

Altogether: Image Captioning via Re-aligning Alt-text

Hu Xu¹, Po-Yao Huang¹, Xiaoqing Ellen Tan¹,
Ching-Feng Yeh¹, Jacob Kahn¹, Christine Jou¹,
Gargi Ghosh¹, Omer Levy¹, Luke Zettlemoyer^{1,2}, Wen-tau Yih¹,
Shang-Wen Li¹, Saining Xie³ and Christoph Feichtenhofer¹

¹Meta FAIR ²University of Washington ³New York University

<https://github.com/facebookresearch/MetaCLIP>

Abstract

This paper focuses on creating synthetic data to improve the quality of image captions. Existing works typically have two shortcomings. First, they caption images from scratch, ignoring existing alt-text metadata, and second, lack transparency if the captioners' training data (e.g. GPT) is unknown. In this paper, we study a principled approach *Altogether* based on the key idea to edit and *re-align* existing *alt-texts* associated with the images. To generate training data, we perform human annotation where annotators start with the existing alt-text and re-align it to the image content in multiple rounds, consequently constructing captions with rich visual concepts. This differs from prior work that carries out human annotation as a one-time description task solely based on images and annotator knowledge. We train a captioner on this data that generalizes the process of re-aligning alt-texts at scale. Our results show our *Altogether* approach leads to richer image captions that also improve text-to-image generation and zero-shot image classification tasks.

1 Introduction

Human social interactions often gravitate towards engaging with individuals who exhibit a higher level of intelligence. This inherent social behavior underscores the aspiration to develop AI agents that surpass the average human intelligence. The pursuit of creating such advanced AI agents hinges significantly on the *quality* of the training data, which ideally encapsulates superhuman intelligence.

However, in the context of image captioning, most existing training data is designed for naive and well-known visual concepts that provide little value to an average user, e.g., a caption “a dog is walking in the park” offer minimal utility to most users unless specific accessibility needs are present, e.g., for individuals with visual impairments. The primary issue with these captions lies in their lack of detail; they fail to convey nuanced information

about the images, such as the breed of the dog or the specific name or location of the park.

Moreover, while alternative text (alt-text) in web-crawled data often contains detailed and concrete visual descriptions, current captioning models generally ignore this information. Instead, these models tend to generate captions solely based on the image content, which misses the opportunity to enhance the relevance and accuracy of the captions.

Additionally, advancements in caption quality often lack transparency and are not easily reproducible. For instance, recent developments such as LLaVA (Liu et al., 2024b) and ShareGPT4V (Chen et al., 2023b) utilize high-quality captions derived from proprietary models like GPT-4V. While these models benefit from high-quality annotations, they are built on processes that are not openly shared. This lack of disclosure presents significant challenges in terms of scalability, intellectual property rights, data integrity and privacy. The use of such proprietary models in industry applications is fraught with risks, particularly when the implementation details remain undisclosed.

This paper presents a principled approach to enhance caption quality and develops a parameter-efficient captioner capable of scaling re-captioning efforts. We assume each image contains information that the caption needs to align with using natural language. Although obtaining the real-world information from an image or generating a perfect ground-truth caption might be challenging, we demonstrate that caption *quality* can be improved relatively by *iteratively refining* captions to better describe the visual content (e.g., adding information on specific objects, colors, spatial relations or more fine-grained named entities).

Our key insight is that the creator who posts an image along with its associated alt-text is likely the most knowledgeable expert regarding the concrete visual concepts within that image (e.g., knowing that the animal is an "iguana" instead of just an "ob-

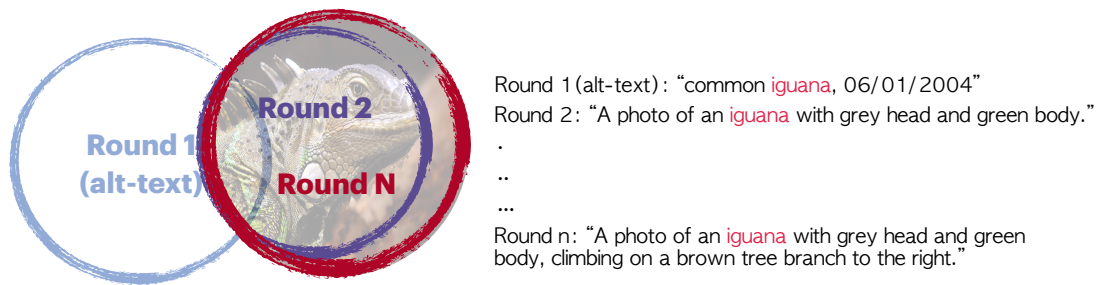


Figure 1: A Venn diagram illustrating caption quality improvement via multiple rounds of *re-aligning* previous captions (starting from alt-text) to the image.

ject," "animal," or "lizard"). It would be difficult for an average annotator to provide similar level of detail within a short annotation timeframe. Instead, these annotators could offer weak yet complementary supervision by either removing non-existent information from the alt-text or describing missing objects using more general concepts ("lizard" instead of "iguana").

Building on this insight, we introduce *Altogether*, an approach to improve image captions through the process of *re-aligning existing alt-texts* with the image content. We instantiate this idea in two forms (i) through human *annotation* to create a fine-tuning dataset and (ii) through a parameter-efficient *captioner* that can re-caption billions of images when fine-tuned for this task.

For annotation (i), we perform multiple rounds of alt-text realignment to preserve concrete visual concepts while adding or removing relevant information, as depicted in Fig. 1. Starting with the initial alt-text, which may partially overlap with the image, subsequent annotation rounds iteratively refine the captions to achieve better alignment with the image’s information. Using this data, we can train a captioner (ii) that is capable of generalizing this process by *reading, grounding, and transforming alt-texts* into dense captions at scale.

We evaluate our re-aligned captions across captioning, generative and discriminative tasks. With a lightweight text decoder, our captioner surpasses alt-texts by 4% in CLIP (Radford et al., 2021) score and outperforms state-of-the-art captioners on a challenging test set, which we annotate based on a subset of the WIT (Wikipedia Image-Text) dataset (Srinivasan et al., 2021). We further evaluate our approach on text-to-image (T2I) generation, where we observe significant improvements in similarity between generated images and text prompts when training latent diffusion models with synthetic captions. For discriminative tasks, we

obtain 1.1% absolute accuracy improvement over 26 zero-shot classification datasets and a 3% gain on retrieval tasks, when using synthetic captions to supplement CLIP training. An interesting observation we make is that generative and discriminative tasks require *widely* different ratios (100% vs. 15%) of synthetic data.

2 Related Work

Synthetic Data and Image Re-captioning. Synthetic data has recently regained popularity (Nguyen et al., 2024; Li et al., 2023b) with DALL·E 3 (Betker et al., 2023) replacing low-quality web data with synthetic data for learning image generators. Since the alt-text of web images serves various purposes and may not fully align with the images they describe, DALL·E mixes alt-texts with synthetic captions to promote better control in image generation. Early work (Chandu et al., 2020) uses sub-selecting content words as skeletons to help generating improved and denoised captions. Another very recent line of concurrent research uses LLMs to fuse or combine alt-texts with captions generated from an off-the-shelf captioner (Lai et al., 2024; Yu et al., 2024). However, the fusion is in language space only and has no access to the image for alignment. The resulting text may include information not present in the image and the fusion behavior of the LLM is unknown for alt-texts. See Table 3 for potential issues of not using vision information.

Dense Captioning. While image captioning is well-studied, generating dense captions precisely aligned with the original images has gained more attention recently. MSCOCO-style captions (Lin et al., 2014) are brief and describe main objects, limiting their value for aligned image-text pairs due to their brevity, general concepts, and constrained image distribution. The DCI dataset (Urbanek et al.,

2023) overcomes the brevity issue but still suffers from the other limitations. DOCCI (Onoe et al., 2024) and ImageInWords (IIW) (Garg et al., 2024) address these challenges for specific datasets using clustering or iterative refinement with object detection tools. Our work proposes a general process to improve caption quality for web images, paving the way for further advancements in this area.

