

ImageInWords: Unlocking Hyper-Detailed Image Descriptions

Roopal Garg¹ Andrea Burns¹ Burcu Karagol Ayan¹
Yonatan Bitton¹ Ceslee Montgomery¹ Yasumasa Onoe¹ Andrew Bunner¹
Ranjay Krishna² Jason Baldridge¹ Radu Soricut¹

¹Google DeepMind, ²University of Washington

Data: <https://github.com/google/imageinwords>

Correspondence: iiw-dataset@google.com

Abstract

Despite the longstanding adage “*an image is worth a thousand words*,” generating accurate hyper-detailed image descriptions remains unsolved. Trained on short web-scraped image-text, vision-language models often generate incomplete descriptions with visual inconsistencies. We address this via a novel data-centric approach with *ImageInWords* (IIW), a carefully designed human-in-the-loop framework for curating hyper-detailed image descriptions. Human evaluations on IIW data show major gains compared to recent datasets (+66%) and GPT-4V (+48%) across comprehensiveness, specificity, hallucinations, and more. We also show that fine-tuning with IIW data improves these metrics by +31% against models trained with prior work, even with only 9k samples. Lastly, we evaluate IIW models with text-to-image generation and vision-language reasoning tasks. Our generated descriptions result in the highest fidelity images, and boost compositional reasoning by up to 6% on ARO, SVO-Probes, and Winoground datasets. We release the IIW-Eval benchmark with human judgement labels, object and image-level annotations from our framework, and existing image caption datasets enriched via IIW-model.

1 Introduction

Today’s state-of-the-art Vision-Language Models (VLMs) are trained using large, noisy web datasets. WebImageText (Radford et al., 2021), ALIGN (Jia et al., 2021), Conceptual Captions (Sharma et al., 2018) and LAION (Schuhmann et al., 2022) rely on alt-text scraped from the internet as an imperfect image caption. Yet alt-text may only mention the photo location (*e.g.* “Europe”), the camera model used (*e.g.* “Canon EOS R6 Mark II”), or is SEO-specific (*e.g.*, “keep calm and carry on”). While data filtering and post-processing can remove noisy text, alt-text ambiguously captures image *content* or *intent* (Wikipedia contributors, 2023a). Therefore, only using image descriptions from the web

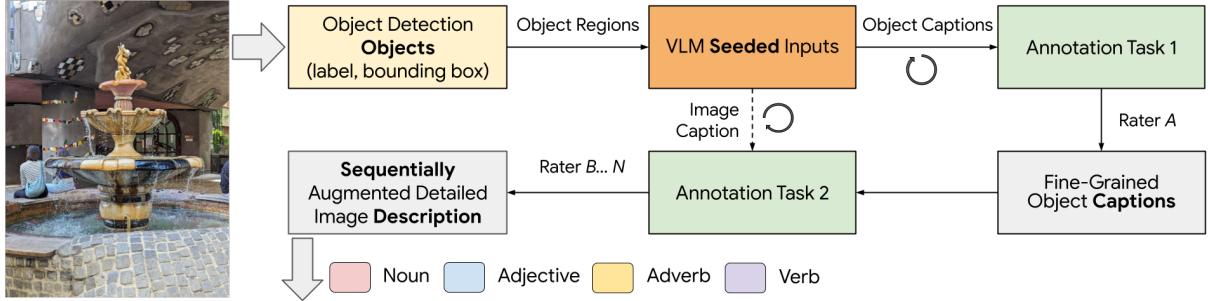
is fundamentally flawed and limits model capabilities (Thrush et al., 2022; Shekhar et al., 2017; Ma et al., 2023; Ray et al., 2023; Hsieh et al., 2024).

To curate better image-text data, recent work has released dense human written (DOCCI (Onoe et al., 2024), DCI (Urbanek et al., 2023)) or model generated caption datasets (PixLore (Bonilla, 2023), DAC (Doveh et al., 2023)). Both have limitations, as using annotators without comprehensive guidelines results in outputs that vary by human attention, bias, and effort (Burghardt et al., 2019; Marshall and Shipman, 2013; Pandey et al., 2022; Ye et al., 2023). In contrast, model-generated captions are cheaper but incomplete and rife with hallucinations (Rohrbach et al., 2019; Dai et al., 2023b).

In this work, we describe *ImageInWords* (IIW), a human-in-the-loop framework for curating hyper-detailed image descriptions, and its resulting annotations. IIW combines the irreplaceable quality of human annotators with seeded metadata from machine generations. First, a VLM generates granular captions for each object in the image to seed our human annotation process, where crowd workers augment and fix the object-level captions to make them richer and hallucination free.

Next, at the image-level, a VLM generates a global caption to seed the final image *description*. Crowd workers consume the image-level seed caption *and* object-level human annotations to fill in contextual gaps. We design guidelines to attend to concepts beyond objects, such as visual perspective, spatial arrangement, and human object interactions. To ensure quality, multiple annotators iterate on a sample sequentially and we also incorporate active learning to produce better VLM seeds (Fig. 1).

With this process, we construct the IIW dataset of 9018 hyper-detailed image descriptions. We find IIW has richer statistics than prior dense description datasets, with an average of 217.2 tokens, 52.5 nouns, 28 adjectives, 5 adverbs, and 19.1 verbs (Tab. 1). We assess quality with human side-by-



An eye-level, vertically-oriented, three-quarter outdoor shot features the multi-tiered, classical Roman fountain at Hundertwasser House in Vienna, Austria, with a backdrop of three people sitting on its rim, a cobble walk area, concrete supports under an arching roof area with inlaid mosaics, arched glass doors, and a bit of sunlight greenery beyond the roof. Four sections make up the fountain, starting with a large, circular concrete basin as the base. The walls of the basin rise to a sitting height, and the face is covered in cobblestone of gray and dark blue shades that are similar in square and rectangle shapes and sizes, but not uniform. This makes for uneven and wavering mortar joints. The basin takes up much of the bottom third of the scene, and the rim alternates between flat sections capped with flat red-blue stone that slightly overhangs the wall, and flowing sections of cobble that angle up from ground-level and flow over the top of the basin wall. One of these flowing cobblestone sections takes up the bottom right corner of the scene and extends left of the horizontal midpoint before transitioning to a flat section. Light gray-blue water fills the basin to a bit below the bottom of the wall caps. From the center of the basin, a wide, jointed stem of what appears to be black, tan and brown ceramic, rises to support the next level of the fountain, which is a ceramic basin of brown, tan, and black segments. Water completely fills this section, and many small streams spill over the sides and fall to the basin below. A shorter, tan hourglass-shaped stem comes up from the center of the second basin to support the third basin of a bronze concrete casting. This basin is smaller than the second, and the lip alternates between short flat sections and higher, double scalloped sections. Thin streams of water fall from the flat sections to the basin below. An ornate, vertically ribbed stem rises from the center to a final, smaller pink ceramic basin. A tan-gold concrete-casting of a cherub-style angel sits on top of this basin. A flat section on the back left side of the bottom basin has two people, who appear to be women, sitting on it and facing away and left. The woman on the left has straight, shoulder-length dark hair and wears a long-sleeved top with navy and white vertical stripes, navy pants, and a large blue bag on her back with the strap slung over her right shoulder and left stomach and chest. To her right is a woman with straight dark hair, a dark top, and light denim jeans with a bit of her bare back between the pants and top. The back of someone in a navy shirt is visible as they sit on a flat section of the basin on the right edge of the scene, facing left. Behind the fountain is a cobbled area, with a large gray rectangular column angled back and right behind the second woman. Several narrow strips of colorful mosaic tiles run gently, curving in a horizontal fashion across the two visible faces of the pillar. Behind the pillar is a gray wall with arched doors on the right and dark brown frames and panels. The arched gray ceiling above has many patches of inlaid mosaic tiles of white, black, gray, and bronze. A bit of sunlight comes in from the right through a large opening, and some green vegetation is visible behind a short wall with an iron vertical railing on top.

Figure 1: ImageInWords Seeded Annotation Framework. Humans enrich and refine outputs *sequentially*, building on prior human or machine inputs. Human annotation starts with fine-grained object captions in Task 1, which are used to compose image-level descriptions in Task 2. VLMs are updated in an active learning loop to produce better object and image-level seeds as annotated data becomes available. UI screenshots are in Appendix B.4.

side (SxS) comparisons to human-written datasets (DCI, DOCCI) and GPT-4V. Our descriptions are rated as more comprehensive, specific, human-like, with fewer hallucinations and better leading sentences at an average of +66% (DCI, DOCCI) and +48% (GPT-4V). We then fine-tune with IIW data and evaluate generated descriptions with the same SxS rubric: IIW model outputs are better by +31% compared to models fine-tuned on prior work.

To better understand IIW models, we also perform text-to-image generation and vision-language reasoning experiments. Images generated with our model’s descriptions are considered a closer reconstruction to the original image than when using other models. For vision-language compositionality, we replace images from ARO (Yuksekgonul et al., 2023), SVO-Probes (Hendricks and Nematzadeh, 2021) and Winoground (Thrush et al., 2022) datasets with generated descriptions. IIW model descriptions help to better reason over attributes, relations, and word order compared to LLaVA-v1.5 and InstructBLIP descriptions.

In summary, our contributions include:

- A human-in-the-loop annotation framework with extensive guidelines, iterative refinement, and VLM active learning that results in state-of-the-art hyper-detailed image descriptions.
- Human SxS on comprehensiveness, specificity,

hallucinations, human-likeness, and tldr-quality. Across these metrics, IIW data is better than recent DCI and DOCCI datasets by +66% and +48% better than GPT-4v, and +31% better when used for fine-tuning than DCI and DOCCI.

- IIW model evaluations with text-to-image generation and vision-language compositional reasoning tasks to complement human SxS. IIW model descriptions generate images most similar to the original image (ranked 1st) and improve distinguishing true image-text pairs given attribute, relation, or word order differences by up to 6%.
- An open source IIW-Eval benchmark of human and model annotations over 2.6k images and their image descriptions, and 1.9k object descriptions. We also release human SxS labels between IIW, DCI, and DOCCI for comparison in future work.

2 Related Work

Image *captioning* has been studied for years, starting with CNN and LSTM encoder-decoder frameworks for generic captions (Vinyals et al., 2015; Anderson et al., 2018), to the more recent Transformer-based VLMs for more difficult captions (Chen et al., 2023b; Li et al., 2023) (e.g., VizWiz (Gurari et al., 2020), NoCaps (Agrawal et al., 2019), TextCaps (Sidorov et al., 2020)). These datasets and many others contain captions with 15 words or

Dataset	Sample Count	Tokens	Tokens	Sentences	NN	ADJ	ADV	VB
		/ Sentence		/ Description				
SVP (Krause et al., 2017)	19,561	11.9	68.5	5.7	17.1	6.7	1.1	5.0
LocNar (Pont-Tuset et al., 2020)	873,107	15.7	41.0	2.6	10.7	1.6	0.4	3.5
DCI ^{extra} ¹ (Urbanek et al., 2023)	7,805	15.8	148.0	9.3	35.3	16.3	3.6	10.5
DOCCI (Onoe et al., 2024)	14,647	19.2	135.7	7.1	34.0	16.6	2.7	9.6
IIW (ours)	9,018	22.1	217.2	9.8	52.5	28.0	5.0	19.1

Table 1: Dataset Statistics Comparing ImageInWords (IIW) to Prior Work. We include the number of descriptions and the average number of tokens, sentences, nouns (NN), adjectives (ADJ), adverbs (ADV), and verbs (VB).

fewer (Desai et al., 2021; Young et al., 2014; Lin et al., 2015; Mao et al., 2016; Plummer et al., 2015; Kazemzadeh et al., 2014; Krishna et al., 2016; Plummer et al., 2015) and may differ by caption grounding level (*e.g.* whole image or region-level captions) or image domain (*e.g.* images taken by people who are blind or images capturing text).

However, few *dense* image *description* datasets exist. PixLore (Bonilla, 2023) used multiple vision-language datasets to generate verbose captions with BLIP-2 (Li et al., 2023). DAC (Doveh et al., 2023) uses a machine-generated approach: pretrained LLMs expand the original image caption and pre-trained VLMs generate captions over smaller image regions. The resulting descriptions are used to fine-tune a VLM model for better compositional reasoning. While model-only approaches are cost effective and avoid the challenges of designing annotation instructions, they risk introducing hallucinations and systematic biases.

DOCCI (Onoe et al., 2024) collects image descriptions with only crowd workers, which we later show can be considerably improved. Closest to IIW is DCI (Urbanek et al., 2023), which uses human annotators to reach denser descriptions. DCI uses the SAM (Kirillov et al., 2023) object detector to generate smaller regions to be described and then composes them into an overall description.

DCI’s available annotations and metadata can be concatenated with additional text to reach 1k+ length. However, filler text and image labels are used to reach this length, and repeated or highly overlapping sentences are often present. As a result, we use their “extra_caption” field for fair comparison as it is the only coherent description available. In contrast to DCI, we also allow crowd workers to update or correct every component of the seeded information. IIW output is then sequentially refined over multiple annotation rounds to produce a single coherent annotation. In comparison to DCI’s “extra_caption” annotation, we collect significantly

better descriptions, as reflected in Tab. 1 statistics.

3 ImageInWords Dataset Collection

The IIW dataset is composed of 9018 (Train: 8573, Test: 445) images that are sampled from a WebLI (Chen et al., 2023b) like dataset and human annotated. Details on the human annotator pool are provided in Appendix B.1. In 3.1, we briefly review our foundational guidelines for crowd workers. Annotation methodology and the types of image-text annotations we collect are described in 3.2 and 3.3.

3.1 Annotation Guidelines

We compile an extensive set of guidelines for human annotators and iterate over them with multiple pilot rounds. Appendix A contains the complete set of guidelines due to space. Annotators are asked to only include details that can be deduced from visual cues, erring on the side of higher precision. To compose coherent descriptions, unnecessary fragmentation of sentences and the use of filler phrases like “*in this image*,” “*we can see*,” and “*there is a*,” should be avoided since they add no visual detail.

While describing the overall image, we instruct annotators to start with a newspaper style TLDR (Too Long Didn’t Read; meant to serve as a succinct summary). Objects should be described in the order of their saliency, noting objects and relationships in a well organized manner. Descriptions should include the overall setting, background, and style, considering the camera angle, overall composition, and rendered text. We also ask to pay special attention to people, apparel, art pieces, locale-specific, and unique attributes with the following as example features: function, shape, size, color, design, pattern, texture, material, condition, opacity, orientation, location, relationship to other components/objects, and text written on objects.

3.2 Annotation Methodology

This section describes the *seeded, sequential* process employed in annotating the IIW dataset. We

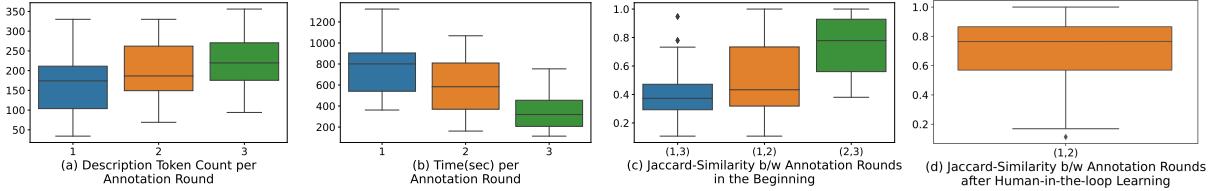


Figure 2: Effects of Sequential Annotation: Over annotation rounds, (a) token count goes up as (b) time spent goes down with (c) higher agreement, measured by Jaccard Similarity ([Wikipedia contributors, 2024](#)). (d) Over time with a constant human annotator pool, each learns from the other via an *implicit feedback loop* and a high agreement rate in round (1,2) can now be observed as was previously only seen in round (2,3) in (c).

highlight that IIW data is meant for supervised fine-tuning rather than pretraining. As a result, our goal was to annotate a small-scale, high quality dataset. Still, we designed the human-in-the-loop process to be as efficient and flexible as possible. The number of sequential annotators and the presence of Task 1 can be adjusted as time and budget permit.

