OVFact: Measuring and Improving Open-Vocabulary Factuality for Long Caption Models

Monika Wysoczańska[1,2]* **Shyamal Buch**[1] **Anurag Arnab**[1] **Cordelia Schmid**[1]

[1]Google DeepMind [2]Warsaw University of Technology

: monika.wysoczanska.dokt@pw.edu.pl

Abstract

Large vision-language models (VLMs) often struggle to generate long and *factual* captions. However, traditional measures for hallucination and factuality are not well suited for evaluating longer, more diverse captions and in settings where ground-truth human-annotated captions are unavailable. We introduce OVFact, a novel method for measuring caption factuality of long captions that leverages open-vocabulary visual grounding and tool-based verification without depending on human annotations. Our method improves agreement with human judgments and captures both caption descriptiveness (recall) and factual precision in the same metric. Furthermore, unlike previous metrics, our reference-free method design enables new applications towards factuality-based data filtering. We observe models trained on an OVFact-filtered (2.5-5x less) subset of a large-scale, noisy (VLM-generated) pretraining set meaningfully improve factuality precision without sacrificing caption descriptiveness across a range of downstream long caption benchmarks.

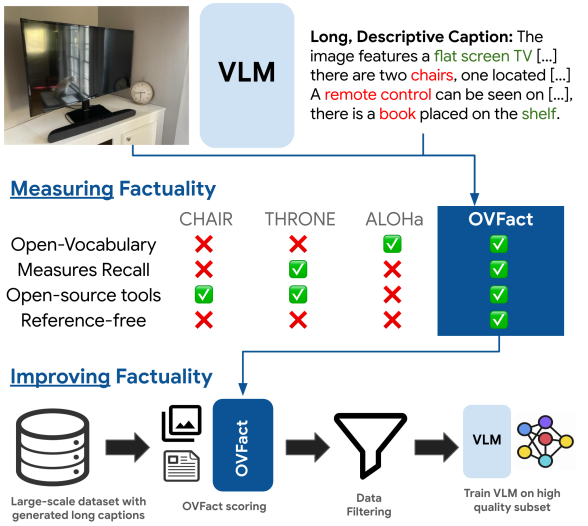

Figure 1: **Measuring and improving factuality.** We propose OVFact, a method for measuring and improving *open-vocabulary factuality* of vision-language models (VLMs) in long captioning. (**top**) Unlike prior work tailored for short captions or specific datasets, OVFact is *reference-free* and does not require annotated ground-truth. (**bottom**) Our flexible design enables a new factuality-based data filtering application, where models trained on OVFact-filtered datasets show improved factuality on downstream benchmarks with 2.5-5x less (higher-quality) training data and without compromising caption descriptiveness.

1 Introduction

Large vision-language models (VLMs) are fundamental to a range of grounded language understanding applications, including multimodal AI assistants and tools (Achiam et al., 2023; Georgiev et al., 2024). These models have grown in capability from generating short sentence descriptions (Socher et al., 2014; Karpathy and Fei-Fei, 2015) to long paragraphs (Beyer et al., 2024; Steiner et al., 2024; Chen et al., 2024b). However, long caption models struggle to maintain *factuality*[1] over these long descriptions (Kaul et al., 2024), including hallucinations of objects that are not present in the input image. As such, there is a crucial need to both measure and improve this limitation.

Prior work for *measuring* factuality for VLMs has been promising but limited. Question-answering approaches (Li et al., 2023; Jiang et al., 2024; Huang et al., 2024; Guan et al., 2024) are indirect, and do not assess if a *particular* output from a model is factual or not (Kaul et al., 2024), as is important in safety-critical settings (Bommasani et al., 2021). On the other hand, metrics for directly assessing caption text (Rohrbach et al., 2018; Kaul et al., 2024; Petryk et al., 2024; Ben-Kish et al., 2024; Qiu et al., 2024) have traditionally been tailored for short captions or specific datasets with limited vocabularies (e.g., MS-COCO (Lin et al., 2014)). Notably, many of these methods

*Work done as a Student Researcher at Google DeepMind

[1]Here, we specifically focus on *object-level* caption factuality: whether noun phrases in VLM-generated text are accurate to the original visual input, and vice-versa, whether key objects are reflected in the description.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 19440–19457
November 4-9, 2025

shown in Fig. 1 also rely on ground-truth human annotations, which means they do not work well when such references are unavailable (Petryk et al., 2024), as is often the case when VLMs are deployed at scale. Finally, these metrics often do not capture "recall", the coverage of detailed objects in the caption (Rohrbach et al., 2018; Petryk et al., 2024). This leads to a conflicting incentive for long caption models to output short, conservative text with fewer objects, reducing the risk of a mistake but sacrificing important details (Xue et al., 2024). The simultaneous lack of these attributes in prior metric designs, highlighted in Fig. 1, also inhibits integrations with techniques that potentially *improve* factuality in long caption models. For example, outputs from strong VLM models have been used to generate large-scale pretraining datasets for long captions (Chen et al., 2024b; Awadalla et al., 2024), but these captions can be prone to factuality errors. Automatically filtering this data to ensure higher quality with previous metrics is not possible since this would require human ground truth. Thus, there is a critical need for a unified, reference-free, open-vocabulary method for both measuring and improving factuality in long captions and models.

In this work, we make the following contribution: (i) we introduce OVFact, a new method for measuring and improving open-vocabulary factuality in vision-language models (VLMs) that output long, descriptive captions. Our method leverages open-vocabulary visual grounding and tool-based verification to operate robustly in settings without human annotations. It combines aspects of both precision (minimizing object hallucinations) and recall (ensuring coverage of diverse objects) in the same metric. We validate the combination of these tools with careful analysis and observe that our method improves agreement with human judgments for a range of VLM model outputs, particularly in reference-free settings. (ii) by addressing the combination of limitations in previous metrics, our method design enables new applications towards factuality-based data filtering. We observe that models trained on an OVFact-filtered subset (with 2.5-5x size reduction) of a large-scale, noisy (VLM-generated) pretraining set (Chen et al., 2024b) meaningfully improve factuality precision without sacrificing caption descriptiveness across a range of downstream benchmarks (Onoe et al., 2024; Pont-Tuset et al., 2020; Rohrbach et al., 2018). These findings are consistent across different model scales and are further validated with human evaluation.

2 Related Work

Object-level factuality in VLMs. VLMs, which are natural extensions of LLMs (Team et al., 2024; Touvron et al., 2023; Achiam et al., 2023), exacerbate existing issues with hallucinations by introducing an additional visual modality for potential errors (Sun et al., 2023; Cui et al., 2023; Bai et al., 2024; Bang et al., 2023). A significant line of work (Li et al., 2023; Jiang et al., 2024; Sun et al., 2023; Huang et al., 2024; Guan et al., 2024; Wu et al., 2024; Wiles et al., 2024) relies on question-answering style verification, where an instruction-tuned VLM is expected to answer whether a particular fact about an image is valid. However, such an approach suffers from several limitations: First, the lack of interpretability makes understanding where the model fails challenging. In the case of binary (yes/no) questions, there is a high probability (50%) of the model answering correctly, but for the wrong reason. Second, the QA abilities of the models do not necessarily translate into captioning capabilities (Kaul et al., 2024).

The alternative approach is to check if the facts mentioned in a model's output align with what is shown in the input images. So far, this has been limited to the presence of objects defined in classic captioning datasets with ground-truth annotations. For instance, the seminal work, CHAIR (Rohrbach et al., 2018), is strictly bound to the MS COCO dataset (Lin et al., 2014) as it requires a dataset with paired captions and object-label annotations. CHAIR extracts nouns from a predicted caption using traditional NLP techniques and verifies their presence in an image using ground-truth object annotations. THRONE (Kaul et al., 2024), ALOHa (Petryk et al., 2024), and VALOR-BENCH (Qiu et al., 2024) improve on CHAIR by employing an LLM to parse generated free-form captions. While those methods can handle concepts outside of MS COCO classes, they can only do so to a limited extent. During parsing, THRONE only considers a closed set of objects specifically defined within a dataset (MS COCO or Object365 (Shao et al., 2019)). VALOR-BENCH (Qiu et al., 2024) only operates on a carefully selected small subset of images from Visual Genome (Krishna et al., 2017). ALOHa relaxes CHAIR's string matching with a text similarity between annotated and predicted objects, yet also assumes access to an exhaustive list of human-