Retrieval Augmented Generation. Realigning alt-texts inherently grounds the captioner on input alt-texts, which is analogous to Retrieval Augmented Generation (RAG) (Lewis et al., 2020; Gao et al., 2023) in terms of taking additional knowledge as input. Image captioning also adopts RAG for caption generation (Ramos et al., 2023; Yang et al., 2023). Our captioner shares similar advantages, such as a parameter-efficient, lightweight model for training and inference at scale, reduced factoid hallucination, and updating knowledge at inference time unavailable during training.

Human Preference Alignment. Image captioning, as an alignment problem between captions and corresponding images, relates to alignment for human preference (Ouyang et al., 2022). However, image captioning alignment is more objective due to the clear target of aligning with information present in the image, whereas human preference alignment is subjective, as preferences can be undefined and vary among individuals.

3 Altogether: Re-aligning Alt-texts

This section presents our method for re-aligning alt-texts to produce dense captions with concrete visual concepts, which we later (§4) instantiate in a parameter-efficient captioner scalable to billions of images. We structure this section into three main parts: (§3.1) revisiting the image captioning task, (§3.2) incorporating re-alignment into existing captioning frameworks, as well as designing annotation tasks (§3.2.1) and learning mechanisms (§3.2.2) for re-aligning alt-texts.

3.1 Image Captioning

We formulate image captioning by predicting caption tokens conditioned on the latent space of an image embedding. The loss function is defined as:

$$L(t, i) = \sum_j \log P(t_j | t_{j-k}, \dots, t_{j-1}; F(i); \Theta), \quad (1)$$

where i represents an image, $F(i)$ its encoding (e.g., CLIP), $t_{j-k:j-1}$ the preceding caption tokens, and Θ the parameters of the captioner. The process involves encoding the image into a latent space and sequentially decoding the caption tokens.

3.2 Re-aligning Previous Captions

To enhance caption accuracy, we condition the captioner on previous captions (e.g., alt-texts),

$$L(t, t', i) = \sum_j \log P(t_j | t_{j-k}, \dots, t_{j-1}; t'_{1:m}; F(i); \Theta), \quad (2)$$

where $t'_{1:m}$ are tokens from the previous caption. This re-alignment aims to refine and better align t' with the image content i .

3.2.1 Annotation

We improve caption quality through *iterative* human annotation, *refining* previous captions (alt-texts) in multiple rounds. Starting with an initial alt-text as caption t , the next round uses:

$$t' \leftarrow t. \quad (3)$$

This iterative process is designed based on the following observations: (i) the creator of alt-texts is possibly the best expert/annotator who can describe the image in fine-grained visual concepts, and it could be very challenging later for an annotator to understand and caption the image at that detail (e.g., identify and specify “iguana” in the caption); (ii) it is also challenging for an annotator to write a detailed caption from scratch, compared to starting from existing information.

In experiments, we show that this iterative process of re-aligning improves the annotated data, captioner, and downstream performance after different rounds of annotation.

3.2.2 Learning

We design a captioner to learn the process of *re-aligning* alt-texts. We build on a simple prefix language model, ClipCap (Mokady et al., 2021), that connects a CLIP encoder and a text decoder via mapping network to implement eq. (1), see Fig. 2.

Mapping Network. The mapping network is a Transformer taking CLIP embeddings as input and produces visual tokens of fixed length (40 is default) that can be fed into a text decoder as the “image prompt”.

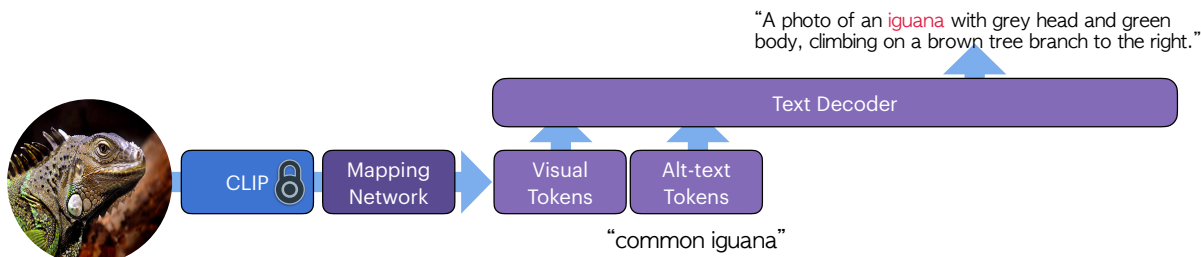


Figure 2: Re-aligning alt-texts: Our captioner takes visual *and* alt-text input. We extract frozen CLIP image embeddings and transform it into a fixed number of visual tokens. Given alt-text, the decoder is able to ground this information, e.g. carrying concrete visual concepts, to generate a better caption that is aligned with the image.

Re-aligning Alt-Texts. To model inputs on alt-texts, we simply append m tokens from alt-texts, after the visual tokens. The training loss is only computed on tokens from generated captions, excluding tokens from both visual and alt-text tokens, as shown in Fig. 2. Note the alt-texts can be empty strings when these are not available.

4 Altogether: Implementation Details

In this section, we first discuss the annotation and training data for our captioning model in §4.1. Then we describe captioner architecture in §4.2.

4.1 Dataset

We use a pre-training + fine-tuning framework to train the captioner, where the goal of pre-training is to learn diverse visual concepts and the later fine-tuning learns to re-align alt-texts as resulting captions.

Pre-training Set For pre-training, we randomly select 22M image-alt-text pairs from the MetaCLIP (Xu et al., 2024) dataset. This data covers long-tailed visual concepts in alt-texts which typically an average human annotator cannot infer from the image content.

Fine-tuning/Annotated Set. We build a fine-tuning set (called *altogether-ft*) to learn and generalize the capability of *re-aligning alt-texts*. We collect 23k images and have 3 rounds of annotation (including alt-texts as the first round). We choose two image sources: 15k images from WIT and 7k images from MetaCLIP (Xu et al., 2024). We use these two sources to ensure rich visual concepts in alt-texts and good coverage on web images in order to mitigate the risk of inference on out-of-domain images. We show the annotation guidelines in Appendix §A and side-by-side comparison of multiple rounds of annotation in Table 14 and Table 15.

4.2 Captioner Architecture

Image Encoder. We choose the pretrained MetaCLIP ViT-H/14 (Xu et al., 2024) as the image encoder, which outputs a single embedding with 1024 dimensions. The image embedding is then transformed into 40 visual tokens via the mapping network to serve as the image prompt for the text decoder. We freeze the image encoder during the training phase and only train the mapping network.

Text Decoder. We adopt a trainable OPT 1.3B (Zhang et al., 2022) as the text decoder for efficient training and inference (e.g., compared to Llama-13B, the throughput of this architecture is $13\times$ faster, see Table 9). We append $m = 128$ tokens from alt-texts after visual tokens and allow a maximum of 256 tokens for generated captions. This extends the total length of decoder to be 424 (40 visual tokens + 128 alt-text tokens + 256 generated tokens). For alt-text tokens, we randomly sample either alt-text or empty text during training. The empty text allows the captioner to generate captions from scratch, in case the alt-texts are not available for the image. We pre-train the captioner for 1 epoch and fine-tune on annotated data for 4 epochs. Detailed hyperparameters are in §F.

5 Evaluation

Our evaluation spans three areas: (i) human annotations, (ii) captions generated from our captioner, and (iii) downstream tasks using our synthetic captions (i.e., text-to-image *generation* and zero-shot image *classification*).

5.1 Annotated Data

We analyze annotations in terms of length (number of words), edit distance (between annotation rounds), and CLIP image-text alignment score.