Seeded Annotation Describing images in detail is highly subjective and complicated. To expedite human annotation, we use PaLI-3 5B outputs to seed the annotation process instead of crowd workers starting from scratch. While VLMs have improved in their ability to capture image details, attempts to generate a consistent rich output still fall prey to hallucinations and recall issues. Our human annotation pipeline ensures that VLM hallucinations can be corrected and missing details filled in.

An initial machine generated caption and high precision, domain specific metadata (*e.g.*, art style or title of a painting) provide a minimal quality and coverage guarantee. As data is collected, the VLMs used for seeding are updated to produce better quality descriptions in an active learning loop (reflected with loops in Figure 1). After batches of 1k samples are annotated, we retrain (*i.e.*, refine-tune) the PaLI-3 5B models with all available annotations (for both Task 1 and Task 2).

We find that these updates significantly improve the baseline model, with early batches shifting PaLI captions from an average of 15 to 150+ words with as few as 3k samples. We do not yet perform specialized sampling for active learning due to the large performance gap between the ImageInWords human annotations and ImageInWords model (as later shown in Tab. 8). However, this could be incorporated in the future if performance saturates.

Sequential Augmentation We further improve framework efficiency with sequential description augmentations. Humans augment a previous crowd worker’s and/or VLM’s outputs instead of starting from scratch. After the first augmentation, both the

machine-generated seed and prior human annotation are provided. The following annotators do not know which is model output versus human written, which can mitigate preference to model outputs.

During the annotation process, it is far more effective in *time* and *quality* to read and augment image descriptions: in Fig. 2 we see that if annotations were done in parallel, we would have 3 competing outputs per image, each with their own style, perspective, and weaknesses, with each containing ~ 170 words and taking ~ 800 seconds. Whereas, in the sequential process, we get a single all-inclusive description that has been verified and augmented by three humans with +20% token count in -30% time. Higher Jaccard similarity over rounds suggests a higher inter-annotator agreement, which also serves as a proxy for quality.

Finally, our framework has an implicit human-to-human learning loop, as each human annotator has the opportunity to read and learn from other perspectives across the annotation rounds, leading to improved individual quality. This is seen in the $\sim 2x$ improved inter-annotator agreement between rounds (1, 2) when comparing (c) and (d) in Fig. 2.

3.3 Annotation Framework

Based on the above guidelines, we present the IIW framework for annotating images across two tasks. The tasks are seeded from VLMs or prior human annotations (Fig. 3), where each can have multiple annotation rounds. Examples are in Appendix B.4.

Task 1: Object-Level Descriptions Similar to Visual Genome ([Krishna et al., 2016](#)), we design this annotation task to capture a (label, bounding box, object description) triplet per salient image object. An object’s label is open vocabulary with no verbosity restrictions, and its description is focused on the object but additionally takes the context of the image into account. The bounding box localizes where the object is in the image (Fig. 3 (left)). To seed the data, we first used an internal object detec-

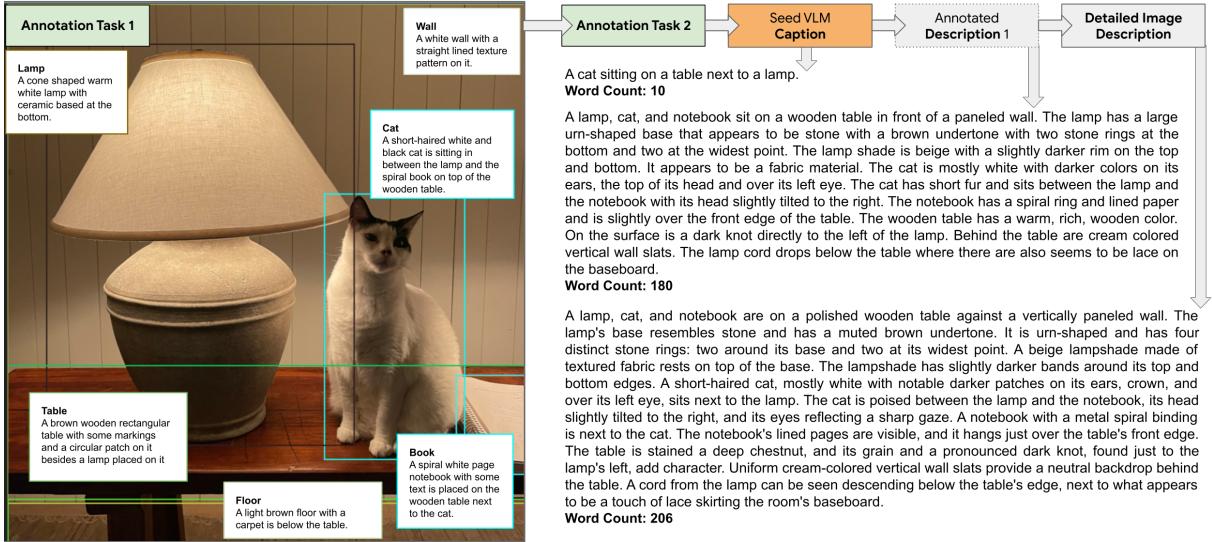


Figure 3: IIW Annotation Tasks. Objects and their attributes are first individually annotated to note the salient objects and focus on coverage of their attributes in Task 1. These outputs, along with a seed VLM caption, are passed to humans to build the initial image-level description. The initial caption is then human augmented and refined in N sequential rounds to attain the final hyper-detailed description in Task 2.

tion (OD) model to obtain a list of (label, bounding box) pairs. Then, object captions are generated by cropping the image to the object bounding box and generating a caption via a periodically fine-tuned PALI-3 5B. Our methodology is agnostic to which VLM, OD (or image-segmentation) model is used.

From the seed list of (label, bounding box, object caption), the annotators are first asked to determine the salient objects and fix the list of (label, bounding box) by editing, removing, adding or merging the object annotations based on their accuracy, importance, and role in the overall image. By limiting the scope to individual objects, annotators can better focus and capture details comprehensively.

Task 2: Image-Level Descriptions Our second annotation task is to form the final hyper-detailed description. Task-1 outputs, optional domain specific metadata (*e.g.*, art style of a painting), and a VLM seed caption are used to hint and help the annotators compose the overall image description.

The bulk of the annotation responsibility falls on the first annotator; note that crowd worker annotation order is randomly assigned per sample and the same annotator is not re-employed for the same sample. This output is then refined and augmented in sequential rounds to mitigate subjectivity and quality drops. Annotators are encouraged to focus on augmentation and only remove things if they are obvious errors, but are free to re-frame information to add new details. We started with 3 annotation rounds and monitored the n-gram Jaccard similarity

between the outputs. Once a 0.8 round-over-round output similarity was achieved, we reduced the numbers of rounds. Optionally, early stopping support could be added to the annotation framework itself to make this instance specific. Over time, we found our similarity threshold can be met between the first two rounds, *i.e.*, (1,2), (Fig. 2) suggesting improved and high individual-annotator quality.

4 IIW Human-Authored Data Eval

To evaluate the IIW annotation framework and resulting human annotations, we start with human SxS evaluations to compare our human annotations to prior work (*e.g.* DCI, DOCCI, GPT-4V). To run a SxS experiment on human-authored description quality, we first need a common pool of human annotated images. For this, we additionally annotate the DCI test set (112) and a comparable number of samples (100) from the DOCCI test set with our IIW annotation framework. We thus have human-authored IIW annotations for direct comparison on images in the DCI and DOCCI datasets, which contribute to our open-source IIW-Eval benchmark.

Our human SxS framework evaluates 5 metrics: Comprehensiveness, Specificity, Hallucinations, quality of the first few line(s) as a TLDR (Too Long Didn't Read; meant to serve as a succinct summary), and Human-Likeness. Comprehensiveness concerns whether a description covers all key information and objects present in an image. Specificity is the degree of detail in which each of

Metric	DCI Test					DOCCI Test				
	DCI		IIW			DOCCI		IIW		
	++	+	-	+	++	++	+	-	+	++
C	3	7	19	30	41	4	6	38	33	19
S	5	3	4	20	68	3	2	8	22	65
H	2	3	48	32	15	0	12	41	34	13
Tldr	3	0	3	20	74	1	4	11	30	54
HL	1	1	14	25	59	1	0	30	46	23

Table 2: Human SxS to Evaluate IIW Human-Authored Data. We report percentages comparing data from prior work with data annotated by the IIW framework on Comprehensiveness (C), Specificity (S), Hallucinations (H), TLDR-quality, and Human-Likeness (HL).

the key objects and details are described in.

We also include TLDR quality as one of our metrics as initial sentences set a precedence for what details to expect, both for the reader and models trained on this data. From a practical perspective, we would like hyper-detailed descriptions to still be useful in a setting that is constrained by input text length; *i.e.*, if we truncate an image description, it should contain the most salient information for vision-language training. While IIW guidelines instruct annotators to include a first sentence which provides an overall summary of the image content, prior work also designed their descriptions to start with either a short caption that summarizes the full image (Urbanek et al., 2023) or have important information covered in earlier sentences (Onoe et al., 2024). As a result, we believe the TLDR metric is reasonable and should be an established practice for hyper-detailed descriptions moving forward.

The evaluation is done on a 5 point scale defined using “substantially better” (++) or “marginally better” (+) ratings on both sides of a “neutral” (-). Higher numbers indicate higher quality across each metric, and our tables report *percentages* for ease of comparison. We emphasize that this is an *extremely challenging* human annotation task, where per image, two text pieces of 100+ words need to be evaluated across 5 metrics in a SxS setting. On average, we observe each comparison takes 15-20 minutes. Details on the annotation setup and UI are in Appendix B.4.

4.1 Human SxS Results

Tab. 2 reports preference percentages for each human-authored test set on our five metrics. Com-

¹We use the extra_caption field of DCI annotations and discuss this in choice in Section 2. All following DCI references refer to the extra_caption description.

paring IIW to DCI and DOCCI, Comprehensiveness is higher by +61% and +42%, Specificity by +80% and +82%, Hallucinations are lower by 42% and 35%, TLDR quality is higher by +91% and +79%, and Human-Likeness improves by +82% and +68%, respectively. This indicates that the IIW human-authored image descriptions on images from DCI and DOCCI are considerably better than those originally published with prior work.

To further quantify the quality of IIW human annotations, we compare with GPT-4V outputs (OpenAI, 2023) in Tab. 3 (right). We use GPT-4V to generate image descriptions on 100 IIW-Eval images. The descriptions are generated with the prompt “Generate a detailed image description” and no other specifications. The results from the Model-Human section of Tab. 3 show that we reach Comprehensiveness (+35%), Specificity (+53%), Hallucination (+59%), TLDR (+70%), and Human-Likeness (+21%) improvements over GPT-4V outputs. Although GPT-4V performs relatively better than the human-authored DCI and DOCCI data when compared to IIW annotations, we assess that considerable future modeling efforts are needed for VLMs to reach IIW human-authored data quality.

5 IIW Model Evaluation

After evaluating IIW human annotations, we turn to quantifying the impact of fine-tuning with IIW data versus fine-tuning with prior work. We fine-tune separate PaLI-3 5B models on DCI, DOCCI and IIW training splits, with their detailed human-authored text as target. Each model is trained with an identical setup (~40 epochs, learning rate 3e-4, batch size 32) and the generic input instruction: “Generate a detailed image description.” More fine-tuning details are provided in Appendix C and D.

As shown in prior work, existing text similarity metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) have been shown to poorly correlate with human judgement as they are heavily dependent on n-gram overlaps, and thus ill-suited for long texts (Kryściński et al., 2019; Caglayan et al., 2020). Prior works DAC, DCI, and DOCCI also are limited by existing image caption metrics, and use LLM summaries of their descriptions or human SxS for evaluation. We report BLEU, ROUGE, CIDEr, BERTScore (Zhang et al., 2020), and BLEURT (Pu et al., 2021) in Appendix D.5 but look to human SxS for more accurate judgements.

We also quantify the richness of the IIW model

Metric	Model Generated												Model-Human							
	LocNar Eval						IIW-Eval						IIW-Eval							
	DCI		IIW		DOCCI		IIW		GPT-4V		IIW		GPT-4V		IIW					
	++	+	-	+	++	++	+	-	+	++	++	+	-	+	++	++	+	++		
Comprehensiveness	7	10	24	32	27	5	22	42	26	5	21	29	36	10	4	3	10	39	29	19
Specificity	6	10	14	24	46	6	14	23	33	24	46	32	12	8	2	6	10	15	35	34
Hallucinations	12	21	43	11	13	9	25	39	21	6	22	29	23	20	6	0	6	29	34	31
TLDR	9	11	9	30	41	6	7	17	42	28	7	15	27	31	20	5	6	8	47	34
Human-Like	11	5	13	32	39	6	12	41	27	14	8	22	60	7	3	6	13	41	27	13

Table 3: Human SxS on Model Predictions. Model Generated compares PaLI-5B fine-tuned with IIW versus prior work DCI and DOCCI and GPT-4V outputs. Model-Human compares GPT-4V model to IIW human-annotations.

outputs via two downstream evaluations which can help us to evaluate IIW model generated descriptions in the absence of better metrics. First, in 5.2, we use generated descriptions from DCI, DOCCI, and IIW fine-tuned models to prompt a Text-to-Image (T2I) model for image reconstruction and evaluate which descriptions result in higher fidelity generated images. Then, in 5.3, we quantitatively show how IIW models can generate descriptions to aid in vision-language reasoning.

5.1 Human SxS Results

Our first evaluation uses the same human SxS setup as in Section 4. We evaluate the IIW, DCI, and DOCCI fine-tuned models on a random sample of LocNar Eval images, which can serve as an unseen test set for each fine-tuning dataset. The results mirror Tab. 2’s human-authored statistics: IIW has gains over (DCI, DOCCI) datasets on Comprehensiveness (+42, +4)%, Specificity (+54, +37)%, TLDR (+51, +57)% and Human-Likeness (+55, +23)% with a relatively small hallucination trade-off (-9, -7)%, largely dominated by marginal rated losses. Overall, compared to DCI and DOCCI, IIW model-generated outputs show a higher average preference from human judgement by +31%.

From Tab. 3 (middle), we see that the IIW PaLI-5B fine-tuned model has clear room for improvement compared to GPT-4V, as expected given its 5B size. It is worth noting that it competes well on the Human-Likeness writing-style metric, and actually excels at learning the TLDR concept, which we built as a distinct feature of our dataset.