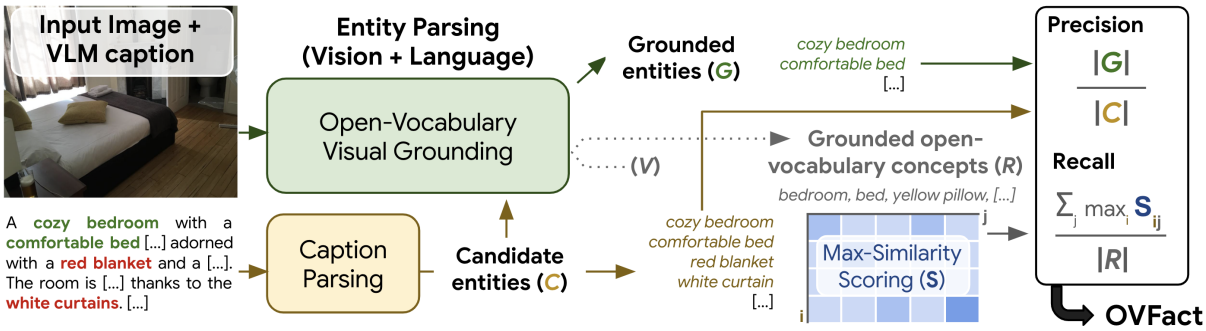

Figure 2: **OVFact overview.** Our *reference-free* method assesses two aspects of long captioning – precision and descriptiveness (recall) – in a unified manner, without requiring ground-truth reference annotations. We process a model output caption into a set of candidate entities $\mathcal{C}$, and assess which subset $\mathcal{G}$ is *groundable* in the input image with *open-vocabulary* detection and segmentation tools. Precision is then measured as a ratio of entities remaining (e.g. above"red blanket" and "white curtains"detected as hallucinated). To measure descriptiveness, we assess a large open-vocabulary concept set $\mathcal{V}$ and identify which concepts are grounded in the image $\mathcal{R}$, then measure their recall within the candidate entities from the VLM $\mathcal{C}$ caption using maximum similarity scoring. Unlike prior work (Petryk et al., 2024; Rohrbach et al., 2018; Kaul et al., 2024), our method can be directly applied to assess factuality in settings where only model-generated caption outputs are available, such as data filtering.

annotated references in the image. In contrast, OpenCHAIR (Ben-Kish et al., 2024) introduces its dataset to measure hallucination by generating images with a text-to-image model (Podell et al., 2023) encompassing concepts different from MS COCO. However, the overall approach still relies on ground-truth references, and since the accompanying captions are in MS COCO style, they are very short and typically describe a single object. In our work, we aim to measure the factuality of captions beyond predefined concepts, datasets, and human annotations, by creating a *reference-free*, open-vocabulary metric that also incorporates both precision and recall. Crucially, our metric being reference-free allows for applicability to filtering of large-scale pretraining datasets, as discussed next.

Data filtering. Data filtering for contrastive image-text pretraining (Radford et al., 2021) on web-scale data has shown it can lead to stronger models and improved training efficiency (Fang et al., 2023; Gadre et al., 2024; Xu et al., 2024; Evans et al., 2024; Udandarao et al., 2024). Meanwhile, data filtering for generative VLMs training has seen less attention, discussing data curation for instruction fine-tuning (Wei et al., 2023; Chen et al., 2024c; Cao et al., 2023). Their primary focus, however, is to improve the instruction-following of the model and not on the factual quality of captioning outputs as we do here. Li et al. (2024) considers data curation for image captioning models by filtering or replacing examples with high training losses on small-scale datasets with human-annotated short captions. In contrast, our proposed metric can operate on large-scale pretraining datasets with noisy (VLM-generated) long captions in terms of their factuality; we show this leads to consistent improvements across a range of measures. Crucially, this application is only possible as our metric is *reference-free* and does not require ground-truth object annotations like prior works (Petryk et al., 2024; Rohrbach et al., 2018; Kaul et al., 2024).

3 Method

We detail OVFact in Sec. 3.1 to *measure* open-vocabulary object factuality and then describe how our proposed metric can be used for data filtering to *improve* a VLM's factuality in Sec. 3.2. Finally, Sec. 3.3 summarizes how OVFact addresses the limitations of existing approaches.

3.1 OVFact

Overview. Our goal is to derive an interpretable, flexible, and reliable method for assessing the factuality of long captions. Because our approach should *generalize to any caption*, regardless of its source dataset or vocabulary, we avoid reliance on dataset-specific terms. We achieve this by verifying the correctness of visual entities and their attributes mentioned in a caption. Specifically, given a pair of image $x \in \mathbb{R}^{H \times W \times 3}$ and caption y, we first parse y to extract a set of candidate entities. We then verify their presence in the image x by running image grounding tools. To avoid promoting non-descriptive captions, we design a recall-based metric that compares candidate entities against reference entities extracted either from ground-truth captions (when reliable) or from grounding tools with a large vocabulary of concepts. The high-level overview of our approach is presented in Fig. 2.

Caption parsing. We start by parsing y to extract candidate entities. To do so, we prompt an LLM to generate a list of objects mentioned in y by specifically instructing an LLM to output all objects with their visual attributes, ignoring abstract concepts with no visual presence, such as *sound* and *atmosphere*, often present in long VLM descriptions. We provide the prompt details in the Appendix A.6. We denote the resulting candidate entity set as $\mathcal{C}$.

Entity grounding. Having obtained the set $\mathcal{C}$, we then validate the presence of each one of the candidate entities $c_i \in \mathcal{C}$ within image y. Note that c is a free-form text; thus, verifying the presence of associated concepts with no prior access to a vocabulary poses a challenge. To address this, we first utilize a state-of-the-art open-vocabulary object detector (Minderer et al., 2024). Specifically, we feed each c_i to a detection model as a separate query. We then define the candidate entity, c_i, as being grounded if its detection confidence score exceeds a threshold value. This yields an initial set of grounded entities, $\mathcal{G}_\mathcal{D}$, where $\mathcal{G}_\mathcal{D} \subset \mathcal{C}$.

In our empirical studies, we observed that long image captions often include much more details about surroundings and "stuff-like" (Kirillov et al., 2019; Forsyth et al., 1996) concepts (e.g., *water*, *sky*, *concrete*) which object detectors typically miss. Thus, we incorporate an *additional* grounding step using an open-vocabulary semantic segmentation model. Each candidate entity c_i is input to a segmentation model to obtain a grounded set $\mathcal{G}_\mathcal{S} \subset \mathcal{C}$. Our final resulting set of all grounded entities is then $\mathcal{G} = \mathcal{G}_\mathcal{D} \cup \mathcal{G}_\mathcal{S}$.

OVFact Scoring. We leverage the parsing and grounding outputs to measure the overall factuality of caption y. Focusing on the specific challenges of long captioning, we assess two key aspects: precision and descriptiveness. While prior metrics (e.g., Rohrbach et al., 2018) are more precision-focused, this is not as suitable for our setting, as this rewards short captions with fewer potential mistakes but few details. By unifying both aspects (in a reference-free manner), we can also apply our single metric for downstream applications (e.g. filtering).

Calculating Precision. We define the precision of a caption as the ratio between the count of the grounded entities $\mathcal{G}$ and candidate entities $\mathcal{C}$:

$$\text{OVFact}_{Prec} = \frac{|\mathcal{G}|}{|\mathcal{C}|} \tag{1}$$

Calculating Recall. To measure descriptiveness, we design an additional metric interpreted as traditional recall. We first obtain an approximate set of entities appearing in y. In an ideal scenario, i.e. when considering fully annotated datasets like MS COCO (Lin et al., 2014), one could use object-level annotations in the dataset as references $\mathcal{R}$. However, MS COCO is the only dataset to date with both human-annotated captions and object detection labels, although the captions are very short and thus not descriptive.

Assuming ground-truth captions are human-annotated and give an exhaustive description of a scene, a possible solution to this problem is extracting $\mathcal{R}$ from ground-truth captions by employing the parsing introduced before, which we propose in our approach. However, if the ground-truth captions are unreliable, e.g. for VLM-generated datasets, we also consider a general case to obtain $\mathcal{R}$ by prompting the grounding tools given a large enough vocabulary of concepts $\mathcal{V}$.