Captioner	CLIP Score	BLEU 1	METEOR	ROUGE	CIDEr	NP F1	NP Precision	NP Recall
alt-text (Round 1)	29.3	5.1	9.5	17.8	4.7	13.5	9.3	36.5
GiT	26.3 (-3.0)	0.0 (-5.1)	2.1 (-7.4)	7.3 (-10.5)	0.0 (-4.7)	1.8 (-11.7)	1.0 (-8.3)	11.3 (-25.2)
BLIPv2	28.0 (-1.3)	0.2 (-4.9)	4.1 (-5.4)	13.0 (-3.8)	0.0 (-4.7)	4.2 (-9.3)	2.5 (-6.8)	14.4 (-22.1)
LLaVAv1.6	27.0 (-2.3)	27.7 (+22.6)	10.5 (+1.0)	20.2 (+2.4)	4.9 (+0.2)	5.8 (-7.7)	5.5 (-3.8)	6.7 (-29.8)
GPT-4V	27.4 (-1.9)	26.7 (+21.6)	10.0 (+0.5)	17.4 (-0.4)	3.7 (-1.0)	4.4 (-9.1)	4.4 (-4.9)	4.9 (-31.6)
GPT-4V-turbo	27.3 (-2.0)	21.4 (+16.3)	9.0 (-0.5)	17.3 (-0.5)	4.4 (-0.3)	4.4 (-9.1)	4.0 (-5.3)	5.5 (-31.0)
GPT-4o	28.3 (-1.0)	18.8 (+13.7)	8.8 (-0.7)	17.7 (-0.1)	4.0 (-0.7)	5.0 (-8.5)	4.3 (-5.0)	7.0 (-29.5)
<i>Altogether</i> ⁽²⁾ w/ alt	33.1 (+3.8)	50.0 (+44.9)	21.5 (+12.0)	37.9 (+20.1)	48.2 (+43.5)	24.0 (+10.5)	24.1 (+14.8)	25.4 (-11.1)
<i>Altogether</i> ⁽³⁾ w/o alt	32.4 (+3.1)	45.7 (+40.6)	18.7 (+9.2)	34.1 (+16.3)	27.7 (+23.0)	19.2 (+5.7)	18.9 (+9.6)	20.9 (-15.6)
<i>Altogether</i> ⁽³⁾ w/ rand alt	29.4 (+0.1)	44.6 (+39.5)	18.0 (+8.5)	33.0 (+15.2)	24.5 (+19.8)	18.7 (+5.2)	18.7 (+9.4)	20.0 (+16.5)
<i>Altogether</i> ⁽³⁾ w/ alt	33.3 (+4.0)	49.6 (+44.5)	21.9 (+12.4)	39.1 (+21.3)	55.6 (+50.9)	25.2 (+11.7)	24.9 (+15.6)	27.3 (-9.2)

Table 1: Evaluation of captioners on a separate test set created from the WIT dataset. We evaluate the CLIP image-text alignment score, captioning metrics which measure alignment of the model captions with *ground-truth* human annotated captions: BLEU / METEOR / ROUGE / CIDEr and noun phrase (NP) F1, precision, and recall. *Altogether*^(2/3) indicates our captioner fine-tuned on round 2/3 annotation; ‘w/o alt’ means captioning from scratch with no alt-text (similar to other baselines), ‘w/ random alt’ means captioning with randomly paired alt-texts and ‘w/ alt’ means captioning via re-aligning alt-texts.

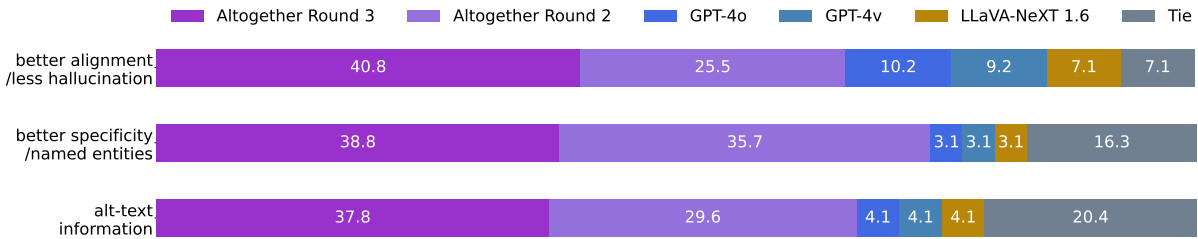


Figure 3: Human evaluation on generated captions on better alignment / less hallucination (“which caption has the best alignment with the image and least hallucination”), specificity (“which caption contains more named entities”) and usefulness of alt-text information (“which caption contain most useful information from alt-texts”).

Annotation	Length	Edit Dist.	Alignment
Round 1 (alt-text)	13.0	-	30.1
Round 2	81.7	403.8	33.7
Round 3	83.2	92.9	33.9

We observe that multiple rounds of annotation (on top of the alt-text) increases the caption length and image-text alignment (CLIP score), with smaller changes in subsequent rounds. This is also reflected by the lower edit distance in the final round. We show further annotation examples in Appendix §B.

5.2 Captioner

Human-annotated Test Set. We believe that existing datasets such as MSCOCO captions are not sufficient for evaluation, since these do not contain fine-grained information, e.g. a caption “a dog sitting in a park” does not contain information about the dog breed or park. Further, existing works (Moon et al., 2023; Onoe et al., 2024) show performance on such benchmarks correlates inversely with caption quality. Therefore, we annotate a test set, consisting of 500 images from the WIT dataset using our 3-round annotation approach and compare our captioner to state-of-the-art captioners.

We use 3 versions of our captioner, after finetuning Round 2/3 annotations, as well as with (w/ alt) and without (w/o alt) feeding alt-text.

We first evaluate the alignment between the images and captions via CLIP score (Hessel et al., 2021) (this metric ignores the ground-truth captions and only uses CLIP similarity as metric). The results are summarized in Table 1, second column. Our *Altogether* captioner improves over alt-texts by 4% on CLIP score and significantly outperforms off-the-shelf captioners such as GiT (Wang et al., 2022; Li et al., 2023a), BLIP2 (Li et al., 2023a) and LLaVA (Liu et al., 2024b,a). It also outperforms proprietary captioners such as GPT-4V (OpenAI, b) and GPT-4o (OpenAI, a). The captions generated by our captioner trained with Round 3 annotation without alt-texts is worse than with alt-texts. This implies that employing alt-texts is important for improving image-text alignment.

Next, we compare the generated captions against the ground-truth provided by the annotators. We use BLEU/METEOR/ROUGE/CIDEr metrics and noun phrase (NP) precision, recall and F1 score. We use spaCy <https://spacy.io> to get two

sets of NPs from generated and ground-truth captions, respectively; then we compute the intersection of these two sets as true positives. We observe that *Altogether* significantly outperforms existing captioners. Non-dense captioners (e.g., GiT or BLIPv2) are struggling to fully describe the image with enough visual concepts (e.g., see BLIPv2’s low scores across all metrics). *Altogether* also outperforms dense captioners (GPT-4V/o or LLaVA1.6), even if our model is not provided with the alt-text. If we provide the model with the alt-text we see a further boost in performance. This can be explained by the long-tailed visual concepts present in alt-texts (Xu et al., 2024), which is difficult for dense captioners to describe purely using image information.

Low Performance of GiT and BLIPv2. We further investigate 0.0 CIDEr scores of GiT and BLIPv2. One reason is from using long-tailed dense captions (averaging over 80 words) as reference to compute CIDEr that penalizing short captions because CIDEr has a length penalty. Also, both GiT and BLIPv2 are trained on the MSCOCO dataset, which typically features captions of less than 10 words focused on common objects. We further fine-tune GiT on *altogether-ft* set for fair comparison, shown in Table 2. GiT is still far left behind *Altogether*, probably because of lacking alt-texts pre-training. Moreover, the WIT dataset includes many out-of-domain images for which these models are not optimized, leading to partial recognition issues (e.g., recognizing “sand on the beach” but failing to detail it further). Occasionally, this mismatch in training and testing also results in the generation of unreadable captions.