5.2 Reconstructing Images with IIW

To complement our SxS analysis, we consider how IIW generated descriptions can empower T2I models to produce more controlled and specific image reconstructions. For this study, we use the PaLI-5B (DCI, DOCCI and IIW) fine-tuned VLMs to

PaLI-ft	Mean Rank ↓				
	1	1-2	1-3	1-4	1-5
DCI	2.05	2.06	1.95	2.00	1.88
DOCCI	1.74	1.79	1.83	1.84	1.86
IIW	1.63	1.69	1.62	1.66	1.66

PaLI-ft	CLIP Image Similarity ↑			
	1	1-2	1-3	1-4
DCI	0.844	0.852	0.855	0.850
DOCCI	0.853	0.862	0.865	0.855
IIW	0.861	0.867	0.870	0.868

Table 4: T2I Reconstruction from Image Descriptions. The original image is compared to images generated from cumulative sentence inputs on relative (Mean Rank) and absolute (CLIP image similarity) metrics.

generate descriptions on 240 images from the LocNar eval set. We then split each image description into sentences as units which are fed as cumulative inputs (*i.e.*, sentence 1, sentence 1-2, sentence 1-3...) to an Imagen model variant (Saharia et al., 2022). By breaking up the description into sentence chunks, we aim to study IIW’s salient description style and also debias our results from description length. We evaluate ~1k generated images across the varied input sentence chunks (over 240 random LocNar images) with a 3-way human ranking evaluation and CLIP similarity between the original and reconstructed image (Radford et al., 2021).

The results in Tab. 4 indicate that IIW’s detailed outputs consistently lead to better T2I reconstruction, with highest mean rank and CLIP similarity regardless of the length of input units. These results confirm that IIW descriptions capture the most visual content with the most detail, and that it is not strictly due to description length, but rather the saliency, comprehensiveness, and specificity in *each* sentence that makes IIW impactful. As input text length is still a limitation in popular VLMs like CLIP, these results provide evidence that using only the first sentence of IIW descriptions can still be

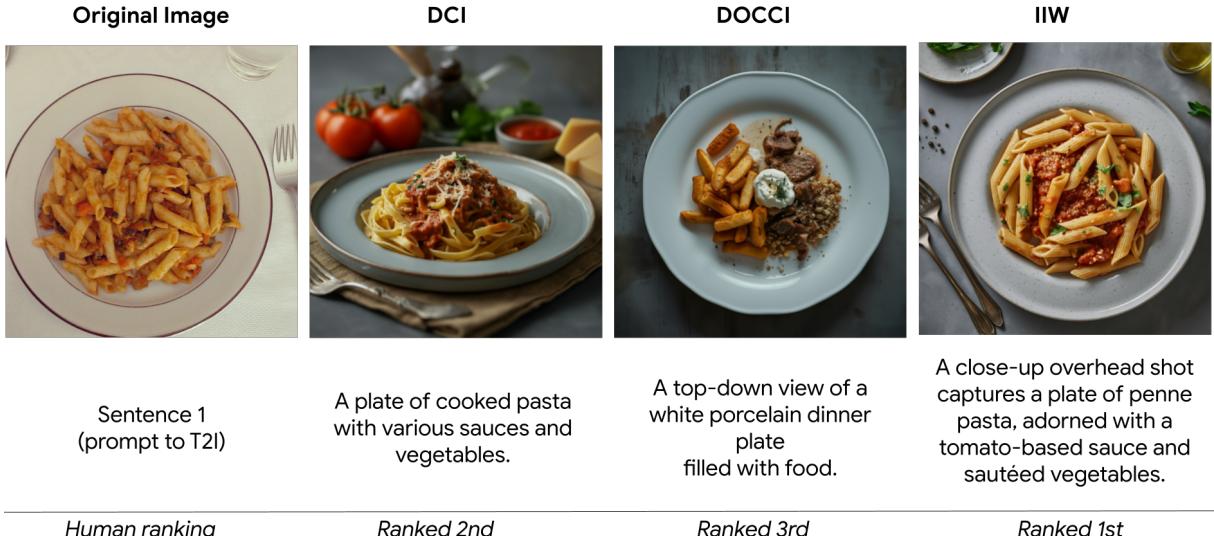


Figure 4: Example T2I Outputs and Human Rankings. We show an example output when the first sentence of the image description from DCI, DOCCI and IIW PaLI-5B fine-tuned models are fed as input to the same T2I model.

useful and performant. In Fig. 4 we show examples of each model’s description’s resulting generated image and associated rank. Additional plots and examples are shared in Appendix D.7.

5.3 Compositional Reasoning with IIW

We look to a second downstream evaluation to quantify the impact of our hyper-detailed image descriptions. Specifically, we use IIW generated descriptions to aid in vision-language compositional reasoning. Probing datasets ARO (Yarom et al., 2023), SVO-Probes (Hendricks and Nematzadeh, 2021), and Winoground (Thrush et al., 2022) modify image captions to no longer match the paired image²: changing visual attributes or relationships, swapping verbs, or shuffling image captions such that they contain the same words but reflect different semantics. This is done to evaluate different types of vision-language reasoning, e.g., visual attribute understanding or verb understanding.

In this experiment we evaluate if IIW descriptions can be used to distinguish the real image caption from the incorrect negative caption in ARO, SVO-Probes, and Winoground datasets using an LLM-only setup. We prompt PaLM2-340B (Anil et al., 2023) to select which of the caption options is true given the image description (see Appendix D.8 for exact input prompts). This essentially replaces the image in these datasets with a generated de-

²SVO-Probes has a negative *image* for each positive image-caption pair. The negative images also have captions, so we use those in our experiments.

Image Desc. Model	ARO VG-A	ARO VG-R	SVO- Probes	Wino- ground
None	56.50	59.94	50.71	49.88
InstructBLIP-7B	83.99	62.73	89.35	65.25
LLaVA-V1.5-7B	84.80	63.71	87.89	63.38
IIW PaLI-3 5B	90.37	66.19	88.66	69.38

Table 5: Vision-Language Compositional Reasoning Accuracy with Image Descriptions. We see if richer IIW descriptions can help distinguish the true matching image caption in ARO (Yuksekgonul et al., 2023), SVO-Probes (Hendricks and Nematzadeh, 2021), and Winoground datasets (Thrush et al., 2022). COCO and Flickr30k Order subsets of ARO are not reported due to a very high language bias baseline of 98%.

scription; the amount the description is able to boost accuracy on these compositional reasoning tests should correlate to the description’s comprehensiveness and specificity. We compare IIW fine-tuned models to two larger (7B) open source models: InstructBLIP-Vicuna-7B (Dai et al., 2023a) and LLaVA-V1.5-7B (Liu et al., 2023) in Tab. 5, with additional models in Appendix D.8.

Our first baseline is the no-image condition (None in the first row of Tab. 5), which simply asks an LLM which image caption is more likely. This serves an important language-bias baseline, and quantifies whether the vision-language compositional reasoning task really requires vision at all. Our results show that SVO-Probes and Winoground have the lowest language bias (baseline performs nearly at random). On the other hand, ARO visual genome attribution and relation subsets are not

IIW-Eval	IIW	#	Annotation Type		
Subset	Source	Images	Task-1	Task-2	SxS
IIW-400	Human	400	1,899	400	200
	Model	—	—	100	—
DCI	Human	112	—	112	112
DOCCI	Human	100	—	100	100
LocNar	Model	1000	—	1000	—
XM3600	Model	1000	—	1000	—
Total		2,612	1,899	2,712	412

Table 6: IIW-Eval Data and Annotation Breakdown.

quite at random baseline; we also note that we do not include the Flickr30k nor COCO order ARO subsets, as the LLM can distinguish the true caption at 98% accuracy without any image description.

When incorporating image descriptions, all models perform significantly better than the language-bias baseline. The IIW model results in the best task performance for ARO Visual Genome Attribution and Relation (VG-A, VG-R) and Winoground, with accuracy gains of nearly 34%, 6%, and 20%, respectively. Moreover, we can further boost performance compared to the InstructBLIP and LLaVA image captions: we improve reasoning accuracy by about 6%, 2%, and 4% compared to the best image description model-based baseline. This reflects the richness of IIW across different parts of speech and comprehensiveness, as more attributes and relationships are captured and can be used to reason about image content. For SVO-Probes, we find smaller differences, with IIW, InstructBLIP, and LLaVA models within ~ 1 point of each other.

6 IIW-Eval Benchmark Release

We release the **IIW-Eval** benchmark (Tab. 6) of human- and model-annotated image descriptions, human SxS results on Human-Human and Model-Human pairs of descriptions. *IIW-400* is a new eval set of 400 images randomly sampled from DOCCI-AAR (Onoe et al., 2024). We re-annotate DCI and DOCCI test samples and enrich two existing datasets with new IIW descriptions: Localized Narratives (LocNar (Pont-Tuset et al., 2020)) and CrossModal-3600 (XM3600 (Thapliyal et al., 2022)). We provide LocNar and XM3600 annotations with significantly improved quality (see statistics in Appendix E). The model generated descriptions may have hallucinations, information recall losses, or non-human like writing style artifacts. By releasing this subset along with human SxS judgements, we encourage the development of new

metrics and evaluation systems to detect them in an automated, scalable manner. It also promotes fair comparison across methods in future work. The dataset is released under a CC BY 4.0 license.

7 Future Work

In future work, robust and effective automatic metrics are needed to evaluate the quality of detailed image descriptions. Next steps may include training model-based metrics or preference models (*i.e.*, autoraters) with human preference data to learn a global quality metric. For additional analysis, we could further break down our current SxS metrics. For example, the human SxS hallucination metric could be broken down to capture fine-grained categories like how many hallucinations are with respect to color, size, or spatial location.

We are working to extend the ImageInWords framework to additional languages and geographically diverse images. In next steps, we note that images need to be sampled globally (across both geographic and cultural identity); this sampling must also be done across different image topics and categories, making equal coverage more complicated. We are currently working on adapting our proposed framework to accommodate locale specific annotators, which are required for cultural specificity. Our continued goal is to make the annotation guidelines holistic, reduce human effort and dependency in the annotation process, and help shift the narrative from captions to descriptions.

8 Conclusion

In this work, we proposed ImageInWords (IIW), a new framework for hyper-detailed image descriptions. Our annotation guidelines and seeded, sequential annotation process lead to human authored descriptions that are strongly preferred over both prior work’s human annotations (+66%) and prior work’s fine-tuned models (+31%). Images reconstructed with IIW generated descriptions were ranked 1st more often, regardless of how much of the image description was used, reflecting higher saliency earlier and better overall quality. Our compositional reasoning evaluation showed IIW generated descriptions to best contain fine-grained visual detail needed to decipher true from false visual attributes and semantics, with accuracy gains of up to 6% over our most performant baselines. Our results collectively demonstrate the quality and utility of IIW image descriptions as state-of-the-art.

Limitations

Finally, we discuss the limitations of our annotation framework and evaluations. In our annotation framework, we define a seeded and sequential annotation process, with both aspects having potential limitations. The quality of the seeded data is of high importance as it will ultimately affect the rest of our human annotation pipeline. Additionally, even with the best possible seeds, they may limit the scope of what our crowd workers write by biasing them towards certain objects or phrases. We employed an active learning loop to iteratively improve the seed generation quality but significant room for improvement still remains. In terms of limitations for the sequential augmentation used, unnecessary time may be spent by annotators if the first annotator output quality is low. By training the annotators through guidelines and feedback and monitoring the initially drafted descriptions, quality can be better ensured so that the framework is as efficient as possible.

With respect to the evaluation of our human annotated data and model generated outputs, we do only perform evaluations on hundreds of samples (as opposed to thousands or more). This is largely due to the cost and time associated with human SxS evaluations for this task, but we note that IIW is rated marginally and substantially better at a much higher rate, which would likely scale to more samples. Our work is also inherently limited by the lack of automated metrics available for long descriptions. We still report standard text similarity metrics in Appendix D.5 and complement them with human SxS, but in future we hope metrics are developed that address the current limitations, as automated metrics can be applied at scale. We note that metric limitations were also faced in prior work, with others opting to use LLM summaries or human SxS for evaluation purposes ([Urbanek et al., 2023](#); [Onoe et al., 2024](#)).

With respect to our trained IIW models, we also note that all results are reported from a single model/run for each evaluation included. In the future, rerunning models with different seeds or aggregating results over different model variants would be beneficial.

While we currently do not plan to open source our models or training set, we do release an evaluation set over images that can serve as a unified benchmark for IIW, recent, and future related work. We also open source the human SxS judgements

and model enriched samples from Localized Narratives and XM3600. We acknowledge that the full annotation framework would take substantial time and effort to rerun from scratch; this is in part due to needing to reproduce the annotation UI and infrastructure for seeding. The framework itself is agnostic to which vision-language models are used for seeding of initial object or image captions, which we hope makes the setup more feasible to reproduce with any open source model of choice. This also becomes increasingly important as new and improved models will continue to be developed, and we'd like our framework to be able to incorporate newer models over time. The number of annotation rounds, annotation volume, and particular set of images can be adjusted to specific use-cases and budget and time constraints.

Lastly, our initial IIW dataset and resulting models are English-only. In the future, we plan to expand our work to have multilingual and multi-cultural coverage over images sampled globally. We also aim to curate images descriptions which are annotated by locale specific annotators to capture regional and cultural nuances, so that we do not strictly have descriptions with a western lens.

Ethics Statement

Our model may have broader societal impact. It may contain unknown biases or stereotypes, or propagate inaccurate or otherwise distorted information. We used a combination of algorithmic methods, manual inspection, and other classifiers for identifying and removing Sensitive Personally Identifiable Information, pornographic, and violence depicting images. Specifically we checked for the presence of: (1) any address, email, or phone number; (2) images with high porn scores; (3) images labeled as portraying abuse; (4) text identified as having certain adult content references. Additionally, we asked human annotators to use an objective and respectful tone while composing the image descriptions. While we made all of these efforts, it is still possible the model may produce some undesirable results.

Additionally, image to text VLMs inherently can have negative impact if the generated image descriptions are inaccurate and/or contain hallucinations. However, our work specifically aims to cover all visual content as comprehensively and accurately as possible to improve data quality and the resulting fine-tuned models.