We then measure the recall of reference entities $\mathcal{C}$ in $\mathcal{R}$. To facilitate the open-vocabulary aspect of our approach (as entities in both sets are free-form texts), we first compute text embeddings for each $c_i \in \mathcal{C}$ and for all $r_j \in \mathcal{R}$; $f(c_i) \rightarrow \vec{c_i} \in \mathbb{R}^D$ and $f(r_j) \rightarrow \vec{r_j} \in \mathbb{R}^D$. We then calculate the similarity between feature representations from the two sets of entities with the cosine similarity $S_{ij} = \frac{\vec{r_j} \cdot \vec{c_i}}{|\vec{r_j}||\vec{c_i}|}$. We define recall by selecting a maximum similarity score for each reference entity and report as a final score an average of all $r_j \in \mathcal{R}$, that is:

$$\text{OVFact}_{Rec} = \frac{1}{|\mathcal{R}|} \sum_{j=1}^{|\mathcal{R}|} \max_{i=1}^{|\mathcal{C}|} S_{ij} \tag{2}$$

Final metric. Finally, to encompass both the precision and descriptiveness of y into one metric (our full OVFact), we calculate our unified F1 score:

$$\text{OVFact}_{F1} = \frac{2 \cdot \text{OVFact}_{Prec} \cdot \text{OVFact}_{Rec}}{\text{OVFact}_{Prec} + \text{OVFact}_{Rec}} \tag{3}$$

3.2 OVFact for Data Filtering

The generalized case of OVFact (Fig. 2) can be completely reference-free. This is particularly beneficial, as it can be applied to any set of image and caption pairs and is not limited to only clean, fully-labelled datasets. This becomes particularly important when considering a growing number of LLM-generated caption datasets (Chen et al., 2024b; Arai et al., 2025). One important application of OVFact is data filtering, where our metric serves as a

scoring function for pruning incorrect samples, e.g. including hallucinated objects.

Consider a dataset of (image x, caption y) pairs, where y is generated at scale using a large VLM, and this data is intended to help train other models, as done in (Chen et al., 2024b). Using the process described in previous sections, our data filtering approach consists of extracting OVFact$_{F1}$ scores for each (x, y) pair. Because our method is *reference-free*, we do not need additional human-generated ground-truth object labels. We then sort the pairs based on their OVFact$_{F1}$ and select the top X% depending on the assumed data pruning ratio. This simple technique can result in a significantly improved performance in factuality for long captioning models without compromising descriptiveness. We demonstrate it experimentally in Sec. 4.3.

3.3 Discussion

Having introduced OVFact, we now revisit closely related work to highlight key differences.

CHAIR (Rohrbach et al., 2018) is a popular metric to measure hallucinations, but is strictly bound to the COCO dataset, both for the instance-level CHAIRi$= \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all objects mentioned}\}|}$ and caption-level CHAIRs $= \frac{|\{\text{sentences with hallucinated object}\}|}{|\{\text{all sentences}\}|}$. Moreover, CHAIR can be artificially improved by simply not predicting any objects. This occurs when a caption genuinely contains no objects or when predicted entities fall outside the COCO vocabulary.

THRONE (Kaul et al., 2024) improves CHAIR's string matching with LLM-based parsing, which can handle free-form captions. However, the parsing *output* is still limited to categories defined in the evaluation set (the LLM is prompted about a predefined list of objects). While THRONE could be extended to datasets other than MS COCO, it would require annotated object detection labels for the final score, rendering it unsuitable for reference-free or object-annotation-free setups. In contrast to CHAIR, THRONE assesses both precision and recall of a caption as we do in this work.

ALOHa (Petryk et al., 2024) also addresses the closed-vocabulary limitation of CHAIR by using similarity scores between text embeddings of reference and candidate entities. As in our approach, a language model is used to parse the candidate entities. However, the reference set is constructed from ground-truth reference captions and additional object detections (Carion et al., 2020). In contrast to our approach, ALOHa performs Hungarian matching (Kuhn, 1955) on the similarity matrix, S, between candidates and references. This difference is significant because Hungarian matching enforces a one-to-one correspondence between candidates and references and is therefore particularly sensitive to having an exhaustive and precise list of references, as illustrated in Fig. 3: If there are not enough references to match with the candidates (i.e., $|\mathcal{R}| < |\mathcal{C}|$), ALOHa will ignore the unmatched candidates, leading to a misleading score. Moreover, when $\mathcal{R}$ is incomplete, the one-to-one matching will force associations between unrelated concepts (in Fig. 3, "sunglasses" is forced to match the reference "cell phone"). Finally, ALOHa favors shorter captions, as fewer candidates have a higher chance of being correctly matched to their references. These limitations of ALOHa are particularly evident when using it as a metric for data filtering, which we show experimentally in the next section.

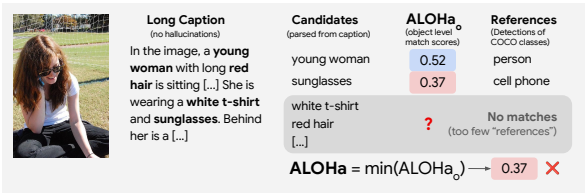

Figure 3: **Limitations of ALOHa in a reference-free setting.** We present an example from ShareGPT4v (i.e., with no access to ground-truth references). ALOHa's reference matching forces comparisons of detected classes with what has been described, an issue exacerbated in a reference-free setting when ground-truth human annotations of objects are unavailable. Our method is more robust to such cases.

4 Experiments

In this section, we first describe our overall experimental setup and technical details in Sec. 4.1. Next, in Sec. 4.2, we discuss the verification of OVFact as a method for *measuring* factuality, including ablation studies of its components. Finally, in Sec. 4.3, we present results for *improving* factuality, applying OVFact for data selection.

4.1 Experimental setup

Implementation details. We use the state-of-the-art open-source LLM Gemma2-27b (Team et al., 2024) for caption parsing and OWL-ViTv2 (Minderer et al., 2024) and OpenSeg (Ghiasi et al., 2022) for entity grounding. We consider an extensive vocabulary $\mathcal{V}$ of open-vocabulary concepts (as discussed in Sec. 3.1). Specifically, we take the union of concepts from standard datasets, i.e. Visual

	DOCCI	Loc. Nar.
Gemma parsing	0.97	0.97
CHAIR parsing	0.02	0.28
POS + CHAIR	0.79	0.58
Entity grounding	0.72	0.79
w/o detection	0.46	0.67
w/o segmentation	0.62	0.64

Table 1: **Validation of OVFact components.** We measure the specificity of each of the steps in our method on 100 DOCCI *qual dev* split and 100 random samples from Localized Narratives *train* set. See Sec. 4.2.

	ALOHa	OVFact$_{Prec}$	OVFact$_{Rec}$
Human agreement	48.6%	72.1%	80.7%

Table 2: **OVFact improves alignment with human judgments.** We extend the analysis by Onoe et al. (2024) on DOCCI over 4 VLMs, observing that OVFact$_{Prec}$ shows higher agreement with human judgment compared to prior work (ALOHa). We highlight that (1) OVFact is *reference-free*, while ALOHa here is given references from ground-truth captions, and (2) our method also captures recall (descriptiveness) and shows high agreement with human judgments (OVFact$_{Rec}$). See Appendix A.1 for further details.

Genome (Krishna et al., 2017), LVIS (Gupta et al., 2019), Open Images (Kuznetsova et al., 2020) and Objects365 (Shao et al., 2019), resulting in a total of 2792 unique concepts. To encode entities for OVFact$_{Rec}$ computation, we use the SigLIP-So400m/14 (Zhai et al., 2023) text encoder. Additional details are in Appendix A.5.

Datasets and Evaluation metrics. To validate the effectiveness of our approach, we conduct experiments on multiple captioning datasets with different levels of complexity.

• **ShareGPT4v** (Chen et al., 2024b) is a large dataset of automatically generated long captions by GPT4v (Achiam et al., 2023). It consists of approximately 100k samples of an average length of 950 characters. We use the ShareGPT4v dataset as the primary *training* dataset for data selection experiments as the VLM-generated captions are noisy, but widely adopted for training (Liu et al., 2024; Lu et al., 2024; Tong et al., 2024).

• **DOCCI** (Onoe et al., 2024) is a recent human-annotated dataset of detailed, high-quality captions (average length of 642 characters), with images sourced from voluntarily contributed, private archives (guaranteed to be no overlap with VLM training sets). Its caption length and diversity mean it is the most challenging published dataset for long, detailed captions. We select its *test* set with 5k samples as our downstream evaluation set.

• **Localized Narratives** (Pont-Tuset et al., 2020) consists of long, human-annotated captions which average 206 characters. We report the downstream evaluation of our method on the COCO *validation* split to complement our analysis further.

• **MS COCO** (Lin et al., 2014) is the standard for measuring hallucinations as it includes bounding box annotations for 80 different object classes. However, captions are short (only 52 characters on average). Following standard practice (Rohrbach et al., 2018; Kaul et al., 2024; Petryk et al., 2024), we report CHAIR on the *val* split to show generalization of our method to prior, standard settings.

4.2 OVFact for Measuring Factuality

To assess the reliability of our proposed method, we validate its performance on manually annotated subsets from the DOCCI *qual dev* split and 100 randomly sampled image-caption pairs from the Localized Narratives *train* split. We manually annotate entities in ground-truth captions and then compute and measure the performance of each step in our approach, which we report in Tab. 1. First, we measure the specificity of the caption parsing stage, which is measured as the ratio of entities extracted by Gemma to annotated entities. We also show an ablation of Gemma parsing by comparing it to string matching on the limited CHAIR vocabulary (CHAIR parsing), as well as to a variant of string matching enhanced with POS Tagging (Honnibal et al., 2020). Our caption parsing demonstrates robust performance, achieving a specificity of 97% on both datasets.

