

Qualitative Analysis of Zero-Shot Caption Generation

Jonas Nasimzada

Department of Computer Science

University of Stuttgart

70569 Stuttgart

st171793@stud.uni-stuttgart.de

Abstract

Vision-language models (VLMs) generate fluent captions, but their robustness under limited visual context is unclear. We compare BLIP-2 (Li et al., 2023), GIT (Wang et al., 2022), LLaVA (Liu et al., 2023), and Molmo (Deitke et al., 2024) on the ManyNames dataset (Silberer et al., 2020) using original and cropped images. Captions are produced zero-shot (no fine-tuning or prompting) and evaluated by human annotators for adequacy, specificity, attributes, relations, hallucinations, fluency, and cross-view consistency. Cropping generally improves accuracy by reducing distractors. GIT is robust and object-centric, but has a lot of OCR hallucination. BLIP-2 is concise but prone to color/object hallucinations. LLaVA offers rich scene narratives with occasional referential drift. Molmo gives fine-grained detail yet raises adequacy risk. Findings show architecture and training drive context sensitivity, motivating crop-aware, hallucination-resistant captioning. All code, captions, and data are available at github.com/JonasNasimzada/VLM-object-naming.

1 Introduction

VLMs have achieved striking progress in image captioning, yet their robustness to changes in visual scope is not well understood. In everyday scenes, the same object can be named differently depending on context, and practical applications often operate on partial views (e.g., zoomed or cropped regions). This raises a central question: how do state-of-the-art VLMs adapt when the available visual context shrinks?

We compare captions for original images and tight crops on ManyNames, a multi-domain dataset with context-dependent names. We evaluate four VLMs (BLIP-2, GIT, LLaVA, Molmo) in a strict zero-shot setting. Human annotators rate adequacy, specificity, attributes/relations, hallucinations, fluency, and cross-view consistency, and

record Original↔Crop deltas (specificity shift, head-noun change, attribute retention). In parallel, a corpus-level analysis tallies top vs. incorrect responses per model/view. A lightweight web app supports annotation (Appendix B).

2 Related Works

The field of image captioning has evolved from rule-based systems to advanced VLMs, such as BLIP-2 or LLaVA, which demonstrate robust zero-shot performance. A substantial body of prior research has examined factors such as caption quality, hallucination, and object naming (Zeng et al., 2024) (Rohrbach et al., 2019). However, the impact of visual context changes, including cropping, remains a largely unexplored area. Our study builds on these foundations, contributing a systematic, human-centered comparison of multiple VLMs under both original and cropped image conditions.

2.1 ManyNames

The ManyNames paper tackles a core limitation in Language & Vision: most recognition datasets assume a single “gold” label per object, whereas humans naturally use multiple, context-dependent names (e.g., duck, bird, animal). The authors released ManyNames, a verified object-naming dataset, based on the Visual Genome dataset (Krishna et al., 2016) that preserves genuine naming variation while filtering out annotation noise, enabling more human-like evaluation of models. ManyNames starts from Visual Genome images and collects up to 36 free-form names for each of ∼25k target objects spanning seven domains (people, clothing, home, buildings, food, vehicles, animals/plants).

2.2 Generative Image-to-Text (GIT)

GIT frames image captioning as conditional language modeling with a minimal encoder-decoder design: a contrastively pretrained image encoder

and a text decoder. Image features are projected into the decoder space and concatenated with token embeddings, enabling autoregressive caption generation without additional modules like region detectors or OCR. Pretraining leverages 0.8B image–text pairs from diverse datasets including COCO (Lin et al., 2015), CC3M/12M (Changpinyo et al., 2021), Visual Genome, ALT200M (Hu et al., 2022), and web sources.

2.3 BLIP-2

BLIP-2 improves efficiency by freezing the image encoder and LLM while introducing a lightweight Querying Transformer (Q-Former) as a bridge. The Q-Former learns to extract relevant visual features through contrastive, generation, and matching objectives, and its outputs are projected into the LLM’s embedding space as soft visual prompts.

2.4 LLaVA

LLaVA employs visual instruction tuning by connecting a frozen CLIP vision encoder (Radford et al., 2021) to an instruction-tuned LLM (Vicuna) via a simple linear projection. Training proceeds in two stages: first aligning modalities on a filtered CC3M subset, then refining on GPT-4-generated instruction data (LLaVA-Instruct-158K (Liu et al., 2023)). This design prioritizes simplicity and data efficiency, enabling conversational captioning and instruction following without heavy architectural modifications.

2.5 Molmo

Molmo tackles image captioning with an autoregressive decoder conditioned on ViT-encoded image tokens (Dosovitskiy et al., 2021) and a lightweight connector, removing the need for task-specific captioning heads. Images are processed via a multi-crop preprocessor (one low-resolution frame and overlapping high-resolution crops), whose pooled features preserve context and are projected into the LLM’s embedding space. These tokens are passed to a decoder-only LLM for caption generation. Key design choices include 2×2 attention pooling with MLP projection, and pretraining on the PixMo-Cap dataset (712k images with rich human-spoken captions refined by an LLM) (Deitke et al., 2024). Training innovations such as text-only dropout (to enforce visual grounding) and length-conditioned captioning (to balance precision and recall) further improve performance.

3 Methodology

A comparative evaluation of four zero-shot image-captioning systems was conducted: BLIP-2, GIT, LLaVA, and Molmo. The evaluation method entailed the generation of two captions per image and per model: one from the full, original image and one from a tight crop centered on the target region, as seen in Figure 1. No fine-tuning, exemplars, or prompt engineering were applied. Each model was used in its native zero-shot mode.