References

- Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. no-caps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957.
- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. *Preprint*, arXiv:1707.07998.
- Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, and Zhifeng Chen et al. 2023. Palm 2 technical report. *Preprint*, arXiv:2305.10403.
- Diego Bonilla. 2023. Pixlore: A dataset-driven approach to rich image captioning. *Preprint*, arXiv:2312.05349.
- Keith Burghardt, Tad Hogg, and Kristina Lerman. 2019. Quantifying the impact of cognitive biases in question-answering systems. *Preprint*, arXiv:1909.09633.
- Ozan Caglayan, Pranava Madhyastha, and Lucia Specia. 2020. Curious case of language generation evaluation metrics: A cautionary tale. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2322–2328, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, and Geoffrey Hinton. 2022. Pix2seq: A language modeling framework for object detection. *Preprint*, arXiv:2109.10852.
- Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, Daniel Salz, Xi Xiong, Daniel Vlasic, Filip Pavetic, Keran Rong, Tianli Yu, Daniel Keysers, Xiaohua Zhai, and Radu Soricut. 2023a. Pali-3 vision language models: Smaller, faster, stronger. *Preprint*, arXiv:2310.09199.
- Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023b. Pali: A jointly-scaled multilingual language-image model. *Preprint*, arXiv:2209.06794.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023a. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.
- Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. 2023b. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. *Preprint*, arXiv:2210.07688.
- Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. 2021. Redcaps: web-curated image-text data created by the people, for the people. *Preprint*, arXiv:2111.11431.
- Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogério Feris, Shimon Ullman, and Leonid Karlinsky. 2023. Dense and aligned captions (dac) promote compositional reasoning in vl models. *Preprint*, arXiv:2305.19595.
- Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind. *Preprint*, arXiv:2002.08565.
- Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. *Preprint*, arXiv:2106.09141.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.
- Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. 2024. Sugarcreepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in Neural Information Processing Systems*, 36.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. *Preprint*, arXiv:2102.05918.
- Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. *Preprint*, arXiv:2304.02643.
- Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings*

- of the IEEE conference on computer vision and pattern recognition*, pages 317–325.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Preprint*, arXiv:1602.07332.
- Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*.
- 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. *Preprint*, arXiv:2301.12597.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.
- Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. 2023. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10910–10921.
- Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.
- Catherine Marshall and Frank Shipman. 2013. Experiences surveying the crowd: Reflections on methods, participation, and reliability. In *Proceedings of the 3rd Annual ACM Web Science Conference, WebSci 2013*, pages 234–243.
- Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, Su Wang, and Jason Baldridge. 2024. DOCCI: Descriptions of connected and contrasting images. In *ECCV*.
- OpenAI. 2023. Gpt-4v(ision) technical work and authors. <https://cdn.openai.com/contributions/gpt-4v.pdf>, 2023. [Online; accessed 19-February-2024].
- Rahul Pandey, Hemant Purohit, Carlos Castillo, and Valerie L. Shalin. 2022. Modeling and mitigating human annotation errors to design efficient stream processing systems with human-in-the-loop machine learning. *International Journal of Human-Computer Studies*, 160:102772.
- 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.
- Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *ECCV*.
- Amy Pu, Hyung Won Chung, Ankur P Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for mt. In *Proceedings of EMNLP*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Arijit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan A. Plummer, Ranjay Krishna, and Kate Saenko. 2023. Cola: A benchmark for compositional text-to-image retrieval. *Preprint*, arXiv:2305.03689.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2019. Object hallucination in image captioning. *Preprint*, arXiv:1809.02156.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Preprint*, arXiv:2205.11487.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig

- Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. *Laion-5b: An open large-scale dataset for training next generation image-text models*. Preprint, arXiv:2210.08402.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. *Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning*. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.
- Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. *Foil it! find one mismatch between image and language caption*. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.
- Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. *Textcaps: a dataset for image captioning with reading comprehension*. Preprint, arXiv:2003.12462.
- Ashish V. Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. 2022. *Crossmodal-3600: A massively multilingual multimodal evaluation dataset*. Preprint, arXiv:2205.12522.
- Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. *Winoground: Probing vision and language models for visio-linguistic compositionality*. Preprint, arXiv:2204.03162.
- Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. 2023. *A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions*. Preprint, arXiv:2312.08578.
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. *Cider: Consensus-based image description evaluation*. Preprint, arXiv:1411.5726.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. *Show and tell: A neural image caption generator*. Preprint, arXiv:1411.4555.
- Wikipedia contributors. 2023a. Alt attribute — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Alt_attribute&oldid=1189330128. [Online; accessed 15-January-2024].
- Wikipedia contributors. 2023b. Automated readability index — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Automated_readability_index&oldid=1145735758. [Online; accessed 22-February-2024].
- Wikipedia contributors. 2023c. Flesch–kincaid readability tests — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Flesch\T1\textendashKincaid_readability_tests&oldid=1192056958. [Online; accessed 22-February-2024].
- Wikipedia contributors. 2023d. Gunning fog index — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Gunning_fog_index&oldid=1181089308. [Online; accessed 22-February-2024].
- Wikipedia contributors. 2023e. Smog — Wikipedia, the free encyclopedia. <https://en.wikipedia.org/w/index.php?title=SMOG&oldid=1192815974>. [Online; accessed 22-February-2024].
- Wikipedia contributors. 2024. *Jaccard index — Wikipedia, the free encyclopedia*. [Online; accessed 24-January-2024].
- Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roei Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. 2023. *What you see is what you read? improving text-image alignment evaluation*. Preprint, arXiv:2305.10400.
- Andre Ye, Sebastin Santy, Jena D Hwang, Amy X Zhang, and Ranjay Krishna. 2023. Cultural and linguistic diversity improves visual representations. *arXiv preprint arXiv:2310.14356*.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. *From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions*. *Transactions of the Association for Computational Linguistics*, 2:67–78.
- Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. *When and why vision-language models behave like bags-of-words, and what to do about it?* Preprint, arXiv:2210.01936.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. *Bertscore: Evaluating text generation with BERT*. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Annotation Guidelines

We now present the full detailed annotation guidelines used for IIW annotations. Our guidelines state that image descriptions should be composed such that they *paint a vivid mental picture* of an actual image in the mind of someone hearing the description and has their eyes closed. In order to reach this level of detail composed in an articulate manner, we compile an extensive set of annotation

guidelines. We iterated over these guidelines with multiple pilot rounds.

The annotators are asked to *operate as if they are instructing a painter to paint with their words* and only include details that can be deduced from visual cues, erring on the side of higher precision. Unnecessary fragmentation of sentences should be avoided to compose writing in a flowy, coherent style, avoiding the use of *filler phrases* like: “*In this image*,” “*we can see*,” “*there is a*,” “*this is a picture of*,” since they add no visual detail and come at a cost of verbosity.

Objects form the lego-blocks of an image. Interactions and spatial arrangements among them help to form the context of the image. In complex multi-object images with dense settings, noting each and every object independently can become cumbersome and highly dependent on the effort the particular human annotator puts in. To define this better and expect a consistent behavior from the annotation outputs, we introduce the notion of *salient objects*. Key objects without which the image would lose its context and meaning are considered *salient*. This can include individual objects or combinations of them depending on the role they play in the image; consider the following 2 cases as examples:

- Three people in the blurry background of an image, with the scene set inside a coffee shop, who play no concrete role individually can be grouped as *people in the background* instead of 3 individual *people* object annotations.
- Two people in the foreground and in-focus, engaged in a conversation in the same scene. The two individuals are likely the focus of the image and hence worth noting individually in detail as separate objects. This is likely what the photographer was attempting to capture.

While annotating each of these *salient objects* in an image, the annotators should consider the following axes as reference (but not limit themselves to this list), paying special attention to features that make them unique or salient:

- **Function** Purpose of the component or the role it plays in the image
- **Shape** Specific geometric shape, organic, or abstract
- **Size** Large, small, or relative size to other objects

- **Color** Specific color with nuances like solid or variegated
- **Design/Pattern** Solid, flowers, or geometric
- **Texture** Smooth, rough, bumpy, shiny, or dull
- **Material** Wooden, metallic, glass, or plastic
- **Condition** Good, bad, old, new, damaged, or worn out
- **Opacity** Transparent, translucent, or opaque
- **Orientation** Upright, horizontal, inverted, or tilted
- **Location** Foreground, middle ground, or background
- **Relationship to other components** Interactions or relative spatial arrangement
- **Text written on objects** Where and how it's written, font and its attributes, single/multi-line, or multiple pieces of individual text

Humans typically associate a set of default features to objects. Consider the following examples:

- *Car* by default is assumed to have 4 of each: tires, door, windows and 1 of each: trunk, hood, steering wheel, roof. Mentioning them separately might not be that useful as it adds no specific visual detail that we did not already know as the norm. Now, if the car is a *coupe*, has a missing window, or contains a door painted with a different color than the overall color, *i.e.*, making it a unique feature, then that would be worth mentioning in the description since it holds specific added visual value.
- *The Golden Gate Bridge* by default is orange. That being said, it does not hurt to include extra detail depending on the use-case. If the annotators do not recognize the bridge as a famous well known entity, then it would make sense to include the color and additional attributes.

When composing the overall image description, start with a newspaper style *tl;dr* sentence that paints a very clear high level picture. Describe the *objects* in order of their *saliency* while noting the description of individual objects and relationships in a coherent manner. Include the overall setting, background, style, and consider:

- **Overall composition** Arrangement of the elements in the image, focal point, balanced, or asymmetrical
- **Lighting** Natural or artificial, light source
- **Color palette** Colors or how they interact with each other
- **Texture** Smooth or rough, shiny or dull
- **Depth of field** Entire image or only a portion of it is in focus, what effect this has on the overall composition
- **Subject matter** Main subject of the image, other elements that are present, how they relate to the subject matter
- **Mood or feeling** Overall mood or feeling of the image

Camera angle (*i.e.*, the position of the camera in relation to the subject) is crucial, as this sets a precedence for what level and kind of information to expect. The choice of camera angle can have a significant impact on the mood and meaning of a photograph. Different camera angles can be used to create different effects and convey different messages, *e.g.*, details about a close-up are different from those of a wide angle shot. Examples of camera angles (see Figure 5):

- **Eye level:** The camera is positioned at the same level as the subject's eyes. This is the most natural and neutral camera angle.
- **High angle:** The camera is positioned above the subject. This angle can make the subject appear smaller, weaker, or less important.
- **Low angle:** The camera is positioned below the subject, anywhere below the eye line, looking up. This angle can make the subject appear larger, stronger, or more important. Sometimes, it is even directly below the subject's feet.
- **Ground level:** The camera is positioned at the ground level. This angle captures what is in the frame at ground level, that is, the feet, or maybe the character lying on the ground.
- **Dutch tilt:** The camera is tilted on its axis. This angle can be used to create a sense of unease or disorientation.

- **Bird's-eye view:** The camera is positioned directly above the subject. This angle can be used to show the subject's relationship to their surroundings.
- **Worm's-eye view:** The camera is positioned directly below the subject. This angle can be used to create a sense of awe or wonder.
- **Top-down view or Overhead shot:** The camera is above the subject and you're taking the photograph downwards from straight above, and not at any kind of angle. It is typically closer to the subject than a bird's eye view (see Figure 5 for comparison).

Some other terms that are sometimes used to describe camera angles and depths:

- **Close-up:** A close-up is a photograph that is taken from a very small distance. Close-ups can be used to show details that would not be visible from a further distance.
- **Medium shot:** A medium shot is a photograph that shows the subject from the waist up or from the knees up. Medium shots are often used to show the subject's body language and facial expressions.
- **Long shot:** A long shot is a photograph that shows the subject from a distance. Long shots can be used to show the subject's relationship to their surroundings.
- **Full shot:** A full shot is a photograph that shows the subject's entire body. Full shots are often used to show the subject's height and stature.
- **Over-the-shoulder shot:** An over-the-shoulder shot is a photograph that is taken from behind one person's shoulder, showing the other person in the foreground. Over-the-shoulder shots are often used to create a sense of intimacy or connection between the two people.
- **Point-of-view shot:** A point-of-view shot is a photograph that is taken from the perspective of the subject. Point-of-view shots can be used to create a sense of immersion in the scene.

When **text** is present, include detail such as whether the text is in a single line or spread along

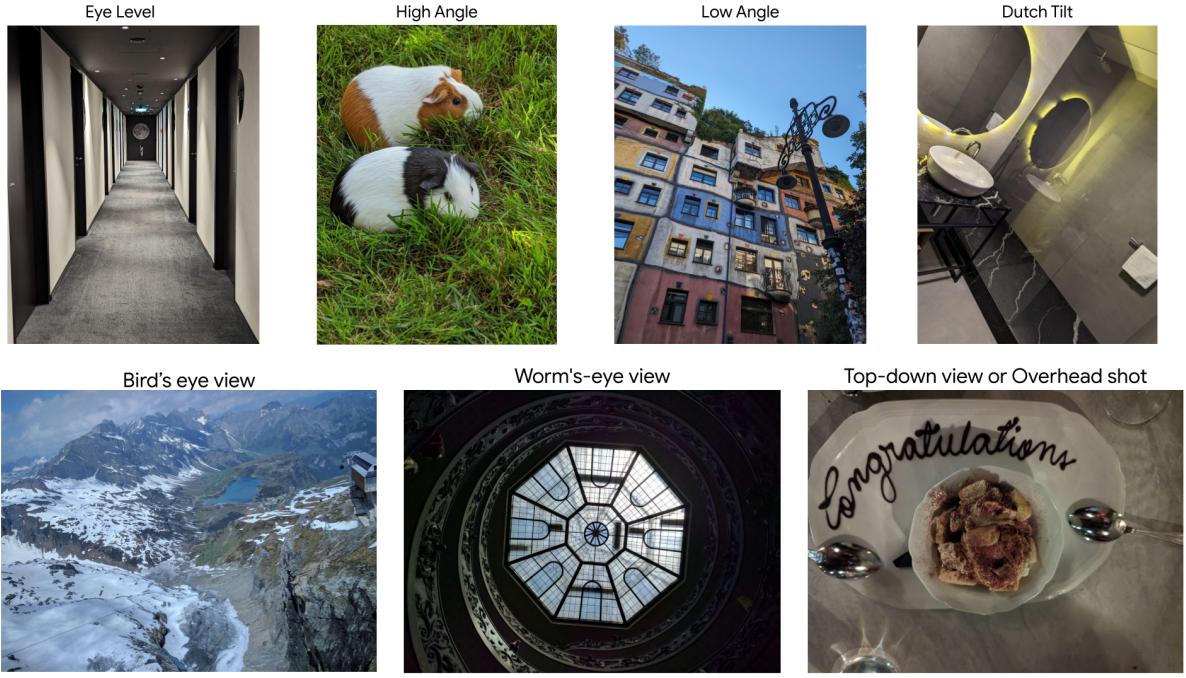


Figure 5: Camera Angles to Consider when Annotating Images. These are important to set a precedence on the level and kind of information to expect in the image description.

multiple lines, if text is in multiple lines whether there is mutual alignment, the features of the *font* such as size, style, color, and orientation (*e.g.*, vertical, horizontal, arched), *casing* (*e.g.*, lower, upper, mixed), and attributes like italics, underlined, bold, written in quotes, clearly visible or blurred. Describe the words if they are written.

If text is written in multiple lines, we should:

- Quote them as individual units that exist on the same line
- Mention its mutual alignment using references like vertically stacked, aligned to the left, *etc.*

For example, in Figure 6, the phrase (“Juice,” “ACROSS THE,” “Universe”) has words “Juice” and “Universe” as capitalized while the phrase “ACROSS THE” is all uppercase, and components are aligned along a diagonal. Information on the font color, type, and shadow effect should be included. As another example from the same image, the phrase (“FREE,” “ARCADE,” “GAMES”) are all upper-cased, vertically stacked and centrally aligned.

If you have a good idea of the font family and are confident, that would be valuable to note.

When **people** are present, special notes should be kept in mind to mitigate different types of bias.

The tone should be *respectful* to the subject and not make assumptions or try to guess their gender, identity, ancestry, where they are from, sexuality, religion, *etc.* We emphasize that the descriptions should be noted in objective, neutral and fair language for related attributes and focus solely on the visual aspects. Consider the following axes with respect to attributes here:

- How much of their body is visible
- Whether the face is fully visible
- Whether they are facing the camera or looking somewhere else
- Where and what they are looking at
- What the person is doing (standing, posing, sitting, running, playing a sport)
- What they are wearing. For each piece, note the clothing item name (dress, pants, short, gloves, shoes), color, pattern (plain, striped), length (if applicable)
- What they are carrying, details about that object (bag, purse, camera)
- Whether they are using any assistance device (wheelchair, cane)
- Whether they have any unique features like marks, tattoos, scars on their body that are



Figure 6: An Example where Quoting *Text* in a Detailed Manner can Enable Precise Reconstruction. The word-casing and alignment attributes of the multi-line phrase (“Juice,” “ACROSS THE,” “Universe”) has words “Juice” and “Universe” as capitalized while the phrase “ACROSS THE” is all upper-cased and all components are aligned along a diagonal. Information on the font color, type, shadow effect should be included. For the phrase (“FREE,” “ARCADE,” “GAMES”) all words are upper-cased, vertically stacked, and centrally aligned.

visible. If applicable, note the respective positions on their body where each is present

- For professions with known gender biases like “nurse,” “doctor,” or “construction worker,” explicitly include the gender (if clearly deducible) and do not operate under the assumption that one gender is more common in that profession.

For any *apparel*, the descriptions should focus on overall style, unique details, silhouette of the garment, how it fits, fabric, color, shades, and tone of the garment. If the branding is visually visible, it should be included while attributes like size should be skipped unless visually verifiable.