Baseline	CLIP Score	BLEU 1	METEOR	ROUGE	CIDEr
GiT (MSCOCO)	26.3	0.0	2.1	7.3	0.0
GiT ⁽³⁾ w/o alt	26.5	17.6	13.5	19.8	0.0

Table 2: Fine-tuning GiT on *altogether-ft* set.

Human Study. We further conduct human evaluation by presenting the images, alt-texts and the captions produced by various models, and asking evaluators about three criteria: Whether the caption (i) is aligned with the image & has fewer hallucinations; (ii) is specific (named entities, detailed description); (iii) carries useful information from the alt-text. We evaluate 5 captioners with random order when presented: LLaVA1.6, GPT-4V, GPT-4o, and our *Altogether* trained with Round

2/3 data. We use 3 evaluators and 100 images from WIT. The results are in Fig. 3. Humans highly prefer *Altogether*, and Round 3 further improves over Round 2, over the three criteria: *Altogether* is also much better in (i) producing *aligned* image captions *without hallucination* (ii) describing images more *specifically*, (iii) we see alt-texts contain useful information and captioning from scratch (LLaVA1.6, GPT-4V/o) struggles to describe this.

To qualitatively understand the behavior of re-aligning alt-texts, we further prompt the captioner with different alt-texts on images from ImageNet, shown in Table 3. We try 3 different styles of alt-text prompting: (i) empty string, (ii) ImageNet class name, (iii) incorrect alt-texts. We can see that *Altogether* can carry over concrete visual concepts and correct the hallucinated / wrong visual concepts in red that captioning from scratch (empty string) has. It further rejects alt-texts that are incorrect (e.g., alt-text “a bird” that is not present the image).

5.3 Text-to-image (T2I) Generation

Setup. We utilize re-aligned (synthetic) captions for training text-to-image generative models. Using synthetic data was shown in DALL-E 3 (Betker et al., 2023) to be highly effective for generating images. We use PixArt-XL2 (Chen et al., 2023a), with DiT-XL backbone. We use CC-12M (Changpinyo et al., 2021) as the training dataset. We train the model from scratch under a controlled setup to compare the performance difference between using original captions and using re-aligned (synthetic) captions as the text inputs. We use the default PixArt training setup and train on CC-12M for 24 epochs on 32 A100 GPUs. Details are in Table 11.

Results. We train T2I models with different mixing ratios p of synthetic captions and original caption. During inference, following the evaluation setup in DALL-E 3, we apply either the original prompt (alt-text) or the descriptive (synthetic) prompt as the text prompt to generate image. We report CLIP scores to evaluate the similarities between the generated images and the corresponding text prompts on a holdout CC-12M set in Table 5.

We compare T2I models trained on alt-texts (Round 1), synthetic caption with and without alt-texts grounding. Similar to DALL-E 3, we first train T2I model with a high mixing ratio $p=0.95$ of synthetic data w/o alt-texts, mixed with original data (alt-texts). Training with synthetic captions improve the CLIP score by 1.3% (29.3 vs 28.0).





Image	Alt-Texts	Generated Captions
	<p>“”</p> <p>“great gray owl, Strix nebulosa”</p> <p>“a bird”</p> <p>“a bird and a dog”</p>	<p>A close-up photo of a Northern Saw-whet Owl (Aegolius nivalis) in a natural habitat...</p> <p>A close-up photo of a Great Gray Owl, Strix nebulosa. The owl is standing on a grassy ...</p> <p>A close-up photo of a Northern Saw-whet Owl (Aegolius nivalis) in a natural habitat...</p> <p>A close-up photo of a Northern Saw-whet Owl (Aegolius nivalis) in a forest...</p>
	<p>“”</p> <p>“conch”</p> <p>“a rock”</p>	<p>A photo of a seashell on a sandy beach. The shell is a light pink color with ...</p> <p>A photo of a conch shell on a sandy beach. The shell is large and has a spiral shape...</p> <p>A photo of a seashell on a sandy beach. The shell is a light pink color with ...</p>
	<p>“”</p> <p>“gyromitra”</p> <p>“a cat”</p>	<p>A photo of a mushroom, specifically a species of the genus Fusarium...</p> <p>A close-up photo of a mushroom, specifically a species of the genus Gyromitra...</p> <p>A photo of a mushroom, specifically a species of the genus Fusarium...</p>
	<p>“”</p> <p>“spider monkey, Ateles geoffroyi”</p> <p>“a bird”</p>	<p>A photo of a white-faced capuchin monkey (Cebus capucinus) sitting on a tree branch...</p> <p>A photo of a spider monkey, Ateles geoffroyi, sitting on a tree branch. The monkey ...</p> <p>A photo of a white-faced capuchin monkey sitting on a tree branch. The monkey has ...</p>

Table 3: Qualitative evaluation for re-aligning different alt-texts as prompts: We mark concepts carried in alt-texts in blue and erroneous captions without grounded in alt-texts in red. The captioner also rejects hallucinated/general visual concepts in alt-texts in brown. This is only possible by performing alignment with text *and* image information.

Then we train a T2I model with 100% ($p=1.0$) synthetic data, generated by *Altogether* with alt-texts prompting. This yields another 0.5 gain on CLIP score. This indicates DALL-E 3’s 5% mixing with original alt-texts is sub-optimal, not necessary and may at risk of increasing mis-aligned data, *if the synthetic caption is already re-aligned* from alt-text. Ablating ratios of mixing existing captions (alt-text) does make a significant difference.

In Table 4, we qualitatively study the re-aligned captions and show this approach promotes fine-grained control and grounding for text-to-image generation with reduced hallucination.

5.4 Classification and Retrieval

Setup. Following the data curation in MetaCLIP (Xu et al., 2024), we collect 5B image-text pairs as CLIP training data. We follow the standard CLIP training setup for evaluating our approach using a ViT-B/32 architecture as in OpenCLIP (Ilharcó et al., 2021) and MetaCLIP (Xu et al., 2024). The training hyperparameters are in Table 12.

We create 3 sets of captions by running inference on the 5B images, with captioners trained with (i) Round 2 annotation, (ii) Round 3 annotation and (iii) Round 3 without alt-texts prompts.

Results. We show the results of CLIP training by zero-shot evaluation on 26 classification tasks in Table 6. We first study the performance of using *only* synthetic captions (ratio of synthetic captions $p=1.0$). Multiple rounds of annotation help to improve accuracy by 1.5% (Round 2 ($p=1.0$) vs Round 3 ($p=1.0$)). Interestingly, the captioner without re-aligning alt-text (w/o alt-text) struggles (44.5% average accuracy), indicating that re-aligning alt-text in the captioner is important.

The next section of Table 6 shows that training with only alt-text performs better than using only synthetic captions above. We believe this is because the captioner is likely not large enough to carry *all* the alt-text information into the synthetic caption. We then mix alt-text and synthetic captions (ablation in Appendix §D) for training CLIP. With a ratio of $p=0.15$ synthetic captions,



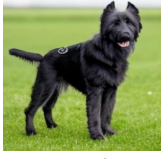
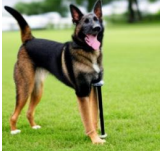


