

Are Large Vision-Language Models Ready to Guide Blind and Low-Vision Individuals?

Eunki Kim^{1,†,*} Na Min An^{2,*} Wan Ju Kang² Sangryul Kim^{3,†}

James Thorne^{4,†} Hyunjung Shim²

¹SKT, ²KAIST AI, ³NAVER, ⁴Theia Insights

{eunkikim, naminan}@kaist.ac.kr

Abstract

Large Vision-Language Models (LVLMs) demonstrate a promising direction for assisting individuals with blindness or low vision (BLV). Yet, measuring their true utility in real-world scenarios is challenging because evaluating whether their descriptions are BLV-informative requires a fundamentally different approach from assessing standard scene descriptions. While the “VLM-as-a-metric” or “LVLM-as-a-judge” paradigm has emerged, existing evaluators still fall short of capturing the unique requirements of BLV-centric evaluation, lacking at least one of the following key properties: (1) *High correlation with human judgments*, (2) *Long instruction understanding*, (3) *Score generation efficiency*, and (4) *Multi-dimensional assessment*. To this end, we propose a unified framework to bridge the gap between automated evaluation and actual BLV needs. First, we conduct an in-depth user study with BLV participants to understand and quantify their navigational preferences, followed by collecting complementary sighted human annotations assessing the quality of the descriptions that affect BLV user safety. The process leads to curating VL-GUIDEDATA, a large-scale BLV user-simulated preference dataset containing image-request-response-score pairs. Our framework shows how preferences from a target user population can be systematically distilled through human studies and scaled into a reward model. We then leverage the dataset to develop an accessibility-aware evaluator, VL-GUIDE-S, which outperforms existing (L)VLM judges in both human alignment and inference efficiency. Beyond BLV evaluation, VL-GUIDE-S also generalizes across diverse image-text alignment benchmarks, suggesting that user-grounded reward modeling can benefit both specialized and general-purpose evaluation. We hope our work lays a foundation for automatic AI judges that advance safe, barrier-free navigation for BLV users while providing a practical framework for building evaluators tailored to diverse user populations.

1 Introduction

“A good guide is better than a hundred maps.” Without reliable assistive guidance, individuals with blindness or low vision (BLV) frequently face unexpected navigational challenges (Brady et al., 2013; Real & Araujo, 2019; Yuan et al., 2024; Prajapati et al., 2024). In complex urban environments, avoiding dynamic obstacles—such as parked vehicles or moving pedestrians—is a matter of physical safety. Surveys indicate that approximately 40% of legally blind individuals experience accidents or falls at least once a month (Manduchi & Kurniawan, 2011; Högner, 2015), with over 90% of these accidents caused by street obstacles (Wilson, 2015; Khan et al., 2023). While AI-based assistive technologies (Wiberg, 2015; Aira, 2015; Microsoft, 2017; TUAT, 2024) have increasingly integrated Large Vision-Language Models (LVLMs) (Yuan et al., 2024; Huh et al., 2023; Zhao et al., 2024; Liu et al., 2025) to provide spatial awareness, a critical gap remains: A standard visual description (e.g., “A car is parked on the right side of the street.”) may be informative to a sighted user,

*Equal contribution

but it lacks the context-rich, spatially grounded navigation cues that a BLV user requires to actively avoid a collision (Bandukda et al., 2019; Kazemi et al., 2023).

Measuring the utility of these models in mobility scenarios is highly challenging because evaluating whether a description is truly *BLV-informative* requires a different modeling perspective than simply assessing a standard scene description. Currently, the automated evaluation of image-text alignment relies on a VLM encoder-based metric or an LVLM judge (Lin et al., 2014; Hossain et al., 2019; Ghandi et al., 2023). However, existing evaluators fail to meet the unique requirements of BLV-centric evaluation, consistently lacking at least one of four crucial attributes: (1) *High correlation with human judgments*, (2) *Long instruction understanding*, (3) *Score generation efficiency*, and (4) *Multi-dimensional assessment*.

Lightweight encoder-based predictors like CLIP-S (Hessel et al., 2021) and BLIP-S (Li et al., 2022) are highly efficient but struggle to process the long, detailed text sequences inherent in step-by-step navigational instructions. Conversely, generative LVLMs (Deitke et al., 2024; Meta, 2024) excel at semantic understanding, but their high inference latency and substantial computational requirements limit their practicality for large-scale evaluation. They also often require prompt-tuning and additional post-processing (Li et al., 2024; Tong et al., 2024; Lambert et al., 2024), including bias subtraction to prevent ranking distortions (Zhu et al., 2025). Finally, while recent scalar-based models like InternLM-XComposer-2.5-Reward (IXCREW-S) (Zang et al., 2025) improve inference speed, they demonstrate weak pointwise correlations with human judgments on image-text alignment benchmarks (e.g., Flickr30K-Exp/CF (Plummer et al., 2015)). Moreover, the model is tightly coupled with a single LVLM backbone, InternLM, thereby limiting architectural modularity and flexibility.

To address the limitations of existing evaluators that do not fulfill the BLV-centric evaluation criteria, we propose a unified framework that bridges the gap between automated LVLM evaluation and actual BLV needs. Since assessment of the unstructured LVLM-generated descriptions imposes significant cognitive overload on BLV users (Chandrasekar et al., 2025; Chen et al., 2025), we distill BLV-specific preferences through direct user studies and then scale those preferences into a dataset sufficiently large for reward-model training and evaluation. Furthermore, sighted human judgments also provide complementary information that cannot be assessed by BLV users alone but impacts safety.

First, by analyzing BLV user preferences across fine-grained dimensions, such as Afraidness and Sufficiency, we isolate exactly what makes an LVLM description useful for safe mobility. We then encapsulate these insights into constructing VL-GUIDEDATA, a novel, large-scale BLV user-simulated preference dataset consisting of 4k image-request-response-score pairs. Unlike existing datasets that evaluate generic image-text alignment (Xu et al., 2024a; Team, 2024) or focus solely on binary Visual Question Answering (VQA) (Yuan et al., 2024; Gurari et al., 2019), our VL-GUIDEDATA centers the nuanced perspectives, formatting preferences, and safety criteria of BLV users navigating complex environments.