We further analyze the specificity of the entity grounding component discussed in Sec. 3.1, calculated as the ratio of correctly grounded entities to all annotated entities in the ground-truth captions. Overall, we observe a difference in the specificity of entity grounding between the two considered datasets, most likely because the DOCCI descriptions are much more detailed, with numerous examples of specific objects. We also analyze in Tab. 1 separate components of our grounding stage and conclude that the use of object detection and segmentation proves crucial for both datasets, producing a specificity improvement of up to 0.26 in DOCCI compared to the use of segmentation alone.

Finally, in Tab. 2, we extend the human study analysis from Onoe et al. (2024) over four state-of-the-art long caption models applied to the DOCCI dataset and observe that our model aligns well with human preferences for both OVFact$_{Prec}$

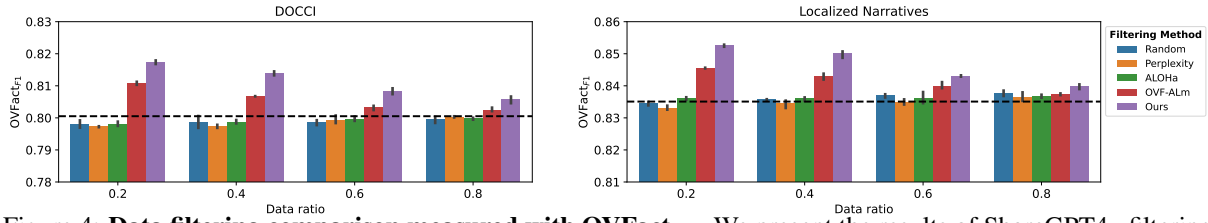

Figure 4: **Data filtering comparison measured with OVFact**$_{F1}$. We present the results of ShareGPT4v filtering for different data ratios, with *zero-shot evaluations* on DOCCI *test* and Localized Narratives *val* sets averaged over 3 training runs. '---' indicates model trained on *full* ShareGPT4v dataset. OVFact significantly improves (↑) on factuality over prior work, particularly as we filter more examples (resulting in a smaller, higher-quality training set). We also report an OVF-ALm baseline, which shows the impact of using ALOHa's (brittle) matching with our OVFact method. We highlight our gains in factuality come without compromising descriptiveness or quality.

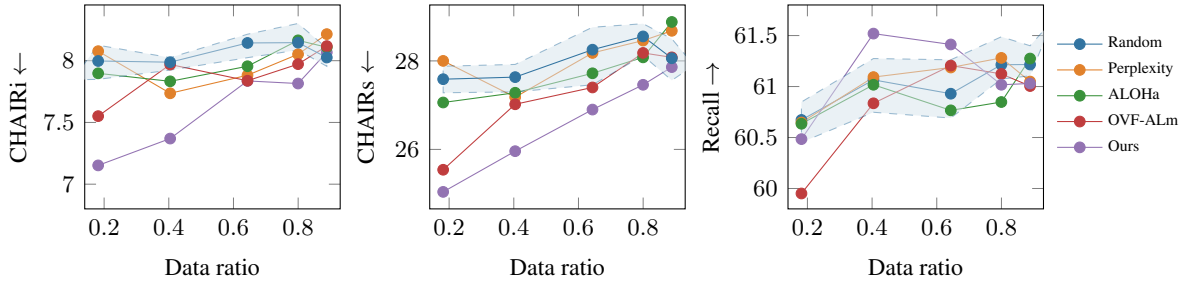

Figure 5: **Generalization to CHAIR benchmark.** We also measure zero-shot evaluation on the CHAIR benchmark, highlighting how our filtering improves on traditional metrics and settings too. Light-blue shading marks variance of the *random* baseline. Our filtering method results in fewer hallucinations at the instance and sentence level (CHAIRi, CHAIRs; **lower** ↓ is better), maintaining or improving recall (**higher** ↑ is better) of generated captions.

and OVFact$_{Rec}$. We also compare the results of ALOHa in the original setup with ground-truth references and observe that OVFact$_{Prec}$ is more aligned with human judgments compared to ALOHa. We provide details and results of this analysis in Appendix A.1.

4.3 OVFact for Improving Factuality

Experimental setup. For our data filtering experiments, we consider how OVFact can be used to score and filter a large-scale, noisy training dataset with generated VLM captions (ShareGPTv, Sec. 4.1), selecting the top $X\%$ with the highest OVFact. We validate our approach by considering how supervised fine-tuning on these filtered sets can improve a representative open-source state-of-the-art VLM, PaliGemma 2 (Steiner et al., 2024), on downstream zero-shot long caption evaluations (details Sec. 4.1). We report here the results with PaliGemma 2 (3B) model, and discuss results with additional sizes in Appendix A.2 and give more training details in Appendix A.5.

Filtering Baselines. To our knowledge, OVFact is the first open-vocabulary, reference-free method for measuring factuality in long captions, and prior methods that require ground-truth human references inherently cannot work in this data filtering setting, where only the noisy VLM-generated captions are available. Nonetheless, in addition to **Random** filtering, we consider the following:

• **Perplexity**: we obtain a perplexity score predicted by the base, pre-trained PaliGemma 2 model for each sample. We then select samples with the lowest perplexity score and discard high-loss samples similarly to Li et al. (2024).

• **ALOHa**: we implement a reference-free variation of ALOHa, where we only rely on a set of detected objects for each image. For consistent comparison, we apply the same base models/tools as in OVFact (Gemma 2, etc.) with vocabulary, matching method, etc. settings from the original paper. Samples with the highest ALOHa "caption-level" score (Petryk et al., 2024) are selected for training.

• **OVFact + ALOHa matching (OVF-ALm)**: to better measure the limitations of ALOHa's matching in our reference-free setting (see Sec. 3.3), we introduce a stronger hybrid baseline OVF-ALm, where initial parsing and grounding are with our OVFact full open-vocabulary concept set $\mathcal{V} \rightarrow \mathcal{R}$, and we apply ALOHa's matching algorithm at the end to obtain the final caption-level score for selection.

Results. Fig. 4 presents the results of applying various data filtering methods and their down-

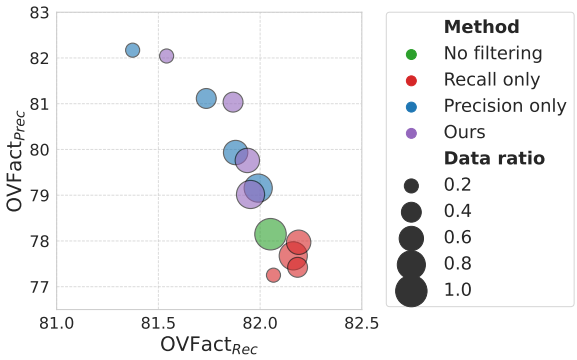

Figure 6: **Data-filtering ablation study**. We train PaliGemma 2 on filtered subsets of ShareGPT4V (with ablations of our method), and evaluate models zero-shot on the DOCCI test set. Note, the top-right corner is the best. Filtering with *full* OVFact (purple) yields models with the best trade-off between precise and descriptive captions.

stream performance on DOCCI and Localized Narratives datasets, evaluated using our proposed metric (OVFact$_{F1}$). Results are averaged over three independent training runs for each method. The dashed line indicates the performance of a model trained on the full ShareGPT4v dataset. Our filtering method consistently outperforms all other methods across all data ratios on both datasets, with the most pronounced improvements observed at lower data ratios, suggesting that the original dataset contains a large amount of noise. Interestingly, both perplexity- and ALOHa-filtering baselines perform comparably to random sampling. ALOHa's strict reliance on reference data makes it unsuitable for data filtering. Our strong hybrid baseline OVFALm gives a positive signal for filtering yet still underperforms our proposed approach due to inherent limitations (Sec. 3.3).

We also analyze the impact of our data selection method on the traditional CHAIR setting to assess generalization. Fig. 5 presents the performance of models trained on varying sizes of ShareGPT4v data, evaluated on both (sentence) CHAIRs and (instance) CHAIRi in a zero-shot manner. In addition, we report recall following standard protocols in prior work (Favero et al., 2024). Our method achieves the best results on CHAIR, particularly when using smaller training datasets. Notably, improved hallucination scores are accompanied by comparable or even higher recall values, demonstrating the effectiveness of our method in balancing the precision and recall of generated captions.

We further analyze the impact of our data selection method on the performance of VLMs with varying parameter sizes and observe that our data selection consistently improves performance across all model sizes, as detailed in Appendix A.2.

