

# Sightation Counts: Leveraging Sighted User Feedback in Building a BLV-aligned Dataset of Diagram Descriptions

Wan Ju Kang<sup>α</sup> Eunki Kim<sup>α</sup> Na Min An<sup>α</sup> Sangryul Kim<sup>α</sup>

Haemin Choi<sup>β,δ</sup> Ki Hoon Kwak<sup>γ,δ</sup> James Thorne<sup>α</sup>

KAIST AI<sup>α</sup> Sungkyunkwan University<sup>β</sup> Yonsei University<sup>γ</sup>

Work done as KAIST AI research intern<sup>δ</sup>

<sup>α</sup>{soarhigh, eunkikim, naminan, sangryul, thorne}@kaist.ac.kr

<sup>β</sup>chm1009@g.skku.edu <sup>γ</sup>kihoon090@yonsei.ac.kr

 <https://hf.co/Sightation>

## Abstract

Often, the needs and visual abilities differ between the annotator group and the end user group. Generating detailed diagram descriptions for blind and low-vision (BLV) users is one such challenging domain. Sighted annotators could describe visuals with ease, but existing studies have shown that direct generations by them are costly, bias-prone, and somewhat lacking by BLV standards. In this study, we ask sighted individuals to assess—rather than produce—diagram descriptions generated by vision-language models (VLM) that have been guided with latent supervision via a multi-pass inference. The sighted assessments prove effective and useful to professional educators who are themselves BLV and teach visually impaired learners. We release SIGHTATION, a collection of diagram description datasets spanning 5k diagrams and 137k samples for completion, preference, retrieval, question answering, and reasoning training purposes and demonstrate their fine-tuning potential in various downstream tasks<sup>1</sup>.

## 1 Introduction

Recent research has seen rapid development in vision-language models (VLM). Seeing the world and the data within has significantly advanced machine intelligence in a variety of tasks (Liu et al., 2024; Zhu et al., 2023; Yang et al., 2024; Qwen et al., 2025; Xu et al., 2024a; Li et al., 2024b), reaching a fast-growing user pool with quicker and easier access.

However, the same cannot be said of blind and low-vision (BLV) individuals. Widely adopted evaluation metrics have been shown to be biased against their preferences (Kapur and Kreiss, 2024) and benchmark studies tend to pursue a larger audience first (Li et al., 2024a,d). Publicly available

<sup>1</sup>Wherever possible, we use color blind safe palettes in figures and tables.

reward models for generic VLMs are scarce (Zang et al., 2025) — let alone for the visually impaired. Vision-language dataset research appears divided between breadth (Tang et al., 2023; Lu et al., 2023), specificity (Masry et al., 2024b,a), and volume (Zhang et al., 2025; Lee et al., 2022).

Perhaps the classroom setting best exemplifies the circumstances BLV individuals face: textual information is combined with images (such as diagrams, graphs, and figures) to help learners fully grasp complex information (Vekiri, 2002; Cheng and Gilbert, 2009; Tippett, 2016; Gates, 2018). VLMs at the command of BLV users must therefore provide select, curated information rather than an indiscriminate narration of data.

Instilling this behavior in VLMs, however, remains challenging primarily due to dataset concerns. The unavailability of large-scale BLV-aligned datasets has prompted previous studies to crowdsource a few expert sighted annotators to *generate* descriptions. The limitation of this approach is twofold: *i*) it does not account for the preference misalignment between the BLV evaluator and the sighted generator (Lundgard and Satyanarayan, 2022); *ii*) it is prone to modeling the generations after the annotator rather than the task, introducing annotator bias into the dataset (Geva et al., 2019). While Kreiss et al. (2022) has illustrated the potential of sighted users as BLV preference estimators for a few specific qualities of generations, whether their findings will generalize to a dataset-scale volume of generations or with other aspects of perceived quality remains unknown.

We construct, what is to the best of our knowledge, the first dataset that addresses the union of aforementioned challenges. We prompt a VLM to generate a guide, which will be input to a second inference pass to latently supervise the second-pass behavior in favor of BLV users. Then, we further invoke the VLM to generate diagram descriptions, saving on crowdsourcing cost and reducing

Dataset	Average Text Length	Validated by BLV?	Applications	Dimensions Assessed
<b>SIGHTATION (Ours)</b> -COMPLETIONS -PREFERENCE -RETRIEVAL -VQA -REASONING	<b>188.3</b> (words)	✓	<ul style="list-style-type: none"> <li>· Completion</li> <li>· Preference alignment</li> <li>· Retrieval</li> <li>· Reward modeling</li> <li>· Question answering</li> </ul>	<ul style="list-style-type: none"> <li>· Factuality</li> <li>· Informativeness</li> <li>· Succinctness</li> <li>· Diversity</li> <li>· Usefulness, in 4 finer aspects</li> <li>· Interpretiveness</li> <li>· Preferred Description</li> <li>· Best Sentence</li> </ul>
VisText (Tang et al., 2023)	74.6	×	Completion	Accuracy, Descriptiveness
MathVista (Lu et al., 2023)	58.0	×	VQA, Reasoning	Correctness
ChartGemma (Masry et al., 2024b)	37.5	×	Completion	Informativeness, Factual Correctness, Structure
DiagramQG (Zhang et al., 2024b)	9.5	×	DQA	Diversity, Object Density
VizWiz-VQA (Gurari et al., 2018)	8.6	✓	VQA	Diversity, Answerability
VizWiz-LF (Huh et al., 2024)	73.2	✓	VQA	Relevance, Helpfulness, Plausibility, Fluency, Correctness

Table 1: The SIGHTATION collection has been validated by teaching professionals who are visually impaired and are experienced instructors at schools for the blind. As the most text-dense diagram description dataset to date, it can be used to drive a variety of training objectives towards BLV accessibility needs. We discuss a few prime examples in Section 4. This table includes only the few most closely related works; we deliver an extended comparison in Table 7.

annotator fatigue. We distribute to sighted annotators a set of assessment tasks, substantially less demanding than a generation task, implying easier recruiting of a sufficiently large annotator population, potentially mitigating annotator bias. Finally, we design the assessment tasks such that they are finer-grained than any prior work we are aware of.

The compilation we named SIGHTATION is the first large-scale BLV-aligned dataset that is validated by BLV professionals and can be used to train on a broad range of objectives. A few statistics to highlight our dataset performance include: preference-tuning a 2B model on our dataset to achieve an average  $1.67\sigma$  increase in the usefulness rated by the BLV group; instruction-tuning a 2B model on our dataset to outperform a 3B model fine-tuned on chart comprehension (Masry et al., 2024b) in 8 out of 11 automatic metrics; contrastive tuning a BLIP-2 (Li et al., 2023) for retrieval purposes to outperform a COCO-tuned BLIP-2 by 65%p on Precision@1.

## 2 Related Work

**Accessibility Studies.** Lundgard and Satyanarayan (2022) found that BLV and sighted reader groups differ significantly on which semantic content they consider as most useful, suggesting that access to meaningful information is strongly reader-specific. VizWiz-VQA (Gurari et al., 2018) contains images

and visual QA pairs produced by blind people encouraging the development of more generalized algorithms that can assist the blind. As an extension, VizWiz-LF (Huh et al., 2024) includes long-form answers from BLV people. VisText (Tang et al., 2023) contains charts and captions that convey different levels of semantic content. As shown in Table 1, VizWiz-VQA and VizWiz-LF were validated by BLV users but only focus on Visual QA (VQA) applications. VisText examines the role of the level of semantic content but was not validated by BLV for dataset purposes. As a diagram description dataset validated by BLV users, SIGHTATION explores diverse use cases, with assessments on various aspects.

**Image Description Tasks and Models.** Wang et al. (2024) presented the QWEN2-VL collection, which includes three open-weights models: 2B, 7B, and 72B. QWEN2-VL matches the performance of GPT-4O and CLAUDE3.5-SONNET (Anthropic, 2024) in multimodal scenarios, surpassing other open-weights VLMs at the time.

GPT-4O (Hurst et al., 2024) accepts multimodal input and generates high-quality outputs including text and codes, showing powerful multimodal understanding capability. Using these VLMs, the image description task aims to generate a descriptive textual context for images of different types (*e.g.*, photographs, illustrations, schematics,

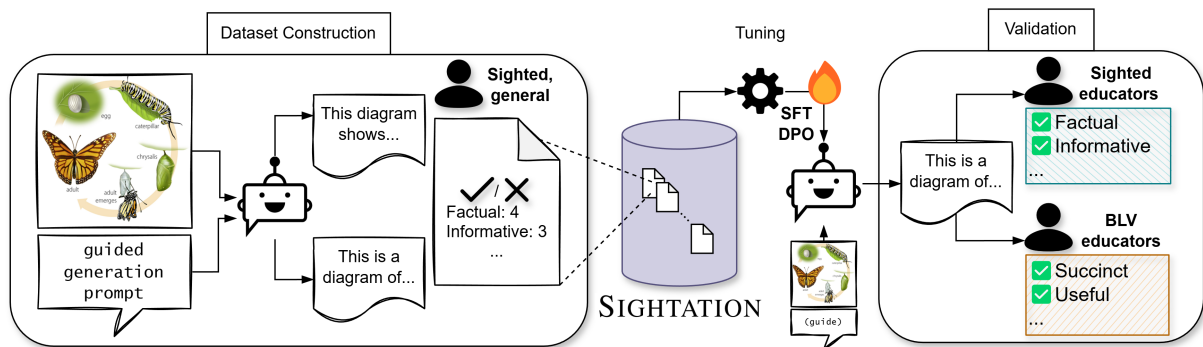


Figure 1: The key benefit of utilizing sighted user feedback lies in their *assessments*, which are based on solid visual grounding. The compiled assessments prove an effective training substance for steering VLMs towards more accessible descriptions. Dataset use and the subsequent validation are described in Sec. 4. A complete list of use cases is provided in Appendix A.

and diagrams). Flickr8K and PASCAL-50S comprise natural images, captions, and human judgments (Hodosh et al., 2013; Vedantam et al., 2015), and Polaris (Wada et al., 2024) incorporated synthetic captions from image captioning models.

ChartGemma (Masry et al., 2024b) contains chart images collected from specialized websites and instruction-tuning data generated from the charts. MathVista (Lu et al., 2023) encompasses diverse visual contexts from natural images to diagrams or plots that require mathematical reasoning. However, Table 1 shows that these datasets have an average text length much shorter than ours, even though charts and mathematical images could be highly information-dense. Complementing the limitation, SIGHTATION provides contexts that top in average text length to date with variants for downstream tasks.

**Human Annotation Efforts.** Human judgment annotations are essential in evaluating image captions, complementary to automatic metrics. Common approaches involve employing annotators to assess captions based on rating scales for specific dimensions of text quality (Gehrmann et al., 2023). How-

ever, it comes with challenges, including subjectivity and consistency issues. Amidei et al. (2019) argues that the evaluation of generated text is intrinsically subjective and relies on different factors including annotator experience, motivation, knowledge, or education. A related line of research (Glockner et al., 2024; Nie et al., 2020) directly addressing this limitation advocates that generations from few-annotator pools fall short in terms of coverage of the distribution of opinions.

### 3 The SIGHTATION Dataset

SIGHTATION is a BLV-specific vision-language dataset for the educational domain. It is built upon the AI2D dataset (Kembhavi et al., 2016): we chose this for two reasons: it contains diagrams from grade school material, requiring no specialized expertise or domain knowledge in our annotator recruiting process; diagrams pose a unique challenge to VLMs in that they often require an understanding of the rendered schematics *and* natural objects.

AI2D contains 5k science diagrams, with 150k annotations, spanning OCR texts and bounding box locations, as well as 15k multiple choice questions. Of these features, we take only the diagrams, to simplify SIGHTATION-like dataset construction in the future. All notation and labeling methods used in this section are summarized in a separate Table 8 to aid comprehension.

#### 3.1 Overview

Different annotator roles can be found in Figure 2. There are a total of 9 aspects to be assessed, and these were inspired by various related studies. In Kreiss et al. (2023), relevance and irrelevance aspects are studied to measure the image information

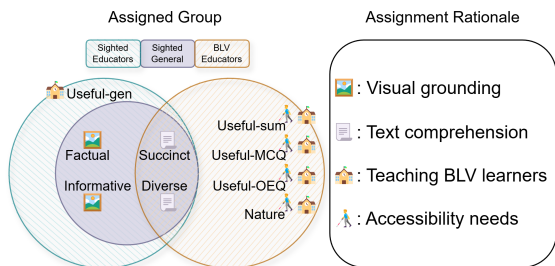


Figure 2: The qualities assessed by their respective groups.

carried in text and the inclusion of extraneous information in the text, respectively. As such, we chose to examine **Informativeness** and **Factuality** dimensions. These both require reliable visual grounding so were assigned to the sighted accordingly. We also opted for some measures to be assessed by all groups. Since brevity (Lundgard and Satyanarayan, 2022) and diverse opinion coverage (Glockner et al., 2024; Nie et al., 2020) have been pointed out as contributors to perceived quality, we chose to incorporate them as the **Succinctness** and the **Diversity** aspects, both of which are assessable with text comprehension alone. Following Tang et al. (2023), we split the use cases for the usefulness measure along typical vision-language comprehension tasks common in the classroom: **Useful-Sum** (summarization), **Useful-MCQ** (multiple-choice questions), and **Useful-OEQ** (open-ended questions). These were assigned to the BLV educators, adept at teaching and knowledgeable in accessibility needs. A general usefulness measure **Useful-Gen** was assigned to the sighted educators to probe their estimate of BLV needs. Finally, a categorical variable, **Nature**, was assigned to the BLV educators to ask for their opinion on how interpretive the text appears.

