

LOW-HALLUCINATION SYNTHETIC CAPTIONS VIA VISUAL CHECKLIST BASED REINFORCEMENT LEARNING FOR VISION-LANGUAGE MODEL PRE-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Current pre-training of Vision-Language Models (VLMs) relies on large-scale, high-quality alt-text datasets. However, alt-text data is typically short yet noisy, with this issue being more pronounced for non-English languages. To address this limitation, this paper proposes a recaptioning model that rewrites original alt-text data into versions with rich details while maintaining low hallucination rates. The key to mitigating hallucinations lies in a reinforcement learning approach that leverages preference data produced via visual checklists. Leveraging this recaptioning model, we construct X-Recap - a dataset comprising 1 billion synthetic image-caption pairs with low hallucinations. We empirically demonstrate that a VLM pre-trained on X-Recap substantially outperforms its counterpart trained on the original alt-text data, achieving an average performance improvement of approximately 4.6% across 15 vision-language tasks. To facilitate further research in the community, 20% of the X-Recap dataset will be released to the public.

1 INTRODUCTION

With the enormous power of large language models, remarkable performance gains have recently been achieved in a variety of tasks in natural language processing (NLP), computer vision (CV), and also in cross-modal fields (Brown et al., 2020; Chung et al., 2022; Chowdhery et al., 2022; Touvron et al., 2023; Dosovitskiy et al., 2021; Alayrac et al., 2022; Liu et al., 2023; Achiam et al., 2023; Wang et al., 2024; Hurst et al., 2024). The Large Language Model (LLM) and the Vision Language Model (VLM) are usually equipped with Transformer (Vaswani et al., 2017) as the backbone and then pre-trained with a tremendous amount of unlabeled data. The strong representation ability of the model, the massive amount of data, and the effective means of training make the foundation models powerful for successfully solving the tasks of vision and language.

While the exponential growth of text data has fueled large language models, the pre-training of vision-language models (VLMs) confronts a critical data bottleneck: the scarcity of high-quality, aligned image-text pairs. The prevailing methods for curating these datasets rely on the extraction of web-based alt-text (Sharma et al., 2018; Changpinyo et al., 2021; Ordonez et al., 2011; Schuhmann et al., 2021; 2022; Byeon et al., 2022; Gadre et al., 2023; Liu et al., 2022; Gu et al., 2022). This approach is fraught with inefficiency; its multistage pipeline of cleaning, filtering, and validation not only causes significant data loss but also yields short yet noisy captions that frequently suffer from misalignment with visual content. Critically, as the finite supply of high-quality, naturally occurring pairs is depleted, scaling these extraction methods leads to diminishing marginal returns, with deteriorating data quality hindering VLM training. This bottleneck highlights an urgent need for novel techniques to generate large-scale, high-quality multimodal datasets.

Although recent exploratory investigations into multimodal synthetic data have shown promise for various cross-modal tasks (Li et al., 2023; Betker et al., 2023; Yu et al., 2024; Li et al., 2024b; Awadalla et al., 2024), their efficacy is consistently undermined by a critical flaw: the prevalence of factual hallucinations. As illustrated in Figure 1, captions generated by previous methods are often rife with erroneous details (highlighted in red) or lack sufficient information. To address this limitation, our work introduces a methodology to substantially mitigate hallucinations in synthetic captions. We

subsequently demonstrate that pre-training on our resulting large-scale, low-hallucination dataset yields significant performance improvements for VLMs.



Alt-text: man performing on stage

Caption-120M: On stage, a man is performing, while a woman wearing a striped shirt stands nearby.

Ours: A person is on stage in front of a large audience. The individual is wearing a striped long-sleeve shirt and dark pants, and they appear to be in a dynamic pose, possibly dancing or performing. The audience is seated in a darkened theater or auditorium with red curtains in the background, and the stage is illuminated by a spotlight or stage light. The person's hand is raised, and the overall scene suggests a live performance or concert.



Alt-text: Illinois Fighting Illini Women's Cream Varsity Sweater

BLIP3-KALE: A woman wears an Illinois Fighting Illini Women's Varsity Sweater in this image. The off-white or cream sweater features a prominent, bold orange 'I' logo on the front, signifying a team or university affiliation. The letter 'I' stands out against the striped sleeves, which are adorned with black and white stripes. The woman underneath wears a light-colored top, completing the look with a neutral expression.