Ablation studies. We study other variants of our filtering method in Fig. 6, where we present a trade-off between OVFact$_{Prec}$ and OVFact$_{Rec}$ measured with OVFact when evaluating zero-shot on the DOCCI test set. In particular, we experiment with *Precision only* approach, where we run our data selection process based on OVFact$_{Prec}$ of original captions in ShareGPT4v. Similarly, we apply the same process but with OVFact$_{Rec}$, which we denote *Recall only*. We compare all the variants against the full ShareGPT4v dataset (*No filtering*). First, we notice that *Recall only* approach results in a slight improvement on OVFact$_{Rec}$ over full data training, yet clearly outperforms other variants in terms of OVFact$_{Prec}$. On the other hand, *Precision only* filtering decreases OVFact$_{Rec}$ with the decrease in training data size. Finally, filtering with OVFact$_{F1}$(full OVFact) achieves a sweet spot between the two aspects, with a model trained with 40% data improving OVFact$_{Prec}$ by 3 points with respect to the model trained with the entire dataset, yet staying almost on par in terms of OVFact$_{Rec}$.

Qualitative examples. We include qualitative analysis of our method in the Appendix A.3, including comparative analysis of both *filtered examples* from ShareGPT4V and of the outputs of the *models* trained on our higher-quality filtered set.

Human evaluation. Since traditional caption metrics (BLEU, CIDEr, etc.) are not well-suited for assessing general quality for long captions, as noted by (Onoe et al., 2024), we also perform a side-by-side comparison of captions generated by a model trained on our higher-quality filtered set (the top 20%, ranked by our method) compared with one trained on the *full* dataset. Our results indicate that humans preferred the outputs from the model trained on the filtered subset **68.2%** of times, even with 5x less training data (see Appendix A.4).

5 Conclusions

We introduced OVFact, a novel method for *measuring* caption factuality that leverages open-vocabulary visual grounding and tool-based verification. Unlike previous metrics, OVFact is not dataset-specific and effectively evaluates the factuality of long, descriptive captions. We further demonstrated how OVFact can be used to *improve* the factuality of VLMs by using it as a metric for

data filtering. We show that models trained on an OVFact-filtered subset (2.5-5x smaller) of a large-scale, noisy (VLM-generated) pretraining set meaningfully improved factuality precision without sacrificing caption descriptiveness when evaluating zero-shot on various downstream datasets.

6 Limitations & Future Work

Potential risks and considerations for broader impacts. Large Vision-Language Models present several significant challenges and risks that require careful consideration (Bommasani et al., 2021; Mitchell et al., 2019; Gebru et al., 2021). These models can potentially propagate harmful biases present in training data (Howard et al., 2024). From a technical perspective, VLMs may hallucinate details or generate false visual interpretations, which could be problematic in high-stakes, safety-critical applications like medical imaging or autonomous systems (Chen et al., 2024a; Rohrbach et al., 2018). Our work is a step towards both measuring and improving the factuality of VLMs. We believe that an automated way to detect and improve hallucinations is an important challenge with a growing amount of AI-generated content and datasets, and that a version of our method can be helpful for settings where AI is deployed at scale, and it would be necessary to assess the factuality of VLM outputs in real-time as part of a larger system. However, our work, as it is presented here, should be considered an academic exploration into this direction, and extensions for deployment settings will require additional work to ensure proper mitigation of potential risks is in place. We detail additional aspects for consideration for both the "measurement" and "improvement" aspects of our work:

Measuring Factuality. Our method is motivated by the limitations of prior work, and succeeds in improving several aspects (e.g., reducing the dependence on ground-truth references noted in Petryk et al. (2024), etc.). However, our OVFact method operates by leveraging base vision and language models as tools, and while we worked to reduce the impact of potential errors in individual tools by composing them together (e.g., combining both detection and segmentation models to leverage their relative strengths), our method inherits and remains sensitive to their intrinsic limitations. We characterize the quantitative performance of these components in Sec. 4.2, but to expand, we observe that challenging visual inputs (e.g., with heavily occluded, very fine-grained, or distant objects or parts) would often prove difficult for tools to properly ground. We anticipate that future models for these tasks, with further refinements in pre-training data and architecture designs, will help to alleviate this, as these are an active area of research (Myers-Dean et al., 2024). Similarly, the pre-training domain for these tools is important: while our datasets were on general images, to apply our general framework to specialized domains (e.g., medical image analysis (Chen et al., 2024a)) more bespoke models will be necessary. Further, the pre-training data for these models will be potential sources of bias inherited by the general framework, and will need to be considered for deployment settings. Finally, we also consider here the limitations in the definition and scope of "factuality" we consider here: as noted in the footnote on the first page, our focus was on *object-level* caption factuality, whether noun phrases in VLM-generated text were both accurate to (precision) and descriptive of (recall) the original visual input. We believe expanding this scope to include additional attributes (e.g., described inter-object relationships and dependencies) and domains (e.g., verbs, motion, adverbial phrases as particularly important in videos, is an exciting direction for future work. Similarly, investigating factuality in the context of external databases of multimodal visual knowledge (e.g., Mensink et al., 2023) would be an interesting direction for future work.

Improving Factuality. We investigate one potential direction to improve factuality: by leveraging our new reference-free, open-vocabulary method, we can apply our method to filtering large-scale pretraining data for a set widely adopted by the community (Liu et al., 2024; Lu et al., 2024; Tong et al., 2024), showing our models achieve substantial gains on factuality and quality, with 2.5 - 5x less training data. However, while this filtered data is higher-quality, our method would need to be integrated as part of a larger framework to facilitate further corrections (e.g., by human annotators) to edit and further refine the captions themselves – this would be an interesting direction for future work. Filtering data has been recently shown in other contexts (contrastive vision-language models, e.g., for CLIP-style models for image retrieval) to unintentionally reduce accuracy in low-resource domains, e.g., in multilingual and multicultural contexts not well-represented by existing (largely English-centric) multimodal evaluations (Pouget

et al., 2024). Exploring how our method could potentially complement this analysis in long captioning settings would be an interesting direction for further study. Finally, there is a growing body of literature on measuring factuality of generative visual models (e.g., Wiles et al., 2025; Lee et al., 2024); while our focus is on long caption generation, these models generate visual data (images, videos) with similar base architectural designs. We believe that exploring connections between our method and related work in that space could prove to be a fruitful direction.

Acknowledgements

We want to thank Yasumasa Onoe for insightful discussions and for sharing the results of human evaluations of DOCCI. We would also like to thank Emanuele Bugliarello for his feedback on the manuscript. This work was partially supported by the National Centre of Science (Poland) Grant No. 2022/45/B/ST6/02817.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Hidehisa Arai, Keita Miwa, Kento Sasaki, Kohei Watanabe, Yu Yamaguchi, Shunsuke Aoki, and Issei Yamamoto. 2025. Covla: Comprehensive vision-language-action dataset for autonomous driving. In *WACV*.

Anas Awadalla, Le Xue, Manli Shu, An Yan, Jun Wang, Senthil Purushwalkam, Sheng Shen, Hannah Lee, Oscar Lo, Jae Sung Park, Etash Guha, Silvio Savarese, Ludwig Schmidt, Yejin Choi, Caiming Xiong, and Ran Xu. 2024. Blip3-kale: Knowledge augmented large-scale dense captions. *Preprint*, arXiv:2411.07461.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *IJCNLP-AACL 2023*.

Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, and Hadar Averbuch-Elor. 2024. Mitigating open-vocabulary caption hallucinations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. ACL.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. 2024. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726*.

Lucas Beyer, Xiaohua Zhai, and Alexander Kolesnikov. 2022. Big vision. https://github.com/google-research/big_vision.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. 2018. JAX: composable transformations of Python+NumPy programs.

Yihan Cao, Yanbin Kang, Chi Wang, and Lichao Sun. 2023. Instruction mining: Instruction data selection for tuning large language models. *arXiv preprint arXiv:2307.06290*.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *ECCV*.

Jiawei Chen, Dingkang Yang, Tong Wu, Yue Jiang, Xiaolu Hou, Mingcheng Li, Shunli Wang, Dongling Xiao, Ke Li, and Lihua Zhang. 2024a. Detecting and evaluating medical hallucinations in large vision language models. *arXiv preprint arXiv:2406.10185*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024b. Sharegpt4v: Improving large multi-modal models with better captions. In *ECCV*.

Ruibo Chen, Yihan Wu, Lichang Chen, Guodong Liu, Qi He, Tianyi Xiong, Chenxi Liu, Junfeng Guo, and Heng Huang. 2024c. Your vision-language model itself is a strong filter: Towards high-quality instruction tuning with data selection. In *ACL*.

Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil

Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. 2023. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*.