Additionally, we introduce VL-GUIDE-S, a human preference predictor specifically designed to capture BLV navigational utility. VL-GUIDE-S satisfies all the above-mentioned four properties required for effective evaluation. It acts as a robust automatic AI judge that rapidly understands long instructions/requests and generates scores consistent with both general sighted judgments and BLV-specific needs. By evaluating across multiple interpretable dimensions rather than providing a single scalar score, VL-GUIDE-S outperforms existing (L)VLM judges in human alignment and inference efficiency.

In summary, our contributions are as follows:

- Comprehensive analysis of BLV user preferences across diverse LVLM-generated scene descriptions, revealing the fine-grained factors that drive navigational utility.
- VL-GUIDEDATA, a large-scale multimodal preference dataset curated from BLV-simulated evaluations and specialized mobility scenarios.
- VL-GUIDE-S, an efficient, multi-dimensional LVLM-based predictor that strongly aligns with both sighted and BLV human judgments.

[†]Work done at KAIST AI.

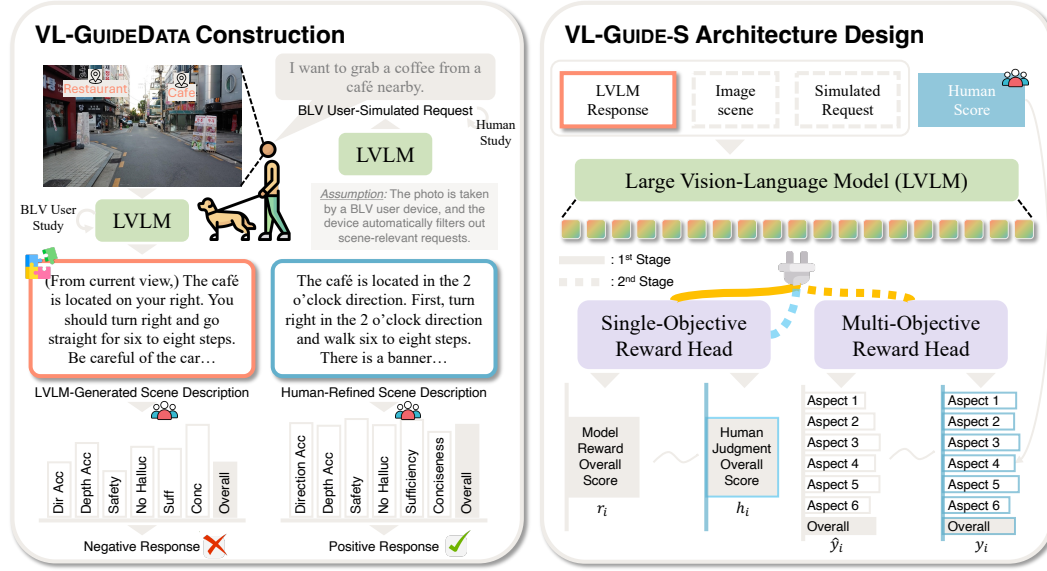


Figure 1: **Overall proposed framework.** (Left) Our final benchmark data consist of a scene, a BLV-simulated request, an LVLM-generated/human-refined scene description, and corresponding scores across multiple aspects. (Right) VL-GUIDE-S is capable of producing both scalar and multiple values reflecting alignment with human judgments by appending reward heads to a frozen LVLM. The last hidden states of LVLM updated during *stage 1* are used to train a ridge regression layer during *stage 2*.

2 Method

To bridge the gap between standard image-text automated evaluation and the specific needs of BLV individuals, we propose a unified framework consisting of two main components (Fig. 1): (1) the curation of a BLV user-aware preference dataset (VL-GUIDEDATA) and (2) the development of a BLV user-aligned human preference predictor (VL-GUIDE-S).

2.1 VL-GUIDEDATA: A BLV User-Aware Preference Dataset

Since collecting real-time, large-scale preference data directly from BLV individuals poses safety concerns and scalability challenges, we adopt a two-phase pipeline. First, we conduct targeted pilot studies with BLV participants to distill their unique navigational requirements into evaluation criteria. We then leverage these insights to guide sighted human annotators in constructing VL-GUIDEDATA that aligns with BLV user-derived guidelines.

(A) Scene Collection and Request Generation. We first construct a dataset of environmental scenes and corresponding navigational requests, assuming a realistic assistive setting in which images are captured by a BLV user’s device and automatically paired with scene-relevant requests. To this end, we filter images from SideGuide (Park et al., 2020) and SideWalk (AIHub, 2019), which contain photos taken by impaired individuals. We ensure realistic and meaningful evaluation settings by curating images that are perceptible to sighted annotators since extremely dark or heavily blurred scenes cannot reasonably support assistive descriptions. We additionally require scenes to contain at least five detectable objects to create more challenging navigation-oriented scenarios and exclude in-vehicle perspectives because our focus is on pedestrian self-mobility assistance for BLV users. This rigorous four-round filtering process by two authors results in 1,152 high-quality scenes (600 from SideGuide, 552 from SideWalk).

To simulate real-world usage, we generate 5 to 10 plausible BLV navigational requests per scene using GPT-4o mini (OpenAI, 2024b) with 3-shot prompting. Then, to ensure that the requests are grounded in the scene, we recruit Institutional Review Board (IRB)-approved 24