3.1 Caption Generation

We use the *ManyNames* dataset and, for each image, consider both the full original view and a cropped view. Cropping coordinates are taken directly from the *ManyNames* dataset to isolate the intended target region. For every image-view pair (original and crop), we generate a caption independently and strictly in zero-shot mode: no fine-tuning, no exemplars, and no prompt engineering. Each captioning model is run with its recommended/default settings as provided by the model authors.

4 Evaluation

To evaluate the results, human annotators (fluent in English, with prior ML experience) assess captions using a lightweight web application that loads the evaluation subset and displays the *Original* and *Crop* side-by-side with captions from a selected model. An illustrative screenshot of the interface is provided in Appendix B. The *ManyNames* dataset is organized into domains, which we mirror in our sampling: *people*, *clothing*, *home*, *buildings*, *food*, *vehicles*, *animals/plants*, *sports/outdoor*, and *other*. We select 30 images at random with an approximately equal distribution across these domains; each selected item contributes both the original and cropped views and includes captions from all models.

For each (image, model, view) we record a set of evaluation metrics: **Adequacy**, **Error type**, **Naming level**, **Attributes correctness**, **Relations correctness**, **Hallucinations**, **Fluency** and **Same-object**

After annotating both views, we also record **Original ↔ Crop delta** metrics once per (item, model) to capture how cropping changed the caption: **Specificity shift**, **Head-noun change** and **Attribute retention**

The full definitions and possible values for all metrics are detailed in Appendix C.

4.1 Quantitative Evaluation

Besides the human annotation study, we also performed a dataset-driven analysis using the Many-Names statistics. For each image, we collected the number of *top responses* (correctly matching the intended label) and *incorrect responses*. These counts were then compared across original and cropped versions for each model. This complementary perspective allowed us to quantify how cropping influences captioning performance at scale, providing a broader context for the qualitative findings.

5 Results

VLM	View	% Top	% Incorrect
BLIP-2	Original	75.2%	24.8%
BLIP-2	Cropped	76.2%	23.8%
LLaVA	Original	69.4%	30.6%
LLaVA	Cropped	70.8%	29.2%
GIT	Original	72.0%	28.0%
GIT	Cropped	73.9%	26.1%
Molmo	Original	66.6%	33.4%
Molmo	Cropped	67.4%	32.6%

Table 1: Percentage of top and incorrect responses per VLM for original and cropped images.

Table 1 reports overall accuracy (“Top”) and error rates for original vs. cropped images. Across models, cropping generally helps (small Top% gains and reduced error rates), suggesting that removing surrounding context reduces distractors and focuses the captioner on the intended target. The domain breakdown in Table 2 shows that improvements are not uniform: vehicles and animals/plants are consistently strong across models, whereas clothing and home remain challenging. For more Visualizations see Appendix D

BLIP-2. BLIP-2 shows the best robustness to cropping (highest Top% on crops; Table 1) regarding correct and incorrect responses. BLIP-2 produces short, terse captions and often detects the main content correctly (see in Appendix, Figure 7), which aligns with solid Top% overall and strong domain scores for vehicles and animals/plants. As illustrated in Table 3 human annotations indicate the presence of frequent hallucinations (particularly

of color, often blue, and of additional objects) and relation mistakes, see Figure 8. Cropping yields a small average gain, but annotators also saw cases where crops added hallucinations when global cues disappeared (Figure 9). Image deltas (Table 4) show mostly stable specificity but lost attributes and occasional head-noun shifts to other concepts. Likely factors are (i) frozen vision encoder + LLM bridging (Q-former) can lean on language priors (e.g., sky/sea \Rightarrow “blue”); (ii) default short decoding promotes compact but sometimes under-grounded outputs; (iii) reduced context in crops weakens relational grounding, increasing guessy color/object insertions.

GIT. GIT leads in several domains (vehicles, animals/plants, clothing; Table 2). Annotators found GIT more reliable than BLIP-2 at naming the central object with short, object-centric captions. A recurring failure mode is an OCR bias: it tries to read text (and sometimes hallucinates it) even when none is present (Figure 10). Table 4 shows both basic \rightarrow superordinate shifts (safer general labels under crops) and instances of hyponym shifts when the crop exposes distinctive details. Likely factors could be (i) encoder-decoder pretraining on very large image-text corpora with abundant embedded text encourages text-reading behavior; (ii) conditional Language Model (LM) objective favors concise, salient-object labeling; (iii) loss of scene context under crops nudges toward more general categories unless fine-grained cues are visible.

LLaVA. LLaVA tends to describe the whole scene: longer captions that weave foreground and background and add plausible context (Figure 8). This style delivers solid domain performance (notably animals/plants) but induces referential errors when the narrative drifts beyond the true target (Table 3). With crops, LLaVA focuses more directly on the target and often shifts superordinate \rightarrow basic (Table 4) (Figure 11), yielding modest Top% improvements (Table 1), though occasional object misinterpretations persist. Likely factors are (i) visual instruction tuning encourages expansive, conversational descriptions; (ii) simple projection from CLIP features to the LLM emphasizes semantic alignment over tight spatial grounding; (iii) cropping reduces the temptation to narrate scene context, improving target specificity.