Ours: The image shows a cream-colored sweater with a large red letter 'I' in the center, outlined in blue. The sleeves have blue and white stripes at the cuffs. This is likely a piece of apparel associated with the Illinois Fighting Illini women's team.



Alt-text: Original new green connector, d45 printhead Mutoh V1204 V1304 V13604Mimaki JV32 JV5 CV300 printer d45 printhead solvent new and original d45 printhead eco solvent d45 print head for epson Roland v540 for mimaki p2 j4 printer

Recap-DataComp-1B: A close-up image of a laptop screen with a black bezel and a visible screen with a white backlight. The screen is detached from the laptop and is placed on a white background. There are two screws visible on the bezel, and the screen is slightly tilted upwards.

Ours: The image shows a printhead assembly, which is a component used in printers. The printhead has a metallic or metal-like central part with a square or rectangular shape, and it is mounted on a black or dark-colored base. There are visible screws securing the printhead to the base. On the left side of the printhead, there is a green connector or interface, which is typically used for electrical connections in printer hardware. The overall design suggests it is a part of a printer used for printing, possibly for models like Mutoh V1204, V1304, V13604, Mimaki JV32, JV5, CV300, or similar printers.



Alt-text: Wireless Adaptor Pci Tp-Link TL-WN861N 300Mbps 3 Antennas

BLIP3-KALE: The PCI wireless network card showcased in the image is a TP-Link TL-WN861N model, which boasts a 300Mbps connection speed and is equipped with three antennas. Mounted on a PCI slot, this device delivers 300Mbps PCI Express wireless connectivity, as indicated by the label's prominent "300Mbps PCI Express Wireless PCI Adapter" text. The card's green color scheme is complemented by TP-Link's branding and certification marks, as well as a visible serial number.

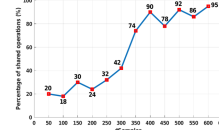
Ours: The image shows a TP-Link TL-WN861N 300Mbps Wireless N PCI Adapter with three black antennas. The adapter is green and has a gold-colored edge connector. The label on the adapter indicates it is a 300Mbps Wireless N PCI Adapter.



Alt-text: Hanging Wall Above the fault plane. Hanging (head) Wall Foot Wall

Recap-DataComp-1B: A digital illustration of a road with a green border labeled 'Hanging Wall'. The road is depicted with a damaged surface, and there is a warning sign that reads 'ABOVE THE FAULT PLANE'. The image is overlaid with a green square that says 'HANGING WALL' and a green triangle with the text 'HANGING WALL'. There is a photo of a woman in a red jacket and a man in a blue jacket, both smiling, with the text 'HANGING WALL' written above them. The image also includes a green rectangle with the text 'HANGING WALL' and a green square with the text 'HANGING WALL'.

Ours: The image is a diagram or educational illustration related to geological fault planes. It includes a title at the top that reads 'Hanging Wall' and a subtitle or description below it that says 'Above the fault plane'. The central part of the image shows a diagram with a green outline that divides the area into two sections: 'Hanging (head) Wall' and 'Foot Wall'. There is a photograph of a person in a red jacket lying on a surface within the 'Hanging (head) Wall' section. The background of the image appears to be a natural setting with some blurred elements, possibly representing a fault or geological formation.



Alt-text: Change in head-level weight sharing. We compute the percentage of shared features within the structures every 50 samples as the search progresses

Ours: The image is a line graph that shows the percentage of shared operations as a function of the number of samples. The x-axis represents the number of samples, ranging from 0 to 600, while the y-axis represents the percentage of shared operations, ranging from 0 to 100%. The graph starts at approximately 20% at 50 samples and shows an overall increasing trend as the number of samples increases. There are notable peaks and troughs, with the highest percentage reaching 95% at around 600 samples. The data points are marked with red squares and connected by a blue line.

Figure 1: Examples of the original caption, previous synthetic caption, and our synthetic caption.