These different subsets were assigned to pursue a synergistic interplay between varying visual abilities, teaching experience, and accessibility requirements. The sighted general group, shown on the left in Figure 1 ensures that the diagram *content* is well-conveyed in the description. Sighted educators, shown on the top right of Figure 1 validate the general group’s assessment whilst also rating the general usefulness of the description to BLV users. Finally, the text-based assessment by BLV educators, shown on the bottom right in the same figure, gauges the alignment of SIGHTATION-tuned descriptions with BLV preferences. A more detailed description of the annotation tasks is in Section 3.3 for the sighted general group and in Section 4.2.1 for the sighted and BLV.

### 3.2 Guided Generation with Latent Supervision

Previous work (Lundgard and Satyanarayan, 2022) has shown that crowdsourced data visualization descriptions written by sighted crowdworkers were not equally useful to the BLV groups as they were to the sighted, in terms of describing low-level numerical elements or high-level insights such as subjective commentary. Building on this, we hypothe-

sized that the key to generating a description that is useful to BLV individuals lies not only in *what* is seen but also in *how* the perceived information is articulated. We hypothesized that introducing auxiliary data such as plausible question-answer pairs, would have a good effect as they assist the description generator with understanding which parts are critical and which are less so.

In implementing this idea, we incorporated a two-pass guided generation process. The first inference pass is to create the guide, which is a VLM-generated set of question-answer pairs in response to an input diagram. We carefully examine the quality of the question and answer pairs we have generated and, in the Appendix A.1, provide a more in-depth analysis of how these pairs differ from those originally included in the AI2D dataset. Then, the second pass generates the diagram description in response to the input diagram *and* the guided generation prompt, as shown on the leftmost part of Figure 1.

We applied this generation process with two models: GPT-4O MINI and QWEN2-VL 72B model, producing four descriptions for each of the 5k diagrams in the AI2D dataset. The working dataset thus contains 20k descriptions.

### 3.3 Annotation Tasks

1k images were randomly sampled from the working dataset. They were then paired with their respective descriptions generated by GPT-4O MINI ( $\text{Desc}^g$  and  $\text{Desc}_{++}^g$ ) and descriptions generated by QWEN2-VL ( $\text{Desc}^q$  and  $\text{Desc}_{++}^q$ ) were distributed to the 30 sighted annotators, to complete three tasks: *i*) preference choice, *ii*) quality rating, and *iii*) best sentence choice. The 1k tuples were partitioned into 10, so that 3 participants perform the annotation on a shared total of 100 tuples.

First, annotators were asked to select pairwise preferred descriptions: one from the GPT pair and the other from the Qwen pair. Second, for all four diagram descriptions, they were asked to rate the description quality across the 4 aspects assigned to them, as in Figure 2, on a 5-point Likert scale.

Lastly, they were asked to pick the best-contributing sentence from each of the four diagram descriptions. Sample screenshots of the annotation interface, along with the annotation guidelines, are provided in Appendix I.

The total number of annotations is 11,804, spanning 998 diagrams and 3,992 descriptions. Further statistics and post-processing steps are found in



Appendix C.

### 3.4 Dataset Construction

In this section, we describe how the annotated tuples are processed for various downstream tasks.

#### 3.4.1 Chat Completion

SIGHTATIONCOMPLETIONS contains instruction-response pairs from two sets: *i*) all the 4k human-annotated descriptions over 1k images, with the base instruction in Appendix G and *ii*) the top 25% highly rated descriptions for each of the 4 aspects annotated. For the latter subset, we augment the base instruction to pair responses that were of high quality in some aspect. We append an aspect-specific suffix outlining the desired quality according to our annotation guidelines in Appendix I. For instance, the aspect suffix for the factuality dimension is: “When generating the diagram description, pay close attention to making it factual. A highly factual description delivers only the facts that are grounded in the diagram.”

With the former set consisting of 4k (diagram, base prompt, description) samples and the latter set consisting of 1k (diagram, augmented prompt, description) samples per aspect, our completions dataset totals 8k samples.

#### 3.4.2 Preference Alignment

SIGHTATIONPREFERENCE also proceeds from the 4k diagram-description pairs, consisting of 4 descriptions for every image. From these 4, we take the 6 possible pairwise combinations and label “chosen” and “rejected” to each contender in the pairwise comparisons as follows.

**In-model Contenders** Within each of the 2 same-model comparisons, (*e.g.*,  $\text{Desc}^g$  versus  $\text{Desc}_{++}^g$ ) we directly take the  $\text{Preference}^{\text{model}}$  annotation to assign “chosen” and “rejected”. This assignment results in  $2 \times 1k = 2k$  chosen-rejected preference pairs.

**Cross-model Contenders** Within each of the 4 cross-model comparisons, (*e.g.*,  $\text{Desc}_{++}^g$  versus  $\text{Desc}^g$ ), we averaged the rating scores per contender and assigned<sup>2</sup> “chosen” to the ratings winner. This assignment results in  $4 \times 1k = 4k$  preference pairs.

**Synthetic Contenders** Additionally, we synthesized an inferior (“rejected”) variant of a description by removing its best sentence. To account for

<sup>2</sup>Ties are technically possible, but the collected annotations did not contain any.

the reduced length, we remove a random non-best sentence from the original description and label this variant “chosen”. This assignment results in  $4 \times 1k = 4k$  preference pairs per annotator. A maximum of three annotators evaluated the same sample, so the preference pairs total 12k. After deduplicating (*e.g.*, annotators selecting the same sentence as the best sentence), we have 10k preference pairs.

Putting together the in-model (2k), cross-model (4k), and synthetic (10k) contenders and their respective labels, SIGHTATIONPREFERENCE spans 16k pairs.

#### 3.4.3 Retrieval

Each row in SIGHTATIONRETRIEVAL contains an image as a retrieval query, accompanied by the top 1, top 5, and top 10 descriptions as the positives, as well as 10 hard negatives. This set contains 1k rows, with a potential well beyond that number. For instance, more than 63 million unique combinations can be derived utilizing 5 random samples from the 10 positives and 5 random samples from the 10 negatives. Further details can be found in Appendix D.

## 4 Performance Analysis

We designed a series of experiments to measure the performance of SIGHTATION as a *dataset*. First, we fine-tuned various models on our dataset. Then, we asked sighted and BLV teachers at schools for the blind to evaluate the generated texts. Additionally, we employ VLM judges and a number of well-known classic metrics to evaluate the descriptions. We report the main findings on the extent and breadth of performance enhancement our dataset can cultivate.

### 4.1 Fine Tuning

We chose to experiment with the QWEN2-VL series (Wang et al., 2024) considering its size variety, state-of-the-art performance at the time of writing, as well as whether the largest variant (72B) could fit on our compute cluster in its default precision, bf16, unquantized. We fine-tuned the 2B and 7B models and performed comparative analyses. Finer details on the tuning configuration are found in Appendix H.

#### 4.1.1 On SIGHTATIONCOMPLETIONS

We conducted supervised fine tuning (SFT) on our completions dataset. The 2B model underwent

full fine tuning, whereas the 7B model underwent parameter-efficient fine tuning (PEFT).

#### 4.1.2 On SIGHTATIONPREFERENCE

For preference alignment tuning, we chose to perform Direct Preference Optimization (DPO, (Rafailov et al., 2024)). Since reward models trained on generic data may not accurately represent BLV preferences, we opted for DPO, a widely used algorithm free of reward models. Before the actual DPO training, as is common in practice, we first subjected the 2B and 7B models to SFT. However, we recognized that sharing the same set of diagrams across the SFT and DPO stages could pose higher overfitting risks. With that in mind, instead of using SIGHTATIONCOMPLETIONS for SFT, we randomly sampled 1k diagrams along with their 4 descriptions from the remaining pool of generated descriptions (*i.e.*, the ones not in SIGHTATIONCOMPLETIONS) and used these to compile 4k completion samples. Afterwards, DPO was run on SIGHTATIONPREFERENCE. At both the SFT and DPO stages, the 2B model was fully fine-tuned, and the 7B model was trained with PEFT.

#### 4.1.3 On SIGHTATIONRETRIEVAL

We performed contrastive training to fine-tune BLIP-2 (Li et al., 2023) for its appeal in image-text matching. To save compute, we trained only parts of the model and with just the top 1 positive and a randomly chosen negative. The training was carried out with InfoNCE loss (Oord et al., 2018), a widely used choice for contrastive objectives.

## 4.2 Evaluation Setup

### 4.2.1 By Teaching Professionals

We recruited 17 specialized educators who teach BLV learners at schools for the visually impaired. 8 of them are themselves blind or have low vision; remaining 9 are sighted. We refer to these groups as the BLV educator group and the sighted educator group, respectively. Their demographics are reported in Tables 17 and 18.

**BLV Educators** Each BLV educator was given 40 diagrams, each with two competing descriptions. They were asked to rate text-based qualities. They were asked to perform a quantitative assessment on the aspect set pictured in Figure 2.

Following Tang et al. (2023); Lundgard and Satyanarayan (2022), we chose to investigate the usefulness of the diagram descriptions, but in three finer manifestations. Specifically, we asked the

BLV educators to assess how useful the description is as a textual aid providing *i*) a summary of the diagram content, *ii*) clues that would be helpful when solving short-answer multiple-choice questions about the diagram, and *iii*) clues that would be helpful when answering long-answer open-ended questions about the diagram.

**Sighted Educators** Each sighted educator was given 40 diagrams, each with two competing descriptions with randomized order of presentation. They were then asked to evaluate the descriptions according to the guidelines for the sighted educator group, found in Appendix I. Their aspect set, also shown in Fig. 2, includes a usefulness estimate to BLV users.

### 4.2.2 By Automatic Metrics

We performed a VLM-as-a-Judge (Dubois et al. (2023), Zheng et al. (2023)) evaluation with QVQ-72B-PREVIEW, where we instruct the VLM to take the **Image**, **Desc**<sup>model</sup>, and **Desc**<sub>++</sub><sup>model</sup> triplet as input and produce a JSON-formatted evaluation with the same aspects as with the human annotation.

As for classic metrics, we collect widely recognized reference-free metrics since the AI2D dataset does not contain references: CLIP score (Hessel et al., 2021), SigLIP score (Zhai et al., 2023), BLIP-2 Retrieval score (Li et al., 2023), Self-BLEU (based on BLEU (Papineni et al., 2002)), PAC score (Sarto et al., 2023), and LongCLIP-B/L (Zhang et al., 2024a). For the retrieval task, we chose to measure recall@*K* and precision@*K* for *K* = 1, 5, 10, as do numerous retrieval studies.

## 5 Results

We report the evaluation results by the BLV educator group, the sighted educator group, VLM judges, and classic metrics. For each group, we discuss the effectiveness of the combined recipe, then with the guided generation ablated, and with the tuning step ablated. Here, we focus on the evaluation by BLV; sighted educator and VLM-as-a-Judge evaluation, as well as classic metric results are found in Appendix E.

### 5.1 Evaluation by BLV Educators

Here, we conduct an analysis of effect size, an intuitive choice for aggregate analysis on different sample sets rated by different evaluators. Figure 3 shows the effect size computed from BLV educators' assessment. The radial axis corresponds to

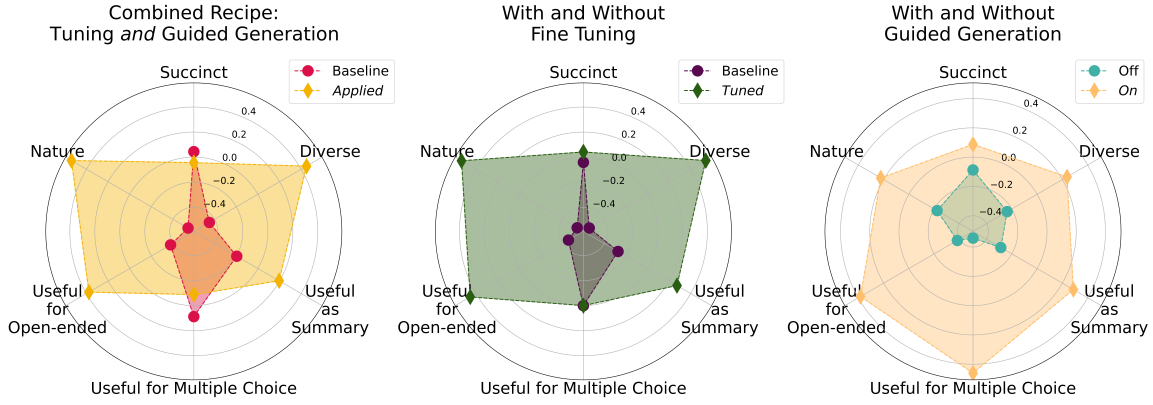


Figure 3: Tuning VLMs on SIGHTATION enhanced various qualities of the diagram descriptions, evaluated by BLV educators, and shown here as normalized ratings averaged in each aspect. The capability of the dataset is most strongly pronounced with the 2B variant, shown above. Full results across 4 models and 22 metrics are reported in Tables 11, 12, 13, and 14.

the mean ratings on each of the two sets of samples under comparison, normalized by their pooled standard deviation ( $\sigma$ ). Naturally, the radial axis is in units of the pooled standard deviation.