Molmo. Molmo generates long, highly detailed captions that cover foreground/background, am-

VLM	View	Food		Vehicles		Home		Buildings		People		Animals/Plants		Clothing	
		%C	%I	%C	%I	%C	%I	%C	%I	%C	%I	%C	%I	%C	%I
BLIP-2	Original	72.6	27.4	92.6	7.4	63.5	36.5	55.0	45.0	76.4	23.6	92.0	8.0	36.4	63.6
BLIP-2	Cropped	73.9	26.1	92.0	8.0	65.6	34.4	58.2	41.8	77.7	22.3	93.7	6.3	44.2	55.8
LLaVA	Original	70.1	29.9	88.5	11.5	58.9	41.1	56.0	44.0	65.3	34.7	87.7	12.3	46.6	53.4
LLaVA	Cropped	70.4	29.6	87.0	13.0	62.2	37.8	56.9	43.1	64.2	35.8	91.0	9.0	57.3	42.7
GIT	Original	76.3	23.7	92.5	7.5	58.5	41.5	50.9	49.1	69.1	30.9	90.8	9.2	41.3	58.7
GIT	Cropped	80.5	19.5	92.8	7.2	59.9	40.1	61.5	38.5	68.5	31.5	92.8	7.2	54.3	45.7
Molmo	Original	68.9	31.1	82.2	17.8	58.7	41.3	56.2	43.8	56.6	43.4	86.8	13.2	58.3	41.7
Molmo	Cropped	65.8	34.2	81.5	18.5	60.4	39.6	57.2	42.8	58.2	41.8	88.2	11.8	58.4	41.6

Table 2: Percentage of correct (%C) and incorrect (%I) responses per domain for each VLM and view. Bold values indicate best-performing model–view pairs for %C (highest) and %I (lowest) in each domain.

Metric	BLIP2	Git	LLaVA	Molmo
Adequacy	0–0.5	0.5–1	1	1
Error Type	Visual, Referential	Linguistic, Visual	Referential, Visual	Linguistic
Naming Level	Basic	Basic	Subordinate	Subordinate
Attribute Correctness	Correct	Correct	Correct	Correct
Relations Correctness	Wrong	Correct	Correct	Correct
Hallucinations	2+	1	0–1	0
Same-Object (Original vs. Cropped)	Uncertain / No	Yes	Yes / Uncertain	Yes

Table 3: Human annotation results per model across evaluation metrics.

bience, and plausible scenario context (Figure 12). Annotators noted good text reading, detection of small or hidden features, and even awareness of cropping (Figure 9). Despite these strengths, Molmo’s Top% trails others (Table 1). Two effects likely contribute: (i) longer outputs raise the chance of any single error under our adequacy rubric; (ii) a preference for subordinate naming increases risk when cues are borderline. Still, Molmo is competitive in home and people under crops (Table 2), and Table 4 indicates largely stable specificity with kept attributes. Likely factors could be (i) multi-crop visual tokenization and length-conditioned captioning encourage exhaustive detail; (ii) training that promotes visual grounding can surface subtle cues (beneficial on tricky images) but also expands the linguistic surface area for errors; (iii) strong fine-grained naming increases adequacy risk when evidence is thin.

Summary. Cropping reduces distractors and modestly improves accuracy across models, but gains depend on model style. BLIP-2’s concise outputs are efficient yet vulnerable to color/object priors when context is limited. GIT is robust and object-centric but exhibits a pretraining-driven text-reading bias. LLaVA’s instruction-following yields

rich scene narratives that become more target-focused with crops, though referential drift remains. Molmo is perceptive and detailed, but length and fine-grained naming can penalize adequacy. Overall, architecture, alignment, pretraining data, and decoding settings together drive how each model behaves under cropping.

6 Conclusions

This paper compared BLIP-2, GIT, LLaVA, and Molmo on ManyNames using original versus cropped images in a zero-shot setup with human evaluation. Cropping generally reduced distractors and modestly improved accuracy, but effects were model-specific: GIT was robust and object-centric, but has a lot of OCR hallucination. BLIP-2 produced concise captions yet hallucinated colors/objects. LLaVA provided rich narratives with occasional referential drift. Molmo delivered fine-grained detail but longer outputs increased adequacy errors for the correct responses in the ManyNames dataset. Future work should couple stronger visual grounding with crop-aware training and hallucination mitigation to achieve reliable, context-robust captioning in practice.

Metric	BLIP2	Git	LLaVA	Molmo
Specificity Shift	Stable	Basic → Superordinate	Superordinate → Basic	Stable
Head-Noun Change	Other	Hyponym (more specific)	Hypernym (more general)	Stable
Attribute Retention	Lost	Kept	Kept	Kept

Table 4: Original ↔ Cropped deltas per model.

References

- Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. *Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts*. *Preprint*, arXiv:2102.08981.
- Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, and 31 others. 2024. *Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models*. *Preprint*, arXiv:2409.17146.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. *An image is worth 16x16 words: Transformers for image recognition at scale*. *Preprint*, arXiv:2010.11929.
- Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. *Scaling up vision-language pre-training for image captioning*. *Preprint*, arXiv:2111.12233.
- 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.
- 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.
- 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.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. *Learning transferable visual models from natural language supervision*. *Preprint*, arXiv:2103.00020.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2019. *Object hallucination in image captioning*. *Preprint*, arXiv:1809.02156.
- Carina Silberer, Sina Zarrieß, Matthijs Westera, and Gemma Boleda. 2020. *Humans meet models on object naming: A new dataset and analysis*. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1893–1905, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. *Git: A generative image-to-text transformer for vision and language*. *Preprint*, arXiv:2205.14100.
- Zequn Zeng, Yan Xie, Hao Zhang, Chiyu Chen, Bo Chen, and Zhengjue Wang. 2024. *Meacap: Memory-augmented zero-shot image captioning*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14100–14110.