Although VLMs offer an efficient pathway to large-scale caption synthesis, their reliability is critically undermined by persistent hallucinations (Bai et al., 2024). This issue is prevalent in prominent datasets such as Recap-DataComp-1B (Li et al., 2024b) and Blip3-KALE (Awadalla et al., 2024), where, as we show quantitatively in Section 4, significant hallucinatory contents (see Table 1). Such erroneous data corrupt the model’s learning process by impeding the acquisition of correct vision-language alignments, making its mitigation a paramount challenge. To address this, we introduce a novel pipeline whose core innovation is the construction of large-scale caption preference data by adjudicating hallucinations against a visual checklist. Constructing large-scale preference captions is a non-trivial challenge that demands a simple, reliable, and scalable production method. To this end, we developed a rigorous model-based framework. Our process first decomposes each caption into fine-grained visual assertions, which a powerful VLM then individually judges for factuality. These assertions act as a visual checklist, enabling the VLM to determine which of two captions contains fewer hallucinations. This "decompose-then-judge" approach is the key to producing high-quality preference data. This preference data then powers an iterative Direct Preference Optimization (DPO) framework. Our approach explicitly aligns the model with preferences for factually grounded descriptions over hallucinatory ones, and we find that it exhibits a distinct scaling law with iterative application. The efficacy of this method is striking: applying our pipeline (detailed in Section 3) to the same dataset used by BLIP3-KALE, we increase the non-hallucination rate from 56.0% to 86.63%. This represents a substantial 30.63 percentage point improvement in data fidelity (metrics in Section 4.1), underscoring the power of the pipeline to create more reliable synthetic captions.

Leveraging our recaption pipeline, we introduce X-Recap, a large-scale synthetic dataset comprising one billion image-text pairs. A primary contribution of this work is the demonstrably high quality of X-Recap, particularly its significantly reduced hallucination rate. We empirically validate the effectiveness of X-Recap as a superior pre-training resource. Systematic comparisons reveal that vision-language models (VLMs) pre-trained on X-Recap consistently outperform those trained on traditional alt-text across various model architectures and data scales. Specifically, our models achieve an average performance gain of at least 4.6% across 15 diverse vision-language tasks under two distinct settings. We attribute this significant improvement to the high information density and factual accuracy of our synthetic data, which enhances the robustness and capabilities of downstream models.

Our contributions are summarized as follows:

1. A novel pipeline for high-fidelity captions. We propose a novel pipeline that first constructs a large-scale preference dataset by adjudicating captions against a visual checklist. This data is then

Dataset	Method	# Visual Details $\uparrow$	# None Hallucinations $\uparrow$	# Hallucination/Visual Details $\downarrow$
Laion2B	CapsFusion-120M	2.58	72.3%	0.14
	w Qwen2-VL-7B	7.00	58.01%	0.094
	w Gemma3-VL-27B	10.99	62.42%	0.052
	w Qwen2-VL-72B	7.11	69.88%	0.065
	Our Method (27B)	6.85	79.66%	0.040
DataComp	Recap-DataComp-1B	6.95	29.7%	0.249
	BLIP3-KALE	6.52	56.0%	0.101
	w Qwen2-VL-7B	6.35	58.02%	0.109
	w Gemma3-VL-27B	6.84	77.64%	0.043
	w Qwen2-VL-72B	8.15	74.67%	0.045
	w Our Method (27B)	6.83	86.63%	0.024

Table 1: **Hallucination Analysis of Synthetic Caption Datasets.** We compare our method (using Gemma3-VL-27B as the baseline) against existing large-scale synthetic datasets on a random sample of 1,000 pairs from each. We report the Non-Hallucination Rate (% of zero-hallucination captions), Hallucination/Detail Rate (ratio of hallucinatory to total details), and average Visual Details per caption. See Section 4.1 for detailed evaluation methodology.

used within an iterative Direct Preference Optimization (DPO) framework to significantly reduce hallucinations, yielding captions with demonstrably higher factual consistency than prior approaches.

2. A superior large-scale dataset for VLM pre-training. We demonstrate that our resulting one billion pair dataset, X-Recap, provides substantial performance gains over existing synthetic data when used for VLM pre-training. Our approach particularly enhances the model’s capabilities, providing the community with a valuable, high-fidelity training resource.