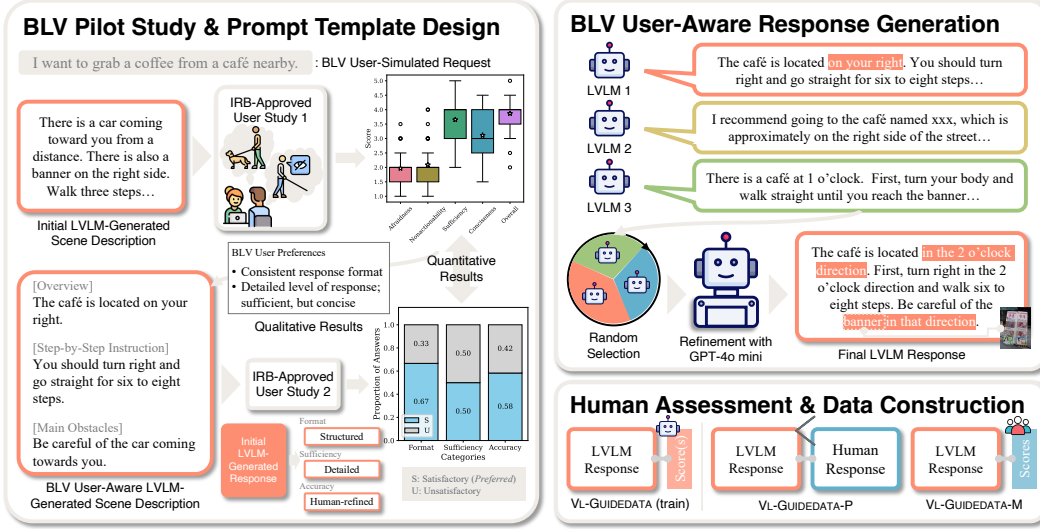


Figure 2: **Main results and process of VL-GUIDEDATA construction.** (Left) LVLm prompts are iteratively refined based on BLV user feedback. While raw responses are deemed safe and actionable, users prefer a consistent, succinct format, validated in a second user study. (Right) To reduce model bias, final responses for the VL-GUIDEDATA benchmark are generated by selecting and refining outputs from three different LVLms. The final LVLm responses are used to construct the training and test set of VL-GUIDEDATA.

sighted annotators to review these requests to filter out unnatural or non-actionable prompts (e.g., “Look in the mirror”). Evaluated by two annotators per pair (62.33% agreement)¹, this process yields 4,979 consensus image-request pairs (4.32±1.42 requests per scene). Note that the quality of the automatically generated requests is later also validated by the BLV participants. The choice of having the BLV users verify—rather than directly generate—the requests is deliberate; we focus on how much BLV users *prefer* certain scene descriptions given the identical requests across participants.

(B) Exploratory LVLm Response Generation (Phase 1). To understand what types of scene descriptions BLV individuals prefer, we initially generate a diverse set of baseline responses. For the image-request pairs, we collect responses from diverse model families, including open-source LVLms known for strong in-context learning (LLaVA-1.6 (Liu et al., 2023), Qwen-VL (Bai et al., 2023), InternLM-XComposer2-VL (Dong et al., 2024), and OpenFlamingo (Awadalla et al., 2023)) alongside GPT-4o mini (OpenAI, 2024b). We apply few-shot prompting on the 7B models and further refine the resulting LVLm responses using GPT-4o mini. This setup enables a controlled evaluation of whether BLV users exhibit a preference for certain model descriptions when the source of generation is not disclosed.

(C) BLV User Pilot Study. We then present the generated subset of LVLm descriptions to BLV participants across two rounds of IRB-approved user studies.

In the initial phase, a panel of six BLV participants evaluates the generated scene descriptions across five key metrics: Fear, Actionability, Sufficiency, Conciseness, and Overall Quality. The first two of the five dimensions, Fear and Actionability, were motivated by prior BLV-centered research (Bandukda et al., 2019), which identified user independence as a key overarching concept in the outdoor experiences of BLV individuals. Also, the study mentions the importance of providing adequate environmental information; hence, Sufficiency is included for our assessment category. We additionally included Conciseness since recent studies have shown that LLMs often exhibit verbosity bias (Saito et al., 2023), raising an important question of whether BLV users prefer increasingly detailed descriptions or concise,

¹The agreement reflects annotation consistency on a double-annotated subset, not on the overall filtering process. Only consensus requests that passed multi-stage filtering and were subsequently approved by BLV participants are retained.

actionable information. Finally, Overall Quality was included to capture participants’ holistic impressions beyond individual dimensions. Overall, these dimensions are selected to provide a comprehensive assessment, measuring not only the *linguistic quality* of the descriptions (Sufficiency, Conciseness) but also their *psychological and practical* impact on BLV users navigating their environments (Fear, Actionability).

The initial results (left panel in Fig. 2) indicate that while the initial LVLM scene descriptions successfully reduce navigational apprehension (Fear ↓, Actionability ↑), *how information is conveyed* needs improvement. Specifically, BLV users strongly prefer structured, deductive formats—such as a brief overview followed by specific details—over unstructured text. Furthermore, participants emphasize the need to balance conciseness with adequate detail, prioritizing concrete directional cues and objective distance metrics (*e.g.*, meters or steps) over subjective, generic warnings (*e.g.*, Watch out for cars or pedestrians.) frequently added by GPT-4o mini.

Drawing on initial BLV user feedback, we develop a structured template to systematically reformat LVLM responses (Fig. 2). To determine whether users demonstrate a significant preference for human-edited content, we also incorporate human-refined outputs into our evaluation set (Appendix). Finally, to validate the hypothesized trade-off between sufficiency and conciseness, we test whether appending specific details to the baseline LVLM responses would yield mixed user preferences. As anticipated, participants strongly prefer template-adherent responses (*Format* in Fig. 2). Surprisingly, they also consistently favor human-edited refinements, even when blinded to the source of the revisions (*Accuracy* in Fig. 2). Consistent with findings from the first round, preferences regarding the balance of detailed versus concise information remained mixed (*Sufficiency* in Fig. 2).

While it is ideal to curate benchmark datasets collected from BLV users, scaling evaluation to a large BLV population is impractical due to recruitment constraints. Moreover, sighted annotators are needed to evaluate properties that require direct visual access to the scene, such as correctness and hallucination detection, which are critical factors that directly impact BLV user safety. Hence, we distill the core variables influencing BLV user preferences to simulate their conditions and scale our evaluation utilizing a more accessible cohort of sighted participants.

(D) Scaled, BLV user-Aware LVLM Response Generation (Phase 2). Guided by the key finding from the BLV pilot study, we construct LVLM descriptions (right panel in Fig. 2) corresponding to scene and BLV-simulated requests. Distinct from the prompts used before the BLV pilot study, we prompt LLaVA-1.6 (Liu et al., 2023), Qwen-VL (Bai et al., 2023), and InternLM-XComposer2-VL (Dong et al., 2024) to adhere to the BLV user-preferred unified template. Then, for each instance, we randomly select one of the three different LVLM responses and refine it by prompting GPT-4o mini (OpenAI, 2024b). This process of using GPT-4o to refine responses from different models rather than generating them from scratch helps to minimize the introduction of bias induced by a single model.

(E) Large-Scale Human Assessment (VL-GUIDEDATA Construction). To construct VL-GUIDEDATA, we employ 25 sighted human annotators (approved by the Institutional Review Board) to evaluate the refined 2.5k image-request-response triplets from the Phase 2 generation. Crucially, since our BLV pilot study reveals that baseline LVLM responses already consistently reduce the fear of BLV users and include BLV user-executable actions, and sighted annotators cannot authentically evaluate subjective BLV experience, we reformulate the pilot metrics into seven visually verifiable dimensions.