The first radar chart in Figure 3 shows the result of comparing  $\text{Desc}^{\text{q2bbbase}}$  and  $\text{Desc}^{\text{q2bdpo}}$ . The latter was rated more than  $1\sigma$  higher in interpretiveness (**Nature**);  $0.8\sigma$  better in diversity and usefulness for open-ended questions;  $0.4\sigma$  units more useful as a summary.

In the middle of the same figure is shown the ablated result of fine tuning, with the guided generation turned on for both sets: a comparison between  $\text{Desc}^{\text{q2bbbase}}$  and  $\text{Desc}^{\text{q2bdpo}}$ . All 6 aspects were judged in favor of the latter, with as large as  $1.2\sigma$  difference in interpretiveness and diversity and  $0.8\sigma$  in usefulness for open-ended questions.

On the right is shown the effect of the guided generation on a SIGHTATIONPREFERENCE-tuned 2B model: a comparison between  $\text{Desc}^{\text{q2bdpo}}$  and  $\text{Desc}^{\text{q2bbbase}}$ . Guided generation yields significant enhancement for the DPO-tuned case, with  $1\sigma$  higher in usefulness for multiple choice questions, followed by approximately  $0.8\sigma$  improvement in usefulness for open-ended questions, an overall improvement in every aspect down to succinctness, with  $0.2\sigma$ . However, as will be discussed with Table 4, this effect by the guided generation is achieved only after the model is fine-tuned on our dataset, implying that a good alignment is a prerequisite for attempting to benefit from test-time prompting.

## 5.2 Evaluation by Sighted Educators

Ratings from sighted educators can be found across Tables 11 to 16. An interesting observation can be made about training effects in Tables 13 and 14. With the base models, sighted educators and BLV educators tended to prefer opposites between **Desc** and **Desc**<sub>++</sub>. However, when training was applied (rightmost column), the two groups' preferences came to a closer agreement.

## 6 Discussion

Tables 2, 3, and 4 show Cohen's  $d$ , which is the size of the effect of the treatment in the respective table. Ratings on **Nature** are not included in the average computation since it is a categorical variable; *i.e.*, a low **Nature** rating simply means the description was perceived to be more straight facts-oriented than commentary-oriented, and not necessarily of a lower quality.

**Combined Effect Size** Table 2 shows the effect size of fine tuning on SIGHTATION *and* applying the guided generation prompt at test time. With the combined recipe applied, the 2B model achieves an average of  $0.36\sigma$  units of improvement, while the 7B model,  $0.58\sigma$  units. Intriguing observations can be made on succinctness. The 2B model exhibited the smallest effect size in this aspect, whereas the 7B model achieved the highest enhancement. This suggests that the combined recipe applied on the smaller model had negligible effect in making its descriptions more succinct. In fact, the combined recipe enhanced **Nature** by a large effect ( $1.08\sigma$ ), implying that, with smaller models, the prime importance of the combined recipe lies in shaping the

Aspect	Combined Effect Size	
	2B	7B
Succinct	-0.09	1.69
Diverse	0.90	0.46
Useful-Sum	0.39	0.53
Useful-MCQ	-0.18	0.20
Useful-OEQ	0.76	0.00
Average	0.36	0.58
Nature	1.08	-2.38

Table 2: Combined recipe effect size on each aspect, measured with BLV assessment.

Aspect	Tuning Effect Size			
	2B	2B+GG	7B	7B+GG
Succinct	0.06	0.08	0.37	-0.11
Diverse	0.87	1.08	-0.06	0.00
Useful-Sum	0.20	0.55	0.14	0.36
Useful-MCQ	0.29	0.00	-0.54	0.00
Useful-OEQ	1.01	0.90	-0.74	-0.19
Average	0.49	0.52	-0.17	0.01
Nature	1.49	1.06	-3.14	-0.31

Table 3: Fine tuning effect size on each aspect, measured with BLV assessment.

Aspect	Guided Generation Effect Size		
	GPT	2B Base	2B DPO
Succinct	0.18	-0.17	0.17
Diverse	-0.13	-0.13	0.47
Useful-Sum	0.48	-0.17	0.57
Useful-MCQ	0.13	-0.20	0.92
Useful-OEQ	0.76	-0.07	0.77
Average	0.28	-0.15	0.58
Nature	0.33	0.08	3.17

Table 4: Guided generation effect size on each aspect, measured with BLV assessment.

descriptions to be far more interpretive. The opposite can be said of the 7B model: the combined recipe greatly ( $1.69\sigma$ ) enhances its succinctness, whilst shaping its descriptions far less interpretive ( $-2.38\sigma$ ) and straight facts-oriented instead. This is in line with 3 separate comments by our BLV educators (B1, B2, and B5) who have, unknowingly of each other’s interview responses, stressed the importance of succinctness: “The description must deliver all visual items in an accurate and consistent manner, with not too long a text and including the key elements.”

**Tuning Effect Size** Table 3 shows the effect size of fine tuning on SIGHTATION. For instance, with guided generation absent, the 2B model still reaps  $0.87\sigma$  units of improvement in the diversity aspect of its descriptions. The improvement margin is even amplified further by applying guided generation on the tuned model, except for usefulness in solving questions. The table shares the observation made on the succinctness-nature relationship conveyed in Table 2, albeit to a lesser extent on the 7B model with guided generation. This set, whose ratings are on the rightmost column of Table 3, showed meaningful effect size only in usefulness as a summary and nature. This implies that larger models are already somewhat capable of capitalizing on the guided generation prompt at test time and carry less reliance on the fine tuning process.

**Guided Generation Effect Size** Table 4 shows that the guided generation yields benefits even to GPT, possibly indicative of the underrepresentation of BLV accessibility needs and preferences in the pre-training data. It is important to note that, for the 2B model, the best effect of guided generation is achieved only *after* the model is tuned on our dataset, again highlighting the BLV alignment capabilities of our dataset, that cannot be mimicked by test-time prompt engineering alone.

**Comparison with Existing Datasets** As can be seen in Tables 1 and 7, no single dataset exactly matches the purpose and design of ours. However, for the sake of impartiality, we performed a number of comparative experiments with subsets of SIGHTATION.

**versus ChartGemma** Table 5 presents results from a completion task against the CHARTGEMMA dataset and model. The 3B language model ChartGemma has been trained on a dataset of the same name for generating captions while our 2B QWEN2-VL was fine-tuned on SIGHTATIONCOMPLETIONS. Despite the size disadvantage, the 2B model outperforms the 3B ChartGemma across many metrics.

**versus COCO** Figure 4 presents results from an image-to-text retrieval task against COCO dataset. BLIP-2 models were cross-validated to examine the retrieval training effectiveness of the datasets. While the COCO-trained model failed to generalize to test-time SIGHTATIONRETRIEVAL, the model trained on SIGHTATIONRETRIEVAL performed similarly well on COCO at test time. Exact statistics are found in Table 6.

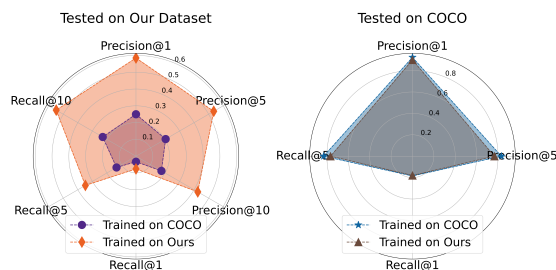


Figure 4: Retrieval performance was measured with 2-way cross validation. On our test set (Left), the COCO-tuned BLIP-2 generalizes poorly, whereas on the COCO test set (Right), the SIGHTATIONRETRIEVAL-tuned BLIP-2 performs on par with the COCO-tuned BLIP-2.



Experiment ID		Assessments for		
Description	Generators	Metrics	Desc <sup>chartgemma</sup>	Desc <sup>q2bsft</sup>
			Experiment 3c	CLIP Score
CHARTGEMMA (3B) vs. FINE-TUNED QWEN2- VL-2B-INSTRUCT	SigLIP Score	0.872	<b>0.940</b>	
	BLIP-2 Retrieval Score	<b>0.511</b>	0.490	
	Self-BLEU	<b>0.305</b>	0.280	
	PAC-Score	0.705	<b>0.716</b>	
	LongClip-B	0.316	<b>0.684</b>	
	LongClip-L	<b>0.559</b>	0.441	
	· VLM-as-a-Judge Evaluation Average		2.951	<b>3.860</b>
	Factuality	3.068	<b>4.119</b>	
	Informativeness	2.848	<b>3.967</b>	
	Succinctness	3.253	<b>3.925</b>	
Diversity	2.635	<b>3.428</b>		

Table 5: A 2B model fine-tuned on SIGHTATIONCOMPLETIONS outperforms a 3B model tuned on a larger dataset. Note that CHARTGEMMA is not meant for conversational use. Hence, for a fair comparison, we did *not* enter our guided generation prompt and instead input only the brief request “Generate a caption” to both models.

2-way Cross-validation of <b>BLIP-2</b>						
Train set	N/A (Pre-trained)		COCO		SIGHTATIONRETRIEVAL (Ours)	
	COCO	Ours	COCO	Ours	COCO	Ours
Recall@1	0.171	0.048	0.185	0.033	0.180	0.076
Recall@5	0.767	0.210	0.831	0.134	0.766	0.348
Recall@10	—	0.340	—	0.229	—	0.549
Precision@1	0.856	0.371	0.924	0.250	0.900	0.585
Precision@5	0.767	0.324	0.831	0.204	0.766	0.535
Precision@10	—	0.263	—	0.175	—	0.425

Table 6: SIGHTATIONRETRIEVAL shows promising potential as a challenging and effective training material for image-to-text retrievers. Two important observations can be made: the model trained on our set generalizes to COCO better than the other direction; our model performs on par with the model that was both trained and tested on COCO.  $K = 10$  values are missing for tests with COCO, since its samples contain only 5 positives each.

## 7 Conclusion

We release SIGHTATION, a suite of datasets showcasing these key characteristics: *i*) produced with BLV-oriented guided generation of VLMs instead of crowdworkers, who pose annotator bias concerns and are bottlenecked by cost and fatigue, *ii*) validated by specialized teaching professionals at schools for the blind, and *iii*) demonstrated across a wide range of use cases, making the most of the invaluable feedback from BLV and sighted groups and inviting continued active endeavor towards accessible language and education.

### Limitations

**Challenges in Supervision and Capturing Details in Diagram** One challenge of our current approach is that the supervision signal predominantly relies on the QA format, leaving the exploration

of alternative supervision substances relatively underdeveloped. In addition, our pipeline does not fully leverage advanced segmentation techniques, which could be crucial for accurately capturing and interpreting complex diagrammatic details. These constraints may affect the system’s performance with diagrams that feature intricate or non-standard layouts. This aspect will be revisited in future research, as it holds the potential to achieve further advancements beyond the performance improvements demonstrated with our current dataset version.

### Ethics Statement

**Potential Risks in Dataset Generation** We acknowledge that during the process of creating our dataset, we utilized various LLMs, and there is a potential ethical risk that unintended biases or unexpected outcomes may have been inadvertently

included. However, once the human labels are applied, the post-processed information minimizes this risk.

**AI Assistant** Also, we hereby acknowledge that we have received assistance with grammar and word choice from LLMs such as chatGPT-4o in preparing this paper. However, all text is ultimately composed in the authors' own words and was originally formulated by them.

## Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No. RS-2024-00457882, AI Research Hub Project and No.RS-2019-II190075, Artificial Intelligence Graduate School Program (KAIST)) and National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.RS-2024-00406715)

## References

- Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. Agreement is overrated: A plea for correlation to assess human evaluation reliability. In *INLG 2019*, Tokyo, Japan. ACL.
- AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1.
- Shreyanshu Bhushan and Minhoo Lee. 2022. Block diagram-to-text: Understanding block diagram images by generating natural language descriptors. In *Findings of AACL 2022*, Online only. ACL.
- Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Sriniwasan, Tianyi Zhou, Heng Huang, et al. 2023. Alpaca: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.
- Maurice Cheng and John K Gilbert. 2009. Towards a better utilization of diagrams in research into the use of representative levels in chemical education. In *Multiple representations in chemical education*, pages 55–73. Springer.
- Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2023. AlpacaFarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36:30039–30069.
- Peter Gates. 2018. The importance of diagrams, graphics and other visual representations in stem teaching. *STEM education in the junior secondary: The state of play*, pages 169–196.
- Sebastian Gehrmann, Elizabeth Clark, and Thibault Selam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.
- Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *arXiv preprint arXiv:1908.07898*.
- Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2024. AmbiFC: Fact-checking ambiguous claims with evidence. *Transactions of the Association for Computational Linguistics*, 12.
- Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.
- Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.
- Mina Huh, Fangyuan Xu, Yi-Hao Peng, Chongyan Chen, Hansika Murugu, Danna Gurari, Eunsol Choi, and Amy Pavel. 2024. Long-form answers to visual questions from blind and low vision people. *arXiv preprint arXiv:2408.06303*.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Rhea Kapur and Elisa Kreiss. 2024. Reference-based metrics are biased against blind and low-vision users' image description preferences. In *NLP4PI 2024*, Miami, Florida, USA. ACL.
- Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer.
- Elisa Kreiss, Cynthia Bennett, Shayan Hooshmand, Eric Zelikman, Meredith Ringel Morris, and Christopher Potts. 2022. Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics. *arXiv preprint arXiv:2205.10646*.