A Example of original and cropped images

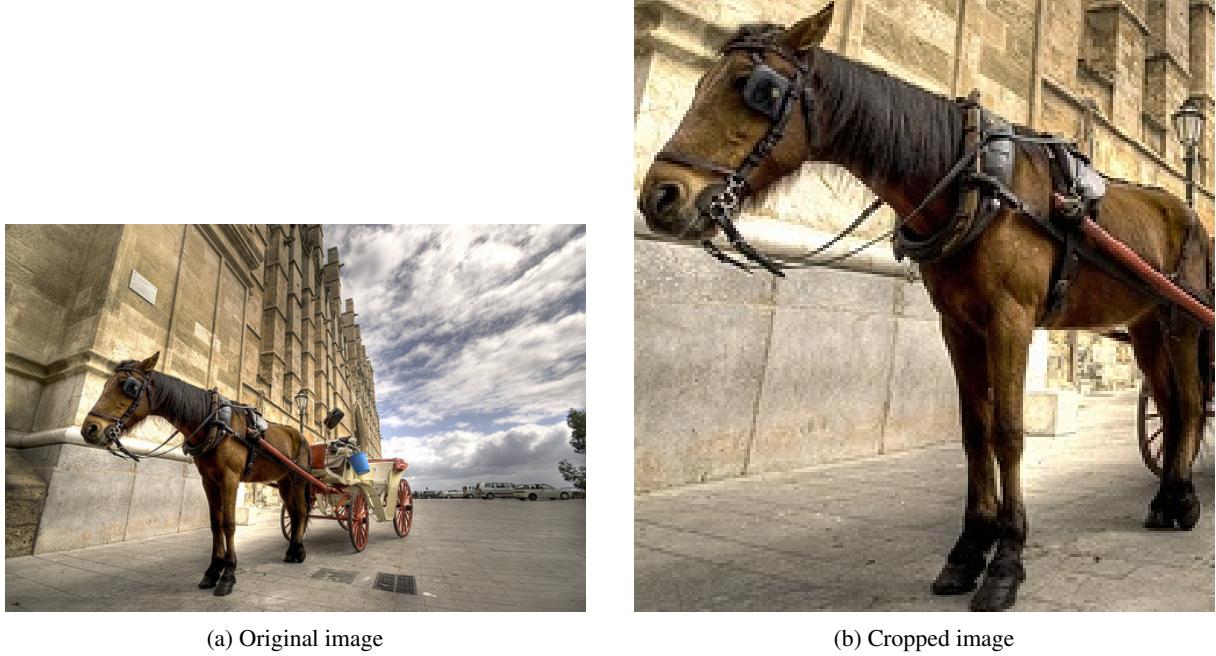


Figure 1: Example of original and cropped images

B Annotation Web Application

Figure 2 shows a screenshot of the lightweight web application used for annotation. The interface presents the *Original* and *Crop* images side-by-side, along with captions from the selected model. Annotators enter per-view metrics and original↔crop deltas directly in the interface. The tool automatically saves progress and allows exporting annotations to CSV or JSON.

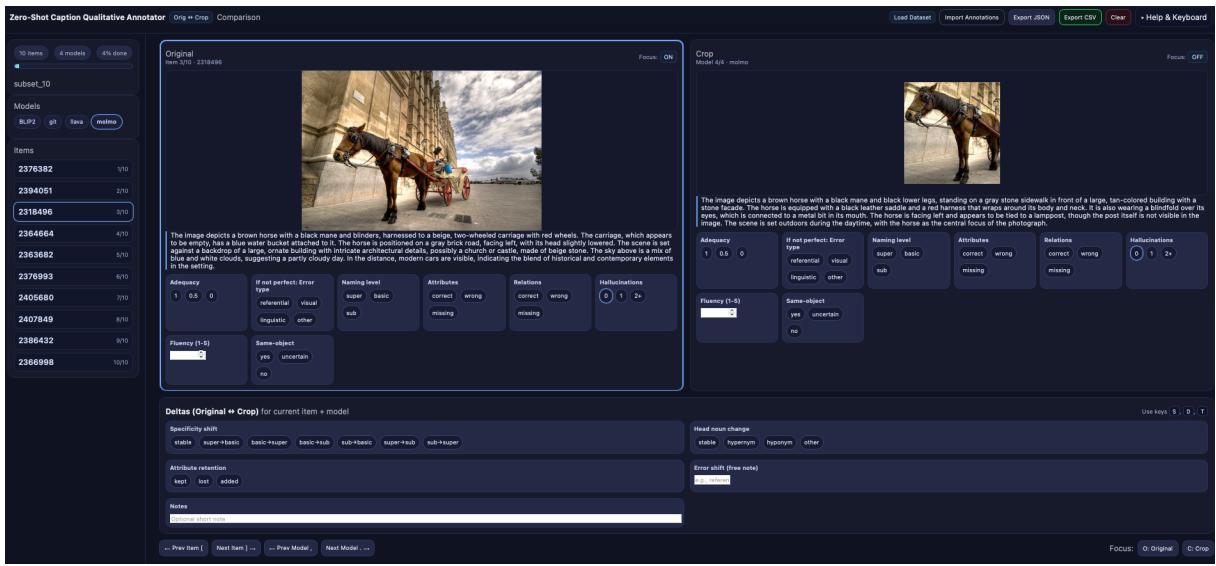


Figure 2: Screenshot of the custom annotation web application.

C Evaluation Metrics

C.1 Per-view Metrics

Adequacy Indicates whether the caption correctly names the intended target.

- **1** – Perfect match.
- **0.5** – Minor issue.
- **0** – Wrong target.