Specifically, while the BLV requirement for “Actionability” assesses the safety from the BLV perspective, we need sighted annotators to assess whether the descriptions factually correspond to the scene because imprecise descriptions could pose significant risk for BLV users. Hence, the safety for BLV users from a sighted human perspective is decomposed into (1) *Direction Accuracy* and (2) *Depth Accuracy*, ensuring the spatial guidance is physically correct and executable based on the image. Similarly, the mitigation of “Fear” is enforced through rigorous visual checks for (3) *Safety* (identifying physical hazards) and the absence of (4) *Hallucination* (preventing fabricated, dangerous guidance). The remaining linguistic

and structural preferences—(5) *Sufficiency*, (6) *Conciseness*, and (7) *Overall Quality*—are retained directly from the BLV pilot study. All aspects are annotated with an averaged scalar score from two to three human judgments on a Likert scale of 1 to 5, except for *Hallucination*, which is evaluated dichotomously as either 0 (presence) or 1 (absence) based on non-related information, inaccurate step-by-step order, or repeated content.

This robust annotation procedure yields 2,112 unique samples to be scored by 2–3 humans for each. Additionally, we collect high-quality, human-refined (positive) captions that perfectly satisfy all seven criteria (Appendix). Filtering any newly added human responses not confirmed by both authors yields two evaluation subsets: VL-GUIDEDATA-M (LVLM responses labeled with pointwise scores across multiple dimensions; $n = 939$) and VL-GUIDEDATA-P (pairwise human/LVLM responses; $n = 722$). We also generate the training set of VL-GUIDEDATA for *stage 2* using GPT-4o mini as a scorer for 1.57k image-request-response pairs.

2.2 VL-GUIDE-S: BLV User-Aware Human Judgment-Aligned Evaluator

Building upon the insights from VL-GUIDEDATA construction, we introduce VL-GUIDE-S, an LVLM-based human judgment predictor designed to quantify *the extent to which LVLMs effectively convey scene information for human users, including individuals with blindness or low vision*. Unlike existing reward models that aggregate multiple scores from conflicting objectives into a single score, VL-GUIDE-S employs a two-stage architecture for robust single- and multi-objective scoring.

Stage 1: Single-Objective Scoring. The goal of *stage 1* is to output a scalar score that captures overall semantic alignment between images and texts, while simultaneously shaping multimodal embeddings useful for the later stage. As illustrated in Fig. 1, we simply append a reward head to a base generative LVLM. This design allows the LVLM to produce a human judgment-aligned multimodal representation in general. Moreover, this architecture inherently supports the comprehension of complex, long-form instructions via the LVLM tokenizer. It also significantly reduces the inference latency compared to text-based generative scoring methods since it eliminates the $\mathcal{O}(k)$ autoregressive decoding overhead required by generative methods to output k tokens, one at a time.

To explicitly align outputs with human judgment, we use Mean Squared Error (MSE) loss: $\min_{\theta} \sum_{i=1}^N (r_i - h_i)^2$ (θ : parameters of the LVLM and appended reward head; h_i : ground-truth human judgment score; r_i : model-predicted scalar score). Initialization on the reward head is conducted using a zero-centered Gaussian distribution with standard deviation $\frac{1}{\sqrt{d+1}}$ (d : hidden dimension) (von Werra et al., 2020). Our modeling design via a reward head stabilizes the optimization process and allows the model to learn a generalized, human-aligned vision-language representation space before we adapt it to the specialized domains.

Stage 2: Multi-Objective Scoring. Optimizing evaluators under a multi-objective scoring setting is particularly challenging. Prior work has shown that gradient-based optimization methods (e.g., MGDA (Zhang et al., 2024b), PCGrad (Yu et al., 2020)) are costly at the scale of LVLMs and often fail to balance inherently conflicting objectives (He & Maghsudi, 2025). Ensemble-based approaches such as EMORL (Kong et al., 2025) partially alleviate these difficulties by aggregating models trained on individual objectives, but still face instability and limited interpretability without post-hoc calibration when deployed at scale. These observations suggest that relying solely on parameter-level updates or model ensembling is sub-optimal, highlighting the necessity for a lightweight training strategy to effectively scale multi-objective scoring.

To address this, we leverage frozen multimodal embeddings from the VL-GUIDE-S architecture trained during *stage 1* to predict scores across multiple dimensions. Let z_i denote the frozen multimodal embedding and y_i be the vector of human scores across K dimensions. We append a lightweight ridge regression layer to predict $\hat{y}_i = Wz_i + b$, optimized via: $\min_{W,b} \sum_{i=1}^N \|y_i - \hat{y}_i\|_2^2 + \alpha \|W\|_F^2$. This formulation enables VL-GUIDE-S to produce instant-

neous, interpretable scores across diverse criteria, extending beyond a simple image-text alignment score.

Our design of reward modeling introduces a new training paradigm for estimating human judgment scores for multimodal data. Existing methods (Wang et al., 2024a; Xu et al., 2024a) adopt a *compose-then-aggregate* strategy: they first predict multi-objective scores and then combine them into a single score via a learned linear or gating head. In contrast, we reverse this design by directly supervising a single holistic reward, and subsequently leveraging the model’s final hidden representation to derive scores across multiple objectives. This design is motivated by the limitation that aggregating predefined dimensions may fail to capture overall quality faithfully and can introduce instability. For example, VisionREW-S exhibits near-zero weights in its aggregation head for certain dimensions, suggesting that some objectives contribute little to the final score. Our approach of projecting the multimodal embedding into dimension-specific scores is inherently more stable and informative, as the trained latent space is already optimized to encode a holistic judgment of humans.

3 Experiments

3.1 Experimental Setup

To demonstrate architectural agnosticity, we instantiate VL-GUIDE-S across several widely used LVLM backbones: Qwen2-VL (2B and 7B), InternLM-XComposer-2.5 (7B), and LLaMA-3.2 (11B). Before adding the layer used for multi-objective scoring, we first train VL-GUIDE-S on Polaris (Wada et al., 2024) and ImgREW (Xu et al., 2024b), which are image-to-text (I2T) and text-to-image (T2I) datasets consisting of overall human judgment scores. Similar to all comparison models in Stage 1, our VL-GUIDE-S has not been exposed to any BLV-specific training data and is evaluated solely on its ability to assess general semantic alignment between images and texts. Then, we fine-tune VL-GUIDE-S using the training split of our data for multi-objective evaluation on our VL-GUIDEDATA. Note that all the comparison models in Stage 2 are also trained on our dataset to ensure fair comparison.