Where applicable use *locale specific names* of objects like *clothing* (e.g., sherwani, kurta, kimono, saree), *food* (e.g., shawarma, dosa, paneer tikka) etc. The aim is to capture the locale specific vocabulary so the downstream models can pick them up

instead of using generic abstract terms.

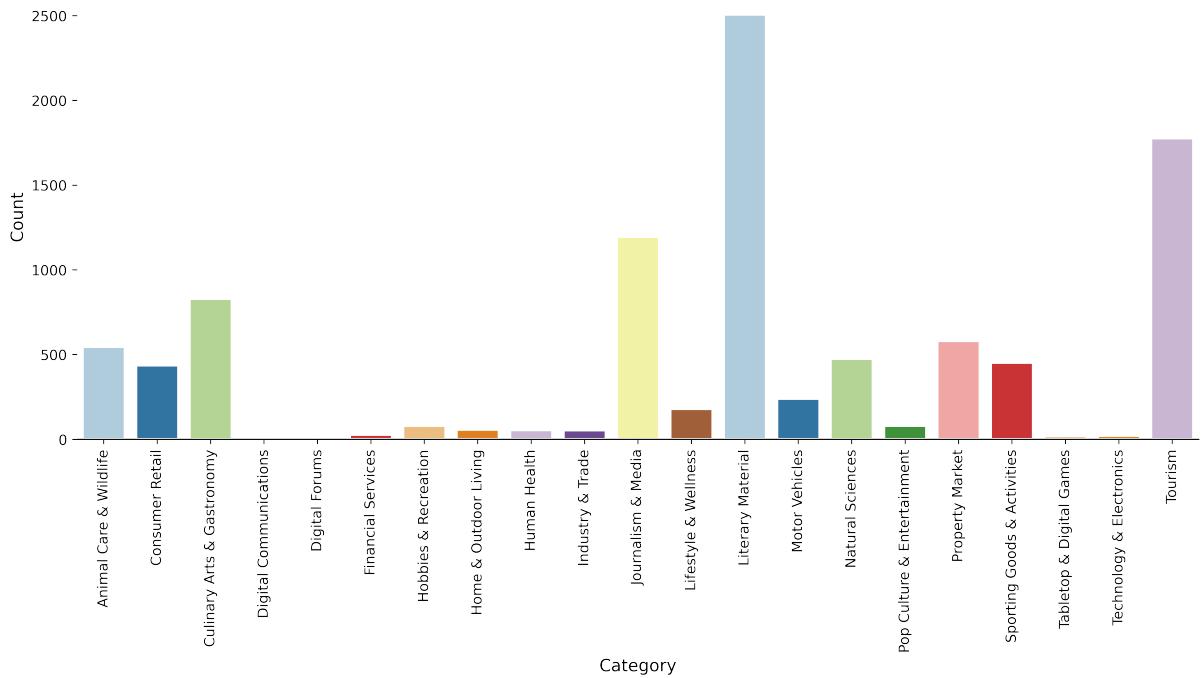
For *art pieces*, include art styles, time periods, mediums, moods, viewpoints, subject matters, cultures as much as possible from the visual cues.

B Dataset Collection

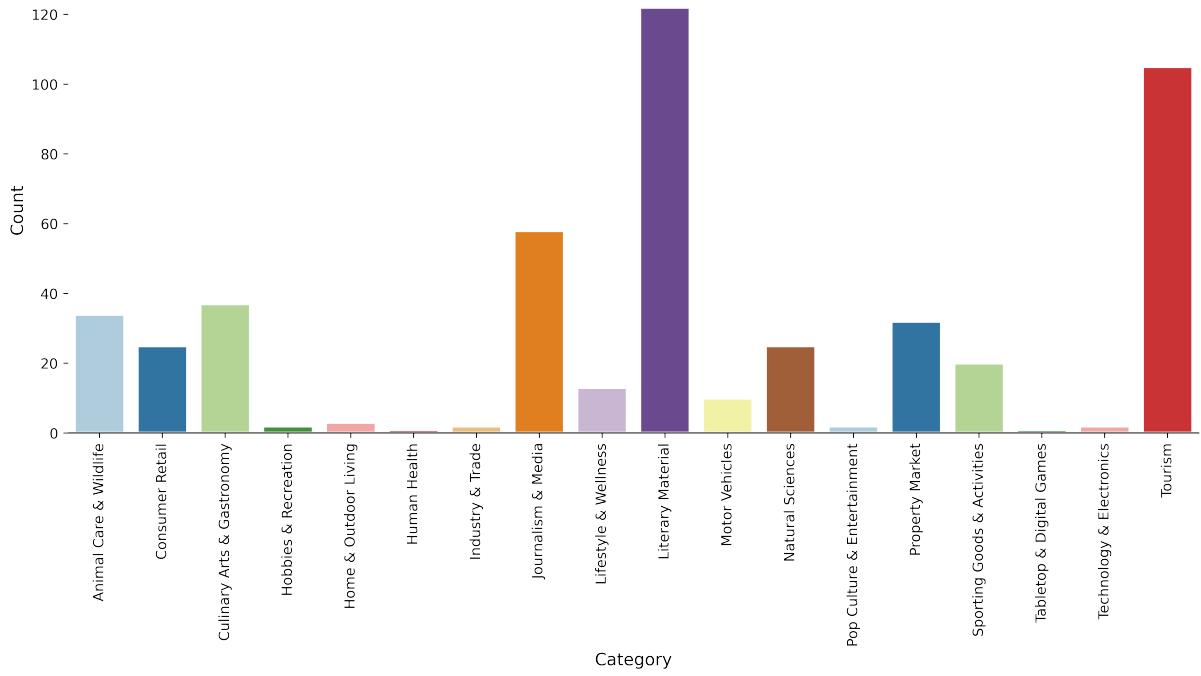
The dataset was sampled to cover a wide range of content. We use an internal image classification system to report the top image categories present across the splits in Figure 7. Getting a more balanced mix remains active work on our part and would be updated in future work.

B.1 Human Annotation Worker Pool

We employed and worked with a fixed human annotator pool comprising of 20+ annotators with mixed backgrounds in creative writing, art, history, photography and related relevant domain subjects to utilize critical domain expertise and perspectives. The pool is based in multiple countries, with



(a) IIW-Train Set Image Category Distribution



(b) IIW-Eval Set Image Category Distribution

Figure 7: Image Category Distribution for the IIW Dataset’s Train and Eval Splits.

a US majority currently. In the future, we plan to intentionally increase diversity in our annotator pool to ensure more locale-specific vocabulary in our image descriptions. The annotators were compensated appropriately taking their skill-set, qualifications, location and the complexity of the task into account. The pool was trained for the

annotation task over a period of month to achieve a sense of consistency on the annotation guidelines as well as the downstream tasks to be covered by the data being collected. The annotators were also communicated clearly on the downstream tasks and data use cases to get a sense of the importance and quality bar needed for this foundation work. For

text-to-image generation rankings, we employed an internal group of six people to rank the images generated by different model-generated image descriptions (*i.e.*, we did not hire crowd workers). People participating are domain experts, familiar with text-to-image generation technology.

B.2 Human Annotation Challenges

Despite the very detailed annotation guidelines we provided to the annotators, there were several challenges during the human annotation process. First, we still found individual instances of random quality or judgment lapses. To circumvent this, we designed our framework to be sequential (*i.e.*, more than one annotator works on each sample). We also found different challenges with respect to each image. For instance, art images require more domain specific expertise to describe an image with appropriate vocabulary. At the start of our annotation process, we observed that annotators had a tendency to use filler words and prefixes such as “*This is a*,” “*There is a*,” or “*This photo was taken with*,” and we provided feedback asking they do not include such phrases.

Another challenge during the annotation process was to encourage annotators to focus on the big picture and write a TLDR first. We also observed some tendency to use slightly subjective language while describing the images, *e.g.* using adjectives that are not explicitly supported by the visual cues. By providing feedback directly to the annotators, pointing to specific samples, and emphasizing that certain language styles do not align with the writing style we were aiming for, we were able to considerably increase the annotation quality and get the desired type of image descriptions from the annotation process.

B.3 Annotation Methodology

Seeded Annotation Considerations to keep in mind:

1. **Quality of the seeding data** is critical. It is counter productive if it’s noisy as the human annotators will take longer to comb signal from the noise than to come up with the information themselves. We recommend to restrict the use of seeding signal to only high precision models.
2. Risk of **biasing** the outputs as the human annotators may take the easy route of relying

on the seed signal more heavily than intended. We suggest to note this point explicitly in the annotation guidelines and spot check the annotations for quality control. Additionally, running annotations with no seeding and comparing the outputs can be helpful to judge the bias being induced.

Sequential Augmentation Considerations to keep in mind:

1. Heavy reliance on the quality of the base dense description from the first annotator. If the quality is not good, the annotator in the next round will spend considerable time fixing the input. There are 2 mitigating steps:
 - (a) Monitor this at the beginning of the annotation project when the annotators are still new to the task using metrics like edit-distance and provide explicit feedback to the annotators as needed.
 - (b) Annotators in each round have the option to start from scratch if they deem the quality from the previous round to be considerably low. Use this as feedback for the annotator from the previous round by presenting them the edited output to learn from.

Human-in-the-Loop Learning Our annotation framework implicitly unlocks a feedback loop for the annotators due to the sequential augmentation process discussed above. Each annotator gets an opportunity to read and learn from each other’s perspective which in turn improves their individual quality. As an example from Figure 8, we demonstrate how Annotator-1 get an opportunity to learn from Annotator-3 for the first image and Annotator-2 gets an opportunity to learn from Annotator-1 in the second image.

Model-in-the-Loop Annotation We employ an active learning loop for the VLMs where after some initial annotation data is available, a model version M_1 can be trained over the base VLM to improve the seed description quality. As more data gets annotated, M_1 can be updated to M_2, M_3, \dots, M_n to reduce the human effort needed.

Advantages:

1. Reduces the dependency on the *human* both in terms of number of annotation rounds and time.

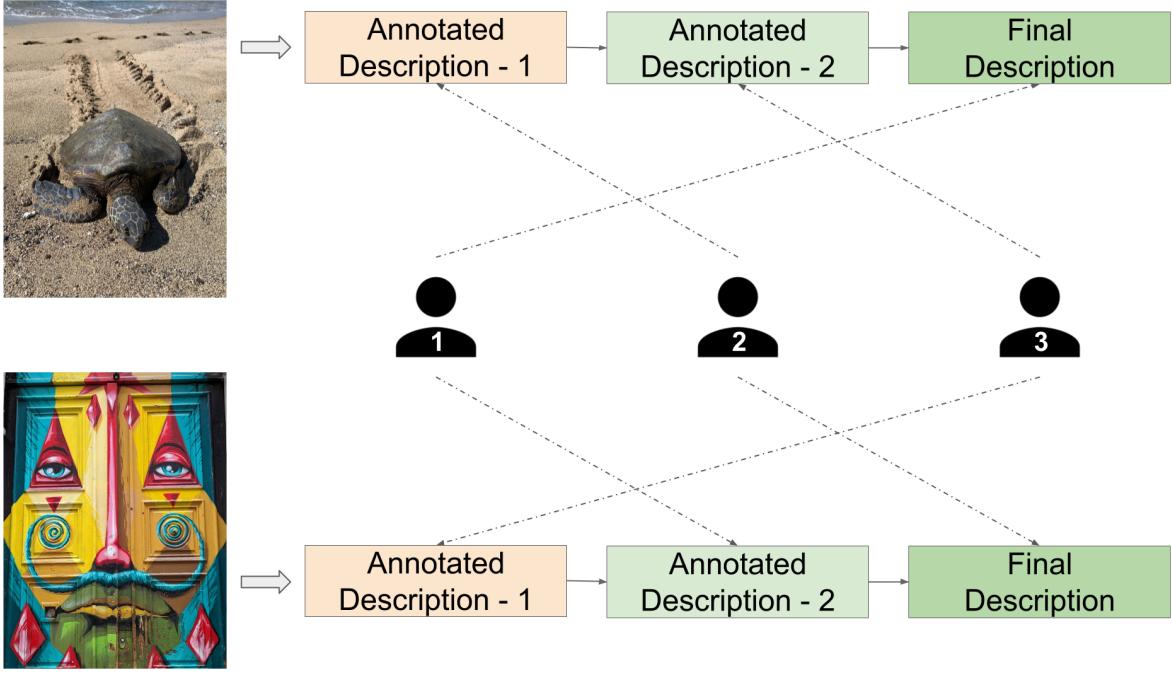


Figure 8: Human-in-the-Loop Learning. Over time with a constant annotator pool, each annotator gets an opportunity to read and learn from others’ perspective via an *implicit feedback loop*. This has shown to improve individual annotator quality as shown in the main paper.

- Provides a way to evaluate current model quality by monitoring the time, volume and patterns of augmentations during the human annotation stage.

Some considerations to keep in mind:

- As discussed above, the effectiveness relies very heavily on the capability of the model, *i.e.*, having high comprehensiveness and low hallucinations.

B.4 Annotation Framework

We now discuss the annotation framework with concrete examples and UI illustrations:

Annotation Task-1: Fine Grained Objects and Attributes In Task-1, the human annotators are presented with seed annotations for the objects from an Object-Detection (OD) model and VLM generated seed captions for each object (see Figure 9). The annotators can then annotate to note the salient objects and their corresponding description (see Figure 10).

Annotators can make the following augmentations to annotate salient objects:

- Edit** make adjustments to the label and/or bounding box. This can include:

- Making the labels more specific, e.g *Animal* to *German Shepherd*
- Enlarging or tightening the bounds of the bounding box by expanding or contracting the seed box.

- Remove** any invalid pre-populated objects or considerably invalid bounding boxes.
- Add** any missing salient object by drawing out a tight bounding box and adding an appropriate fine-grained label to it.
- Merge** if object(s) are fragmented and/or pre-populated as two or more objects, the annotators can remove the individual objects and create a new single object.

- Closely placed objects of the same/similar label/type which individually hold low value but can be described as a collection to hold a higher context value should be combined, *e.g.*, five identical cups in an image lined up next to each other do not need to be tagged as separate objects. If there are attributes that separate one or more of them from the others, we expect the annotators to split them in groups and proceed accordingly.



Figure 9: IIW Annotation UI for Task-1 with VLM seeds. We illustrate the seed object-detection objects and VLM generated object-level captions with object cropped image bytes as input.



Figure 10: IIW Annotation UI for Task-1 after human augmentation. We illustrate the human augmented salient objects and their human-authored descriptions. The annotations are built on seed information from Figure 9. This example demonstrates how humans can alter the seed annotations based on the annotation guidelines, which can include merging, deleting, editing and adding new salient objects and then describing each.

– Sub-components of a larger object should not be explicitly tagged unless there is something unique and/or worth mentioning about them. Think *does missing this detail create a different men-*

tal picture than the actual image?, e.g., doors, windows, or tires of a *Car* can be omitted unless there is something unique about them, as they are standard expectations from a *Car* object.

For each (label, bounding box) pair, we ask the annotators to generate a detailed description focused on the object in the context of the image considering the several axes as reference (see Appendix A).

Annotation Task-2: Overall Image Description

In Task-2, human annotators are presented with the annotations from Task-1 and a seeded VLM description (see Figure 11) which is then refined by human annotators in sequential rounds to produce the final hyper-detailed description (see Figure 12).

C IIW Fine-Tuning Tasks

We define seven tasks with the IIW Task-1 and Task-2 annotations to fine-tune two IIW based VLM model variants of PaLI-3 5B (Chen et al., 2023a). Our models include *IIW Combined*, trained on a mixture of all seven tasks and *IIW-Task-2* based aka *IIW Model*, which is only trained on the final most detailed image description output. The seven tasks can be grouped into three categories: image region, salient objects, and detailed description based tasks, see Figure 13 for illustration.