- Elisa Kreiss, Eric Zelikman, Christopher Potts, and Nick Haber. 2023. Contextref: Evaluating reference-less metrics for image description generation. *arXiv preprint arXiv:2309.11710*.
- Dong Won Lee, Chaitanya Ahuja, Paul Pu Liang, Sanika Natu, and Louis-Philippe Morency. 2022. Multimodal lecture presentations dataset: Understanding multimodality in educational slides. *arXiv preprint arXiv:2208.08080*.
- Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024a. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Lei Li, Yuanxin Liu, Linli Yao, Peiyuan Zhang, Chenxin An, Lean Wang, Xu Sun, Lingpeng Kong, and Qi Liu. 2024b. Temporal reasoning transfer from text to video. *arXiv preprint arXiv:2410.06166*.
- Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024c. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*.
- Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, et al. 2024d. Vlrwardbench: A challenging benchmark for vision-language generative reward models. *arXiv preprint arXiv:2411.17451*.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.
- Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.
- Alan Lundgard and Arvind Satyanarayan. 2022. Accessible visualization via natural language descriptions: A four-level model of semantic content. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1073–1083.
- Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of ACL 2022*, Dublin, Ireland. ACL.
- Ahmed Masry, Mehrad Shahmohammadi, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2024a. ChartInstruct: Instruction tuning for chart comprehension and reasoning. In *Findings of ACL 2024*, Bangkok, Thailand. ACL.
- Ahmed Masry, Megh Thakkar, Aayush Bajaj, Aaryaman Kartha, Enamul Hoque, and Shafiq Joty. 2024b. Chartgemma: Visual instruction-tuning for chart reasoning in the wild. *arXiv preprint arXiv:2407.04172*.
- Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? *arXiv preprint arXiv:2010.03532*.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. **Qwen2.5 technical report**. *Preprint*, arXiv:2412.15115.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.
- Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2023. Positive-augmented contrastive learning for image and video captioning evaluation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6914–6924.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tiejun Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.
- Benny J Tang, Angie Boggust, and Arvind Satyanarayan. 2023. Vistext: A benchmark for semantically rich chart captioning. *arXiv preprint arXiv:2307.05356*.

- Christine D Tippett. 2016. What recent research on diagrams suggests about learning with rather than learning from visual representations in science. *International Journal of Science Education*, 38(5):725–746.
- Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.
- Ioanna Vekiri. 2002. What is the value of graphical displays in learning? *Educational psychology review*, 14:261–312.
- Yuiga Wada, Kanta Kaneda, Daichi Saito, and Komei Sugiura. 2024. Polos: Multimodal metric learning from human feedback for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13559–13568.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2024a. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024b. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *arXiv preprint arXiv:2406.08464*.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhua Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.
- Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, et al. 2025. Internlm-xcomposer2. 5-reward: A simple yet effective multi-modal reward model. *arXiv preprint arXiv:2501.12368*.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.
- Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2024a. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, pages 310–325. Springer.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Wenqi Zhang, Hang Zhang, Xin Li, Jiashuo Sun, Yongliang Shen, Weiming Lu, Deli Zhao, Yueting Zhuang, and Lidong Bing. 2025. 2.5 years in class: A multimodal textbook for vision-language pretraining. *arXiv preprint arXiv:2501.00958*.
- Xinyu Zhang, Lingling Zhang, Yanrui Wu, Muye Huang, Wenjun Wu, Bo Li, Shaowei Wang, and Jun Liu. 2024b. Diagramqg: A dataset for generating concept-focused questions from diagrams. *arXiv preprint arXiv:2411.17771*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.



## A Our Complete Dataset Collection

We describe the rest of the dataset collection.

### A.1 SIGHTATIONVQA

In constructing **Desc<sub>++</sub>** for comparison with **Desc**, we discovered that the quality of the Question–Answer pairs directly determines the quality of the resulting context. To clarify why we invested significant effort in carefully designing these question answer pairs, we employed an LLM as a judge to evaluate and classify them according to different quality levels. To measure the quality of the Question Answer pairs, we used the VLM-as-a-Judge prompt using GPT-4o model. The prompt itself is found in Appendix G.

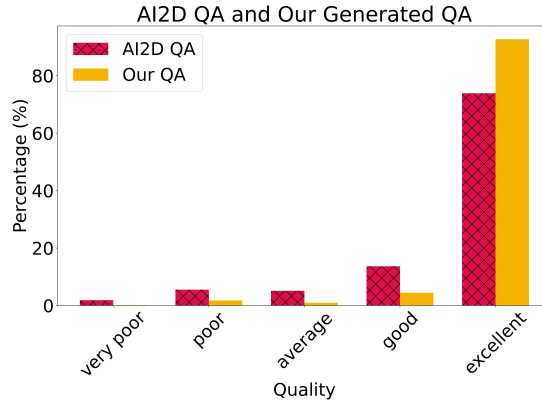


Figure 5: Percentage distribution of the quality of question-answer pairs in AI2D and SIGHTATIONVQA

Following [Chen et al. \(2023\)](#) and [Xu et al. \(2024b\)](#), we compared two sets of QA pairs with GPT-4o. Our generated QA sets are with up to six QA pairs for each of 4,903 diagrams, producing a total of 29,438 QA pairs (sometimes exceeding six pairs per diagram). As can be seen Figure 5, we found that 92.66% of these our generated QA pairs were rated “Excellent”, while 4.47% were deemed “Good”, underscoring their high quality. By contrast, the QA pairs sourced from the AI2D dataset, though numerous, included a large portion of masked or minimally informative queries. After filtering out these masked questions, we were left with 9,708 self-contained questions spanning 3,099 diagrams, where 73.86% received an “Excellent” rating and 13.65% were deemed “Good”. This comparison reveals that our generated QA pairs provide a more robust and contextually relevant foundation, reinforcing the value of our meticulous QA design in constructing effective **Desc<sub>++</sub>**.

### A.2 SIGHTATIONREASONING

Employing **Desc** and **Desc<sub>++</sub>**, we constructed SIGHTATIONREASONING, a reasoning dataset that consists of reasoning path and reasoning QA pairs. The prompts used for the construction of reasoning datasets are found in Appendix G. To verify the quality of contents as a reasoning dataset, 10% of the samples were randomly selected to be manually inspected.

**Reasoning Path** The reasoning path explains the logical flow or deployment of the contents in a diagram such as cause-effect relationships, step-by-step processes, explanations of phenomena, comparisons of contrasts, or dependencies between components. Employing 1k diagram images and descriptions in SIGHTATION, the reasoning path was identified and generated by QVQ-72B-Preview. The reasoning path extracted from **Desc** and **Desc<sub>++</sub>** is denoted as **RPath** and **RPath<sub>++</sub>** respectively. Consequently, one diagram possesses two reasoning paths, resulting in 2k paths in total.

**Reasoning QA** The reasoning QA encompasses five types of QA pairs that require a logical understanding of diagram contents and reasoning capabilities: Causal, Process, Conditional, Explanatory, and Reverse. Similarly to the reasoning path data, **RQA** and **RQA<sub>++</sub>** were generated by QVQ-72B-PREVIEW using 1k diagram images and descriptions. As a result, one diagram contains 10 reasoning QA pairs in which **RQA** and **RQA<sub>++</sub>** respectively include 5 pairs. While SIGHTATIONVQA covers the visual

Dataset	Average Text Length	Validated by BLV?	Applications	Dimensions Assessed
<b>SIGHTATION (Ours)</b> -COMPLETIONS -PREFERENCE -RETRIEVAL -VQA -REASONING	<b>188.3</b> (words)	✓	· <b>Completion</b> · <b>Preference alignment</b> · <b>Retrieval</b> · <b>Reward modeling</b>	· <b>Factuality</b> · <b>Informativeness</b> · <b>Succinctness</b> · <b>Diversity</b> · <b>Usefulness</b> , in 4 finer aspects · <b>Interpretiveness</b>
VisText (Tang et al., 2023)	74.6	×	Completion	Accuracy, Descriptiveness
MathVista (Lu et al., 2023)	58.0	×	VQA, Reasoning	Correctness
ChartGemma (Masry et al., 2024b)	37.5	×	Completion	Informativeness, Factual Correctness, Structure
CBD (Bhushan and Lee, 2022)	114.5	×	Summarization	Adequacy, Fluency, Coherence
VizWiz-VQA (Gurari et al., 2018)	8.6	✓	VQA	Diversity, Answerability
VizWiz-LF (Huh et al., 2024)	73.2	✓	VQA	Relevance, Helpfulness, Plausibility, Fluency, Correctness
DiagramQG (Zhang et al., 2024b)	9.5	×	DQA	Diversity, Object Density
ScienceQA (Lu et al., 2022)	119.7	×	VQA, Reasoning	Correctness
ChartQA (Masry et al., 2022)	13.0	×	VQA	Syntactic Diversity
Flickr8K (Hodosh et al., 2013)	11.8	×	Description	Diversity
PASCAL-50S (Vedantam et al., 2015)	8.8	×	Description	Factuality, Literality, Generality
Polaris (Wada et al., 2024)	11.5	×	Description	Fluency, Relevance, Descriptiveness
Multimodal Arxiv (Li et al., 2024c)	49.7	×	Description, VQA, Reasoning	Factual Alignment, Visual Clarity, Unambiguous Textual Information, Question and Option Relevance, Comprehensive Integration, Equitable Content
MMMU (Yue et al., 2024)	53.2	×	VQA, Reasoning	Difficulty, Knowledge, Reasoning

Table 7: Extended related work.

structure and details of a diagram, the reasoning QA in SIGHTATIONREASONING consists of more knowledge-intensive questions that require logical thinking, paving the way for the reasoning applications of SIGHTATION.

**Evaluation** The reasoning path of SIGHTATIONREASONING can be used as an overall representation of "logical flow" or "relationships between instances" in a diagram when understanding it, which was emphasized in the BLV educator questionnaire. To make a model employ this information when responding to reasoning questions and evaluate the reasoning paths, we fed QWEN2-VL-7B-INSTRUCT with **RPath** and **RPath<sub>++</sub>** separately and asked it to solve 10 questions in **RQA** and **RQA<sub>++</sub>**. The similarity score between the gold answers and generated answers was calculated using BERTScore (Zhang et al., 2019), and the scores for the two cases both resulted in 0.975, verifying the equal usefulness of **RPath** and **RPath<sub>++</sub>**.

## B Further Related Work

In Table 7, we extend Table 1 for a more comprehensive view of neighboring datasets. To the best of our knowledge, there exists no dataset to date surpassing our contribution in terms of the breadth of use cases and granularity of validation with BLV individuals.

## C Details on the Annotations

### C.1 Logistics

All experimentation was reviewed and approved by the Institutional Review Board. Recruiting the sighted general group was done via an online forum. Each sighted general group annotator was paid

Mission : Describe below diagram for BLV users

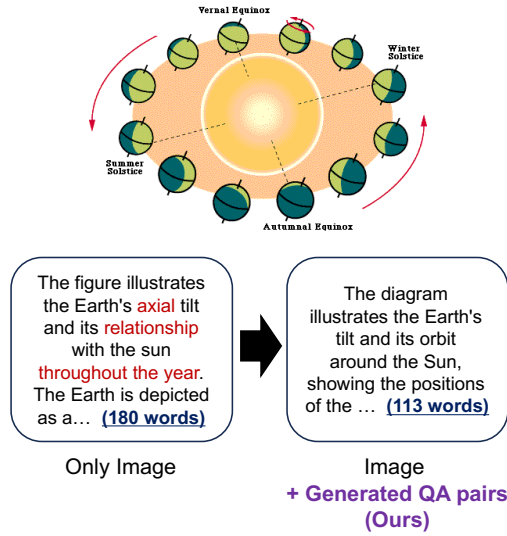


Figure 6: Less can be more for BLV users. Our approach streamlines details to highlight the core information while emphasizing key details to increase information density and maximize information efficiency per unit length.