Error type Recorded if adequacy $\neq 1$, specifying the cause of the mismatch:

- **Referential** – failure to identify the correct object in context.
- **Visual** – visual misclassification of the object.
- **Linguistic** – grammatical or wording error altering meaning.
- **Other** – any error not covered above.

Naming level Granularity of the head noun:

- **Superordinate** – very general category.
- **Basic** – default lexical level (e.g., *dog*).
- **Subordinate** – fine-grained term (e.g., *dalmatian*).

Attributes correctness Evaluation of descriptive properties (color, number, material, size):

- **Correct**
- **Wrong**
- **Missing**

Relations correctness Evaluation of spatial or interaction terms (e.g., on, under, holding, next to):

- **Correct**
- **Wrong**
- **Missing**

Hallucinations Number of non-visible objects or attributes mentioned:

- **0**
- **1**
- **2+**

Fluency Subjective rating of naturalness and grammaticality:

- Scale: **1–5** (1 = poor, 5 = excellent).

Same-object Primarily for crops: whether the caption refers to the same target as the original:

- **Yes**
- **Uncertain**
- **No**

C.2 Original ↔ Crop Delta Metrics

Specificity shift Change in naming level between original and crop captions:

- **Stable**
- **super**→**basic**
- **basic**→**super**
- **basic**→**sub**
- **sub**→**basic**
- **super**→**sub**
- **sub**→**super**

Head-noun change Semantic relationship between head nouns in the two captions:

- **Stable**
- **Hypernym** – crop noun more general.
- **Hyponym** – crop noun more specific.
- **Other** – non-hierarchical change (e.g., *man*→*jacket*) or part–whole relation.

Attribute retention Whether attributes are preserved or altered between views:

- **Kept**
- **Lost**
- **Added**

D Additional Visualizations

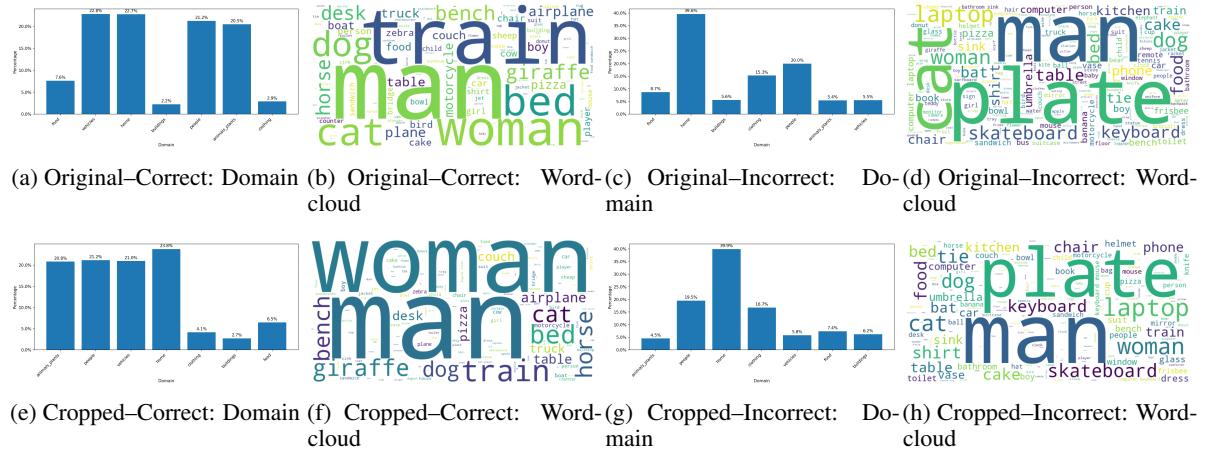


Figure 3: Per-domain distributions and wordclouds for BLIP-2 across Original/Cropped and Correct/Incorrect subsets.

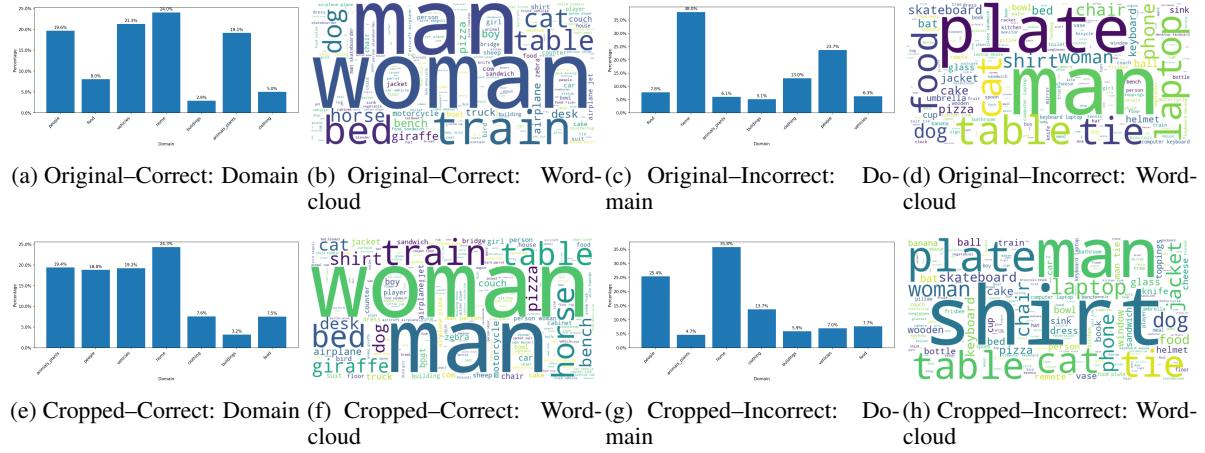


Figure 4: Per-domain distributions and wordclouds for LLaVA across Original/Cropped and Correct/Incorrect subsets.