As we later discuss, we generally find the IIW (Task 2 only) Model to be preferred over the IIW Combined variant, but include details on the additional training tasks and resulting ablations here for completeness. All results in the main paper use the IIW Model.

C.1 Image Region Tasks

Using one object at a time from the list of (*label, bounding box, description*) Task 1 annotations, we perform three region-based tasks. We use normalized bounding boxes in [$ymin, xmin, ymax, xmax$] format as in Pix2Seq (Chen et al., 2022). Our first task is description-label grounding. In multi-object dense images, a label in itself is not enough to uniquely identify an object. Thus, we create a grounding task with (*image, label, description*) inputs that are tasked to predict the corresponding normalized bounding box coordinates.

Our second image region task is label prediction, in which we predict an open vocab label for the object with input (*image, bounding box*). Lastly, we perform object description generation, which produces descriptions for each object in the image given (*image, bounding box, label*).

C.2 Salient Objects Tasks

Our next category of fine-tuning tasks concerns the salient objects in an image. We target the aggregated list of (*label, bounding box*) object features per image from Task 1. Our first task is label generation, in which given an image, we aim to generate a text list of the salient object labels. The object labels are sorted alphabetically for consistency, but in future work ordering by saliency would be useful. Our second object-level task is *grounded* label generation. The task is to generate the list of (*label, bounding box*) pairs per object in the image; we similarly sort the list alphabetically with respect to label name.

C.3 Detailed Description Tasks

Finally, our last fine-tuning tasks relate to the sequentially annotated descriptions from Task 2. We perform description elaboration in addition to direct description generation. Given the image and description from the N^{th} sentence, description elaboration trains the model to elaborate the current description to the final description. We also create synthetically corrupted versions of the final description to serve as additional training samples. Specifically, we randomly drop $X\%$ of sentences. Sentences are dropped starting from the last sentence so that the structure of the overall text piece is maintained (as opposed to random sentence removal). For final description generation, given the image, a VLM learns to generate the final most hyper-detailed description available from the entire annotation framework. This final task (and not description elaboration), is the only task used to train the IIW model (whereas all are used for the IIW Combined ablation).

D Experiments

D.1 Seeded Annotation SxS

We additionally run a human SxS evaluation to compare the effects of seeding in the IIW annotation framework. In Table 7, we compare descriptions written without and with VLM seeding on a subset of IIW-400 (50 samples). There is a trend across all metrics that seeding improves description quality, as seen with marginal or substantial gains across comprehensiveness (+54%), specificity (+48%), TLDR quality (+28%), and human-likeness (+25%). The hallucinations metric is primarily neutral with a slight preference to seeded descriptions (+9%). This is somewhat ex-



Figure 11: IIW Annotation UI for Task-2 with seed VLM description. This VLM has been fine-tuned in an active learning mode as data was collected iteratively. The seed caption from the same VLM (PaLI-5B) without the IIW fine-tuning is “*a pink bicycle with a basket of flowers on it.*” The seed annotation is then refined and augmented by human annotators, see Figure 12 for the final resulting description.



Seed VLM Description

A detailed close-up captures a pink bicycle model, meticulously crafted from metal, perched on a shelf adorned with burlap. The bicycle, positioned on the left side of the frame, boasts a black seat and handlebars, complemented by a silver wire basket filled with pink shells. The wheels, a glossy black, are secured by sturdy metal stands. The bicycle model is set against the backdrop of a large, cream-colored vase, embellished with a blue and white floral motif. The vase is situated on a glass shelf, adding a touch of intrigue to the composition. The lighting, seemingly artificial, casts a soft glow on the scene, enhancing the overall visual appeal.

Annotator-1 Detailed Description

Annotator-2 Detailed Description

Annotator-1 Detailed Description

Close-up shot at a slightly high angle features a small statue or trinket of a pink bicycle placed on an indoor surface. The bicycle is positioned to its left side with the front slightly skewed towards the viewer. It appears old, as some of the pink paint on the frame is worn away, as well as the gray paint on the handlebars and spokes, and the black paint on the tires. There is a metal basket on the front that has a few light pink flowers inside. The seat is black and the kickstand is down on the bike's left side. The bike is set on a glass surface that is covered in a tan rattan cloth, which folds up between the bike's wheels. A muted yellow vase stands behind and to the right of the bike. It has a blue floral design along its curved body. Behind and to the left of the bike, there is a wooden frame that is painted in a white faded paint, revealing the darker wood beneath it. It is unclear if the frame is for a mirror or window. An indiscernible white object can be seen in either the mirror or window. There is also an odd glare in the upper right of the frame, of which the light source is unknown. It creates three short panels.

Annotator-2 Detailed Description

A close-up, eye-level indoor shot of a small pink bicycle figurine placed on a rattan cloth on a clear glass surface. The bicycle is positioned with its left side entirely visible, with the front slightly skewed towards the viewer. It appears old, as some of the pink paint on the frame is worn away, as well as the gray paint on the handlebars and spokes and the black paint on the tires. There is a metal basket on the front that has a few light pink, fake mini flowers inside. The seat is black, and the kickstand is down on the bike's left side. The bike is set on a dusty glass surface framed in black that is covered in a tan rattan cloth that folds up between the bike's wheels. A muted yellow vase stands behind and to the right of the bike. It has a blue floral design along its curved body. Behind and to the left of the bike, there is a wooden frame that is painted in a faded white paint, revealing the darker wood beneath it. It is unclear if the frame is for a mirror or a window. An indiscernible white object can be seen in either the mirror or window. There is also an odd glare in the upper right of the frame, whose source is unknown.

Figure 12: IIW Final Annotation UI for Task-2. We illustrate the human annotations available from Task-1 as the human annotators hover over the salient objects in the image. The annotators can additionally switch between hiding all salient objects to view the image properly. Task-2 annotations start with the seed caption from the VLM and is then refined by human annotators in sequential rounds, building on top of the previous round’s output.

pected, and affirms that despite model-generated outputs having a potential risk for hallucinations, the humans are able to correct and improve on them. Thus, the SxS confirms seeding is advantageous to the IIW annotation framework.

D.2 IIW Human versus IIW Model SxS

In Table 8, we perform a SxS evaluation on a subset of IIW-400 (on 100 samples). This compares data from the human authored IIW annotation framework to descriptions generated by the IIW fine-

tuned model. Across all metrics there is an extremely high preference to the human annotated data, with significant and marginal gains: comprehensiveness (+78%), specificity (+91%), fewer hallucinations (+31%), TLDR quality (+58%), human-likeness (+52%). This confirms the quality of data produced by the IIW human-in-the-loop annotation framework, and demonstrates the need for more modeling efforts to bridge the gap between the IIW human authored versus model generated description quality. For example, larger capacity models

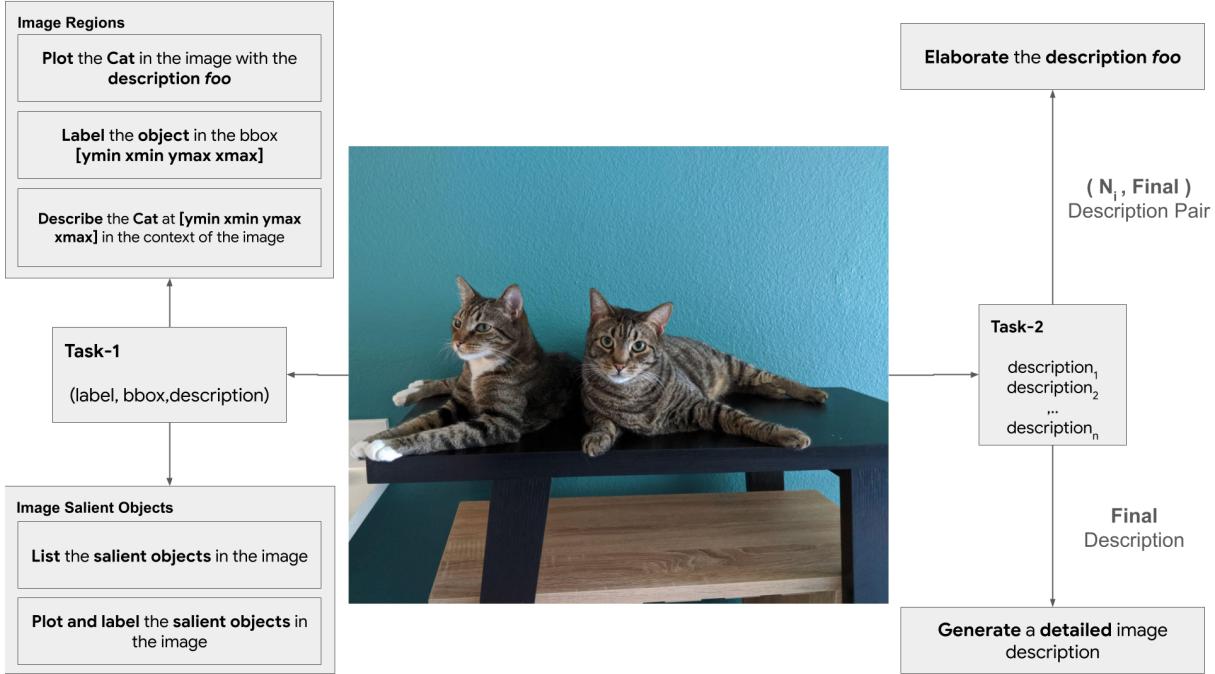


Figure 13: IIW based VLM Fine-tuning Tasks. We show tasks based on data collected from Task-1 and Task-2 per the IIW annotation framework. Different tasks enable the fine-tuning to focus on the image at (object, attribute), (image, objects) or (image, hyper-detailed description) levels.

Metric	IIW-400				
	Unseeded		Seeded		
	++	+	-	+	++
Comprehensiveness	6	8	18	45	23
Specificity	10	6	20	39	25
Hallucinations	4	16	51	23	6
TLDR	4	27	10	43	16
Human-Likeness	10	12	31	33	14

Table 7: Human SxS to Evaluate Gains from Seeding the Annotation in the IIW Annotation Framework. We report rounded percentages comparing 50 IIW-400 samples annotated by the IIW framework with and without machine-generated seeding on Comprehensiveness, Specificity, Hallucinations, TLDR quality, and Human-Likeness.

may be needed.

D.3 Automatic Readability Measurements

In addition to our human SxS comparisons, we use a suite of *readability* metrics to quantify writing style differences between DCI, DOCCI, and IIW. We run heuristics based readability metrics over both human-authored and model-generated descriptions representing each style, and present the results in Table 9. Each metric roughly estimates the level of education needed to understand a piece of written text using different units, *e.g.* education

years or grade-level. While they are proxy signals, a pattern across all can be seen as a clear indication of a more mature and articulate writing style for IIW in comparison with the other alternatives.

For the metrics, we used spaCy (Honnibal et al., 2020) (v3.0.0rc2) to tokenize the text and the implementation in Github’s [py-readability-metrics repo](#) (v1.4.1) to calculate the scores. We also include the readability metric distributions in Figure 14. The distributions further demonstrate a more mature writing style in both the IIW human-authored dataset and fine-tuned model generated outputs.

D.4 Side-by-Side (SxS) Evaluation Framework

We demonstrate the Human SxS annotation UI to show the input (see Figure 15) and the corresponding human responses (see Figure 16) across the 5 metrics, each on a 5 point scale. The metrics are defined as:

- **Comprehensiveness:** The description should capture all of the important elements of the image, including objects, people, locations, actions, relationships between objects, *etc.*
- **Specificity:** The description should use precise and descriptive language to avoid vagueness and ambiguity. *E.g.* “3 apples” and “Taj

Metric	IIW-400				
	IIW-Human		-	IIW-Model	
	++	+	-	+	++
Comprehensiveness	40	43	12	4	1
Specificity	79	14	5	2	0
Hallucinations	6	46	33	17	4
TLDR	29	43	14	10	4
Human-Like	27	32	34	6	1

Table 8: Human SxS to Evaluate IIW Fine-tuned PaLI-3 5B Model Predictions when compared to IIW Human-authored Data on IIW-400 using 100 samples.

Dataset	Human Authored				Model Generated			
	ARI↑	FK↑	GF↑	SMOG↑	ARI↑	FK↑	GF↑	SMOG↑
DCI	5.8	5.7	8.1	8.1	2.9	3.7	6.2	6.9
DOCCI	7.5	7.1	9.5	8.7	6.4	6.6	8.7	8.2
IIW	10.4	9.5	11.8	11.5	9.3	9.0	11.3	11.7

Table 9: Readability Metrics on Human and Model Annotated Data. We include ARI ([Wikipedia contributors, 2023b](#)), Flesch Kincaid (FK) ([Wikipedia contributors, 2023c](#)), Gunning Fog (GF) ([Wikipedia contributors, 2023d](#)), and SMOG ([Wikipedia contributors, 2023e](#)) metrics. They approximate the grade level needed to comprehend the text and results indicate a more mature writing style in IIW human-authored and model generated outputs.

Mahal” are more specific than “some apples” and “a white marble structure,” respectively.

- **Hallucinations:** The description should be factually correct and avoid making assumptions or interpretations that are not visually supported by the image.
- **First few line(s) as tldr:** The first few line(s) should paint a high level picture of what to expect in the image and create a succinct summary.
- **Human-Like:** The descriptions should feel as if an educated person wrote them and should be free from artifacts hinting that a machine generated them (*e.g.* stuttering, repeating facts, fragmented chain of thought, *etc.*).

The 5 metrics are defined to capture 3 broad umbrella metrics of precision, recall and writing-style. An *overall metric* score can further be computed by taking an average of the 3 umbrella metrics. Each can be defined as follows:

$$\text{Recall} = \text{avg}(\text{Comprehens., Specific.})$$

$$\text{Precision} = \text{Hallucination}$$

$$\text{Writing Style} = \text{avg}(\text{TLDR, Human Like})$$

$$\text{Overall} = \text{avg}(\text{Rec., Prec., Writing Sty.})$$

D.5 Additional Automatic Metrics

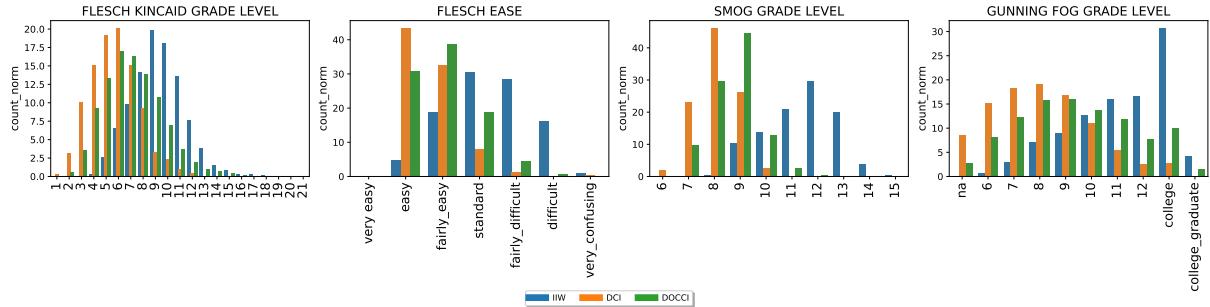
We include evaluations of model-generated outputs with automated text similarity metrics for completeness, but note that common text similarity metrics are ill-suited for long texts and more recent image-text metrics are often length limited. We report these results simply to emphasize the limitations of these metrics when measuring the quality of hyper-detailed image descriptions. Using standard automatic metrics, Table 10 illustrates how fine-tuned models largely perform better in replicating their own style.