3.2 Comparison Baselines

We benchmark VL-GUIDE-S against a comprehensive suite of existing evaluators:

- **VLM-based Metrics:** CLIP-S/RefCLIP-S (Hessel et al., 2021), LongCLIP-S (Zhang et al., 2025), PAC-S/RefPAC-S (Sarto et al., 2024), Polos (Wada et al., 2024), BLIP-S (Li et al., 2022), and ImgREW-S (Xu et al., 2024b).
- **LVLM-based Generative Judges:** Molmo-7B (Deitke et al., 2024), LLaVA-Critic-7B (Xiong et al., 2024), Qwen2-VL-7B (Wang et al., 2024b), InternVL2-8B (Chen et al., 2024), and G-VEval (Tong et al., 2024).
- **Scalar Reward Models:** IXCREW-S (Zang et al., 2025) for single objective scoring and VisionREW-S (Zhang et al., 2024a) for multiple objective scoring.

3.3 Benchmarks and Metrics

Evaluations are conducted on standard general-domain datasets (PASCAL (Xu et al., 2019), FOIL (Shekhar et al., 2017), Flickr-Exp/CF (Plummer et al., 2015), Polaris (Wada et al., 2024), and ImgREW (Xu et al., 2024b)) to ensure baseline competence. To measure true BLV-centric utility, we assess various methods on Sighting (Kang et al., 2025) and VL-GUIDEDATA (test). For multi-objective scoring evaluation, we benchmark predictors on VisionREW (Xu et al., 2024a), Align-anything (Team, 2024), and VL-GUIDEDATA. Metrics used to quantify human-alignment correlation include Pairwise Accuracy (P-Acc) for pairwise ranking datasets and Kendall’s τ_b or τ_c for pointwise datasets.

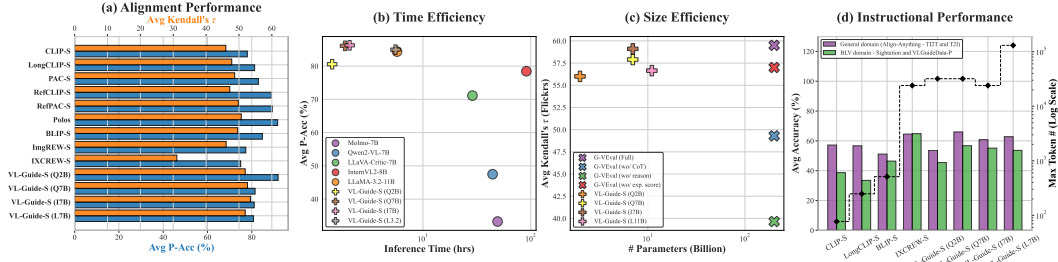


Figure 3: **Single-objective performance results.** (a) VL-GUIDE-S consistently achieve competitive or superior alignment with human judgments. (b) Ours variants significantly reduce inference time while maintaining high performance. (c) Our open-source VL-GUIDE-S achieves comparable human correlation scores to proprietary GPT-4o-based evaluator. (d) VL-GUIDE-S is capable of processing long texts while showing high performance in both general and BLV user domains.

3.4 Results

3.4.1 Alignment with Human Judgments for a Single Objective

General image-text datasets. VL-GUIDE-S significantly outperforms conventional VLM/LVLM-based evaluators and scalar reward models (Fig. 3a). For instance, VL-GUIDE-S (instantiated with the InternLM-7B) achieves the highest Kendall’s rank correlations of 59.3 and 39.5 on FlickrExp and FlickrCF, surpassing IXCREW-S by 28.7% and 65.1%, respectively ($p < 0.0001$). Furthermore, VL-GUIDE-S (Qwen-2B-S) attains near-perfect P-Acc of 98.0% on FOILR1 (5–10% gains over IXCREW-S, $p < 0.005$). These results confirm that our novel architecture yields robust and accurate scores highly aligned with human judgment across diverse image-text pairs.

Simultaneously, VL-GUIDE-S achieves higher efficiency even when compared with the original generative version of LVLM (Fig. 3b). For example, while InternVL2-8B takes 90 hours for 14k samples from the Polaris dataset, inference time is greatly reduced by 97.8% for ours. VL-GUIDE-S demonstrates performance on par with the GPT-4o-based judge, G-VEval, while operating without external API costs (Fig. 3c).

Instructional image-text datasets. Human instructions, particularly navigation requests from BLV users, are inherently complex, often requiring the model to interpret extended, sequential, and multi-step directives. While general image-text datasets typically involve captions that can be evaluated under a single binary variable $y \in \{0, 1\}$ indicating correctness, instructional datasets introduce additional contextual variables $c \sim p(c)$ that influence evaluation criteria. Consequently, the judgment score is better characterized as a conditional distribution $p(y | x, c)$ as used in VL-GUIDE-S, rather than a marginal $p(y | x)$ (x : multimodal inputs) used by VLM encoder-based metrics (e.g., CLIP-S) with limited context windows. As shown in Fig. 3d, while IXCREW-S achieves slightly higher raw accuracy, this is attributed to its overfitting on preference-matching scores rather than granular instruction following. In contrast, VL-GUIDE-S consistently achieves generally high performance in both pointwise and pairwise ranking across general and BLV user-aware domains.

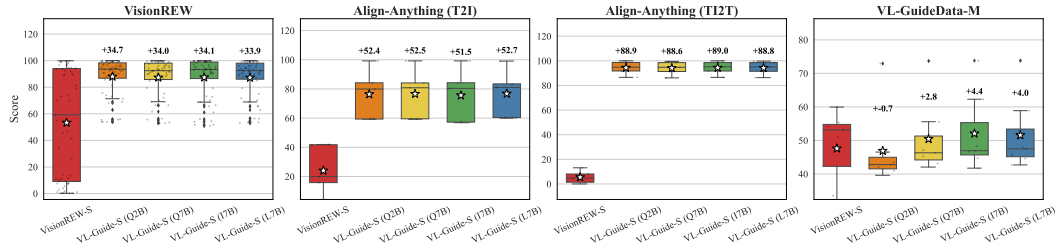


Figure 4: **Multi-objective performance results.** The proposed VL-GUIDE-S consistently outperforms the baseline across diverse datasets (per-dimension results in the Appendix). Furthermore, VL-GUIDE-M demonstrates robust scalability as model architecture size increases.