Notation	Description
$(\cdot)^{model}$	The description <b>Desc</b> generated by (or an annotation on a generation from) a $model \in \{g, q\}$ , for GPT-4O MINI and QWEN2-VL, respectively. Later overloaded with narrower descriptors, such as base, sft, and sft+dpo to refer to the baseline/tuned models.
$(\cdot)^{anchor}$	The conditioning input at the description generation stage. $anchor \in \{None, ++\}$ , for the one-pass image-only conditioning and the two-pass image+QA conditioning, respectively.
<b>Preference</b> <sup>model</sup>	Preference annotation between two <b>Desc</b> <sup>model</sup> 's on different conditioning inputs. Value takes either of the $anchor$ set $\{None, ++\}$
$Aspect^{model}_{anchor}$	Rating annotation in terms of $Aspect \in \{\mathbf{Factuality}, \mathbf{Informativeness}, \mathbf{Succinctness}, \mathbf{Diversity}, \mathbf{Usefulness-Gen}, \mathbf{Usefulness-Sum}, \mathbf{Usefulness-MCQ}, \mathbf{Usefulness-OEQ}, \mathbf{Nature}\}$ , for a description generated by $model$ conditioned on $anchor$ . Value is an integer ranging from 1 to 5, on the 5-point Likert scale.
<b>Best</b> <sup>model</sup> <sub>anchor</sub>	Best sentence annotation. Value is a substring of <b>Desc</b> <sup>model</sup> <sub>anchor</sub> .

Table 8: Notations

an approximate equivalent of USD80 for completing the assigned task. Recruiting the educators was done by directly corresponding with the schools for the blind. A sighted educator was compensated an approximate equivalent of USD80. A BLV educator was compensated an approximate equivalent of USD80 to USD160, depending on the number of samples completed.

## C.2 Annotations Statistics

**Preliminaries** Of the 1,000 diagrams distributed to the annotators, 956 have been annotated by three annotators; 41 by two; 1 by a single annotator; and 2 by none. We collected annotations on 3,992 diagram-description pairs, each with at most 3 annotations.

**Internal Consistency** In Table 9, we report the Cronbach’s alpha value for each assessment group. The statistic is widely interpreted as the reliability of a set of survey items.

Group	Cronbach’s $\alpha$
Sighted General	0.70
Sighted Educators	0.94
BLV Educators	0.80

Table 9: Our survey items are considered of acceptable ( $\geq 0.7$ ) to excellent ( $\geq 0.9$ ) reliability.

**Point-Biserial Correlation** We examine the relationship between the binary variable, **Preference**, and the 5-point scale ratings per aspect.

Group	Aspects				
	Factuality	Informativeness	Succinctness	Diversity	Usefulness-Gen
Sighted General	0.36***	0.37***	0.31***	0.34***	0.43***
Sighted Educators	0.25***	0.30***	0.30***	0.34***	—

Table 10: Correlation values between preference choice and aspect ratings were found to be moderately positive and statistically significant. (\*\*\*:  $p < 0.001$ )

**Cohen’s  $d$**  Cohen’s  $d$  is a widely used statistic to measure the size of the effect of a treatment. It is the difference in the means of the treatment and control groups, normalized by the pooled standard deviation. By guidelines set forth by Cohen himself, values over 0.2 are typically considered a small effect size; 0.5, medium; and 0.8, large.

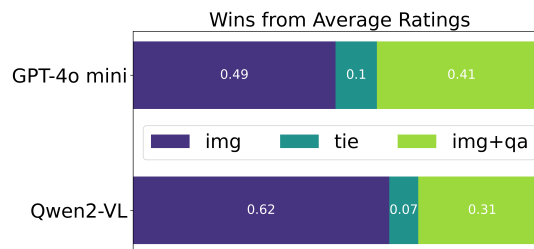


Figure 7: Win rates by *model*.

### C.3 Annotations Post-processing

**Preference Choice** We aggregate the multiple annotations on the basis of majority. That is, for the three-annotation samples, a 3:0 or 2:1 is considered a “victory” and the victor **Desc** wins that sample. For two-annotation samples with differing preferences, a tie is recorded. The overall win-loss statistics normalized against the number of diagrams (998) is shown in Figure 7.

#### Rating Assessment

**Best Sentence Choice** The best sentence for each context was manually selected by BLV annotators after listening to the context. We analyzed people’s preferences by examining the position and length of the best sentence within each context.

**Position** The normalized position of the best sentence is shown in Figures 8-9. To calculate the relative position, both the context and the best sentence were tokenized at the word level, and the position of the overlapping best sentence within the context was identified. This position was then normalized to



a value between 0 and 1 by dividing it by the total length of the context. Furthermore, since some BLV annotators could not select a best sentence within the context, a filtering step was applied by setting an overlap threshold of 0.9 to account for such cases.

Figures 8-9 illustrate that the best sentences in each context are predominantly positioned at the beginning and end. This pattern can be attributed to cognitive biases, specifically primacy bias and recency bias. Primacy bias refers to the tendency to place greater importance on the first pieces of information encountered in a sequence, while recency bias reflects the tendency to prioritize the most recently encountered information. Consequently, these biases increase the likelihood that preferred sentences will be selected from the beginning and end of the context.

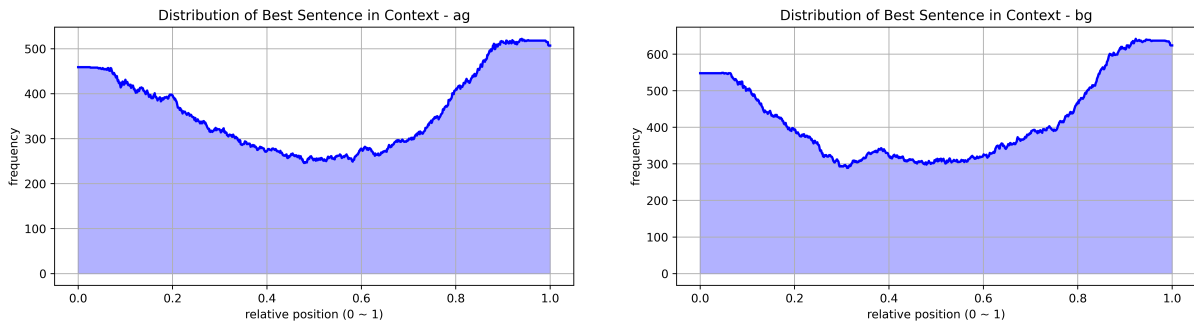


Figure 8: Descriptions generated by GPT-4O MINI

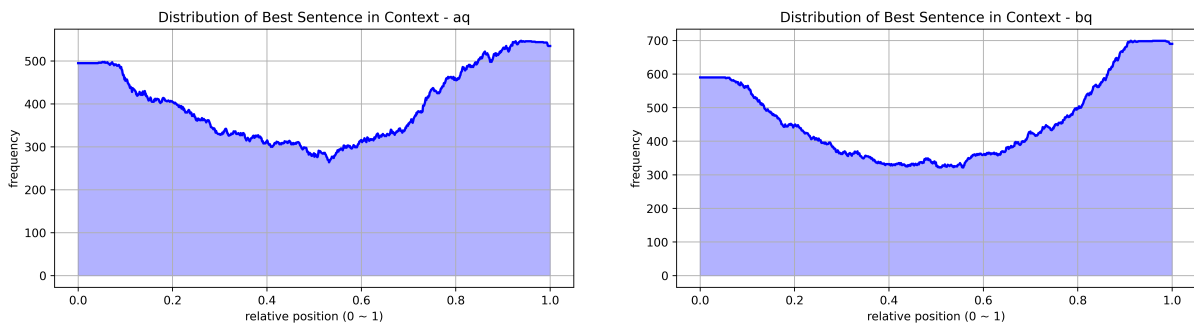


Figure 9: Descriptions generated by QWEN2-VL

**Length** The length of the best sentence in each context is presented in Figure 10. The length was determined by counting the total number of words in the best sentence. As shown in Figure 10, the best sentences across different contexts predominantly consist of 20 to 30 words, exhibiting a similar distribution pattern.

## D Retrieval Dataset Construction

The winner among the four human-annotated descriptions was assigned as the top 1 positive in terms of preference and average rating. The top 5 set contains all 4 human-annotated descriptions and 1 synthesized description; the top 10 set is a superset of the top 5, joined by 5 more synthetic descriptions. The synthetic descriptions are perturbed versions of the human-annotated descriptions, each missing a random, non-best sentence. The 10 hard negatives for an image were selected among the combined pool of top 1 descriptions for other images, sorted by cosine similarity in the embedding space. The embeddings were computed by a widely used sentence transformer, ALL-MPNET-BASE-V2(Song et al., 2020).

## E Detailed Results

We report the VLM-as-a-Judge evaluation and classic metric results in Tables 11, 12, 13, and 14.

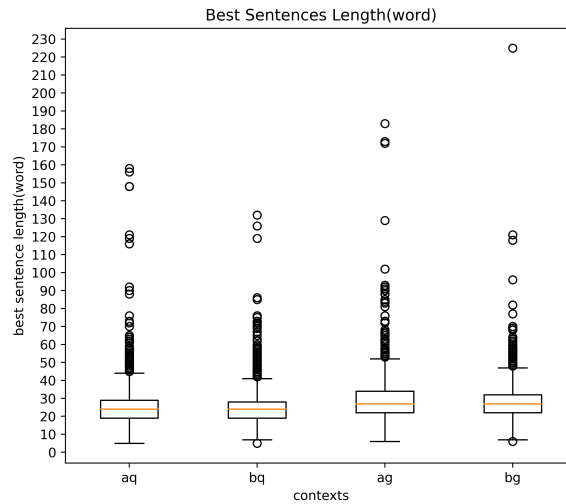


Figure 10: boxplot of best sentence length

### E.1 Evaluation by Automatic Metrics

**QVQ-72B-PREVIEW** On GPT and Qwen 72B generations, the VLM judge did not reveal significant difference between the two anchors, and the little differences present aligned with assessments by the sighted general group, as can be expected from a general-purpose VLM.

It is important to note that even a state-of-the-art VLM fails to capture the BLV perspectives in text evaluation.

**Classic Metrics** To our surprise, almost all instances of classic metric evaluations resulted in a win for the ++ anchor. However, the numbers from classic metrics evaluation are more of a shortcoming on the part of the classic metrics, rather than an accurate portrayal of the effectiveness of our proposed latent supervision. This is because our “gold” ground truths from BLV educators show that, while the QA-guided generation does manifest in ways beneficial to BLV individuals, classic automatic metrics poorly represent the assessment space covered by BLV, such as with the **Diversity** and **Usefulness-OEQ** aspects.

Experiment ID		Assessments for	
Description	Generators	Metrics	
			Desc Desc <sub>++</sub>
Experiment 1a GPT-4O MINI vs. GPT-4O MINI		CLIP Score	0.476 <b>0.524</b>
		SigLIP Score	<b>0.921</b> 0.914
		BLIP-2 Retrieval Score	0.495 <b>0.505</b>
		Self-BLEU	0.256 <b>0.268</b>
		PAC-Score	0.699 <b>0.703</b>
		LongCLIP-B Score	<b>0.507</b> 0.493
		LongCLIP-L Score	<b>0.531</b> 0.469
		· VLM-as-a-Judge Evaluation Average	<b>4.080</b> 4.033
		Factuality	4.433 <b>4.445</b>
		Informativeness	<b>4.200</b> 4.166
		Succinctness	4.108 <b>4.146</b>
		Diversity	<b>3.578</b> 3.375
		· Sighted General Group Average	<b>3.983</b> 3.962
		Factuality	<b>4.128</b> 4.093
		Informativeness	<b>4.367</b> 4.032
		Succinctness	3.556 <b>4.040</b>
		Diversity	<b>3.879</b> 3.685
		· Sighted Educator Group Average	3.22 <b>3.35</b>
		Factuality	<b>3.35</b> 3.30
		Informativeness	3.43 3.43
		Succinctness	2.78 <b>3.53</b>
		Diversity	<b>3.18</b> 3.08
		Usefulness to BLV	3.35 <b>3.40</b>
		· BLV Educator Group Average	2.98 <b>3.17</b>
		Succinctness	2.43 <b>2.55</b>
		Diversity	<b>3.23</b> 3.15
		Usefulness, Summary	2.95 <b>3.33</b>
		Usefulness, Multiple-choice Questions	3.20 <b>3.28</b>
		Usefulness, Open-ended Questions	2.88 <b>3.13</b>
		Nature of Context	2.98 3.17

Table 11: The full evaluation on descriptions by GPT. Nature of Context values are not in bold because it is a categorical variable.