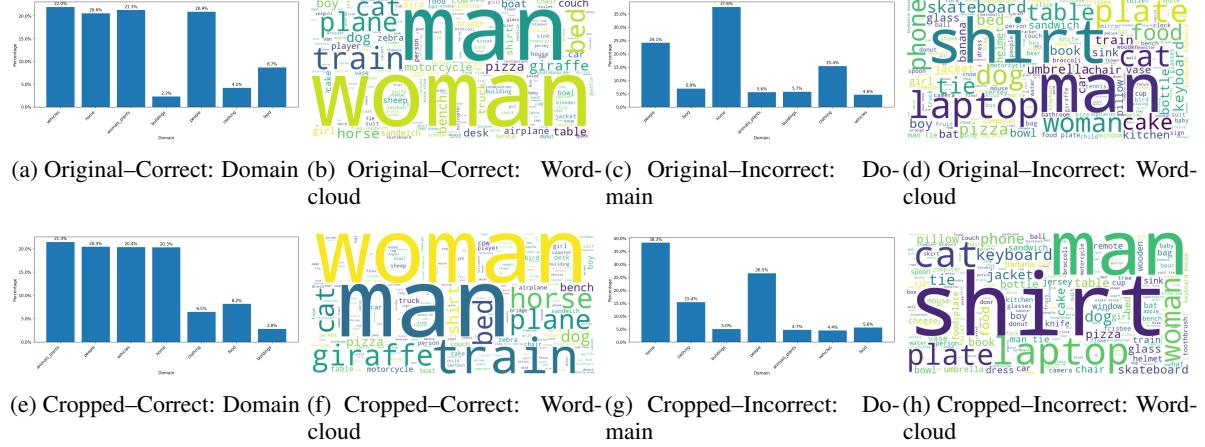


Figure 5: Per-domain distributions and wordclouds for GIT across Original/Cropped and Correct/Incorrect subsets.

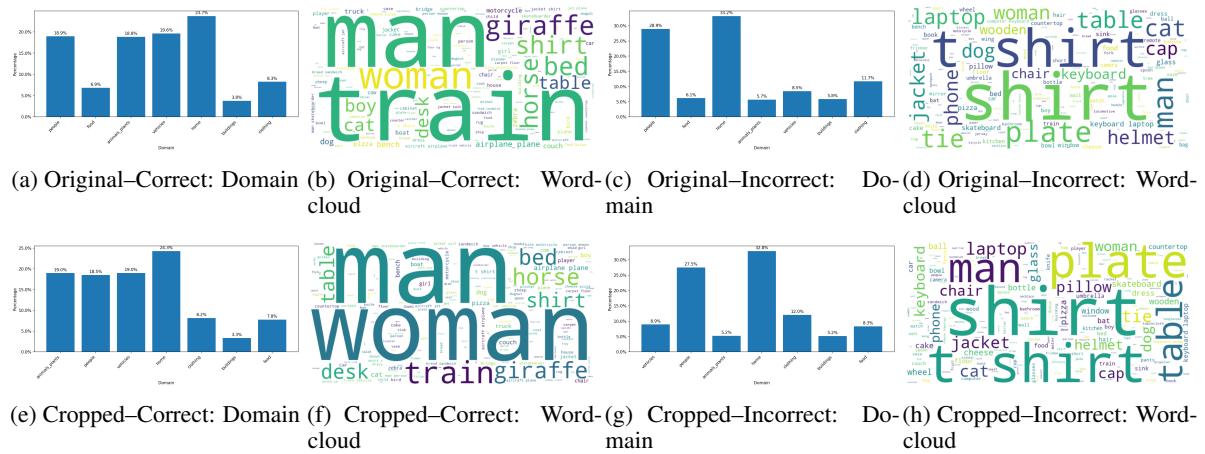


Figure 6: Per-domain distributions and wordclouds for Molmo across Original/Cropped and Correct/Incorrect subsets.

E Example Caption



Original Captions:

BLIP-2	a horse pulling a carriage
GIT	a horse is pulling a carriage with a horse on it.
LLaVA	The image shows a horse pulling a carriage down a street. The horse is positioned on the left side of the scene, and the carriage is located on the right side. There are also a few people in the scene, with one person standing near the horse and another person standing further away. The overall atmosphere of the image suggests a historical or traditional setting.
Molmo	The image depicts a brown horse with a black mane and blinders, harnessed to a beige, two-wheeled carriage with red wheels. The carriage, which appears to be empty, has a blue water bucket attached to it. The horse is positioned on a gray brick road, facing left, with its head slightly lowered. The scene is set against a backdrop of a large, ornate building with intricate architectural details, possibly a church or castle, made of beige stone. The sky above is a mix of blue and white clouds, suggesting a partly cloudy day. In the distance, modern cars are visible, indicating the blend of historical and contemporary elements in the setting.



Cropped Captions:

BLIP-2	a blue horse pulling a carriage
GIT	a horse with a bridle that says "the horse".
LLaVA	The image shows a horse pulling a carriage down a street. The horse is wearing blinders, which are a type of horse harness designed to prevent the horse from being distracted by its surroundings. The carriage is carrying a passenger, and the scene captures the interaction between the horse and the street environment.
Molmo	The image depicts a brown horse with a black mane and black lower legs, standing on a gray stone sidewalk in front of a large, tan-colored building with a stone facade. The horse is equipped with a black leather saddle and a red harness that wraps around its body and neck. It is also wearing a blindfold over its eyes, which is connected to a metal bit in its mouth. The horse is facing left and appears to be tied to a lamppost, though the post itself is not visible in the image. The scene is set outdoors during the daytime, with the horse as the central focus of the photograph.

