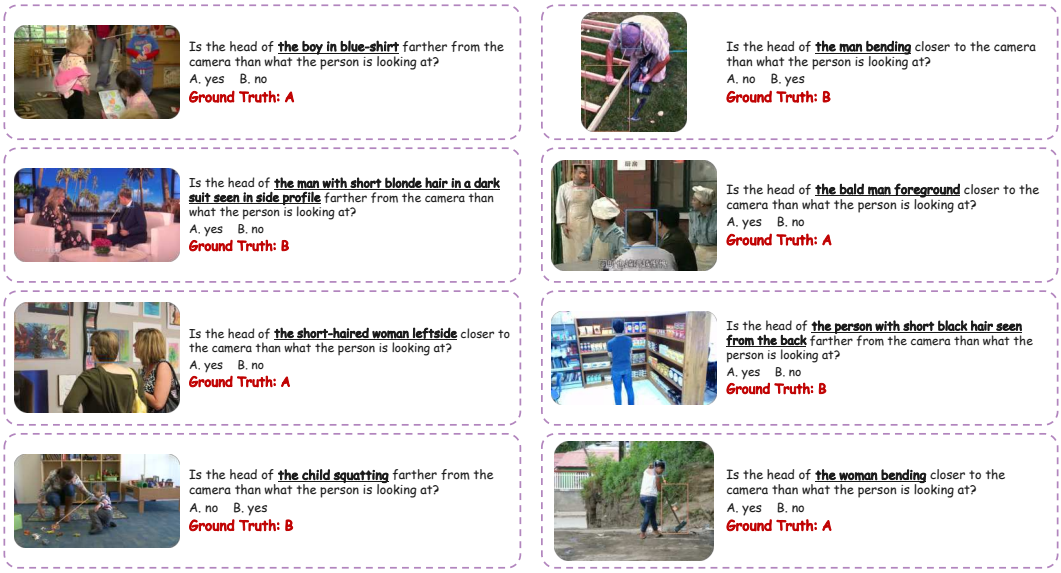


Figure 46: Sample cases of GDO in HoliGaze-Corpus.

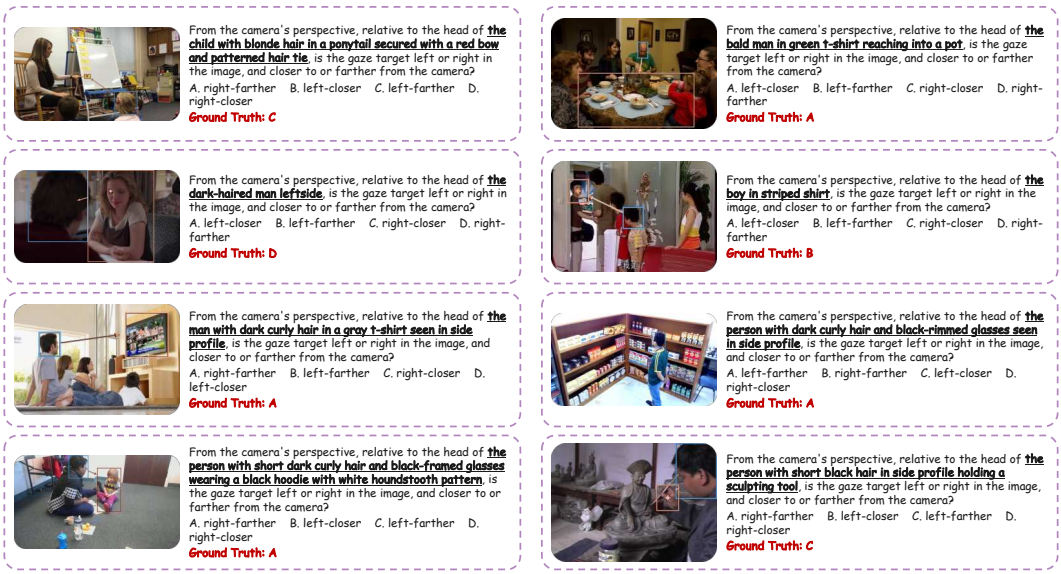


Figure 47: Sample cases of GDR in HoliGaze-Corpus.