Experiment ID		Assessments for	
Description	Generators	Metrics	
			Desc Desc <sub>++</sub>
Experiment 1b QWEN2-VL-72B- INSTRUCT vs. QWEN2-VL-72B- INSTRUCT		CLIP Score	0.451 <b>0.549</b>
		SigLIP Score	0.911 <b>0.932</b>
		BLIP-2 Retrieval Score	0.494 <b>0.506</b>
		Self-BLEU	0.260 <b>0.274</b>
		PAC-Score	0.709 <b>0.716</b>
		LongCLIP-B	0.443 <b>0.610</b>
		LongCLIP-L	0.468 <b>0.532</b>
		· VLM-as-a-Judge Evaluation Average	<b>4.094</b> 3.916
		Factuality	<b>4.483</b> 4.428
		Informativeness	<b>4.239</b> 3.952
		Succinctness	4.026 <b>4.072</b>
		Diversity	<b>3.629</b> 3.210
		· Sighted General Group Average	<b>4.002</b> 3.850
		Factuality	3.982 <b>4.060</b>
		Informativeness	<b>4.233</b> 3.782
		Succinctness	3.889 <b>4.035</b>
		Diversity	<b>3.905</b> 3.523
		· Sighted Educator Group Average	4.01 <b>4.13</b>
		Factuality	4.05 4.05
		Informativeness	<b>4.38</b> 4.13
	Succinctness	3.80 <b>4.48</b>	
	Diversity	3.80 <b>3.83</b>	
	Usefulness to BLV	4.03 <b>4.15</b>	

Table 12: The full evaluation on descriptions by the 72B model. Due to limited recruiting, BLV annotators were not given this set.



Fine-tuning QWEN2-VL-2B-INSTRUCT		Pairwise Assessments for Desc <sup>q2b</sup> vs. Desc <sup>q2b</sup>					
Metrics (Scores) by		Desc <sup>base</sup>	Desc <sup>base</sup>	Desc <sup>sft</sup>	Desc <sup>sft</sup>	Desc <sup>sft+dpo</sup>	Desc <sup>sft+dpo</sup>
CLIP Score		0.442	<b>0.558</b>	0.466	<b>0.534</b>	0.451	<b>0.549</b>
SigLIP Score		0.916	<b>0.941</b>	0.911	<b>0.931</b>	0.914	<b>0.940</b>
BLIP-2 Retrieval Score		0.491	<b>0.509</b>	0.493	<b>0.507</b>	0.491	<b>0.509</b>
Self-BLEU		0.274	<b>0.278</b>	0.285	<b>0.291</b>	0.277	<b>0.281</b>
PAC-Score		0.711	<b>0.718</b>	0.706	<b>0.710</b>	0.712	<b>0.718</b>
LongCLIP-B		0.419	<b>0.581</b>	0.452	<b>0.548</b>	0.445	<b>0.555</b>
LongCLIP-L		0.417	<b>0.583</b>	0.454	<b>0.546</b>	0.459	<b>0.541</b>
· VLM-as-a-Judge Evaluation Average		3.307	<b>3.509</b>	<b>3.732</b>	3.663	3.334	<b>3.519</b>
Factuality		3.426	<b>3.783</b>	3.926	<b>3.974</b>	3.431	<b>3.784</b>
Informativeness		3.394	<b>3.567</b>	<b>3.854</b>	3.715	3.438	<b>3.577</b>
Succinctness		3.346	<b>3.662</b>	3.707	<b>3.774</b>	3.347	<b>3.659</b>
Diversity		<b>3.062</b>	3.025	<b>3.442</b>	3.188	<b>3.118</b>	3.054
· Sighted Educators Group Average		3.91	<b>3.95</b>			4.34	<b>4.49</b>
Factuality		3.95	<b>4.03</b>			4.42	<b>4.66</b>
Informativeness		4.03	<b>4.05</b>			4.39	<b>4.50</b>
Succinctness		<b>3.98</b>	3.90			4.37	<b>4.50</b>
Diversity		3.65	<b>3.80</b>			4.18	<b>4.32</b>
Usefulness to BLV		3.93	<b>3.98</b>			4.34	<b>4.50</b>
· BLV Educators Group Average		<b>3.33</b>	3.25		—	2.62	<b>3.17</b>
Succinctness		<b>3.45</b>	3.33			3.15	<b>3.30</b>
Diversity		<b>3.18</b>	3.10			2.03	<b>2.53</b>
Usefulness, Summary		<b>3.53</b>	3.40			2.88	<b>3.45</b>
Usefulness, Multiple-choice Questions		<b>3.15</b>	3.10			2.88	<b>3.73</b>
Usefulness, Open-ended Questions		3.15	<b>3.21</b>			2.28	<b>3.00</b>
Nature of Context		3.33	3.25			2.50	3.00

Table 13: Evaluation of the 2B model from baseline to SFT to DPO. Note that human evaluation results are unnormalized values on the 5-point Likert scale, so direct comparisons are meaningful only within the pairwise shaded columns. SFT versus SFT samples were not distributed due to limited annotator resources. Nature of Context values are not in bold because it is a categorical variable.

Fine-tuning QWEN2-VL-7B-INSTRUCT		Pairwise Assessments for Desc <sup>7B</sup> vs. Desc <sup>7B</sup>										
Metrics (Scores) by	Desc <sup>base</sup>		Desc <sup>base</sup> <sub>++</sub>		Desc <sup>sft</sup>		Desc <sup>sft</sup> <sub>++</sub>		Desc <sup>sft+dpo</sup>		Desc <sup>sft+dpo</sup> <sub>++</sub>	
	Desc	Desc <sub>++</sub>	Desc	Desc <sub>++</sub>	Desc	Desc <sub>++</sub>	Desc	Desc <sub>++</sub>	Desc	Desc <sub>++</sub>	Desc	Desc <sub>++</sub>
CLIP Score	0.423	<b>0.577</b>	0.411	<b>0.589</b>	0.407	<b>0.593</b>	0.407	<b>0.593</b>	0.407	<b>0.593</b>	0.407	<b>0.593</b>
SigLIP Score	0.922	<b>0.952</b>	0.918	<b>0.944</b>	0.923	<b>0.952</b>	0.923	<b>0.952</b>	0.923	<b>0.952</b>	0.923	<b>0.952</b>
BLIP-2 Retrieval Score	0.490	<b>0.510</b>	0.489	<b>0.511</b>	0.490	<b>0.510</b>	0.490	<b>0.510</b>	0.490	<b>0.510</b>	0.490	<b>0.510</b>
Self-BLEU	0.268	<b>0.274</b>	0.275	<b>0.282</b>	0.268	<b>0.275</b>	0.268	<b>0.275</b>	0.268	<b>0.275</b>	0.268	<b>0.275</b>
PAC-Score	0.713	<b>0.720</b>	0.706	<b>0.714</b>	0.711	<b>0.718</b>	0.711	<b>0.718</b>	0.711	<b>0.718</b>	0.711	<b>0.718</b>
LongCLIP-B	0.419	<b>0.581</b>	0.452	<b>0.589</b>	0.417	<b>0.583</b>	0.417	<b>0.583</b>	0.417	<b>0.583</b>	0.417	<b>0.583</b>
LongCLIP-L	0.417	<b>0.583</b>	0.486	<b>0.514</b>	0.412	<b>0.588</b>	0.412	<b>0.588</b>	0.412	<b>0.588</b>	0.412	<b>0.588</b>
· VLM-as-a-Judge Evaluation Average												
Factuality	<b>3.951</b>	3.652	<b>4.021</b>	3.758	<b>3.948</b>	3.642	<b>4.021</b>	3.758	<b>3.948</b>	3.642	<b>4.021</b>	3.758
Informativeness	<b>4.271</b>	4.157	<b>4.371</b>	4.261	<b>4.289</b>	4.161	<b>4.371</b>	4.261	<b>4.289</b>	4.161	<b>4.371</b>	4.261
Succinctness	<b>4.101</b>	3.645	<b>4.161</b>	3.770	<b>4.100</b>	3.642	<b>4.161</b>	3.770	<b>4.100</b>	3.642	<b>4.161</b>	3.770
Diversity	<b>3.946</b>	3.892	<b>3.974</b>	3.964	<b>3.904</b>	3.858	<b>3.974</b>	3.964	<b>3.904</b>	3.858	<b>3.974</b>	3.964
	<b>3.486</b>	2.913	<b>3.576</b>	3.036	<b>3.498</b>	2.906	<b>3.576</b>	3.036	<b>3.498</b>	2.906	<b>3.576</b>	3.036
· Sighted Educators Group Average												
Factuality	<b>4.37</b>	3.97	<b>4.82</b>	4.56	<b>4.00</b>	3.95	<b>4.37</b>	3.97	<b>4.00</b>	3.95	<b>4.37</b>	3.97
Informativeness	<b>4.67</b>	3.87	<b>4.15</b>	3.87	<b>4.13</b>	3.95	<b>4.67</b>	3.87	<b>4.13</b>	3.95	<b>4.67</b>	3.87
Succinctness	3.95	<b>4.15</b>	3.88	<b>4.00</b>	3.88	<b>4.00</b>	3.88	<b>4.00</b>	3.88	<b>4.00</b>	3.88	<b>4.00</b>
Diversity	<b>4.23</b>	3.64	<b>3.88</b>	3.70	<b>3.88</b>	3.70	<b>4.03</b>	3.95	<b>3.88</b>	3.70	<b>4.03</b>	3.95
Usefulness to BLV	<b>4.37</b>	3.97	<b>4.03</b>	—	<b>4.03</b>	3.71	<b>4.03</b>	—	<b>4.03</b>	3.71	<b>4.03</b>	—
· BLV Educators Group Average	<b>3.87</b>	3.82	<b>3.82</b>	—	<b>3.82</b>	3.71	<b>3.82</b>	—	<b>3.82</b>	3.71	<b>3.82</b>	—
Succinctness	4.30	<b>4.55</b>	4.48	<b>4.65</b>	4.48	<b>4.65</b>	4.48	<b>4.65</b>	4.48	<b>4.65</b>	4.48	<b>4.65</b>
Diversity	4.20	4.20	<b>4.13</b>	3.90	<b>4.13</b>	3.90	<b>4.13</b>	3.90	<b>4.13</b>	3.90	<b>4.13</b>	3.90
Usefulness, Summary	4.15	<b>4.55</b>	4.25	<b>4.35</b>	4.25	<b>4.35</b>	4.25	<b>4.35</b>	4.25	<b>4.35</b>	4.25	<b>4.35</b>
Usefulness, Multiple-choice Questions	<b>4.40</b>	4.20	<b>4.15</b>	3.95	<b>4.15</b>	3.95	<b>4.15</b>	3.95	<b>4.15</b>	3.95	<b>4.15</b>	3.95
Usefulness, Open-ended Questions	3.80	3.80	<b>3.70</b>	3.58	<b>3.70</b>	3.58	<b>3.70</b>	3.58	<b>3.70</b>	3.58	<b>3.70</b>	3.58
Nature of Context	2.35	1.60	2.23	1.85	2.23	1.85	2.23	1.85	2.23	1.85	2.23	1.85

Table 14: Evaluation of the 7B model. Note that human evaluation results are nominal values on the 5-point Likert scale, so direct comparisons are meaningful only within the pairwise shaded columns. As with the 2B case, SFT versus SFT samples were not distributed due to limited annotator resources. Nature of Context values are not in bold because it is a categorical variable.

## F Annotator Demographics and Interviews

### F.1 Demographics

#### F.1.1 BLV Educators

Please refer to Table 17.

#### F.1.2 Sighted Educators

Please refer to Table 18.

## G Prompts

### Prompts for Context Generation

**Generating Desc** You are a helpful expert who is knowledgeable in various fields of academia. You are skilled in reading, interpreting, and understanding academic papers and figures contained therein. You are tasked with elaborating on the given information, which consists of a figure image. Write an informative and explanatory text in one paragraph under 200 words that describes the basic characteristics of the figure and incorporates important information. You may attempt to internally identify implicit points of curiosity for someone who is trying to understand the given figure, and then include explanations for those points in your response. Avoid mere reiteration of the given information as much as possible. You need not specify the origins of various parts of your response. [Optional: Aspect Suffix]

**Generating Desc<sub>++</sub>** You are a helpful expert who is knowledgeable in various fields of academia. You are skilled in reading, interpreting, and understanding academic papers and figures contained therein. You are tasked with elaborating on the given information, which consists of a figure image and several question-answer pairs that have been derived from the figure. Write an informative and explanatory text in one paragraph under 200 words that describes the basic characteristics of the figure and incorporates important information from the question-answer pairs. You may attempt to internally identify implicit points of curiosity for someone who is trying to understand the given figure, and then include explanations for those points in your response. Avoid mere reiteration of the given information as much as possible. You need not specify the origins of various parts of your response. Here is the reference information: [QA\_PAIRS: vqas] [Optional: Aspect Suffix]

### Aspect Suffixes

**Factuality** When generating the diagram description, pay close attention to making it factual. A highly factual text delivers only the facts that are grounded in the diagram.

**Informativeness** When generating the diagram description, pay close attention to making it informative. A highly informative text describes all of the diagram, holistically.

**Succinctness** When generating the diagram description, pay close attention to making it succinct. A highly succinct text is concise and to the point.

**Diversity** When generating the diagram description, pay close attention to making it diverse. A highly diverse text captures a variety of perspectives from the diagram and employs multiple effective ways of getting the diagram message across.

### Prompt for Question-answer Pair Generation

Please generate six question-and-answer pairs based on the provided image to aid in creating a comprehensive context. This context should include all essential details, allowing BLV (Blind and Low Vision) users to rely on the generated text instead of viewing the image (*e.g.*, accessing information audibly). The question-and-answer pairs should cover both the main structure and finer details present in the image.