3.4.2 Alignment with Human Judgments for Multiple Objectives

When multiple evaluation objectives are considered, the model estimates a joint or factorized distribution over dimensions, *i.e.*, $p(\mathbf{y} | x, c)$ or $\prod_k p(y_k | x, c)$, where each y_k corresponds to a specific evaluation criterion. Although the objectives are not strictly independent, this factorization follows from the chain rule $p(\mathbf{y} | x, c) = \prod_k p(y_k | x, c, y_{<k})$, and approximating the conditional dependencies by $p(y_k | x, c, y_{<k}) \approx p(y_k | x, c)$ yields the tractable form $\prod_k p(y_k | x, c)$, which serves as an efficient surrogate to the full joint distribution.

Fig. 4 shows that VL-GUIDE-S outperforms the existing multi-objective reward model, VisionREW-S, in both general and our BLV user-aware preference datasets. Similar to the single-objective results, VL-GUIDE-S architecture substantially reduces inference time (99.1%–99.7%; 51 days *vs.* 4–11 hrs). These results suggest that VL-GUIDE-S successfully outputs interpretable, highly accurate assessments across the distinct, fine-grained dimensions necessary for multimodal inputs with additional human/BLV user instructions.

While we adopt a ridge regression layer to map multimodal embeddings to multi-dimensional scores, this design is not restrictive. As noted in Tab. 1, replacing it with alternatives such as a Multi-Layer Perceptron (MLP) or a Random Forest regressor yields comparable performance, but introduces additional inference overhead. These results highlight that ridge regression offers the most efficient and scalable choice for multi-dimensional assessment.

Layer Type	Acc. (↑)	Complexity (↓)
MLP – <i>Baseline</i>	88.6	$O(\sum_{l=1}^L n_{l-1} n_l)$
Random Forest	88.2 (−0.4)	$O(T \cdot D)$
Ridge Regression	87.2 (−1.4)	$O(d)$

Table 1: **The effect of layer types on VisionREW performance (averaged across 59 dimensions) of VL-GUIDE-S.** Ridge regression yields the lowest inference complexity while maintaining competitive accuracy (d : input feature dimension; L : # layers; n_l : # hidden units in layer l , T : # trees; D : maximum tree depth.).

4 Related Work

4.1 AI for Blind and Low-Vision Individuals

AI has increasingly been deployed to assist BLV individuals, with recent advancements in multimodal LVLMS showing potential to transform accessibility (Dong et al., 2024; Zong et al., 2025; Liu et al., 2024). Applications such as Be My Eyes (Wiberg, 2015) and BLV navigation frameworks like WalkVLM (Yuan et al., 2024) highlight the real-world utility of these barrier-free AI systems. Yet, measuring their true efficacy requires highly representative datasets. Prior datasets, such as VizWiz (Gurari et al., 2018; Tseng et al., 2022) and BIV-Priv-Seg (Tseng et al., 2024), collect images directly from BLV users. While insightful, the reliance on low-resolution or degraded imagery (Bigham et al., 2010) fails to reflect modern imaging capabilities and artificially restricts the evaluation of advanced AI models (Chiu et al., 2020; Olson et al., 2021). To address this, we focus on high-quality outdoor and indoor scene datasets (Park et al., 2020; AIHub, 2019) that better represent modern assistive camera inputs.

Furthermore, existing AI systems and datasets predominantly focus on vision-centric tasks, such as standard obstacle detection (Park et al., 2020; Xia et al., 2023; Tang et al., 2023) or VQA (Yuan et al., 2024; Gurari et al., 2019; Yang et al., 2024), where outputs are treated as strictly correct or incorrect. These binary formulations do not capture the complexity of generating rich, navigational scene descriptions in dense environments filled with signage and overlapping spatial obstacles. To bridge this gap, we introduce our multimodal BLV preference dataset, VL-GUIDEDATA. Unlike previous datasets, it contains fine-grained BLV navigational preferences alongside qualitative reasoning, shifting the focus from objective scene recognition to subjective, real-world BLV utility.

4.2 Automated Evaluators

Model-based metrics or judges aim to assess image-text alignment by approximating human preferences. Traditional VLM encoder-based metrics are typically divided into reference-based approaches (*e.g.*, Polos (Wada et al., 2024), RefPAC-S (Sarto et al., 2024)), which require

costly ground-truth human annotations, and reference-free approaches (e.g., CLIP-S (Hessel et al., 2021), BLIP-S (Li et al., 2022)), which score alignment based solely on the image-caption pair. While pointwise and pairwise ranking datasets (Xu et al., 2024b; 2019; Tripathi et al., 2025) have enabled nuanced analyses of these metrics, they predominantly focus on general domain alignment rather than specialized user needs. While recent methods like IIT-DAS (Zur et al., 2024; Kreiss et al., 2022) effectively enhance CLIPScore to align with BLV user preferences, it remains limited to the CLIP architecture and web-based images. Although CLIP-style architectures are highly efficient, they are inherently not suitable for comprehending the *instructional* texts from the users.

To overcome the limitations of standard evaluators, recent work has shifted toward the “LVLM-as-a-judge” paradigm, exploring *generative* LVLMs for evaluation (Deitke et al., 2024; Xiong et al., 2024; Meta, 2024). For instance, (Tong et al., 2024) utilizes GPT-4o (OpenAI, 2024b) with Chain-of-Thought (CoT) prompting, while scalar-based reward models have emerged by training linear projection heads on top of LVLMs (Zang et al., 2025). However, existing multi-objective reward models, such as VisionREW-S (Xu et al., 2024a) and MPS (Zhang et al., 2024c), lack in inference efficiency and accessibility. Crucially, because current evaluation datasets lack human judgment scores tailored to the navigational criteria of BLV users (Team, 2024; Kang et al., 2025), existing LVLM judges fail to perform multi-dimensional assessments that consistently correlate with BLV preferences. We overcome this limitation by introducing VL-GUIDE-S, an automated judge designed to evaluate LVLM-generated scene descriptions across multiple dimensions aligned with BLV needs.