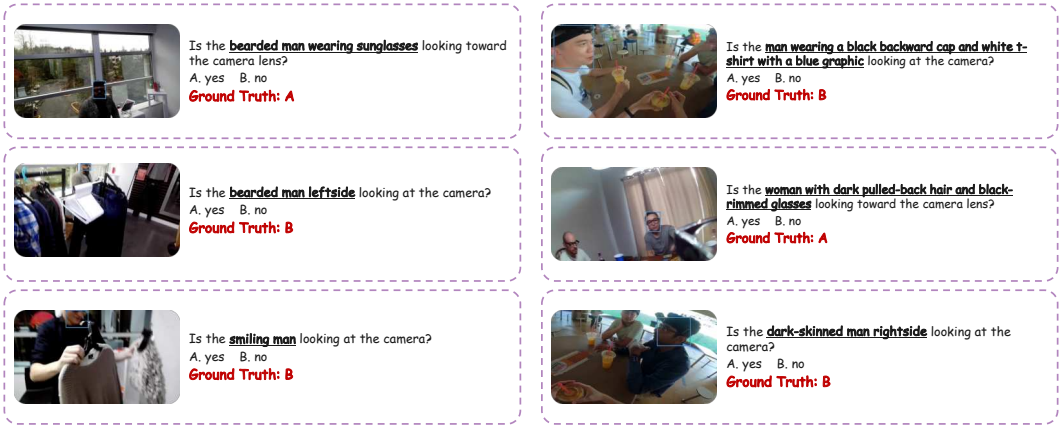


Figure 48: Sample cases of GCD in HoliGaze-Corpus.

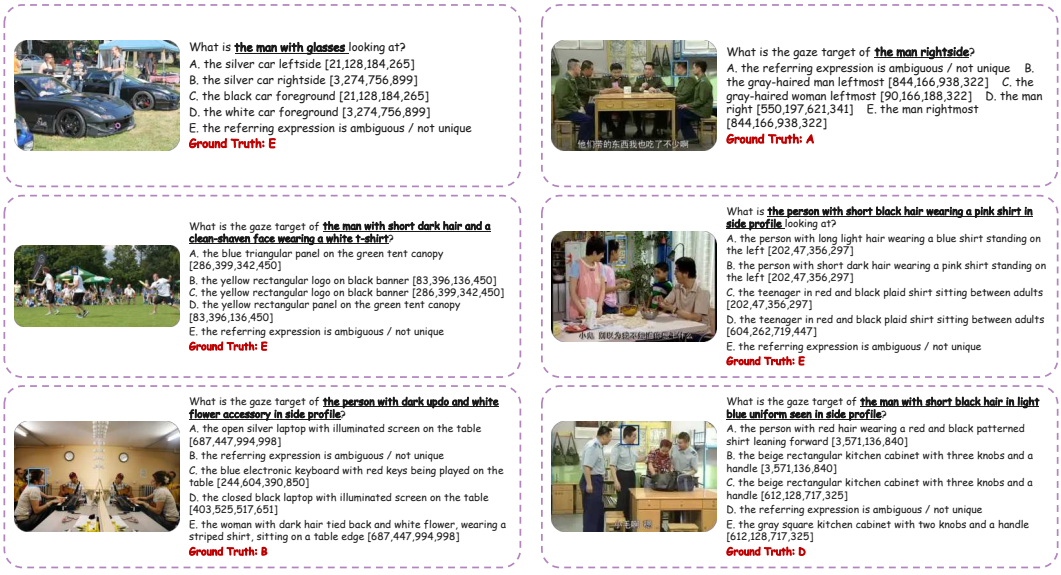


Figure 49: Sample cases of ARR-MCQ in HoliGaze-Corpus.



Figure 50: Sample cases of ARR-OpenQA in HoliGaze-Corpus.

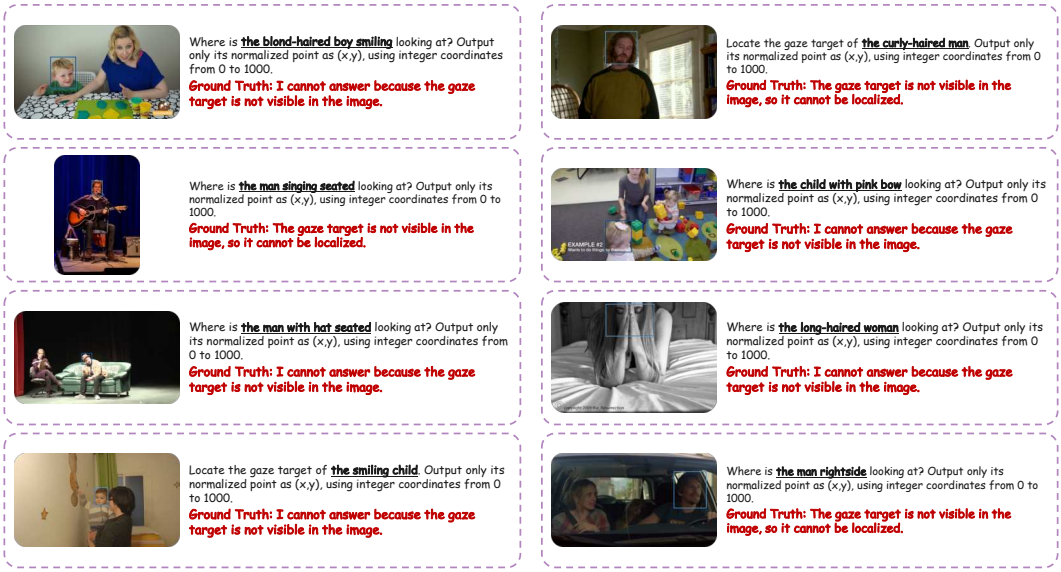


Figure 51: Sample cases of PV in HoliGaze-Corpus.