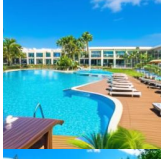


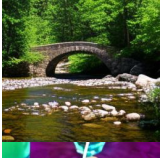


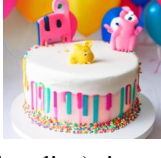
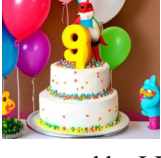
Prompt	Original	<i>Altogether</i>
A hummingbird in mid-air , hovering above a bright red flower. The bird is mostly green with a black head and a long, pointed beak. Its wings are spread wide and blurred due to the fast movement. The flower is a bright red color with five petals and a yellow center . The background is a blurred green, with hints of other leaves and flowers visible.		
A Belgian Malinois dog wearing a prosthetic leg . The dog is standing on a grassy field with a blurred background. The prosthetic leg is made of metal and has a rubber sole. The dog is looking directly at the camera with its mouth open, as if it's smiling. The dog's fur is a mix of brown and black .		
Three potted plants, each placed in a woven rattan basket, isolated on a white background. The plants are of different sizes and species, with one being a tall, leafy plant with a thick stem, another being a shorter, bushy plant with a thin stem, and the third being a small, round plant with a thin stem . The baskets are made of natural-colored wicker and have a braided design.		
A beautiful, modern resort with a large swimming pool and a stunning view of the sea . The pool is surrounded by a wooden deck with lounge chairs and umbrellas , and there are palm trees and other greenery around the pool area. In the background, you can see the blue sea and a few boats sailing on it. The resort buildings are visible in the background, with a mix of modern and traditional architecture.		
A scenic view of a river flowing through a forest. There is a small stone bridge with a few trees growing on either side. The bridge is made of large, rough-hewn stones and has a distinctive arched shape. The river water is clear and shallow, with a few rocks and branches visible beneath the surface. The forest in the background is dense and green, with tall trees stretching up towards the sky.		
Two tacos on a white plate, with a violet background. Each taco has a crispy corn tortilla shell filled with shredded meat, topped with sliced avocado, shredded lettuce , and a sprinkle of red cabbage. There's a dollop of creamy sauce on top of each taco. There are two glasses of drinks , one with a pink straw and the other with a yellow straw , placed on either side of the plate.		
A colorful birthday cake topped with a large number 9 made of fondant and decorated with colorful sprinkles. There are also several small fondant decorations on top of the cake, including a yellow chick, a pink pig, and a blue bird . The cake is placed on a white cake stand and surrounded by colorful balloons.		

Table 4: **Text-to-Image Generation**. In each group, *left*: Text prompt; *middle* (baseline): image generated by LDM trained with original captions; *right*: image generated by LDM trained with *Altogether* synthetic captions (Round 3). Hallucinations and errors generated by **baseline**, **Altogether** or **both** are marked with colors. As observed, an LDM trained with *Altogether* data follows text instruction *closer* and *improves image-prompt alignment* in complex scenes and specialized entities (e.g. “a Belgian Malinois dog”).

Training Data	Inference Prompt	
	Original	Synthetic
alt-texts (Round 1)	27.0	28.0
<i>Altogether</i> ⁽³⁾ , w/o alt-texts, $p=0.95$	27.1 (+0.1)	29.3 (+1.3)
<i>Altogether</i> ⁽³⁾ , w/ alt-texts, $p=0.75$	27.2 (+0.2)	29.6 (+1.6)
<i>Altogether</i> ⁽³⁾ , w/ alt-texts, $p=0.95$	27.3 (+0.3)	29.8 (+1.8)
<i>Altogether</i> ⁽³⁾ , w/ alt-texts, $p=1.0$	27.3 (+0.3)	29.8 (+1.8)

Table 5: Evaluation of text-to-image generation on CC-12M: CLIP similarity scores between prompts (original or synthetic) and generated images.

we see a +1.1% improvement over 26 classification tasks (Table 6), showing how re-align can provide complementary information for CLIP training. Finally we train a large ViT-H/14 model with

mixed *Altogether* captions and observe 73.2% average accuracy compared to the 72.4% with the same model in MetaCLIP (Xu et al., 2024).

Finally, we evaluate on zero-shot text-to-image retrieval tasks from DataComp (Gadre et al., 2023). Results are in Table 7. Mixing alt-text with synthetic captions leads to +3% for retrieval on ViT-B and even larger gains over MetaCLIP ViT-H/14.

Discussion. An interesting observation is that image generation and classification require different amount of mixing ratios for synthetic captions—the optimal mixing ratio is $\sim 100\%$ for T2I generation whereas as low as $\sim 15\%$ for CLIP classification. The root cause may stem from very different definitions of these two problems: T2I needs fully

	Average	ImageNet	Food-101	CIFAR10	CIFAR100	CUB	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	MNIST	FER-2013	STL-10	EmoSAT	RESCIS45	GTSRB	KITTI	Country211	PCAM	UCF101	Kinetics700	CLEVR	HaeflinMemes	SST2
ViT-B/32																											
<i>Altogether</i> ⁽²⁾ ($p=1.0$)	52.3	51.5	68.7	90.2	70.4	47.5	57.8	67.0	13.2	37.7	67.2	88.4	51.6	64.0	43.0	95.4	50.0	57.0	44.8	15.2	8.6	54.2	54.1	37.8	23.9	51.5	50.0
<i>Altogether</i> ⁽³⁾ (w/o alt-text)	44.5	39.8	47.4	88.6	65.7	14.8	50.0	54.4	4.9	29.8	54.2	79.2	30.4	71.9	25.7	89.6	39.3	54.2	37.9	23.9	5.1	53.5	47.4	31.5	15.0	54.9	49.2
<i>Altogether</i> ⁽³⁾ ($p=1.0$)	53.8	52.8	70.0	90.4	71.4	47.7	57.4	67.5	14.7	41.5	69.1	88.4	50.6	62.9	42.1	94.7	56.1	55.1	48.8	33.0	8.9	57.2	56.8	38.7	23.0	52.0	48.9
Alt-text (Round 1)	59.3	68.1	84.4	93.1	74.5	66.5	67.2	77.9	27.9	59.4	90.7	91.7	72.0	25.1	45.1	97.0	45.8	63.3	37.0	30.1	18.8	63.3	67.5	47.7	19.1	55.9	52.4
<i>Altogether</i> ⁽²⁾ ($p=0.15$)	60.3	67.9	84.1	92.1	75.3	66.7	67.1	78.2	25.1	58.8	89.4	92.5	70.3	37.4	40.2	95.7	55.0	67.3	38.3	31.9	18.0	59.7	67.4	48.0	33.1	56.2	52.9
<i>Altogether</i> ⁽³⁾ ($p=0.15$)	60.4	68.2	84.3	92.7	75.6	67.0	67.1	77.8	25.6	62.6	89.1	92.6	71.2	36.7	44.5	96.8	53.2	63.8	38.6	35.9	18.8	58.2	68.1	48.2	24.2	53.5	55.1
ViT-H/14																											
MetaCLIP	72.4	80.5	94.2	98.0	86.4	83.4	74.1	90.0	50.2	72.4	95.4	95.6	85.1	72.7	55.2	99.4	66.3	74.6	62.5	38.2	37.2	65.8	82.2	64.1	30.1	59.3	69.2
<i>Altogether</i> ⁽³⁾ ($p=0.15$)	73.2	82.1	95.0	97.8	87.1	88.6	74.6	93.1	63.2	73.0	95.9	95.9	86.8	86.1	54.6	99.5	70.3	76.0	57.9	28.1	43.3	50.1	85.4	65.4	32.5	58.3	62.5

Table 6: Results on 26 CLIP zero-shot classification tasks. First section: Training with pure ($p=1.0$) synthetic captions from our captioners that were trained after different rounds of annotations. Second section: Mixing in alt-text during training (ratio of $p=0.15$). Third section: Comparison of a large ViT-H/14 model trained on our synthetic captions with mixed alt-text outperforms MetaCLIP (Xu et al., 2024) (72.4 vs. 73.2 average accuracy).

	Avg. retrieval	Flickr	COCO	IN Dist. Shift	VTAB
ViT-B/32					
Alt-text (Round 1)	52.6	72.9	46.6	52.3	55.3
<i>Altogether</i> ⁽³⁾ ($p=1.0$)	46.1	69.0	42.8	41.7	47.8
<i>Altogether</i> ⁽³⁾ ($p=0.15$)	55.6	76.0	48.9	52.5	55.9
ViT-H/14					
MetaCLIP	60.4	85.0	57.5	66.1	64.6
<i>Altogether</i> ⁽³⁾ ($p=0.15$)	65.7	87.6	60.7	67.3	66.2

Table 7: Zero-shot retrieval evaluation.

aligned captions to have text controlling the generated images in every detail; whereas the problem of CLIP only needs to recognize a single class name from a long-tailed vocabulary.