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Talfan Evans, Nikhil Parthasarathy, Hamza Merzic, and Olivier J Henaff. 2024. Data curation via joint example selection further accelerates multimodal learning. *arXiv preprint arXiv:2406.17711*.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. 2023. Data filtering networks. *arXiv preprint arXiv:2309.17425*.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *CVPR*.

David A Forsyth, Jitendra Malik, Margaret M Fleck, Hayit Greenspan, Thomas Leung, Serge Belongie, Chad Carson, and Chris Bregler. 1996. Finding pictures of objects in large collections of images. In *ECCV*.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2024. Datacomp: In search of the next generation of multimodal datasets. *NeurIPS*.

Yunhao Ge, Xiaohui Zeng, Jacob Samuel Huffman, Tsung-Yi Lin, Ming-Yu Liu, and Yin Cui. 2024. Visual fact checker: enabling high-fidelity detailed caption generation. In *CVPR*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*.

Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. 2022. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*.

Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Phillip Howard, Avinash Madasu, Tiep Le, Gustavo Lujan Moreno, Anahita Bhiwandiwalla, and Vasudev Lal. 2024. Socialcounterfactuals: Probing and mitigating intersectional social biases in vision-language models with counterfactual examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11975–11985.

Kung-Hsiang Huang, Mingyang Zhou, Hou Pong Chan, Yi Fung, Zhenhailong Wang, Lingyu Zhang, Shih-Fu Chang, and Heng Ji. 2024. Do LVLMs understand charts? analyzing and correcting factual errors in chart captioning. In *ACL*.

Chaoya Jiang, Wei Ye, Mengfan Dong, Jia Hongrui, Haiyang Xu, Ming Yan, Ji Zhang, and Shikun Zhang. 2024. Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models. In *ACM Multimedia 2024*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, CJ Taylor, and Stefano Soatto. 2024. Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models. *CVPR*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR*.

Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. 2019. Panoptic segmentation. In *CVPR*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*.

Harold W. Kuhn. 1955. *Naval Research Logistics Quarterly*.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov,

et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. 2024. Holistic evaluation of text-to-image models. *NeurIPS*.

Wenyan Li, Jonas Lotz, Chen Qiu, and Desmond Elliott. 2024. The role of data curation in image captioning. In *ACL*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *EMNLP*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *CVPR*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *NeurIPS*.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. 2023. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. In *ICCV*.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. 2024. Scaling open-vocabulary object detection. *NeurIPS*.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*.

Josh Myers-Dean, Jarek Reynolds, Brian Price, Yifei Fan, and Danna Gurari. 2024. Spin: Hierarchical segmentation with subpart granularity in natural images. In *European Conference on Computer Vision*.

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*.

Suzanne Petryk, David M Chan, Anish Kachinthaya, Haodi Zou, John Canny, Joseph E Gonzalez, and Trevor Darrell. 2024. Aloha: A new measure for hallucination in captioning models. In *ACL*.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *ECCV*.

Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas Peter Steiner, Xiaohua Zhai, and Ibrahim Alabdulmohsin. 2024. No filter: Cultural and socioeconomic diversityin contrastive vision-language models. *NeurIPS*.

Haoyi Qiu, Wenbo Hu, Zi-Yi Dou, and Nanyun Peng. 2024. Valor-eval: Holistic coverage and faithfulness evaluation of large vision-language models. In *ACL*.

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 *ICML*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *EMNLP*.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Jing Li, Xiangyu Zhang, and Jian Sun. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*.

Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al. 2024. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *NeurIPS*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vishaal Udandarao, Nikhil Parthasarathy, Muhammad Ferjad Naeem, Talfan Evans, Samuel Albanie, Federico Tombari, Yongqin Xian, Alessio Tonioni, and Olivier J Hénaff. 2024. Active data curation effectively distills large-scale multimodal models. *arXiv preprint arXiv:2411.18674*.

Lai Wei, Zihao Jiang, Weiran Huang, and Lichao Sun. 2023. Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4. *arXiv preprint arXiv:2308.12067*.

Olivia Wiles, Chuhan Zhang, Isabela Albuquerque, Ivana Kajić, Su Wang, Emanuele Bugliarello, Yasumasa Onoe, Chris Knutsen, Cyrus Rashtchian, Jordi Pont-Tuset, et al. 2024. Revisiting text-to-image evaluation with gecko: On metrics, prompts, and human ratings. *arXiv preprint arXiv:2404.16820*.

Olivia Wiles, Chuhan Zhang, Isabela Albuquerque, Ivana Kajic, Su Wang, Emanuele Bugliarello, Yasumasa Onoe, Pinelopi Papalampidi, Ira Ktena, Christopher Knutsen, Cyrus Rashtchian, Anant Nawalgaria, Jordi Pont-Tuset, and Aida Nematzadeh. 2025. One slice is not enough: In search of stable conclusions in text-to-image evaluation. In *ICLR*.

Xiyang Wu, Tianrui Guan, Dianqi Li, Shuaiyi Huang, Xiaoyu Liu, Xijun Wang, Ruiqi Xian, Abhinav Shrivastava, Furong Huang, Jordan Lee Boyd-Graber, Tianyi Zhou, and Dinesh Manocha. 2024. AutoHallusion: Automatic generation of hallucination benchmarks for vision-language models. In *EMNLP*.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. 2024. Demystifying CLIP data. *ICLR*.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. 2024. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *ICCV*.

A Appendix

A.1 OVFact and Human Alignment

We include here the details and the results of the human alignment analysis. We extend the analysis of human studies in DOCCI (Onoe et al., 2024). We compare 4 models: InstructBLIP (Dai et al., 2023), LLaVA-1.5 (Liu et al., 2023), GPT4v (Achiam et al., 2023) and PaLI-5B (Chen et al., 2023) fine-tuned on DOCCI. We conduct side-by-side (S×S) comparisons between all 6 pairs of models (the original study focused on 3 of these), where participants are asked to provide a preference between model A or model B in two areas: Precision and Descriptiveness (Recall). We randomly swap models while displaying them to ensure that there is no side bias. Annotations are made by four different annotators, and each sample has at least two judgments. Participants select one of the three options: *A is better*, *Neutral*, *B is better*. We then compare for each sample whether human judgments for Recall correspond to OVFact_{Rec} (OVFact_{Rec} (model A) > OVFact_{Rec} (model B)), and the same for OVFact_{Prec}, whether they correspond to human preference on the Precision scale. In other words, if we present outputs from two models A and B (randomly shuffled for each $A \times B$ pair), does our metric (the difference between OVFact(A) and OVFact(B)) match with what humans prefer, for precision and recall?

Additionally, we compare with other baselines, particularly ALOHa, against Human preference in Precision, though notably we provide ALOHa access to references parsed from human-annotated ground-truth captions (not available to our reference-free metric) to make it a strong comparison point. We also consider the CLIP-Image-Score proposed in (Ge et al., 2024), which measures factuality with an indirect approach: an image generation model is applied to the output caption under consideration, and CLIP similarity is measured between the original image and this generated one – we implement this metric with the same image generation model (Stable Diffusion XL) as in the original paper. Tab. A1 details the results, which complement Tab. 2 from the main paper. We present S×S comparisons between all combinations of pairs of discussed VLMs, and observe consistent improvements throughout the range of output styles and capabilities, indicating the general applicability of our method. We highlight that our method, with its grounding and explicit precision/recall measures, matches substantially better with human judgment than both ALOHa and CLIP-Image-Score[2].

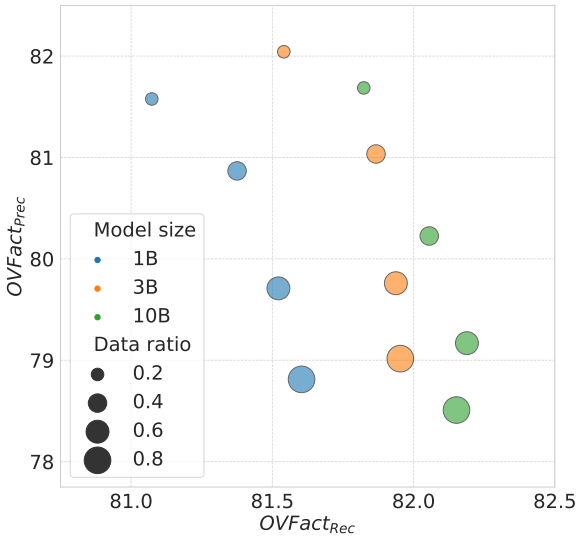

Figure A1: **Model size study** as a Precision-Recall curve. We analyze the impact of applying our data filtering strategy to various PaliGemma2 model sizes on DOCCI *test* measured with OVFact. Note that the best-performing models are closest to the top-right corner. Our method consistently improves models' performance across different model sizes.