Figure 7: Example Captions



Original Captions:

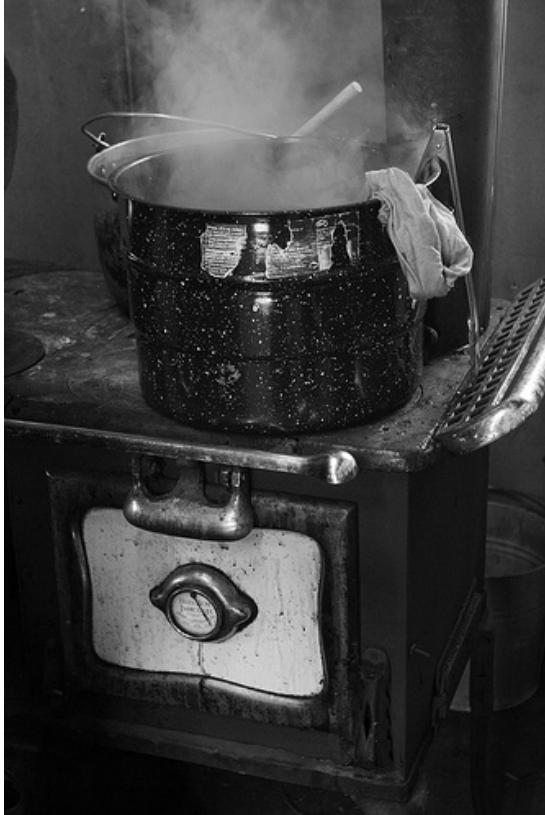
BLIP-2	a woman with blue skin and a black coat talking on a cell phone
GIT	a woman talking on a cell phone while walking down a street.
LLaVA	The image shows a woman talking on her cell phone while standing on a street corner. She is wearing a black jacket and appears to be engaged in a conversation. The scene also includes a bicycle parked nearby, a handbag, and a couple of other people in the background. The overall atmosphere suggests an urban environment where people are going about their daily activities.
Molmo	In this detailed photograph, a woman is captured mid-conversation while standing on a city street. She is dressed in a black coat and a light tan scarf, with her dark hair pulled back into a bun. A black purse hangs over her right shoulder, and she holds a cell phone to her ear with her left hand. The woman is positioned in front of a one-way street sign, which points to the left, and a building with a maroon awning that reads "East Side Vision" in white letters. Below the awning, the address "159 First Avenue" is displayed, along with the phone number "212-228-0950" and the phrase "eyes examined." A red bicycle is parked against the building, and a crosswalk is visible in the background. The scene is set during the daytime, with the woman as the central focus of the image.



Cropped Captions:

BLIP-2	a woman in a black coat
GIT	a woman is walking and looking at her cell phone.
LLaVA	The image shows a woman wearing a black coat and holding a handbag. She is standing on a street corner, possibly waiting for someone or something. The woman appears to be dressed in winter clothing, which suggests that the weather is cold. The scene captures a moment of everyday life in an urban environment.
Molmo	The image captures a woman standing on a city street, engaged in a phone conversation. The photograph is cropped just below her nose, revealing only her lips and chin. She is dressed in a black wool coat with a collar and two front pockets, and her left hand is holding a cell phone to her ear. Her right hand is tucked into her coat pocket. She wears a light gray, thick scarf wrapped around her neck, and a black purse hangs from her right shoulder. The background features a gray road with a white crosswalk and a manhole cover visible in the bottom left corner. The overall scene is set against a muted urban backdrop, emphasizing the woman's attire and the everyday nature of her phone call.

Figure 8: BLIP-2 color hallucinations & LLaVA Scene description



Original Captions:

BLIP-2	a pot of boiling water
GIT	a pot with a lid that says " no smoking " on it.
LLaVA	The image shows a large pot or kettle sitting on top of a stove. The pot is filled with water, and steam is rising from it, indicating that it is being heated. The stove appears to be an old-fashioned one, and the scene suggests that the pot is being used for cooking or heating purposes.
Molmo	This black and white photograph captures a vintage kitchen scene centered around a large, black pot with a silver handle, positioned on a small, old-fashioned stove. The stove, which appears to be from the early 1900s, features a white door with a silver handle and a temperature gauge. The pot, which is steaming vigorously, has a white cloth draped over its right side and a spoon protruding from its center. The pot's lid is set aside, leaning against the back wall. The stove's surface is cluttered with various items, including a towel and a rag, adding to the rustic charm of the scene. The background is dark, emphasizing the contrast and details of the pot and stove.



Cropped Captions:

	a woman in a kitchen
BLIP-2	
GIT	a pot with the word " p " on it
LLaVA	The image shows a large pot or kettle, possibly made of metal, placed on a stove or a burner. The pot is filled with water, and steam is rising from it. The scene suggests that the pot is being used for cooking or heating purposes.
Molmo	This black and white photograph captures a large, cylindrical pot, likely made of metal, with a black exterior adorned with white speckles. The pot, which appears to be a pressure cooker or a large cooking vessel, is positioned on a concrete floor. A white towel is draped over the right side of the pot, and a metal handle extends from the top, partially obscured by the steam billowing out. The pot's surface shows signs of wear, with patches of paint peeling off, revealing the metal underneath. In the background, a second pot is visible, though it is mostly hidden by the steam rising from the main pot. The overall scene suggests intense cooking activity, with the steam indicating that the contents are at a high temperature.