6 Conclusion

This paper presents *Altogether*, a principled way of improving image captions by re-aligning existing alt-text to images. Re-aligning alt-text allows concrete visual concepts to be carried into the resulting caption. In experiments, we show that a lightweight captioner trained to perform this task can generate captions with significantly better captioning performance than alternatives. We further observe that the resulting captions can be used for improving both text-to-image generation and zero-shot recognition across a broad set of tasks.

7 Limitations

We observe the following limitations in this work:

1. Evaluating captions with rare and specific concepts is challenging for the following reasons.
 - (i) Re-aligned alt-texts can contain superhuman information (think e.g., a very specific model type of a car or boat is not known to the majority of people). It is challenging to verify correctness, even by human evaluators.
 - (ii) There is no perfect metric to quantify the

overall quality of alt-texts and complementary information added via re-aligning.

(iii) Lack of external high-quality ground-truth captions (that describe both alt-text and complementary information well). Note a higher quality benchmark can evaluate a lower quality caption, but not the reverse. For example, existing literature reports that benchmarks such as MSCOCO or Flickr contain only well-known visual concepts and are negatively correlated with human evaluation (IIW (Garg et al., 2024)) or higher quality benchmarks (AnyMAL (Moon et al., 2023)).

2. Due to limited compute, we cannot evaluate image generation at a larger scale.
3. Current synthetic captioning can improve alignment but cannot go beyond the concrete visual concepts described in alt-texts to improve challenging benchmarks such as ImageNet classification.
4. Working on large multimodal language models faces various constraints, including being competitive without using data from proprietary models (the community is actively distilling information from models such as GPT-4V), which leads to lack of transparency (black-box LLMs). In this work we aim to show a principled way of improving image captions with maximally preserving transparency. We will make our code, models and data available for future use.

Acknowledgments

We thank Xian Li, Ping Yu, Yuandong Tian, Chunting Zhou, Armen Aghajanyan and Mary Williamson for the insightful discussion.

References

- James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3):8.
- Khyathi Raghavi Chandu, Piyush Sharma, Soravit Changpinyo, Ashish Thapliyal, and Radu Soricut. 2020. Denoising large-scale image captioning from alt-text data using content selection models. *arXiv preprint arXiv:2009.05175*.
- 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. In *CVPR*.
- Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. 2023a. Pixart- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis. *Preprint*, arXiv:2310.00426.
- Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023b. Sharegpt4v: Improving large multimodal models with better captions. *arXiv preprint arXiv:2311.12793*.
- Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Musmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. 2023. *Datacomp: In search of the next generation of multimodal datasets*. *Preprint*, arXiv:2304.14108.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
- Roopal Garg, Andrea Burns, Burcu Karagol Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Baldridge, and Radu Soricut. 2024. Imageinwords: Unlocking hyper-detailed image descriptions. *arXiv preprint arXiv:2405.02793*.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.
- Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. *Openclip*. If you use this software, please cite it as below.
- Zhengfeng Lai, Haotian Zhang, Bowen Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, Yinfei Yang, and Meng Cao. 2024. *Veclip: Improving clip training via visual-enriched captions*. *Preprint*, arXiv:2310.07699.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K \ddot{u} ttler, Mike Lewis, Wen-tau Yih, Tim Rockt \ddot{a} schel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.
- Wenyan Li, Jonas F Lotz, Chen Qiu, and Desmond Elliott. 2023b. The role of data curation in image captioning. *arXiv preprint arXiv:2305.03610*.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll \acute{a} r, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. *Llava-next: Improved reasoning, ocr, and world knowledge*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.
- Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, et al. 2023. Anymal: An efficient and scalable any-modality augmented language model. *arXiv preprint arXiv:2309.16058*.
- Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. 2024. Improving multimodal datasets with image captioning. *Advances in Neural Information Processing Systems*, 36.
- Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, et al. 2024. Docci: Descriptions of connected and contrasting images. *arXiv preprint arXiv:2404.19753*.

- OpenAI. a. Gpt-4o. <https://openai.com/index/hello-gpt-4o>. Accessed: 2024-05-13.
- OpenAI. b. Gpt-4v. https://cdn.openai.com/papers/GPTV_System_Card.pdf. Accessed: 2023-09-25.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjheva. 2023. Smallcap: lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2840–2849.
- Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449.
- Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. 2023. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. *arXiv preprint arXiv:2312.08578*.
- 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. *arXiv preprint arXiv:2205.14100*.
- 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*. In *The Twelfth International Conference on Learning Representations*.
- Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, et al. 2023. Re-vilm: Retrieval-augmented visual language model for zero and few-shot image captioning. *arXiv preprint arXiv:2302.04858*.
- Qiyang Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wang, and Jingjing Liu. 2024. Capsfusion: Rethinking image-text data at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14022–14032.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

A Annotation Guidelines

This section details our annotation guidelines. We highlight the overall goal and good practice for annotation first, then show the detailed instructions for annotators in Fig. 5. Our annotations aim to enhance the *alignment* between image and existing captions. We use the metadata of the image (i.e., alt-text attributes) as the starting point. The alt-text is considered to contain ground truth information of the image but only partially describes the image. The goal of our annotation is to significantly improve image-caption alignment and make the caption *just* right: e.g., do not mention missing objects in the image or information beyond the image content.

Good Practices

- We use short prompts as the starting points of captions: such as “a photo of”, “a painting of”, “a sculpture of”, instead of verbose prompts such as “This is an image showing ...”. We provide a recommended list of starting prompts in Table 8.
- We provide annotation steps to guide the annotator’s workflow during annotation. See “Annotation Steps” in Fig. 5.
- We further provide a checklist to help annotators confirm if they follow each step of the guidelines well. Fig. 6 provides a screenshot of our annotation interface.
- We leverage two vendors for annotation and ask each vendor to rewrite/criticise the other vendor’s annotation from the previous round. We split the data to annotate between the two vendors, and swap the data in the next round.

“a photo of”
“a product photo of”
“a low resolution photo of”
“a cropped photo of”
“a close-up photo of”
“a black and white photo of”
“a blurry photo of”
“a rendering of”
“a sculpture of”
“a painting of”
“a cartoon of”

Table 8: Recommended starting prompts for captioning annotation.

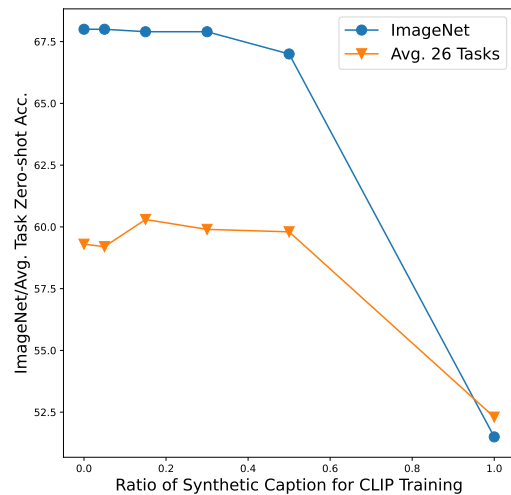


Figure 4: Zero-shot classification accuracy on ImageNet and averaged 26 CLIP tasks with different ratio of mixing synthetic captions during training of various CLIP ViT-B/32 models.

B Side-by-side Comparison of Multiple Rounds of Annotation

We show side-by-side comparison of annotations in Table 14 for WIT images and Table 15 for MetaCLIP images (images are not shown).

C Altogether Evaluated on MSCOCO

The *Altogether-ft* fine-tuning set is very different in style from the popular captioning dataset MSCOCO. As a reference, we also report performance on MSCOCO 2017 as the reference caption in Table 13.