A.2 Model size study

In the main paper, our analysis was primarily on the 3B version of PaliGemma 2. We analyze the impact of our data selection method on the performance of Vision-Language Models (VLMs) with varying parameter sizes. Figure A1 presents the trade-off between OVFact_{Prec} and OVFact_{Rec} measured on the DOCCI test set for several PaliGemma2 variants. Across all model sizes, our data selection method improves performance. As the amount of training data increases, we observe a consistent trend: precision decreases while average similarity (OVFact_{Rec}) increases. This trade-off is most pronounced for the smallest model, suggesting the importance of balancing precision and recall, especially for models with limited capacity. Interestingly, the largest (10B parameter) model consistently achieves slightly higher OVFact_{Rec} scores than smaller models, but often at the cost of slightly lower OVFact_{Prec}.

[2]Image generation models have known issues with *faithful* generation for long input prompts (Wiles et al., 2025) and these results suggest this kind of indirect, multi-stage generation approach may not (yet) be well-suited for the large-scale filtering applications explored here.

	Agreement rate % with Human Judgments				
	Precision			*Recall*	
S×S model comparison	ALOHa	CLIP-Image-Score	OVFact$_{Prec}$	CLIP-Image-Score	OVFact$_{Rec}$
GPT4v × LLaVA-1.5	55.6	66.7	88.9	75.0	91.7
GPT4v × PaLI-5B	46.2	38.5	53.8	60.0	73.3
GPT4v × InstructBLIP	25.0	58.3	50.0	53.3	80.0
LLaVA-1.5 × PaLI-5B	58.8	58.8	82.3	70.0	90.0
LLaVA-1.5 × InstructBLIP	50.0	40.0	70.0	25.0	75.0
InstructBLIP × PaLI-5B	54.5	54.5	90.9	30.0	69.2
Overall (Tab. 2)	48.6	52.7	72.1	55.4	80.7

Table A1: **OVFact improves agreement with side-by-side (S×S) human evaluations**. Here, we show the extended analysis from Tab. 2. We extend the initial human study from Onoe et al. (2024), which compared the outputs of a fine-tuned PaLI-5B (Chen et al., 2023) against each of 3 other state-of-the-art models (GPT4v (Achiam et al., 2023), InstructBLIP (Dai et al., 2023), and LLaVa-1.5 (Liu et al., 2024)). We report the agreement rate (%) of our method vs. the judgment of human annotators for which model's output has higher precision or recall (see Sec. A.1), and show results for all 6 pairs of the 4 models. Our model significantly improves on capturing precision compared to the SOTA metric in prior work (ALOHa (Petryk et al., 2024)), even when provided references from ground truth (GT) captions. In contrast, our metric is fully *reference-free*. We also report the agreement between our method and human recall judgments, demonstrating a high level of agreement. We also include the results for CLIP-Image-Score (Ge et al., 2024), which incorporates image generation models to indirectly measure hallucination - we find our method substantially improves over this technique as well.

A.3 Qualitative examples

Qualitative: Data filtering. We present in Fig. A2 qualitative examples of our method, applied to data filtering. In particular, we present two examples from the ShareGPT4v dataset, showing how our method improves over our strongest baseline (our hybrid OVF-ALm method, which illustrates the brittle nature of the ALOHa matching method). For each example, we show the matching results for OVF-ALm to contextualize why this method incorrectly includes or does not include that example. We also show the recall-precision breakdown of OVFact score and the detected hallucinations by our method. The left image is an example that was *filtered out by our method* from the training mixture at a 0.4 data ratio but was kept in the baseline OVF-ALm (our strongest baseline, with ALOHa matching). Our method correctly identifies "tree" as a hallucination, while the compared baseline misses it. The right image is an example with a clean caption within the 0.4 data ratio mixture of Ours, and is (incorrectly) not included in the baseline filtered set due to enforced one-to-one matching between "brown white fur" and "dog collar".

Qualitative: Model outputs. We also present in Fig. A5 example outputs from two trained PaliGemma 2 models, one on the *full* training set, and one on our OVFact-filtered training set, with 5x less training data; i.e., a model trained on the full ShareGPT4v vs. our subset. Both models depicted are run zero-shot on the Localized Narratives downstream dataset. We observe that our qualitative model outputs (on both DOCCI and Localized Narratives) consistently have improved factuality precision, without compromising descriptiveness/recall. We also provide S×S quantitative analysis with human evaluation in Sec. A.4.

A.4 Models trained with OVFact-filtered data: Traditional Metrics and Human Evaluations

As mentioned in the main results section, our method improves factuality without compromising descriptiveness. In the main paper, we describe results across challenging long-caption downstream benchmarks and also show consistent results with prior/standard metrics (CHAIR) on specific domains (MS-COCO).

Traditional metrics. Regarding traditional overall quality metrics, as noted by Onoe et al. (2024), metrics that may provide a signal for short captions (e.g., BLEU, CIDEr, etc.) do not provide a meaningful signal for long, descriptive captions when compared with human preferences. Thus, we do two things to evaluate models trained on our filtered subsets: (1) we report perplexity to ensure we are not sacrificing this for our other metrics gains, and (2) we provide human preference analysis of a model trained on our split vs. a model trained on the full dataset. Our gains in factuality reported in the main paper do not come at undue cost to

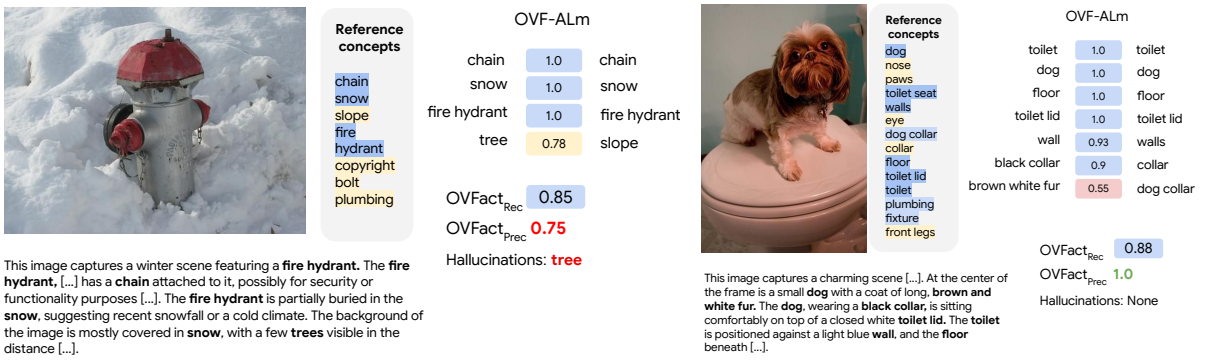

Figure A2: **Qualitative example of our filtering approach.** We present two examples from the ShareGPT4v dataset, showing how our method improves over our strongest baseline (our hybrid OVF-ALm method, which illustrates the brittle nature of the ALOHa matching method). For each example, we show the matching results for OVF-ALm to contextualize why this method incorrectly includes or does not include that example (Note that the OVF-ALm "score" is the minimum value of the set as with ALOHa, i.e., 0.78 on the left and 0.55 on the right). We also show the recall-precision breakdown of OVFact score and the detected hallucinations by our method. **(left)** Example that was *filtered out by our method* from the training mixture at a 0.4 data ratio but was kept in the baseline OVF-ALm (our strongest baseline, with ALOHa matching). Our method correctly identifies "tree" as a hallucination, while the compared baseline misses it. **(right)** Example with a clean caption within the 0.4 data ratio mixture of Ours, and is (incorrectly) not included in the baseline filtered set due to enforced one-to-one matching between "brown white fur" and "dog collar". See Sec. A.3.

Figure A3: **Qualitative example of trained model outputs.** We present an example comparison of two trained PaliGemma 2 models, **(top)** one on the *full* training set, and **(bottom)** one on our OVFact-filtered training set, with 5x less training data; both models are run zero-shot here on the Localized Narratives downstream dataset. We observe our model outputs consistently have lower rates of hallucination, without compromising descriptiveness. See Sec. A.3.

perplexity (a traditional metric for assessing the overall diversity of the caption, more suitable for long captions) across both DOCCI and Localized Narratives.

Specifically, at the 20% ratio (numbers below are (DOCCI, Localized Narratives) zero-shot evaluations):

- Random filtering: $(\mathbf{2.605 \pm 0.004}, \mathbf{2.955 \pm 0.010})$
- Perplexity filtering: $(2.601 \pm 0.002, 2.938 \pm 0.006)$
- ALOHa filtering: $(2.603 \pm 0.003, 2.958 \pm 0.001)$
- OVF-ALm filtering: $(2.610 \pm 0.003, 2.944 \pm 0.008)$
- Ours: $(2.610 \pm 0.003, 2.956 \pm 0.001)$

We emphasize that in the above, our goal is to show that the perplexity for all models is effectively *in the standard deviation window of the random filtering baseline* – in other words, our metric does *not* sacrifice perplexity to achieve all the gains we observe in factuality precision and recall. We also reiterate that developing strong metrics for long caption settings remains an open problem, one in which we hope OVFact can play a supporting role. Finally, we note that even for the perplexity metric, the perplexity filtering does not seem to gain an advantage – while prior work (Li et al., 2024) observed promising results applying perplexity/loss filtering, this was only in the setting of small-scale, short sentences with human-generated captions; in our setting, with model-generated noisy data at a large-scale, it seems this signal is not as clear and does not result in models that will then generalize well zero-shot to new downstream datasets.