Experiment ID		Assessments for	
Description Generators	Metrics	Desc <sup>q72bbase</sup>	Desc <sup>q7bdpo</sup> <sub>++</sub>
Experiment 3a QWEN2-VL-72B- INSTRUCT vs. FINE-TUNED QWEN2-VL-7B- INSTRUCT	CLIP Score	0.390	<b>0.610</b>
	SigLIP Score	0.911	<b>0.952</b>
	BLIP-2 Retrieval Score	0.487	<b>0.513</b>
	Self-BLEU	0.260	<b>0.275</b>
	PAC-Score	0.709	<b>0.719</b>
	LongCLIP-B Score	0.388	<b>0.612</b>
	LongCLIP-L Score	0.445	<b>0.555</b>
	· VLM-as-a-Judge Evaluation Average	<b>4.095</b>	3.650
	Factuality	<b>4.477</b>	4.238
	Informativeness	<b>4.262</b>	3.586
	Succinctness	<b>3.990</b>	3.894
	Diversity	<b>3.652</b>	2.880
	· Sighted Educators Group Average	<b>3.21</b>	3.01
	Factuality	<b>3.30</b>	3.28
	Informativeness	<b>3.33</b>	2.95
	Succinctness	2.95	<b>3.18</b>
	Diversity	<b>3.13</b>	2.68
	Usefulness to BLV	<b>3.35</b>	2.98
	· BLV Educators Group Average	3.69	<b>4.33</b>
	Succinctness	3.60	<b>4.55</b>
Diversity	3.60	<b>3.90</b>	
Usefulness, Summary	3.95	<b>4.30</b>	
Usefulness, Multiple-choice Questions	3.70	<b>4.55</b>	
Usefulness, Open-ended Questions	3.70	<b>4.45</b>	
Nature of Context	3.60	4.25	

Table 15: The smaller model outperforms a larger variant across many metrics. It is also important to note that the VLM judgments align better with sighted educators than with BLV educators. Further analysis is found in Section 5. This tendency is especially strong with the pairwise comparison between 72B- and 7B-generated descriptions. Nature of Context values are not in bold because it is a categorical variable.

Experiment ID		Assessments for		
Description	Generators	Metrics	Assessments	
			Desc <sup>q7bbase</sup> Desc <sup>q2bdpo</sup>	
Experiment 3b	QWEN2-VL-7B-INSTRUCT	CLIP Score	0.486 <b>0.514</b>	
		SigLIP Score	0.922 <b>0.940</b>	
		BLIP-2 Retrieval Score	0.500 0.500	
		Self-BLEU	0.268 <b>0.281</b>	
		PAC-Score	0.713 <b>0.718</b>	
		LongCLIP-B Score	0.316 <b>0.684</b>	
		LongCLIP-L Score	<b>0.559</b> 0.441	
		· VLM-as-a-Judge Evaluation Average		<b>3.921</b> 3.545
		Factuality	<b>4.203</b> 3.935	
		Informativeness	<b>4.046</b> 3.592	
vs. FINE-TUNED QWEN2-VL-2B-INSTRUCT		Succinctness	<b>3.942</b> 3.709	
		Diversity	<b>3.493</b> 2.945	
		· Sighted Educators Group Average		<b>4.75</b> 4.44
		Factuality	<b>4.75</b> 4.50	
		Informativeness	<b>4.65</b> 4.38	
		Succinctness	<b>4.88</b> 4.40	
		Diversity	<b>4.80</b> 4.63	
		Usefulness to BLV	<b>4.65</b> 4.28	
		· BLV Educators Group Average		4.13 <b>4.32</b>
		Succinctness	4.05 <b>4.15</b>	
Diversity	4.08 <b>4.15</b>			
Usefulness, Summary	3.85 <b>4.13</b>			
Usefulness, Multiple-choice Questions	4.53 <b>4.58</b>			
Usefulness, Open-ended Questions	4.23 <b>4.35</b>			
Nature of Context	4.08 4.50			

Table 16: The 2B model performs on par with the 7B variant. Again, VLM judgments align better with sighted educators than with BLV educators. Further analysis is found in Section 5. Nature of Context values are not in bold because it is a categorical variable.



ID	Sex	Age	Teaching Experience (years)	Onset Age	AI Use, Generic	AI Use, Accessibility
B1	M	54	28	16	ChatGPT, Gemini	SenseReader
B2	F	46	21	Congenital	ChatGPT	SenseReader
B3	M	47	5	9	ChatGPT, Gemini	SenseReader
B4	M	51	26	14	SeeingAI, ChatGPT, Adot, Perplexity, Adot	SenseReader, NVDA, VoiceOver
B5	M	20	1	Congenital	SeeingAI, ChatGPT	SenseReader, NVDA
B6	M	46	19	—	—	SenseReader
B7	M	44	21	Congenital	Be_My_Eyes, SeeingAI, ChatGPT, Claude	SenseReader, VoiceOver
B8	M	45	19	Congenital	Be_My_Eyes, SeeingAI, ChatGPT	SenseReader, VoiceOver

Table 17: BLV Teachers Information. All the BLV teachers in our study were of blindness level 1, the severest.

ID	Sex	Age	Teaching Experience (years)	AI Use - Generic
S1	M	39	6.5	ChatGPT
S2	M	51	20	ChatGPT, wrtn
S3	M	48	21	ChatGPT
S4	F	40	13	ChatGPT
S5	F	56	33	—
S6	F	49	20	ChatGPT
S7	M	49	20	Gemini
S8	F	49	24	ChatGPT, Claude
S9	M	44	14	—
S10	F	50	20	ChatGPT

Table 18: Sighted Teachers Information.

### Prompts for Reasoning Path Generation

You will be provided with a diagram, along with two descriptions of it. As an expert and experienced educator, you are tasked to examine your descriptions to identify common reasoning paths, such as cause-effect relationships, step-by-step processes, explanations of phenomena, comparisons of contrasts, and dependencies between components.

The identified reasoning paths should be under 25 words. Please provide the reasoning paths that you examined in the following JSON format:

```

{"Context1": {"ReasoningPath": text},
 "Context2": {"ReasoningPath": text}}

```

DO NOT return anything other than the JSON above.

## Prompts for Reasoning QA Generation

You will be provided with a diagram, along with two descriptions of it. As an expert and experienced educator, you are tasked to examine your descriptions to generate reasoning question and answer pairs of five categories such as:

- Causal Reasoning: "Why does [event] happen?"
- Process Reasoning: "What happens after [event]?"
- Conditional Reasoning: "What if [condition] changes?"
- Explanatory Reasoning: "Explain the role of [component] in the process."
- Reverse Reasoning: "Given [outcome], what might have caused it?"

Please provide the reasoning question and answer pairs that you generated in the following JSON format:

```
{
  "Context 1": {
    "Causal": {
      "Question": text,
      "Answer": text
    },
    "Process": {
      "Question": text,
      "Answer": text
    },
    "Conditional": {
      "Question": text,
      "Answer": text
    },
    "Explanatory": {
      "Question": text,
      "Answer": text
    },
    "Reverse": {
      "Question": text,
      "Answer": text
    }
  },
  "Context 2": {
    "Causal": {
      "Question": text,
      "Answer": text
    },
    ..
    "Reverse": {
      "Question": text,
      "Answer": text
    }
  }
}
```

Each generated question should be under 15 words and each corresponding answer should be under 25 words.

DO NOT return anything other than the JSON above.

### Prompt for VLM-as-a-Judge Evaluation of Description Pairs

You will be provided with a diagram, along with two descriptions of it. As an expert and experienced educator, you are tasked to evaluate each description based on the following qualities on a 5-point Likert scale. For each statement, give a score corresponding to how strongly you agree with the given statement: 1 (Strongly Disagree), 2 (Disagree), 3 (Neutral), 4 (Agree), or 5 (Strongly Agree).

- **Diversity:** The description captures a variety of perspectives from the diagram and conveys multiple effective ways of getting the diagram message across.
- **Succinctness:** The description is concise and to the point, avoiding unnecessary details.
- **Factuality:** The description is accurate and reflects solely the information presented in the diagram.
- **Informativeness:** The description covers the diagram holistically, and it effectively conveys the main trends and insights of the diagram.

Please provide your ratings in the following JSON format:

```
{
  Context 1: {
    Diversity: score,
    Succinctness: score,
    Factuality: score,
    Informativeness: score
  },
  Context 2: {
    Diversity: score,
    Succinctness: score,
    Factuality: score,
    Informativeness: score
  }
}
```

DO NOT return anything other than the JSON above.

## Prompt for VLM-as-a-Judge Evaluation of Question-Answer Pairs

### Instruction

You need to rate the quality of the given Question and Answer in relation to a diagram. Specifically, assess whether the Q&A correctly references and interprets the information presented in the diagram. Consider the Q&A's clarity, specificity, and coherence as they pertain to the diagram's content. You must take into account not only the Answer but also whether an appropriate Question has been provided for the given diagram at the same time. The rating scale is as follow:

- very poor: The Q&A is unclear, vague, or incoherent in relation to the diagram. It lacks essential information or misinterprets the diagram's content.
- poor: The Q&A is somewhat unclear or omits important details from the diagram. It requires significant clarification or correction to align with the diagram.
- average: The Q&A is moderately clear and specific. It may need additional details or minor clarifications to fully match the diagram's information.
- good: The Q&A is clear, specific, and mostly well-formed in referencing the diagram. It provides sufficient context to understand how the diagram supports the question and answer.
- excellent: The Q&A is very clear, specific, and well-articulated. It precisely references and fully aligns with the diagram, containing all necessary details and context.

### Output Format

Given the user's diagram, question, and answer, you must:

Provide an assessment that briefly explains the strengths and/or weaknesses of how the Q&A relates to the diagram. Output your rating (one of: very poor, poor, average, good, excellent) by filling in the placeholders below.

```
[  
{  
  "explanation": "[...]",  
  "input_quality": "[very poor/poor/average/good/excellent]"  
},  
...  
]
```

### Notes

- DO NOT return anything else other than the JSON above.
- Number of item in above list should be same as the number of given QA pairs. Also the order for the explanation and input quality should be same as input QA's order

## H Fine-tuning Configurations

Parameter	SFT Config (Qwen2-VL-2B-Instruct)	DPO Config (Qwen2-VL-2B-Instruct)
<b>Script Arguments</b>		
Dataset Name	SIGHTATIONCOMPLETIONS	SIGHTATIONPREFERENCE
<b>Training Configurations</b>		
Output Directory	anonymous	anonymous
Evaluation Strategy	steps	steps
Train Batch Size	1	1
Evaluation Batch Size	1	1
Gradient Accumulation Steps	8	8
Training Epochs	1	1
Save Total Limit	5	5
bfloat16 Enabled	true	true
Evaluation Steps	10	10
Label Names	["labels"]	["labels"]
Load Best Model at End	true	true
Metric for Best Model	eval_loss	eval_loss
Use Liger	true	true
Max Sequence Length	1024	1024
Remove Unused Columns	false	true
Dataset Kwarg	skip_prepare_dataset: true	skip_prepare_dataset: false
Gradient Checkpointing	true	true
Gradient Checkpointing Kwarg	use_reentrant: false	use_reentrant: false
Dataset Num Processors	8	8
Torch Compile	true	—
DDP Find Unused Parameters	—	true
<b>Model Config</b>		
Use PEFT	false	false
Model Path	Qwen/Qwen2-VL-2B-Instruct	Qwen/Qwen2-VL-2B-Instruct
Torch Dtype	bfloat16	bfloat16
Attention Implementation	flash_attention_2	flash_attention_2

Table 19: SFT and DPO configurations for Qwen2-VL-2B-Instruct. Tuning was performed on 4 ×A6000 GPUs.

Parameter	SFT Config (Qwen2-VL-7B-Instruct)	DPO Config (Qwen2-VL-7B-Instruct)
<b>Script Arguments</b>		
Dataset Name	SIGHTATIONCOMPLETIONS	SIGHTATIONPREFERENCE
<b>Training Configurations</b>		
Output Directory	anonymous	anonymous
Evaluation Strategy	steps	steps
Train Batch Size	1	1
Evaluation Batch Size	1	1
Gradient Accumulation Steps	8	8
Training Epochs	1	1
Save Total Limit	5	5
bfloat16 Enabled	true	true
Evaluation Steps	10	10
Label Names	["labels"]	["labels"]
Load Best Model at End	false	false
Metric for Best Model	eval_loss	eval_loss
Use Liger	true	true
Max Sequence Length	1024	1024
Remove Unused Columns	false	true
Dataset Kwarg	skip_prepare_dataset: true	skip_prepare_dataset: false
Gradient Checkpointing	true	true
Gradient Checkpointing Kwarg	use_reentrant: false	use_reentrant: false
Dataset Num Processors	8	8
DDP Find Unused Parameters	true	true
<b>Model Config</b>		
Use PEFT	true	true
Model Path	Qwen/Qwen2-VL-7B-Instruct	Qwen/Qwen2-VL-7B-Instruct
Torch Dtype	bfloat16	bfloat16
Attention Implementation	flash_attention_2	flash_attention_2
LoRA Rank (r)	16	16
LoRA Alpha	16	16
LoRA Dropout	0.1	0.1
LoRA Target Modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj

Table 20: SFT and DPO configurations for Qwen2-VL-7B-Instruct. Tuning was performed on 4 ×A6000 GPUs.