D Ratio of Mixing Synthetic Captions for CLIP Training

We ablate different mixing ratios of synthetic captions vs. ImageNet zero-shot accuracy, and average accuracy across the 26 CLIP datasets in Fig. 4 and notice that a high ratio of synthetic caption can reduce the performance significantly. A good trade-off ratio is around 15%, which allows synthetic caption to complement alt-text, which is our default value throughout the paper. This is likely due to two reasons: (i) human annotation optimizes alignment and is conservative on alt-texts when it concerns ambiguous image information. For example, a “\$18/night room” in alt-texts could still supervise an image having a room of poor condition but is at risk of having mis-aligned description

Decoder	Seq. Len.	Imgs per Second	GPU Days for 1B Imgs	Days on 256 GPUs for 3B Imgs
Llama 2 13B Chat (w/o alt-texts)	296	2.6	4452	52.2
OPT 1.3B (w/o alt-texts tokens)	296	19.7	589	6.8
OPT 1.3B (w/ alt-texts tokens)	424	15.6	740	8.6

Table 9: Throughput of different text decoders measured on NVIDIA A100 80GB GPUs.

Hyperparameter	
Arch.	ClipCap(Mokady et al., 2021)
Frozen Encoder	MetaCLIP (Xu et al., 2024)
Resolution	224×224
CLIP Embedding Size	1024
Visual Tokens	40
Trainable Decoder	OPT 1.3B
Attention	Flash Attention 2
Batch Size	512
Learning Rate	1e-3
Minimal Learning Rate Ratio	0.1
Warm-up	2k
Pre-training Data	MetaCLIP 22M
Pre-training Steps	44k
Fine-tuning Data	WIT 15k + MetaCLIP 7k
Fine-tuning Steps	96
Temperature	0.2
Top-p sampling (nucleus sampling)	0.7

Table 10: Hyperparameters of captioner training.

Hyperparameter	PixArt- α
Arch.	DiT-XL
Activation Function	GELU
Training Data	CC12M
Image Size	256
Batch Size	8192
Learning Rate	2.0e-5
Warm-up	1000
Training Epochs	24

Table 11: Hyperparameters of text-to-image generation training.

Hyperparameter	ViT-B/32	ViT-H/14
Activation Function	QuickGELU	GELU
Seen Pairs	12.8B	51.2B
Batch Size	32768	120k
Learning Rate	5.0e-4	4.0e-4
Warm-up	2k	2k

Table 12: Hyperparameters of CLIP training.

Baseline	CLIP Score	BLEU 1	METEOR	ROUGE	CIDEr
COCO annotation	30.37	-	-	-	-
<i>Altogether</i> ⁽³⁾ w/o alt	33.69	17.5	17.3	19.0	0.0

Table 13: Altogether evaluated on MSCOCO.

on price, so an annotator may remove that from alt-text; and (ii) existing benchmarks such as classification/retrieval test specific (object) classes instead of whole image alignment.

E Throughput of Different Text Decoders

To scale captioner inference to billions of images, we ablate the throughput of different decoder setups in Table 9. We note that using such an LLM is 13.2× slower than OPT (2.6 vs. 19.7 images per second).

F Hyperparameters

We detail the hyperparameters of the captioner in Table 10, downstream text-to-image training in Table 11 and CLIP training in Table 12, respectively.

Goal The goal of this task is to enhance the alignment in-between image and caption via caption editing, leveraging the metadata of the image (i.e. alt-text attributes). The collected data will be used to train a rewrite model for caption generation. The factoid knowledge and concrete visual concepts in alt-text is expected to be added to improve the caption and *no extra* personal knowledge from annotators are expected as part of the improved caption.

Task Description We provide a pair of (*image, alt-text*) to annotators, and ask annotators to *leverage the provided alt-text as factoid knowledge* and rewrite to improve the alignment between the *caption* and the image. A better alignment means: 1) *removing any nonfactual parts* in the caption; 2) *adding missing information* into the caption (object shown in the image but not mentioned in caption). If the image-caption pair is 90% aligned, make it 99% aligned.

Annotation Steps

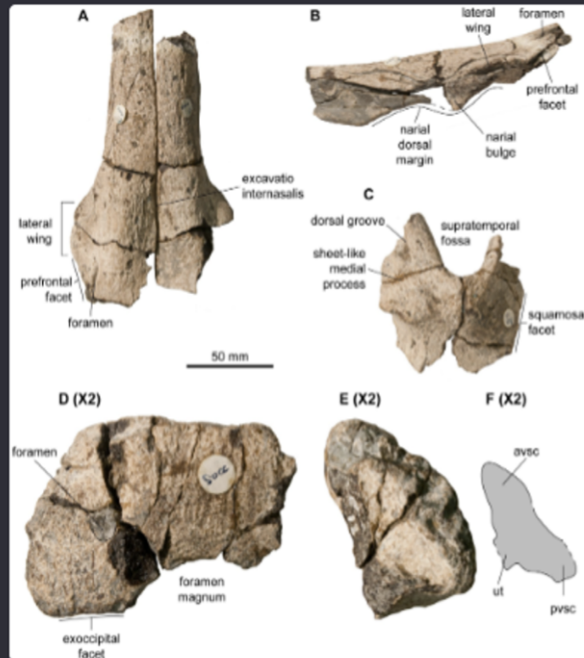
1. Copy and paste the “Previous Caption” to the box of “Rewritten caption”.
2. A concise starting prompt to describe what the image is about, such as “a photo of”, “a product photo of”, depends on types of images, rather than “This image shows. . .”
3. Use alt-text as much as possible if appropriate (mostly in 1st sentence) to improve the factuality of the caption.
 - Paraphrasing is encouraged, but please do not change the meaning of the alt-text.
 - Using concrete visual concepts in alt-texts as much as possible: write “Bentley” (alt-texts) as “a Bentley” instead of “a car”.
 - Alt-texts with metadata such as filenames/dates or “photographed by ABC” can be ignored.
 - Using external tool (e.g., Google) is encouraged to help understand the alt-text.
4. Remove/Edit any hallucinated parts in the caption (anything that’s either not exists in the image or wrongly described, e.g., wrong color)
5. Remove sentence describing theme/feeling of caption, e.g. “overall this image gives an impression of xxx” or imaginative description “this boy must have a bright future.”
6. To the extent the image contains people, please DO NOT provide any information about that person’s
 - racial or ethnic origin (including skin color, hair color, apparent nationality or citizenship);
 - Sexual orientation;
 - Political affiliation;
 - Health condition or disability;
 - Religion;
 - Membership in a Trade Union;
 - Facial features, expression or emotion (e.g. smiling/crying as well as “mood”), hair color (e.g., “dark haired”, “blonde-haired”, etc.);
 - DO NOT add any identifying information about people or objects such as names, address and emails.
7. Add in visible missing details if there’s any.
 - When less certain / in case of blurry image, use vague and general terms to describe the objects such as “This may be NYC” rather than “This is NYC”; or “animal” instead of “dog”/“cat” (when it’s hard to judge detailed type).
 - Transcribe any readable characters in the image.
8. Check the overall structure (deductive structure etc) of the rewritten caption.
 - Make sure everything in the caption is factual.
 - Check the structure of caption (see the next section).

Structure of Caption

1. Caption structure
 - Objects: A good dense caption should follow a “deductive structure” where it typically starts with a general statement, followed by subjects, secondary objects, background, and concluding with minor details.
 - Order of objects: Similar to how a human would usually read images e.g., “left to right”, “top to bottom”, or “near to far” order. Once done with describing the most salient objects, for secondary objects and backgrounds that are hard to sort by saliency, we can arrange secondary objects and background elements in a similar way, depending on the image structure.
 - The default spatial terms is based on viewer’s angle (3rd person); if 1st person view angle is needed, explicitly write down that angle: “on her left is a cute dog”;
 - Describe spatial relation from big to small, from main to accessory: ” ... a cake. There’re 4 cherries on the cake.”
 - Count objects of the same type when it is less than or equal to 10; for more than 10 objects, annotator may use the word “many x”.
 - Long paragraph: Please split a long paragraph into shorter and coherent paragraphs, and organize them with a clear logical order for easier understanding.
2. Caption length
 - Conciseness, correlates with “complexity” of the image. Though we want to have detailed descriptions, we also want to have the details being described in a concise way. If there is only one object present in the image, we shouldn’t have a long paragraph.