In addition to reporting BLEU-4, ROUGE-1, and ROUGE-2 automatic metrics, we include CIDEr ([Vedantam et al., 2015](#)), BERTScore ([Zhang et al., 2020](#)), and BLEURT ([Pu et al., 2021](#)) metrics in Table 11. We include BERTScore and BLEURT as they are newer, model-based metrics which have been shown to correlate more closely with human judgements. CIDEr, like BLEU and ROUGE metrics are not limited by sequence length. BERTScore and BLEURT have a maximum sequence length of 512 (we specifically use the “wwm_cased_L-24_H-1024_A-16” BERT checkpoint and the latest BLEURT-20 model), but for our descriptions, they likely fit under this maximum length, with only outliers being truncated.

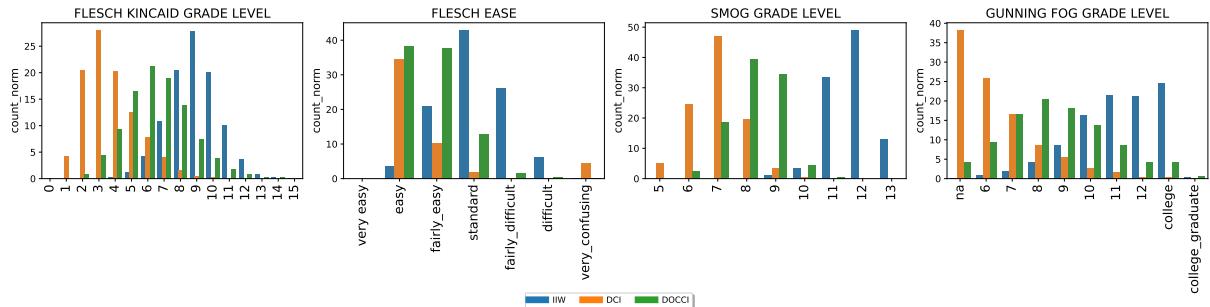
CIDEr and BERTScore generally show the same trend of each fine-tuned model performing best on the same test domain (*i.e.*, DCI fine-tuned mod-

PaLI-ft	DCI Test (112)			DOCCI Test (5k)			IIW Test (445)		
	bleu-4	rouge-1	rouge-2	bleu-4	rouge-1	rouge-2	bleu-4	rouge-1	rouge-2
DCI	4.97	35.38	12.70	5.24	39.55	12.95	2.30	31.70	8.58
DOCCI	4.24	34.60	10.70	8.68	45.50	17.07	3.50	36.10	10.02
IIW	3.02	31.59	8.02	4.60	38.10	10.06	5.66	38.57	11.73

Table 10: Cross Dataset Automatic Metric Evaluation of Fine-tuned Models.



(a) Distribution on the Human Authored Datasets from DCI, DOCCI and IIW.



(b) Distribution on the Fine-tuned Model Generated Outputs from DCI, DOCCI and IIW.

Figure 14: Distribution-based Readability Metrics. We compare both human authored and model generated outputs from IIW and prior work to show the distribution of Education based units reflected in the writing style. IIW outputs from both the human annotators and the model produce a more mature style across the metrics.

els perform best on DCI test set, DOCCI models perform best on DOCCI test set, and so on. One anomaly occurs with CIDEr on the DCI test set, where PaLI models fine-tuned with DOCCI slightly outperform the DCI trained model (4.91 versus 4.57). Due to how low the metric values are, these differences may not be significant. When evaluating the DCI, DOCCI, and IIW test sets with BLEURT, we instead find a slight preference for IIW models. Across all three datasets, BLEURT shows PaLI-IIW variants perform better or similarly to the same-domain test set. Thus, newer metrics may reveal IIW fine-tuned models generalize better than models fine-tuned on other datasets.

D.6 IIW Fine-tuned Model Ablations

As an IIW ablation study, we fine-tune a separate PaLI-5B model, *IIW-Combined*, using all the data from Task 1 and Task 2 as a mixture of 7 training tasks, defined in Appendix C. Table 11 and 12 show

that this has no clear significant gains on Task-2’s final description eval set. This currently remains a less explored area and we aim to investigate this in future work to further improve the model on Task-2 evaluations.

D.7 Reconstructing Images with IIW Descriptions

For reconstructing images sentence-by-sentence, we fed the T2I model the *first sentence*, *first two sentences*, *first three sentences*, etc. as prompts from each of the three datasets (DCI, DOCCI and IIW). Figure 17 showcases the prompts and the T2I model outputs from three descriptions along with the original image.

We then asked human annotators to rank the generated images by how similar they are to the original image. The image most similar to the original image is ranked number 1. We allowed generated images to be ranked the same if they are very sim-



A

This is a close up photo of a lizard. The lizard is sitting on a large green leaf. The lizard is looking to the left of the image. The lizard's head is turned towards the left side of the image. The lizard's eyes are closed. The lizard's neck is visible. The lizard's tail is visible. The lizard's legs are visible. The lizard's claws are visible. The lizard's feet are visible. The leaves in the background are blurry and out of focus. The leaves in the foreground are blurry and out of focus.

B

A vivid close-up captures a lizard perched on a verdant leaf, set against a softly blurred backdrop of lush green foliage. The lizard's body is adorned with a striped pattern of light brown and tan scales, punctuated by a distinctive black spot on its head. The lizard's eyes are closed, adding a sense of tranquility to its demeanor. The leaf it's perched on is a vibrant green, its edges tinged with a reddish hue, adding a pop of color to the otherwise monochromatic scene. The lizard's claws are visible, adding to the overall composition of the shot.

Figure 15: Human SxS Annotation UI. Annotators are shown the input image and two input image descriptions to evaluate side-by-side. The input descriptions could be from any combination of (human, model) sources. This information is not shared with the annotators and the sources are randomly flipped and marked as *A* or *B* to prevent any source or order based bias.

PaLI-ft	DCI Test (112)			DOCCI Test (5k)			IIW Test (445)		
	CIDEr	BERT	BLEURT	CIDEr	BERT	BLEURT	CIDEr	BERT	BLEURT
DCI	4.57	0.60	0.41	4.71	0.61	0.42	0.75	0.56	0.40
DOCCI	4.91	0.58	0.39	11.09	0.65	0.45	2.40	0.59	0.41
IIW	1.87	0.56	0.41	4.52	0.59	0.46	4.04	0.61	0.45
IIW Comb.	0.61	0.56	0.43	4.15	0.59	0.46	1.77	0.60	0.46

Table 11: Additional Automatic Metric Results. We report CIDEr, BERTScore (referred to as BERT in table due to space), and BLEURT metrics for all fine-tuned models. We compare DCI, DOCCI, IIW, and IIW Comb. (Combined).

ilar. Figure 18(a) shows the reconstruction rank counts for all the sentence counts and Figure 18(b) shows the rank counts when we use sentence 1, sentence 1 and 2, sentence 1, 2 and 3, and sentence 1, 2, 3, and 4. Sentences from IIW descriptions are ranked first much more frequently than sentences from DCI and DOCCI descriptions. Specifically, for the first sentence, the difference is most notable, supporting our claim that IIW descriptions are higher quality earlier on and IIW first sentences are designed to capture a TLDR.

D.8 Compositional Reasoning with IIW Descriptions

In our downstream evaluation of ARO, SVO-Probes, and Winoground compositional reasoning benchmarks with IIW descriptions, we formulate a

new LLM-only method of evaluation. We prompt a LLM (*e.g.*, PaLM 2) to determine which is the true matching caption given the generated image description and the image caption options to select from. We define the LLM prompt which includes an image description as:

“Given the following image description and image caption options, choose the most likely OPTION number :

IMAGE-DESCRIPTION : <DESCRIPTION>

OPTIONS : <CHOICES>

RESPONSE : ”

where we fill in the <DESCRIPTION> from each VLM description model (*e.g.*, either our IIW

	A is substantially better	A is marginally better	Neutral	B is marginally better	B is substantially better	
Hallucination ?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Both of them say the eyes are close, but the visible eye is open. A has the tail in view when it is not, and also has all the foreground leaves out of focus. There is a stripe on the lizard, but B calls it a "striped pattern".
Comprehensiveness ?	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	B does not give the direction the lizard is facing, however it gives more description of the lizard and leaves.
Specificity ?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	B provides colors on the lizard, scales, and the color of the leave's edges.
First few line(s) as tl;dr ?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	A provides the lizard but nothing else.
Human Like ?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	A repeats sentence stems such as "The lizard" and "The leaves". B has an overall better flow.

Figure 16: Human SxS Annotation UI responses for the input image and two image description pairs (see Figure 15). The annotators respond to the 5 metrics independently on a 5 point scale. They are additionally asked to justify their choices which can be used to sanity check and perform quality sweeps.

PaLI-ft	DCI Test (112) bleu-4 rouge-1 rouge-2			DOCCI Test (5k) bleu-4 rouge-1 rouge-2			IIW Test (445) bleu-4 rouge-1 rouge-2		
IIW	3.02 31.59 8.02			4.60 38.10 10.06			5.66 38.57 11.73		
IIW Combined	2.95 30.63 7.30			4.76 38.25 10.48			5.40 37.64 11.62		

Table 12: Ablation Results Comparing IIW Variants on Automatic Metrics.

fine-tuned model, InstructBLIP or LLaVA) and the list of <CHOICES> are from the corresponding evaluation dataset, respectively. Choices are enumerated in a list-like fashion, and we ask the model to generate the number of the most likely caption.

We define a different prompt for the language bias baseline, which serves as a sanity check that the image/image description is truly needed for these datasets. It provides a lower bound for comparison, too. While the prompt is different as we do not input any image description, we try to make it as similar as possible to the above image description based prompt. We set the language bias prompt to:

“Given the following image caption options, choose the most likely OPTION number :
 OPTIONS : <CHOICES>
 RESPONSE : ”

where <CHOICES> are filled in in the same

format as previously described.

Importantly, when filling in the caption choices, we deterministically swap the index of the “answer,” *i.e.*, the true matching caption, among the choices list in the prompt. This is done to ensure an equal distribution and reduce any order bias (*e.g.*, a LLM may be more prone to believing the first option is the correct option).

To obtain the image description which is then fed into the LLM, we prompt our fine-tuned models with “Generate a detailed image description.” For the InstructBLIP and LLaVA models, we define similar prompts given the prompts used in their published papers: “Write a long and detailed description for the photo.” and “Provide a detailed description of the given image” for InstructBLIP and LLaVA, respectively.

We process the LLM outputs as classes, (*e.g.*, when choosing between image caption choices [1] and [2], LLM responses are ‘1’ or ‘2’) and calculate accuracy with respect to the true image caption class. If the LLM does not produce a valid class,

it's considered an incorrect prediction. Note that this task set up is different from how VLM models are typically evaluated on these reasoning datasets: prior work considers a sample to be correctly reasoned about if the image-text similarity of the true image caption is higher than the image-text similarity of the incorrect image caption. Due to the long length of our descriptions, we cannot compute image-text similarity reasonably with models like CLIP without significantly truncating our image descriptions. In future work, once input length limitations are mitigated, dual-encoder VLMs like CLIP can be fine-tuned with our rich data, which will help to improve VLM reasoning.

Note that ARO and Winoground datasets are built with positive and negative captions for each image. SVO-Probes differs in that it originally contained a positive and negative *image* for each positive caption. For our experiments, we need a true and false caption associated with an image. A large portion ($\sim 90\%$) of the SVO-Probes negative images also serve as separate samples (where they are considered positive images, with associated captions). Thus, we can pull these captions to serve as the negative caption for the original sample.

For the remaining $\sim 10\%$, we use the negative triplet (the S, V, O triplet specifying the subject, object, and verb, with one of them being modified) to automatically flip the negative S, V, or O in the positive caption. Ten of these samples did not have negative triplets in the dataset, so they were removed. Lastly, there were 114 samples with positive captions not containing the S, V, or O that needed to be swapped to form the negative caption. This happens as a result of SVO triplets containing root forms of the words, which were not spelled the same way in the caption. For example, an SVO may be “man,lie,beach” with the caption stating “A man lying on a beach.” Due to the verb tense differences, it would require additional processing to match “lie” to “lying.” We remove these edge cases for simplicity.

Finally, we include more vision language compositional reasoning results with different PaLI fine-tuned models in Table 13. Here we additionally include the models fine-tuned with DCI and DOCCI datasets. The IIW descriptions still result in highest reasoning accuracy for ARO VG-A and are comparable with DOCCI on Winoground. Trends also stay the same with SVO-Probes, with DOCCI performing similarly to IIW, but InstructBLIP performing slightly better (by less than 1 accuracy

point). Finally, we find that DOCCI performs best on VG-R, which might be result of its dataset being designed to explicitly contain connected and contrasting images, which might more frequently capture similar images that only differ by the visual relationship between objects.

While performance differences between DCI, DOCCI, and IIW are smaller, this could be an artifact of the reasoning datasets; ARO, SVO-Probes, and Winoground are all built upon short caption datasets, so the utility and quality differences between DCI, DOCCI, and IIW are not fully captured by these probing datasets.

E Enriching Image Caption Datasets

As discussed in the main paper, we enrich 1k samples from two existing image caption datasets, namely, Localized Narratives and CrossModal (XM) 3600, with new image descriptions generated by IIW fine-tuned models. The goal of releasing these enriched versions is to provide longer, hyper-detailed image descriptions that can be used for evaluation purposes in future work. The enriched versions not only allow for finer-grained, full coverage evaluations of the content in images (via new metrics or probing datasets), but also may enable autorater models which learn from the precision and recall errors in the generated descriptions.

In Table 14, we report the language statistics on the original 1k samples from each dataset and the enriched versions. It is clear that the IIW descriptions are significantly longer and richer, as we have higher counts of tokens, sentences, and each part of speech.

F Percentages Reported in the Main Paper

We re-quote and define all analysis percentages reported in the main paper for clarity on how they were calculated in Tables 15-17. The reference location is defined by the section, paragraph, and line it appeared in. We only include paragraph number for multi-paragraph sections, and only include line number if the same percentage occurs more than once within a paragraph. For example, “S4.3 P2 L3” means Section 4, Paragraph 2, Line 3. Most percentages were rounded to the nearest point in the main paper.