5 Discussion

In this paper, we introduce a framework for evaluating LVLMs tailored to the needs of BLV users. By grounding evaluation in BLV preferences collected through a series of human interviews, we enable both accurate and scalable assessment. Through VL-GUIDEDATA and VL-GUIDE-S, we demonstrate that our framework captures nuanced BLV preferences across diverse scenarios. More broadly, our results suggest that user-grounded reward modeling offers a practical approach for distilling specialized user preferences into scalable automatic evaluators. Although our setup does not fully capture unconstrained real-world BLV interactions, it provides a practical approximation grounded in direct user feedback. While developed for BLV accessibility, the framework can naturally extend to other communities whose evaluation criteria differ from those of standard annotator pools.

Several directions remain for future work. Extending VL-GUIDE-S to continuous, real-time settings (e.g., frame-by-frame guidance from egocentric wearable cameras) would better reflect practical deployment. Incorporating user-adaptive evaluation by modeling factors such as mobility aids (e.g., cane *vs.* guide dog), navigation proficiency, and environmental context (e.g., indoor *vs.* outdoor or crowded *vs.* sparse) could further improve personalization. We believe that AI judges grounded in real human feedback (An et al., 2023; Song et al., 2024) will enable more reliable assistive systems, ultimately supporting safe, trustworthy, and barrier-free navigation for individuals who are blind or have low vision.

Acknowledgments

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Korea government (MSIT) (No. RS-2025-00520207); Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (Nos. RS-2024-00457882, 2022-0-01045, 2022-0-00680, RS-2019-II190075); a grant partly supported by both IITP (MSIT) and Korea Evaluation Institute of Industrial Technology (KEIT) (MOTIE) (No. RS-2025-02217259); and the next-generation CultureTech technology development (R&D) project for cultural space AX transformation through the Korea Creative Content Agency (KOCCA) grant funded by the Ministry of Culture, Sports and Tourism (MCST) in 2026 (No. RS-2026-25508598).

References

- AIHub. Ai hub: Sidewalk dataset, 2019. URL <https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=189>.
- Aira. About us and our values, 2015. URL <https://aira.io/aira-about-us/>.
- Na Min An, Hyeonhee Roh, Sein Kim, Jae Hun Kim, and Maesoon Im. Reinforcement learning framework to simulate short-term learning effects of human psychophysical experiments assessing the quality of artificial vision. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2023. doi: 10.1109/IJCNN54540.2023.10191870.
- Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models, 2023. URL <https://arxiv.org/abs/2308.01390>.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report, 2023.
- Maryam Bandukda, Aneesha Singh, Nadia Berthouze, and Catherine Holloway. Understanding experiences of blind individuals in outdoor nature. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–6, 2019.
- Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samuel White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23rd annual ACM symposium on User interface software and technology*, pp. 333–342, 2010.
- Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P Bigham. Visual challenges in the everyday lives of blind people. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 2117–2126, 2013.
- Prashant Chandrasekar, Mariel Couvillion, Ayshwarya Saktheeswaran, and Jessica Zeitz. Llm impact on blv programming, 2025.
- Meng Chen, Akhil Iyer, and Amy Pavel. Surfacing variations to calibrate perceived reliability of mllm-generated image descriptions, 2025.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024.
- Tai-Yin Chiu, Yinan Zhao, and Danna Gurari. Assessing image quality issues for real-world problems. in 2020 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3643–3653, 2020.
- Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024.
- Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model, 2024. URL <https://arxiv.org/abs/2401.16420>.
- Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. Deep learning approaches on image captioning: A review. *ACM Computing Surveys*, 56(3):1–39, 2023.

- Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018.
- Danna Gurari, Qing Li, Chi Lin, Yanan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P Bigham. Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 939–948, 2019.
- Qiang He and Setareh Maghsudi. Pareto multi-objective alignment for language models. *arXiv preprint arXiv:2508.07768*, 2025.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021.
- N Högner. Challenges in traffic for blind and visually impaired people and strategies for their safe participation. *Klinische Monatsblätter für Augenheilkunde*, 232(8):982–987, 2015.
- MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, 51(6):1–36, 2019.
- Mina Huh, Yi-Hao Peng, and Amy Pavel. Genassist: Making image generation accessible. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–17, 2023.
- Wan Ju Kang, Eunke Kim, Na Min An, Sangryul Kim, Haemin Choi, Ki Hoon Kwak, and James Thorne. Sighting counts: Leveraging sighted user feedback in building a blv-aligned dataset of diagram descriptions. *arXiv preprint arXiv:2503.13369*, 2025.
- Homa Kazemi, Mohammad Kamali, Reza Salehi, and Hossein Mobaraki. Recognizing the viewpoint and experience of blind people in navigation and daily traffic. *Function and Disability Journal*, 6(1):0–0, 2023.
- Wasiq Khan, Abir Hussain, Bilal Muhammad Khan, and Keeley Crockett. Outdoor mobility aid for people with visual impairment: Obstacle detection and responsive framework for the scene perception during the outdoor mobility of people with visual impairment. *Expert Systems with Applications*, 228:120464, 2023.
- Lingxiao Kong, Cong Yang, Susanne Neufang, Oya Deniz Beyan, and Zeyd Boukhers. Emorl: Ensemble multi-objective reinforcement learning for efficient and flexible llm fine-tuning. *arXiv preprint arXiv:2505.02579*, 2025.
- Elisa Kreiss, Cynthia Bennett, Shayan Hooshmand, Eric Zelikman, Meredith Ringel Morris, and Christopher Potts. Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics, 2022.
- Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.
- Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, et al. Vrewardbench: A challenging benchmark for vision-language generative reward models. *arXiv preprint arXiv:2411.17451*, 2024.

- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL <https://arxiv.org/abs/2304.08485>.
- Ruiping Liu, Jiaming Zhang, Angela Schön, Karin Müller, Junwei Zheng, Kailun Yang, Kathrin Gerling, and Rainer Stiefelham. Objectfinder: Open-vocabulary assistive system for interactive object search by blind people. *arXiv preprint arXiv:2412.03118*, 2024.
- Xingyu Liu, Biao Wang, Wayne Zhang, Ziqian Liao, Ziwen Li, Amy Pavel, Xiang’Anthony’ Chen, et al. Cosight: Exploring viewer contributions to online video accessibility through descriptive commenting, 2025.
- Roberto Manduchi and Sri Kurniawan. Mobility-related accidents experienced by people with visual impairment. *AER Journal: Research and Practice in Visual Impairment and Blindness*, 4(2):44–54, 2011.
- Meta. Llama 3.2, 2024. URL <https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/>.
- Microsoft. Seeing ai - talking camera app for the blind, 2017. URL <https://www.seeingai.com/>.
- Lauren Olson, Chandra Kambhampettu, and Kathleen McCoy. Towards using live photos to mitigate image quality issues in vqa photography. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 1–3, 2021.
- OpenAI. Gpt-4o mini, 2024b. URL <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>.
- Kibaek Park, Youngtaek Oh, Soomin Ham, Kyungdon Joo, Hyokyoung Kim, Hyoyoung Kum, and In So Kweon. Sideguide: a large-scale sidewalk dataset for guiding impaired people. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10022–10029. IEEE, 2020.
- Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.
- Devanshi Prajapati, Prapti Bordoloi, Rushil Sheth, and Ankit K Sharma. A survey on navigation assistance system for visually impaired and blind people. In *International Conference on Information and Communication Technology for Intelligent Systems*, pp. 569–577. Springer, 2024.
- Santiago Real and Alvaro Araujo. Navigation systems for the blind and visually impaired: Past work, challenges, and open problems. *Sensors*, 19(15):3404, 2019.
- Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. Verbosity bias in preference labeling by large language models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023. URL <https://openreview.net/forum?id=magEgFpK1y>.
- Sara Sarto, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Positive-augmented contrastive learning for vision-and-language evaluation and training. *arXiv preprint arXiv:2410.07336*, 2024.
- Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. FOIL it! find one mismatch between image and language caption. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 255–265, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1024. URL <https://aclanthology.org/P17-1024/>.

- Junha Song, Tae Soo Kim, Junha Kim, Gunhee Nam, Thijs Kooi, and Jaegul Choo. Is user feedback always informative? retrieval latent defending for semi-supervised domain adaptation without source data. In *European Conference on Computer Vision*, pp. 72–92. Springer, 2024.
- Wu Tang, De-er Liu, Xiaoli Zhao, Zenghui Chen, and Chen Zhao. A dataset for the recognition of obstacles on blind sidewalk. *Universal Access in the Information Society*, 22(1):69–82, 2023.
- PKU-Alignment Team. Align anything: training all modality models to follow instructions with unified language feedback. <https://github.com/PKU-Alignment/align-anything>, 2024.
- Tony Cheng Tong, Sirui He, Zhiwen Shao, and Dit-Yan Yeung. G-veval: A versatile metric for evaluating image and video captions using gpt-4o. *arXiv preprint arXiv:2412.13647*, 2024.
- Tuhina Tripathi, Manya Wadhwa, Greg Durrett, and Scott Niekum. Pairwise or point-wise? evaluating feedback protocols for bias in llm-based evaluation. *arXiv preprint arXiv:2504.14716*, 2025.
- Yu-Yun Tseng, Alexander Bell, and Danna Gurari. Vizwiz-fewshot: Locating objects in images taken by people with visual impairments. In *European Conference on Computer Vision*, pp. 575–591. Springer, 2022.
- Yu-Yun Tseng, Tanusree Sharma, Lotus Zhang, Abigale Stangl, Leah Findlater, Yang Wang, and Danna Gurari. Biv-priv-seg: Locating private content in images taken by people with visual impairments. *arXiv preprint arXiv:2407.18243*, 2024.
- TUAT. Sullivan a, 2024. URL <https://play.google.com/store/apps/details?id=tuat.kr.sullivan.pro&hl=ko>.
- Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. Trl: Transformer reinforcement learning. <https://github.com/huggingface/trl>, 2020.
- Yuiga Wada, Kanta Kaneda, Daichi Saito, and Komei Sugiura. Polos: Multimodal metric learning from human feedback for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13559–13568, 2024.
- Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*, 2024a.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.
- Hans Jørgen Wiberg. Be my eyes - see the world together, 2015. URL <https://www.bemyeyes.com/>.
- M Wilson. Who put that there: The barrier to blind and partially sighted people getting out and about, 2015.
- Haiying Xia, Cong Yao, Yumei Tan, and Shuxiang Song. A dataset for the visually impaired walk on the road. *Displays*, 79:102486, 2023.
- Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models. *arXiv preprint arXiv:2410.02712*, 2024.

- Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*, 2024a.
- Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024b.
- Yingyue Xu, Dan Xu, Xiaopeng Hong, Wanli Ouyang, Rongrong Ji, Min Xu, and Guoying Zhao. Structured modeling of joint deep feature and prediction refinement for salient object detection, 2019. URL <https://arxiv.org/abs/1909.04366>.
- Bufang Yang, Lixing He, Kaiwei Liu, and Zhenyu Yan. Viassist: Adapting multi-modal large language models for users with visual impairments. *arXiv preprint arXiv:2404.02508*, 2024.
- Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836, 2020.
- Zhiqiang Yuan, Ting Zhang, Jiawei Zhang, Jie Zhou, and Jinchao Zhang. Walkvlm: Aid visually impaired people walking by vision language model. *arXiv preprint arXiv:2412.20903*, 2024.
- Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, et al. Internlm-xcomposer2. 5-reward: A simple yet effective multi-modal reward model. *arXiv preprint arXiv:2501.12368*, 2025.
- Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip, 2025.
- Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024a.
- Qi Zhang, Peiyao Xiao, Shaofeng Zou, and Kaiyi Ji. Mgda converges under generalized smoothness, provably. *arXiv preprint arXiv:2405.19440*, 2024b.
- Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8018–8027, 2024c.
- Yi Zhao, Yilin Zhang, Rong Xiang, Jing Li, and Hillming Li. Vialm: A survey and benchmark of visually impaired assistance with large models. *arXiv preprint arXiv:2402.01735*, 2024.
- Xiao Zhu, Chenmian Tan, Pinzhen Chen, Rico Sennrich, Yanlin Zhang, and Hanxu Hu. Charm: Calibrating reward models with chatbot arena scores. *arXiv preprint arXiv:2504.10045*, 2025.
- Yongshuo Zong, Ondrej Bohdal, and Timothy Hospedales. VL-ICL bench: The devil in the details of multimodal in-context learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=cpGPPLLYYx>.
- Amir Zur, Elisa Kreiss, Karel D’Oosterlinck, Christopher Potts, and Atticus Geiger. Updating clip to prefer descriptions over captions, 2024.