Figure 5: Annotation guideline.

Image 1



Alt Text 1

Skull roof bones of the *Acamptoneustes densus* holotype: GLAHM 132588
Skull roof of *Acamptoneustes densus* (GLAHM 132588, holotype). A: articulated nasals in dorsal view. B: left nasal in lateral view. C: right supratemporal in dorsal view. D: supraoccipital magnified two times with respect to the other bones, in posterior view (D) and in left anterolateral (otic) view (E,F). Note the lateral wing of the nasal forming an overhang on the naris, the narial process of the nasal, the long and straight squamosal facet of the supratemporal, and the weakly arched shape of the supraoccipital. Abbreviations: avsc: impression of the anterior vertical semicircular canal; pvsc: impression of the posterior vertical semicircular canal; ut: utriculus.

Previous Caption

A photo of a group of skull roof bones of the *Acamptoneustes densus* holotype; a type of dolphin like reptile from the Cretaceous period. There are displays from A-F showing the different bones with a drawing in the far bottom right corner showing where these bones may have come from. The bones are on a white background and have various parts labeled.

⚠ If the image is broken, distorted, or otherwise compromised, please skip this task and move to the next image.

Rewritten caption

i When re-writing the image caption ensure you are closely following the Instructions & Guidance document.

Submission Checklist

- I confirm the caption begins with Recommended Starting Prompts or similar
- I confirm I edited with Alt-Text as much as possible if appropriate
- I confirm I removed wrong / hallucinated description in the new caption
- I confirm I removed theme/feeling/ambience descriptions in the new caption
- I confirm no facial expressions were described in the new caption
- I confirm the final caption follows an overall "deductive structure" structure.

Figure 6: Annotation interface.

Image	Alt-Text (Round 1)	Round 2	Round 3
	A wooden gilded statue of the bodhisattva Avalokiteśvara from the Chinese Song Dynasty (960-1279), from the Shanghai Museum. Français: Statue en bois doré de bodhisattva Avalokiteśvara. Dynastie Song (960-1279), Musée de Shanghai.	A photograph of a wooden gilded statue of the bodhisattva Avalokiteśvara from the Chinese Song Dynasty . The image depicts a statue of a seated figure placed on a black pedestal. The figure is adorned with elaborate attire and jewelry, suggesting it might represent royalty or divinity. The clothing and accessories feature intricate designs and patterns. Unfortunately, the face of the statue is obscured by a beige-colored rectangle. The background has a dark red hue, creating contrast with the lighter color of the statue.	A photo of a wooden gilded statue of the bodhisattva Avalokiteśvara from the Chinese Song Dynasty (960-1279), from the Shanghai Museum . The image depicts a statue of a seated figure placed on a black pedestal. The figure is adorned with elaborate attire and jewelry, suggesting it might represent royalty or divinity. The clothing and accessories feature intricate designs and patterns. Unfortunately, the face of the statue is obscured by a beige-colored rectangle. The background has a dark red hue, creating contrast with the lighter color of the statue.
	Alwalkeria Illustration of Alwalkeria maleriensis by Karkemish	A 3D render of a dinosaur, specifically an Alwalkeria . It is depicted in a running pose, with its left foot forward and its right foot back. The dinosaur's body is a mix of orange, yellow, and brown colors, with black stripes on its back and tail. It has a long, curved tail and sharp teeth. It is isolated in a white background.	A 3D render of a dinosaur, specifically an Alwalkeria . It is depicted in a running pose, with its left foot forward and its right foot back. The dinosaur's body is a mix of orange, yellow, and brown colors, with black stripes on its back and tail. It has a long, curved tail and sharp teeth. It is isolated on a white background.
	Juvenile male in Tasmania, Australia A male crescent honeyeater (Phylidonyris pyrrhopterus) in Lindisfarne, Tasmania, Australia	A close-up photo of a juvenile male crescent honeyeater bird perched on a gray picket fence. The bird has a dark grey plumage, pale gray underparts, and yellow wing patches. The bird also has a black bill, a pair of dark eyes, and gray legs. A few leaves and branches of a tree are partly visible in the blurring background.	A photo of a juvenile male crescent honeyeater, (Phylidonyris pyrrhopterus) . The bird is perched on a gray picket fence in Lindisfarne, Tasmania, Australia . The bird has a dark grey plumage, pale gray underparts, and yellow wing patches. The bird also has a black bill, a pair of dark eyes, and gray legs. The background has a few leaves and branches of a tree that are blurred.

Table 14: Examples of annotated images from WIT dataset in multi-rounds.

Alt-Text (Round 1)	Round 2	Round 3
Vintage 1992 University of Miami Hurricanes Orange Bowl T-shirt	A photo of a green t-shirt with a straight hem. The t-shirt is made of a thick, high-quality fabric that is soft to the touch. The color is a deep green, almost a forest green, with a slight sheen to it. The image is taken on a yellow background.	A photo of a Vintage 1992 University of Miami Hurricanes Orange Bowl T-shirt. The T-shirt is made of a thick fabric. The color of the T-shirt is a deep green, almost a forest green color. The t-shirt is depicted on a yellow background.
Aqua Recessed swimming Pool 11x11cm MR 16 IP68 Stainless Steel AISI 316	A product photo of Aqua Recessed swimming Pool 11x11cm MR 16 IP68 Stainless Steel AISI 316 light fixture . The image shows a round, stainless steel submersible outdoor pool light fixture with a flat, glass lens. The light is recessed into the fixture and surrounded by a ring of four metal flanges. The flanges have small holes drilled in them. The light fixture is secured to the ground with a large bolt in the center. The light source is not visible, but it appears to be an LED or other small light source. The image is on a white background, and the light fixture is the only object in the image.	A product photo of Aqua Recessed swimming Pool 11x11cm MR 16 IP68 Stainless Steel AISI 316 light fixture . The image shows a round, stainless steel submersible outdoor pool light fixture with a flat, glass lens. The light is recessed into the fixture and surrounded by a ring of four metal flanges. The flanges have small holes drilled in them. The light fixture is secured to the ground with a large bolt in the center. The light source is not visible, but it appears to be an LED or other small light source. The image is on a white background, and the light fixture is the only object in the image.
North Carolina Tar Heels Team Logo Gray Adjustable Hat GS	a product photo of a North Carolina Tar Heels Gray Adjustable Hat GS . The hat is a gray and blue snapback hat with a blue logo of interlocking letters "NC" on the front. The hat has a blue flat bill and a blue adjustable snapback closure on the back. The logo is surrounded by a white outline, creating a sharp contrast with the gray background. The image consists of two photos of the same hat, a smaller one in the top left section that shows the back of the hat, and a bigger one in the bottom right section showing the front of the hat. The background of the image is white.	A product photo of a North Carolina Tar Heels Gray Adjustable Hat GS . The hat is a gray and blue snapback hat with a blue logo of interlocking letters "NC" on the front. The hat has a blue flat bill that contains a label sticker that is hard to see, and a blue adjustable snapback closure on the back. The logo is surrounded by a white outline, creating a sharp contrast with the gray background. The image consists of two photos of the same hat, a smaller one in the top left section that shows the back of the hat, and a bigger one in the bottom right section showing the front of the hat. The background of the image is white.
Data Visualization with Python and Matplotlib	A photo of image features a graph created using Matplotlib, a widely-used data visualization library for Python . The graph showcases three circles arranged in a spiral-like pattern. The innermost circle contains two distinct -shaped images in yellow and blue, while a quarter shape is prominently orange in color. Across the image is the text "Matplotlib". The entire composition is set against a grey background.	A photo of image representing data visualization using Python and Matplotlib . The image showcases three circles arranged in a spiral-like pattern. The innermost circle contains two distinct -shaped images in yellow and blue, while a quarter shape is prominently orange in color. Across the image is the text "Matplotlib". The entire composition is set against a grey background.

Table 15: Re-aligned alt-texts from MetaCLIP (Xu et al., 2024) images.