Human evaluation (Model outputs). Finally, we validate our method on a stronger metric: human preference. We run a similar evaluation as we did with the metric analysis, but this time, we compare our model trained on our OVFact-filtered data (5x less than the full set) against a model trained on the full set. We use images from DOCCI *qual test* dataset and ask 3 independent annotators to give

their preference. In this side-by-side comparison, model outputs for the same image are compared, and the order of model A (left) and model B (right) in the UI is randomized to prevent order bias and we repeat this for the DOCCI evaluation set with at least two human evaluations per image like before. This time, we care about model preference, and the question posed to evaluators asked about general quality with both descriptiveness and accuracy. We observe that humans select our model over the model trained on the full dataset with a strong preference rate of **68.2%**. This indicates that our metric improvements correlate with overall quality improvements and that models trained on our filtered subset have higher quality with 5x less training data.

A.5 Additional experimental details

We implement our approach in the Big Vision (Beyer et al., 2022) codebase in JAX (Bradbury et al., 2018) and choose models for our experiments and base model tools with non-restrictive open-source licenses (all models are used in accordance with intended usage). Our implementation details consist of two key aspects: (1) the model training for the data filtering experiments and (2) our open-source models (tools) that are used for our metric. We elaborate on key details below and plan to release further supporting material to facilitate the adoption of our method in the broader community.

We train all our PaliGemma 2 (Steiner et al., 2024) models with input images of size 448 x 448, following the best practices outlined in the original paper. Our main experiments in the paper focus on the PaliGemma 2 model with 3B parameters (SigLIP-So400m encoder with 14x14 pixel patches, yielding 1024 tokens per image, and a 2B parameter LLM decoder), but we show results for other model sizes in Section A.2. We train our model with a batch size of 128 (image, caption) examples for 5 epochs over the full input large-scale training set. Following (Steiner et al., 2024), we use the Adam optimizer (Kingma and Ba, 2015) with default hyperparameters (b2 = 0.999, grad clip norm = 1.0, etc.) throughout and adjust the learning rate for the different model sizes (e.g., learning rate 1e-6 for 3B). The model card for PaliGemma 2, detailing its pre-training data mixture, is available through the official model release on GitHub and Huggingface, which can be found in the paper (Steiner et al., 2024).

```
You are an assistant that parses visually present objects from an image caption. Given an image
caption, you list ALL the objects visually present in the image or photo described by the captions
in a python list. Strictly abide by the following:
1. Include all visual attributes and adjectives that describe the object, if present.
2. Do not include objects that are mentioned but have no visual presence in the image, such as
light, sound, or emotions.
3. Always give the singular form of the object, even if the caption uses the plural form.
4. Return only a python list.
I will give you some examples. Image caption: A cozy room illuminated by a warm fireplace and soft
candles. A large painting hangs on the wall, while a small painting adorns the opposite side of
the fireplace. A wooden desk sits in the center of the room, with a book open on its surface. A
white candle burns brightly on the desk, casting long shadows across the room. A blue and white
vase sits on the fireplace, while a white candle burns on the table. The room is filled with a
sense of tranquility and elegance. Answer: ['room', 'fireplace', 'candle', 'large painting',
'small painting', 'wooden desk', 'book', 'white candle', 'blue and white vase', 'table', 'white
candle', 'wall']. Image caption: A row of slot machines in a casino, illuminated by a light
fixture on the ceiling. The machines are made of wood and have a brown frame. The screens are
colorful and the buttons are red. The chair is blue and the ceiling fan is on. The wall is white
and the ceiling is brown. The slot machine is in a wooden cabinet and the coin slot is on the
front of the machine. Answer: ['slot machine', 'casino', 'ceiling', 'wooden machine', 'brown
frame', 'colorful screen', 'red button', 'blue chair', 'ceiling fan', 'white wall', 'brown
ceiling', 'slot machine', 'coin slot', 'wooden cabinet']
```

Figure A4: **Gemma 2 prompt.** We provide the prompt details above for the Gemma 2 parsing step, which maps the input text y to the candidate set $\mathcal{C}$.

For our tools, as discussed in Sec. 4.1, we use state-of-the-art open-source models to ensure the broader community can also run our method with consistent results. We leverage for our LLM Gemma2-27b (Team et al., 2024) for caption parsing (see Sec. A.6 for prompt), and OWL-ViTv2 (Minderer et al., 2024) and OpenSeg (Ghiasi et al., 2022) for entity grounding (open-vocabulary detection and segmentation). As per Sec. 3.1, to assess descriptiveness (recall) in a reference-free manner, we consider an extensive vocabulary $\mathcal{V}$ of open-vocabulary concepts. Specifically, we take the union of concepts from standard datasets, i.e. Visual Genome (Krishna et al., 2017), LVIS (Gupta et al., 2019), Open Images (Kuznetsova et al., 2020) and Objects365 (Shao et al., 2019), resulting in a total of 2792 unique concepts. To encode entities for OVFact$_{Rec}$ computation, we use SigLIP-So400m/14 (Zhai et al., 2023) text encoder. We also highlight that our OVFact is intended to be a general framework that can continue to incorporate other base models as these continue to improve.

A.6 LLM prompt

We present the LLM prompt we use for caption parsing in Fig. A4. As discussed in the main paper, our goal for the model is to ensure that it captures key, *groundable* objects from the input text y (VLM / model caption) and to output a final candidate set $\mathcal{C}$ for later use in our method.

A.7 Discussion on efficiency

We design our method with efficiency and scalability in mind since we explore a novel application of our method to large-scale pretraining datasets.

We note that the grounding tools (OWL-ViT, OpenSeg) we chose in this work have a "late fusion" architecture, which means that they are extremely scalable in terms of the number of text queries.

Number of text concepts	1	30	300	2,792 (Ours)	300,000 (>100x Ours)
Runtime in *miliseconds*	7.35 ± 0.12	7.54 ± 0.13	7.57 ± 0.32	7.58 ± 0.24	12.93 ± 0.26

Table A2: **Runtime scaling experiments of OWL-ViT w.r.t. number of prompted text queries (concepts).** We validate that our method scales well with an increased number of concepts V.

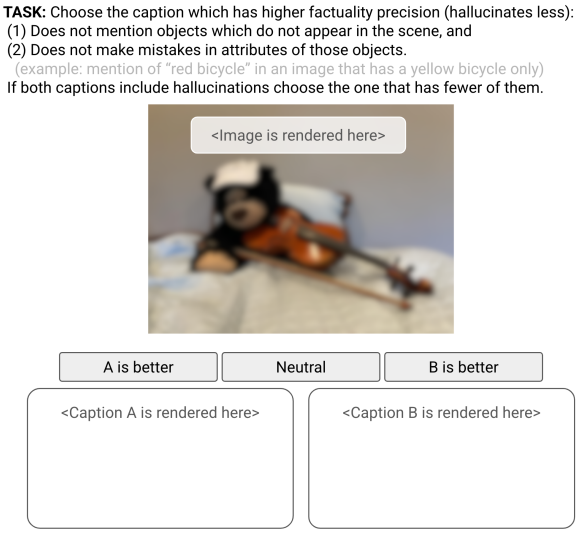

Figure A5: **UI interface example.** We show a screenshot of our UI for our human studies of S×S comparisons (blurring the image and removing the captions). We extend the analysis from (Onoe et al., 2024). See Sec. A.1 and Sec. A.4 for discussions of human evaluations of our method: we show consistent improvements over key baselines, for both measuring agreement and better quality model outputs for models trained on our filtered subset.

This is because the input image and text queries can be embedded separately, text embeddings are only used at the "end" of the network (the bulk of visual processing is query-agnostic), and text embeddings can be cached and reused across different inference calls. To illustrate by example, we focus on our OWL-ViT model in detail below.

This is confirmed by the table below, where increasing the number of text queries from 1 to 2792 (the total size of our reference set), increases the runtime by only 3% (< 1 millisecond). Note that the timing was performed on an NVIDIA A100 GPU using an image resolution of 448 x 448. While we precomputed text embeddings for the above table, we note that it only takes 0.84 ms to compute all 2792 text embeddings, which is an order of magnitude less than the rest of the network (and which needs to be only done once).