Figure 9: BLIP-2 added hallucinations when global cues disappeared & Molmo detection of small or hidden features



Original Captions:

BLIP-2	a bench sitting in the grass
GIT	a park bench with the number 4 on it
LLaVA	The image shows a park bench situated in a grassy area. The bench is positioned in the foreground, and the background features a lush green field. The scene appears to be set in a park or a similar outdoor space.
Molmo	The image depicts a serene park scene centered around a wooden bench with a distinctive design. The bench, made of dark brown wood, features a unique backrest composed of five horizontal slats, while the seat is a single solid piece. The bench's legs are intricately curved, resembling the shape of a 'W' or a 'U,' and are connected by a horizontal bar at the bottom. The bench is positioned on a lush, green lawn dotted with small white flowers, possibly dandelions. In the background, a dense array of trees with dark green leaves creates a natural canopy, allowing only a few beams of sunlight to filter through. The overall atmosphere is tranquil and inviting, suggesting a perfect spot for relaxation or quiet contemplation.



Cropped Captions:

BLIP-2	a bench in the grass
GIT	a bench with a metal frame that says " no one " on it.
LLaVA	The image shows a park setting with a long bench situated in the grass. The bench is positioned in the foreground, and the grassy area extends behind it. The scene appears to be a peaceful and relaxing environment, perfect for spending time outdoors.
Molmo	This image captures a detailed view of a park bench situated in a grassy area. The bench is constructed from dark brown wood, featuring five horizontal slats for the seat and five for the backrest. The legs of the bench are made of metal, with a distinctive design that includes a curved, almost heart-shaped cutout in the middle, adding an artistic touch to the sturdy structure. The bench is positioned on a lush green lawn, with the grass appearing well-maintained and vibrant. The background is simple, focusing solely on the bench and the grass, emphasizing the serene and peaceful atmosphere of the park setting.

Figure 10: GIT OCR hallucinations



Original Captions:

BLIP-2	a sheep is standing next to a boat
GIT	a boat is being loaded onto a boat that says " 519 ".
LLaVA	The image shows a boat being loaded onto a trailer. The boat is placed on the trailer, and there are sheep nearby, possibly grazing or resting in the area. The scene suggests that the boat is being transported, and the sheep are sharing the same space.
Molmo	The image depicts a small, silver boat with the number 5514 inscribed on its side, resting on a rusted metal trailer with two large black tires. The boat, which appears to be a fishing boat, features a small cabin with two windows and a red life preserver on top. The trailer is situated on a patch of dry, brown grass, indicative of a hot, arid environment. In the foreground, three sheep are grazing; one stands while the other two lay down. The background reveals a large, green hillside covered in lush vegetation, and a telephone pole is visible to the left. The sky above is a mix of blue and white, filled with clouds.



Cropped Captions:

BLIP-2	a boat is parked on a dock
GIT	crop a boat with the number 3788 on it
LLaVA	The image shows a boat, possibly a fishing boat, sitting on a trailer. The boat is parked on the trailer, which is connected to a truck. The scene suggests that the boat is being transported or towed to a different location. The boat is positioned on the trailer, and the truck is visible in the background.
Molmo	The image depicts a small, silver fishing boat, likely a commercial vessel, positioned on a black metal trailer. The boat, which appears to be around 15 to 20 feet long, features a flat bottom and a cabin with three windows. The boat's identification number, 55104, is prominently displayed in black on its side. The trailer, equipped with a black metal railing and a wooden post, is situated on a grassy field with a backdrop of green hills and a partly cloudy sky. The boat's cabin has a small window at the front and two larger windows on the sides, with an orange life preserver mounted on the top right corner. The scene is set outdoors, with the boat and trailer ready for transport.

Figure 11: LLaVA superordinate → basic shift



Original Captions:

BLIP-2	a man with blue paint on his face
GIT	a man is cutting his hair with scissors.
LLaVA	The image shows a man with a spiky haircut, standing in a room with a corkboard on the wall. He is holding a pair of scissors in his hand, possibly getting ready to cut his hair or trimming his spiky hair. The man appears to be focused on his task, and the room seems to be a comfortable and personal space for him.
Molmo	In this detailed photograph, a young man with spiky, dark brown hair is captured in the act of styling his hair with a black comb in his right hand. He is wearing a gray t-shirt and has a nose ring in his left nostril. His expression is a subtle smirk as he looks directly at the camera. The setting appears to be a kitchen, evidenced by the white ceiling, an orange wall, and a window behind him through which green trees are visible. To his right, there is a cork board pinned with various papers, including blue and white document and a green one. Below the cork board, a white refrigerator is partially visible, along with a white countertop and a white backsplash. The scene is well-lit, suggesting it is daytime, and the overall ambiance is casual and homely.



Cropped Captions:

BLIP-2	a man with blue paint on his face
GIT	a man is cutting his hair with scissors.
LLaVA	The image shows a man with a spiky haircut, who is holding a pair of scissors in his hand. He appears to be in a room with a window, and there is a clock on the wall. The man is wearing a necklace, and the overall scene suggests that he might be in the process of cutting or styling his hair.
Molmo	The image depicts a young man, likely in his late teens or early twenties, standing in a room with peach-colored walls and a window behind him. He has spiky, dark brown hair that he is in the process of cutting with a pair of black scissors held in his right hand. His hair is styled in a somewhat messy, spiky manner, and he is looking directly at the camera with a slight smirk. He is wearing a gray t-shirt and a distinctive necklace with a silver pendant. On his right wrist, he sports a black and brown bracelet. The background features a cork board on the right side, adorned with various papers, including a blue and white sign. The window behind him reveals a glimpse of green trees outside.

Figure 12: Molmo Scene description