Component	Configuration
<b>Model</b>	BLIP-2 (Salesforce/blip2-itm-vit-g)
<b>GPUs</b>	Text model on CUDA:0, Vision model on CUDA:1
<b>Dataset</b>	SIGHTATIONRETRIEVAL
<b>Loss</b>	InfoNCE (temperature = 0.07)
<b>Batch Size</b>	1 (with gradient accumulation steps = 4)
<b>Epochs</b>	5
<b>Optimizer</b>	AdamW (Text LR: 5e-5, Vision LR: 2e-5)
<b>Gradient Clipping</b>	Max norm = 1.0
<b>Scheduler</b>	Linear warmup (10% of steps)
<b>Frozen Layers</b>	All except: layernorm, projection, encoder layers 10-11 (Vision); layernorm, projection, encoder layers 10-11, crossattention (Text)
<b>Checkpoints</b>	Best and per-epoch saved to anonymized path

Table 21: Training configurations for BLIP-2 image-text retrieval.



## I Guidelines

### Annotation Guidelines for the Sighted General Group (1/4)

Please carefully read the guidelines below and ensure accurate labeling. Your responses are considered high-quality data and can have critical implications for the experiment. Pay special attention to the Caution section.

**Annotation Guidelines** Thank you for contributing to this project. In the following paragraphs, we will walk you through the project description, your tasks, and annotation examples.

**Project** Our project targets the visually impaired. People who are Blind or have Low Vision (BLV) do not always benefit from the latest AI developments in the same way or extent as sighted users. In this pilot study, we would like to first assess exactly how much state-of-the-art models may assist sighted users, so that we may gain insights into (i) what state-of-the-art models can and cannot do and (ii) what modifications might be necessary to alter their assistive information to cater specifically for BLV users.

**Task** Each task you are about to complete consists of:

- 1 image
- 2 image description pairs, each containing two texts

Given these, you are tasked with:

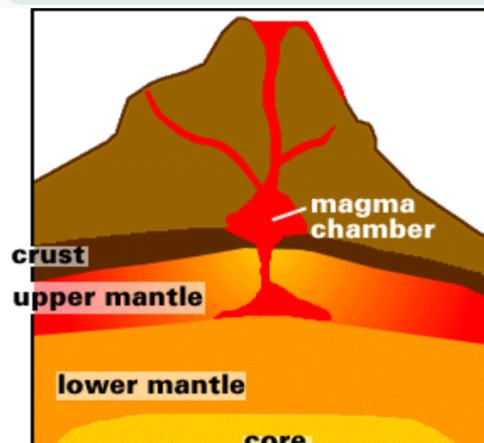
- Selecting the overall “winner” for each pair
- Rating the qualities of each text
- Copying and pasting the best contributing sentence for each text

Please complete the following with respect to the image below and its description pairs 1 and 2:

• For each pair, select the text that is the better overall description of the given image.

• Rate each text (left and right) with respect to the qualities listed.

• Copy and paste the overall best contributing sentence from each text.



## Annotation Guidelines for the Sighted General Group (2/4)

Here is a detailed instruction on each task.

### Selecting the overall winner

- Have a look at both texts. In your opinion, which is the better description of the text?
- This may be based on general impression or your internal criteria.

▼ Description Pair 1

text on the left<sup>[1]</sup>

text on the right<sup>[2]</sup>

The image illustrates a simplified geological cross-section of the Earth, focusing on the structure that contributes to volcanic activity. At the top of the image, there is a depiction of a mountainous terrain, representing the Earth's crust, which is marked with brown, illustrating solidified rock. Below the crust, the sections of the mantle are color-coded: the upper mantle, shown in orange and yellow shades, transitions into the lower mantle, represented with deeper hues of red and orange, reflecting the increasing temperature and pressure as one descends. Central to this image is a red magma chamber, prominently displayed, indicating a reservoir of molten rock. The magma is shown ascending through a fissure, potentially leading to volcanic eruptions at the surface. Surrounding the magma chamber are labels identifying key geological layers—the crust, upper mantle, lower mantle, and the core beneath, which remains unlabelled but implied to be a solid inner core and molten outer core. This diagram effectively conveys the relationship between these geological layers and the process of magma movement that can lead to volcanic activity.

The diagram presents a cross-sectional view of the Earth, illustrating its four main layers, which include the crust, upper mantle, lower mantle, and core. At the top, the crust is depicted as the outermost layer, providing a solid surface layer where we live. Directly beneath the crust lies the upper mantle, characterized by its semi-solid state, which supports tectonic activity. Below this, the lower mantle extends deeper, exhibiting more intense temperatures and pressure due to the weight of the layers above it. At the center of this structure is the core, composed of iron and nickel, contributing to the Earth's magnetic field. In the middle of the illustration, a 'magma chamber' is highlighted; this area represents where molten rock accumulates beneath the Earth's surface, often the source of volcanic eruptions. The striking red lines in the diagram signify the movement of magma as it rises through the layers, typically indicating volcanic activity as it approaches the surface. The image's color palette transitions from darker shades at the crust to brighter hues towards the core, illustrating the increasing temperature and different material compositions of each layer, effectively conveying the dynamic nature of Earth's geology.

## Annotation Guidelines for the Sighted General Group (3/4)

### Rating the qualities of each text

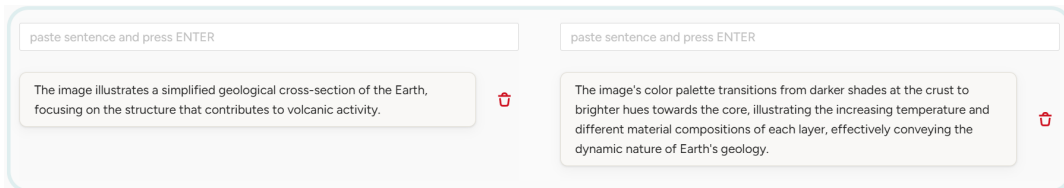
- We break down what makes an image description a “good” one into a few qualities.
- Do not feel overly pressured to justify or rationalize your choice of the overall winner. In fact, you may treat the “overall winner” choice task completely independent of the quality rating task.
- On a scale from 1 to 5, how well does the text exhibit the following qualities?
  - Factuality:
    - \* Does the text contain facts about the image content?
    - \* You may give a low score if the text contains statements that cannot be inferred from the image. These extraneous claims can include knowledge from the “world external to the image”, even though that knowledge itself may be true.
    - \* You may give a low score if the text contains wrong statements about the image.
  - Informativeness:
    - \* Does the text holistically describe the image content and help you become better informed about it?
    - \* You may give a low score if some parts of the image seem “left out” in the text description.
  - Succinctness:
    - \* Does the text describe the image content in a concise yet helpful way?
    - \* Judgments based solely on text length should be avoided.
    - \* Instead, think of the “density” of information contained in the text.
    - \* You may give a low score if the text contains redundant/repeated information, inefficient sentence structures, and/or overly simple vocabulary that tend to make the text feel “sparse”.
  - Diversity:
    - \* Does the text help you understand the image in various ways?
    - \* There may be multiple effective descriptors about one image. There may be multiple perspectives and/or approaches to understanding one image. Do you think the given text addresses these?
    - \* You may give a low score if the text feels too focused on small parts of the image or views the image in an overly specific (possibly contrived) perspective to lay out the description.

☆☆☆☆☆	← Factuality →	☆☆☆☆☆
☆☆☆☆☆	← Informativeness →	☆☆☆☆☆
☆☆☆☆☆	← Succinctness →	☆☆☆☆☆
☆☆☆☆☆	← Diversity →	☆☆☆☆☆

## Annotation Guidelines for the Sighted General Group (4/4)

### Copying and pasting the best sentence for each text

- For each text, drag (highlight) the sentence that has best contributed to each text overall.
- The best sentence does not have to function as a one-sentence summary.
- This should be from the English text, not the Korean translation.
- Paste the sentence into the text field and press enter to submit.
- You will see your submitted best sentence on the display. Press the trash can button (Delete) to change your mind and submit a different sentence.



The image shows a screenshot of a web-based annotation interface. At the top, there are two text input fields, each containing the placeholder text "paste sentence and press ENTER". Below each input field is a light blue rounded rectangle containing a short paragraph of text and a red trash can icon to its right. The first paragraph reads: "The image illustrates a simplified geological cross-section of the Earth, focusing on the structure that contributes to volcanic activity." The second paragraph reads: "The image's color palette transitions from darker shades at the crust to brighter hues towards the core, illustrating the increasing temperature and different material compositions of each layer, effectively conveying the dynamic nature of Earth's geology."

### Caution

- Text length should not be a criterion for your assessment.
- Best sentence should be copied from the English paragraph.

## Evaluation Guidelines for the Sighted Educator Group (1/2)

Thank you again for joining our experiments.

Attached is a spreadsheet containing 40 images, each of which has two descriptions, produced by various AI models in response to our request to generate a context for the input diagram.

As annotators, you are tasked with evaluating the quality of these texts.

**Selecting the Preferred Text** Have a look at the diagram in the “Image” column, along with the two contexts written in the “Context1” and “Context2” columns. In the “Preferred Text” column, enter your choice as 1 or 2.

- This preference may rely on your own personal criteria.

## Evaluation Guidelines for the Sighted Educator Group (2/2)

**Quantitative Assessment** Give a score ranging from 1 to 5 for each quality listed below. For each statement, enter your assessment on a scale from a 5 if you “Strongly Agree” to a 1 if you “Strongly Disagree”.

- **Factuality:** The text delivers only facts that are grounded on the diagram.
  - Even if a piece of knowledge in the text is factual, you may give a low score if that knowledge cannot be inferred from the diagram.
  - Wrong textual descriptions of the diagram content also have low merit.
- **Informativeness:** The text describes all of the diagram, holistically.
  - You may give a low score to a context leaving out parts of the diagram.
- **Succinctness:** The text is concise and to the point.
  - Please assess whether the context conveys an appropriate “density” of information.
  - You may give a low score if a context seems repetitive.
  - Please avoid scoring based on apparent text length.
- **Diversity:** The text captures a variety of perspectives from the diagram and employs multiple effective ways of getting the diagram message across.
  - There may be multiple different ways to understand a diagram. Please assess whether these ways have been put together in the given context.
- **Usefulness:** The text is helpful to BLV.
  - As an experienced educator for learners with visual impairments, please evaluate how useful and helpful the text would be.

**Qualitative Assessment** In the “Reason” column, please justify your preference choice (*i.e.*, the 1 or 2 selection) with a brief explanation. Simple comments, as long as they tell us the textual quality your choice was based on, may still prove helpful for our research. See examples below.

- “contains various descriptions of ants”
- “written more logically”
- “Context 1 contains a more realistic definition of a food web.”
- “2 is more concise.”
- “Context 1 lacks a description of the artery.”
- “While both texts faithfully address the rotation and revolution movements of the Earth, Context 1 describes in-depth how they manifest as different natural phenomena on the planet.”

## Evaluation Guidelines for the BLV Educator Group

Thank you again for joining our experiments.

Attached is a spreadsheet containing 80 images, each with a description that has been produced by various AI models in response to our request to generate a context for the input diagram.

The spreadsheet comes with 81 rows and 8 columns. The table headers are as follows: Image, Context, Succinctness, Diversity, Usefulness (Summary), Usefulness (Multiple-choice Questions), Usefulness (Open-ended Questions), and Nature of Context. Apart from the header row, the remaining 80 rows each contain 1 image and 1 description.

As annotators, you are tasked with evaluating these texts, based on the text alone.

**Quantitative Assessment** Give a score ranging from 1 to 5 for each quality listed below. For each statement, on a scale from a 5 if you “Strongly Agree” to a 1 if you “Strongly Disagree”.

- **Succinctness:** The text is concise and to the point.
  - Please assess whether the context conveys an appropriate “density” of information.
  - You may give a low score if a context seems repetitive.
  - Please avoid scoring based on apparent text length.
- **Diversity:** The text captures a variety of perspectives from the diagram and employs multiple effective ways of getting the diagram message across.
  - There may be multiple different ways to understand a diagram. Please assess whether these ways have been put together in the given context.
- **Usefulness (Summary):** The text serves as a good summary.
  - Please assess how well the text helps you formulate an idea of the diagram content.
- **Usefulness (Multiple-choice Questions):** The text would be useful in solving short-answer, multiple-choice questions based on the diagram.
  - Suppose you are to solve short-answer multiple-choice questions that have been constructed from the diagram. How well would the context help you answer these questions?
- **Usefulness (Open-ended Questions):** The text would be useful in solving descriptive, essay type questions based on the diagram.
  - Suppose you are to answer an open-ended long-answer question that has been constructed from the diagram. How well would the context help you answer such a question?
- **Nature of Context:** The text is rich in interpretive detail.
  - On a scale of 1 to 5, if the text appears to lay out plain and straightforward facts from the diagram, give a score of 1. If it rather contains interpretive descriptions, and/or reasoned explanations, give a score of 5.