Image Description Model	ARO		SVO-Probes	Winoground
	VG-A	VG-R		
None (Language Bias Baseline)	56.50	59.94	50.71	49.88
InstructBLIP-Vicuna-7B	83.99	62.73	89.35	65.25
LLaVA-V1.5-7B	84.80	63.71	87.89	63.38
PaLI-3 + DCI 5B	88.19	66.47	86.50	64.62
PaLI-3 + DOCCI 5B	89.70	68.85	88.73	69.50
PaLI-3 + IIW 5B	90.37	66.19	88.66	69.38
PaLI-3 + IIW Combined 5B	89.46	64.88	87.78	66.88

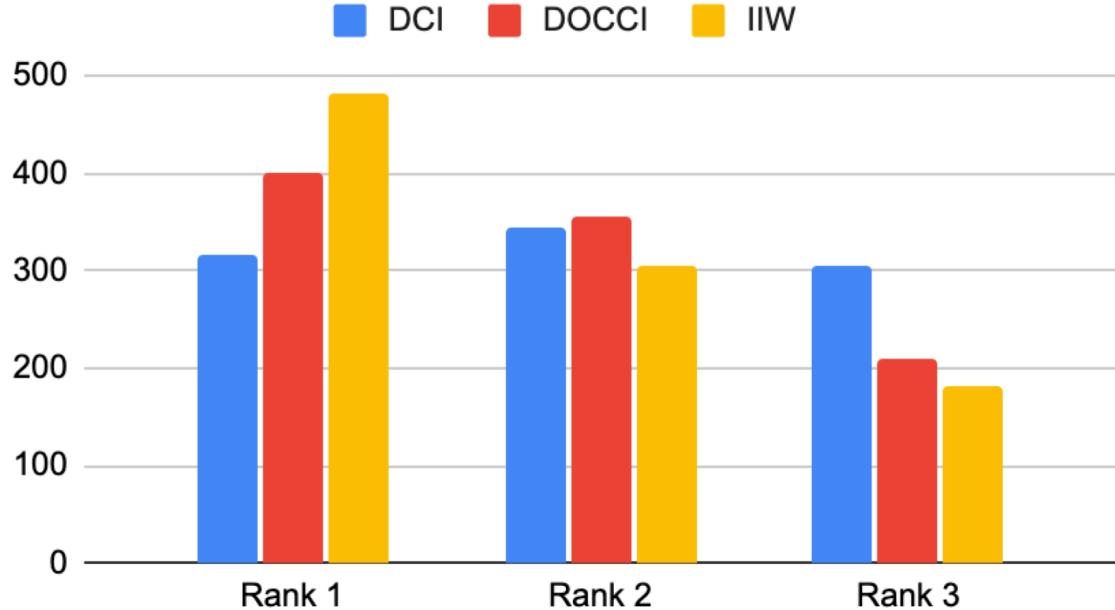
Table 13: VL Compositional Reasoning Accuracy with Image Descriptions. We evaluate whether rich descriptions can distinguish the true matching image caption in ARO (Yuksekgonul et al., 2023), SVO-Probes (Hendricks and Nematzadeh, 2021), and Winoground (Thrush et al., 2022) datasets. The COCO and Flickr30k Order subsets of ARO are not reported due to a very high language bias baseline of 98%.

Dataset	Sample Count	Tokens	Tokens	Sentences	NN	ADJ	ADV	VB
		/ Sent.		/ Desc.				
LocNar (Pont-Tuset et al., 2020) IIW Enriched	1000	14.35	30.56	2.12	8.02	1.09	0.16	2.39
		22.19	128.87	5.80	32.37	16.02	1.82	11.44
XM3600 (Thapliyal et al., 2022) IIW Enriched	1000	10.40	10.40	1.00	3.45	1.08	0.04	0.61
		22.25	130.56	5.86	33.18	15.82	1.72	11.87

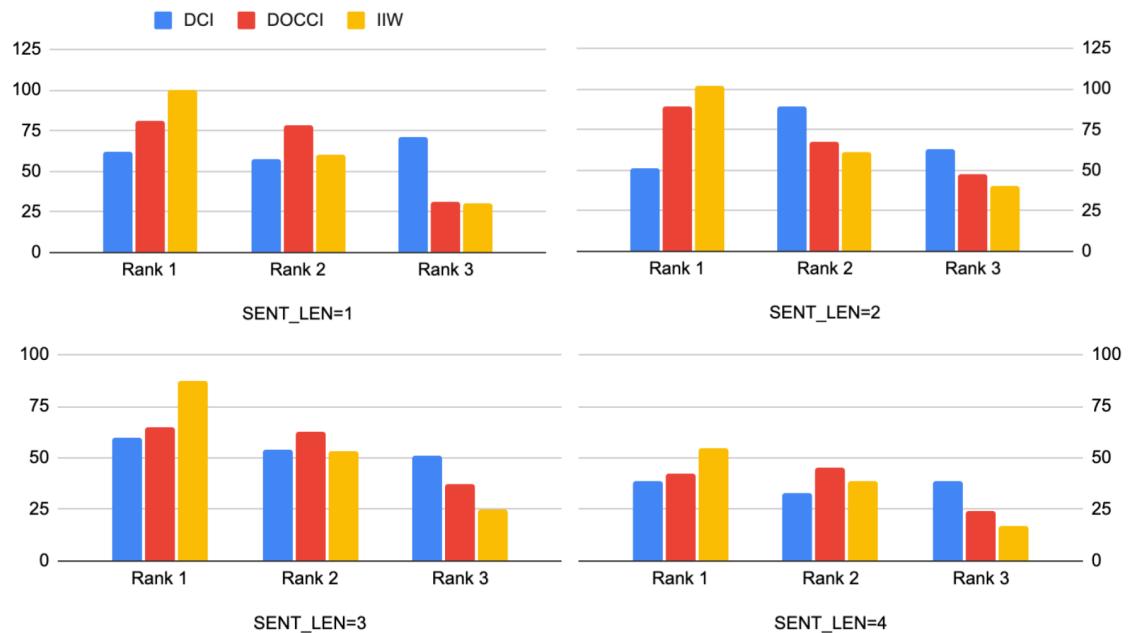
Table 14: Dataset Statistics Comparing ImageInWords (IIW) Descriptions of Prior Work to their Original Annotations. We include the number of samples (*i.e.*, subset of captions/descriptions that we enrich) and the average number of tokens, sentences, nouns (NN), adjectives (ADJ), adverbs (ADV), and verbs (VB). Language statistics are averages reported per description unless otherwise noted.

Original Image	DCI	DOCCI	IIW
			
Sentence 1 (prompt to T2I)	A close up photo of a large sphere hanging from the ceiling.	A medium-close-up view of a yellow lamp that is hanging from the ceiling.	A detailed close-up captures a spherical light fixture suspended from a black cord, set against a stark black backdrop.
Human ranking	Ranked 3rd	Ranked 2nd	Ranked 1st
			
Sentence [1-2] (prompt to T2I)	A close up photo of a large sphere hanging from the ceiling. The background of the image is very dark with nothing visible at all.	A medium-close-up view of a yellow lamp that is hanging from the ceiling. The lamp is made up of a grid that is made up of thin black lines that run vertically and horizontally, these lines make up the grid.	A detailed close-up captures a spherical light fixture suspended from a black cord, set against a stark black backdrop. The fixture's outer sphere is adorned with a network of squares, each encased within a larger sphere.
Human ranking	Ranked 3rd	Ranked 2nd	Ranked 1st
			
Sentence [1-3] (prompt to T2I)	A close up photo of a large sphere hanging from the ceiling. The background of the image is very dark with nothing visible at all. The focus of the image is on the large sphere.	A medium-close-up view of a yellow lamp that is hanging from the ceiling. The lamp is made up of a grid that is made up of thin black lines that run vertically and horizontally, these lines make up the grid. Inside the grid, there is a yellow light that is made up of lines that run vertically and horizontally.	A detailed close-up captures a spherical light fixture suspended from a black cord, set against a stark black backdrop. The fixture's outer sphere is adorned with a network of squares, each encased within a larger sphere. The sphere's interior is bathed in a warm, yellowish glow, punctuated by a solitary light source at its heart.
Human ranking	Ranked 3rd	Ranked 2nd	Ranked 1st
			
All sentences (prompt to T2I)	A close up photo of a large sphere hanging from the ceiling. The background of the image is very dark with nothing visible at all. The focus of the image is on the large sphere. The sphere is lit up bright yellow in the center with a bright white light shining out from the center of it. All of the sections of the sphere are made of black frames. The frame around the sphere is very intricate and meticulous.	A medium-close-up view of a yellow lamp that is hanging from the ceiling. The lamp is made up of a grid that is made up of thin black lines that run vertically and horizontally, these lines make up the grid. Inside the grid, there is a yellow light that is made up of lines that run vertically and horizontally. Along the top portion of the globe, there is a black line that runs vertically and horizontally. The bottom portion of the lamp is made up of thin black lines that run vertically and horizontally. The area where the light is shining is completely pitch black.	A detailed close-up captures a spherical light fixture suspended from a black cord, set against a stark black backdrop. The fixture's outer sphere is adorned with a network of squares, each encased within a larger sphere. The sphere's interior is bathed in a warm, yellowish glow, punctuated by a solitary light source at its heart. The sphere's surface is etched with swirling patterns, adding a dynamic element to its design.
Human ranking	Ranked 3rd	Ranked 2nd	Ranked 1st

Figure 17: T2I Outputs and Human Ranking Evaluations. We show example T2I results where the first sentence, first two sentences, ..., all the sentences of the image descriptions from DCI, DOCCI and IIW models are fed sequentially as inputs, *i.e.*, at each step an additional sentence chunk is fed to the T2I model.



(a) Reconstruction Rank Counts across Inputs over All Cumulative Sentence Chunks.



(b) Reconstruction Rank Counts across Inputs of Specific Cumulative Sentence Chunks.

Figure 18: T2I Human Rank Distributions. We illustrate bar plots for the image reconstruction evaluation results using image descriptions from fine-tuned PaLI-5B models on three datasets (DCI, DOCCI, IIW). Images reconstructed from IIW descriptions are consistently ranked better than other descriptions.

Percent	Reference	Equation and Explanation
+66%	Abstract, Intro P5, Conclusion	Average difference of IIW preference vs. other dataset preference, averaged over DCI and DOCCI datasets and averaged over the five metrics corresponding to (comprehensiveness, specificity, hallucinations, tldr, human-likeness). Differences of IIW marginally and substantially better - other dataset marginally and substantially better for (comprehensiveness, specificity, hallucinations, tldr, human-likeness) metrics from Table 2 correspond to DCI (61, 80, 42, 91, 82) and DOCCI (42, 82, 35, 79, 68). The final average preference over the five metrics and two datasets is 66.2%.
+48%	Abstract, Intro P5	Average difference of IIW preference vs. GPT-4V outputs, averaged over the five metrics corresponding to (comprehensiveness, specificity, hallucinations, tldr, human-likeness). Differences of IIW marginally and substantially better - GPT-4V marginally and substantially better for (comprehensiveness, specificity, hallucinations, tldr, human-likeness) metrics from Table 3 correspond to (35, 53, 59, 70, 21). The final average preference over the five metrics is 47.6%.
+31%	Abstract, Intro P5, S5.1 P1, Conclusion	Average difference of IIW model output preference vs. other fine-tuned model output preference, averaged over DCI and DOCCI fine-tuned models and averaged over the five metrics corresponding to (comprehensiveness, specificity, hallucinations, tldr, human-likeness). Differences of IIW marginally and substantially better - other dataset marginally and substantially better for (comprehensiveness, specificity, hallucinations, tldr, human-likeness) metrics from Table 3 correspond to DCI (42, 54, -9, 51, 57) and DOCCI (4, 37, -7, 57, 23). The final average preference over the five metrics and two datasets is 30.9%.
20% more	S3.2 P6	The median increase in token count from annotation round 1 to round 3: $(205-170)/170 = 20\%$.
30% less	S3.2 P6	The median decrease in time spent annotating from round 1 to round 3 compared to if three individual round 1s occurred: $((800*3)-(800+600+300))/(800*3) = 30\%$.
+61%	S4.1 P1	The amount IIW is more comprehensive than DCI in Table 2: $(30+41) - (3+7) = 61\%$.
+42%	S4.1 P1 L4	The amount IIW is more comprehensive than DOCCI in Table 2: $(33+19) - (4+6) = 42\%$.
+80%	S4.1 P1 L5	The amount IIW is more specific than DCI in Table 2: $(20+68) - (5+3) = 80\%$.

Table 15: Percentages from the Main Text. We reference each percentage and define how they were calculated for clarity.

Percent	Reference	Equation and Explanation
+82%	S4.1 P1 L5	The amount IIW is more specific than DOCCI in Table 2: $(22+65) - (3+2) = 82\%$.
42%	S4.1 P1 L5	The amount IIW contains fewer hallucinations than DCI in Table 2: $(32+15) - (2+3) = 42\%$.
35%	S4.1 P1 L6	The amount IIW contains fewer hallucinations than DOCCI in Table 2: $(34+13) - (0+12) = 35\%$.
+91%	S4.1 P1 L6	The amount IIW contains better TLDR than DCI in Table 2: $(20+74) - (3+0) = 91\%$.
+79%	S4.1 P1 L7	The amount IIW contains better TLDR than DOCCI in Table 2: $(30+54) - (1+4) = 79\%$.
+82%	S4.1 P1 L7	The amount IIW is more human-like than DCI in Table 2: $(25+59) - (1+1) = 82\%$.
+68%	S4.1 P1 L8	The amount IIW is more human-like than DOCCI in Table 2: $(46+23) - (1+0) = 68\%$.
+35%	S4.1 P2	The amount IIW is more comprehensive than GPT-4V outputs in Table 3: $(29+19)-(3+10) = 35\%$.
+53%	S4.1 P2	The amount IIW is more specific than GPT-4V outputs in Table 3: $(35+34) - (6+10) = 53\%$.
+59%	S4.1 P2	The amount IIW is contains fewer hallucinations than GPT-4V outputs in Table 5: $(34+31) - (0+6) = 59\%$.
+70%	S4.1 P2	The amount IIW contains better TLDR than GPT-4V outputs in Table 3: $(47+34) - (5+6) = 70\%$.
+21%	S4.1 P2	The amount IIW is more human-like than GPT-4V outputs in Table 3: $(27+13) - (6+13) = 21\%$.
+42%	S5.1 P1	The amount IIW is more comprehensive than DCI in Table 3: $(32+27) - (7+10) = 42\%$.
+4%	S5.1 P1	The amount IIW is more comprehensive than DOCCI in Table 3: $(26+5) - (5+22) = 4\%$.
+54%	S5.1 P1	The amount IIW is more specific than DCI in Table 3: $(24+46) - (6+10) = 54\%$.
+37%	S5.1 P1	The amount IIW is more specific than DOCCI in Table 3: $(33+24) - (6+14) = 37\%$.

Table 16: Percentages from the Main Text. We reference each percentage and define how they were calculated for clarity.

Percent	Reference	Equation and Explanation
+51%	S5.1 P1	The amount IIW contains better TLDR than DCI in Table 3: $(30+41) - (9+11) = 51\%$.
+57%	S5.1 P1	The amount IIW contains better TLDR than DOCCI in Table 3: $(42+28) - (6+7) = 57\%$.
+55%	S5.1 P1	The amount IIW is more human-like than DCI in Table 3: $(32+39) - (11+5) = 55\%$.
+23%	S5.1 P1	The amount IIW is more human-like than DOCCI in Table 3: $(27+14) - (6+12) = 23\%$.
-9%	S5.1 P1	The amount IIW contains fewer hallucinations than DCI in Table 3: $(11+13) - (12+21) = -9\%$.
-7%	S5.1 P1	The amount IIW contains fewer hallucinations than DOCCI in Table 3: $(21+6) - (9+25) = -7\%$.
34%	S5.3 P4	The accuracy improvement on VG-A from using IIW over the language bias baseline: $(90.37) - (56.50) = 33.87\%$.
6%	S5.3 P4	The accuracy improvement on VG-R from using IIW over the language bias baseline: $(66.19) - (59.94) = 6.25\%$.
20%	S5.3 P4	The accuracy improvement on Winoground from using IIW over the language bias baseline: $(69.38) - (49.88) = 19.5\%$.
6%	Abstract, S5.3 P4, Conclusion	The accuracy improvement on VG-A from using IIW over the next best baseline LLaVA: $(90.37) - (84.80) = 5.57\%$.
2%	S5.3 P4	The accuracy improvement on VG-R from using IIW over the next best baseline LLaVA: $(66.19) - (63.71) = 2.48\%$.
4%	S5.3 P4	The accuracy improvement on Winoground from using IIW over the next best baseline InstructBLIP: $(69.38) - (65.25) = 4.13\%$.

Table 17: Percentages from the Main Text. We reference each percentage and define how they were calculated for clarity.