2 RELATED WORK

Vision-Language Model. Vision-Language Models (VLMs) can be broadly classified into two paradigms. The first is the dual-encoder architecture, exemplified by models such as CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023b), which learn aligned cross-modal representations through contrastive training and excel at retrieval and zero-shot classification. The second, more recent paradigm integrates a vision encoder with a Large Language Model (LLM), a direction pioneered by Flamingo (Alayrac et al., 2022) and popularized by LLaVA (Liu et al., 2023). Driven by the remarkable success of LLMs (Brown et al., 2020; OpenAI, 2023; Grattafiori et al., 2024; Guo et al., 2025), this architecture has become the focus of extensive research and refinement (Wang et al., 2024; Wu et al., 2024; McKinzie et al., 2024; Chen et al., 2024b; Liu et al., 2024). Despite their architectural differences and the advanced capabilities of modern LLM-based VLMs, both paradigms share a fundamental dependency: a reliance on large-scale, high-quality image-text pairs for effective training. This shared requirement has created a critical bottleneck. The supply of naturally occurring, high-quality image-text data is finite and increasingly exhausted (Villalobos et al., 2022), making it difficult to keep pace with the ever-growing scale of state-of-the-art models. To break this impasse and enable continued progress, synthetic data has emerged as a promising solution.

Synthetic data. Building on its success in enhancing large language models (Zhang et al., 2024; Shao et al., 2024; Zhu et al., 2024; Yang et al., 2024), synthetic data generation is increasingly being explored within the multimodal domain. Some methods, such as BLIP-2 (Li et al., 2023), utilize image captioning models to synthesize short captions as substitutes for alt-text, with CapsFusion (Yu et al., 2024) leveraging LLMs for refinement. However, captions generated by these approaches often remain simplistic, providing minimal added information value beyond the original alt-text. Recap-DataComp-1B (Li et al., 2024b) and BLIP3-KALE (Awadalla et al., 2024) attempt to transform alt-text using existing vision language models, but face the significant challenge of hallucination inherent in current VLMs. Other works focusing on dense captioning, including Allava (Chen et al., 2024a) and Dense-Fusion-1M (Li et al., 2024c), generate localized descriptions, but the resulting datasets can also suffer from notable hallucinations. These limitations in existing synthetic VLM data highlight the need for more sophisticated generation techniques. In summary, generating high-quality multimodal synthetic data presents challenges related to controlling hallucination. Although VLM hallucination is a studied topic (Bai et al., 2024), a comprehensive synthetic visual captioning method capable of simultaneously producing diverse and rich content with minimal hallucination is still needed.

Our work introduces a novel approach for generating high-fidelity synthetic captions, culminating in a large-scale dataset with a demonstrably low hallucination rate. We provide comprehensive empirical evidence that this dataset serves as a highly effective resource for large-scale VLM pre-training. This contribution offers a viable and scalable alternative to traditional data sources, effectively addressing the dual challenges of scarcity of high-quality real-world data and the diminishing returns of web-scraped datasets.

3 METHODOLOGY

In this section, we detail our synthetic data generation pipeline. We begin by outlining its two core stages: an initial Supervised Fine-Tuning (SFT) phase, followed by an Iterative Direct Preference Optimization (DPO) framework. As a key part of our DPO description, we will elaborate on the novel method used to construct large-scale preference data. The section concludes with the full training procedures and hyperparameter configurations.

3.1 SUPERVISED FINE-TUNING

A significant limitation of previous recaption models (Li et al., 2023; Yu et al., 2024) is their tendency to produce generic descriptions that lack specificity. For example, a caption might read "a man performing on stage" (Figure 1), which, due to its limited informational value, offers minimal benefit in training powerful vision-language models. In contrast, our work places a significant emphasis on integrating rich, factual information. To achieve this, we leverage GPT-4o to construct a dataset of captions with high information density, which serves as the training data for our Supervised Fine-Tuning (SFT) stage. This approach compels our model to generate descriptions with a much higher degree of specificity and detail.

Please describe the content of the image in detail with one paragraph, let's proceed step by step:
 Step 1: Please describe the content of the image accurately and specifically in one paragraph, avoiding subjective comments (such as "gives a sense of tranquility"). **Entities/knowledge** directly related to the image content (**such as "Eiffel Tower" instead of "a metal tower"**) should be included in the description. Ensure semantic correctness and fluency of the sentences, and the length of the returned description text should be more than 50 words.
 Step 2: Please refer to the image title to improve the description from Step 1. **If the image title contains specific place names, personal names, IP names, etc., and these can be inferred from the image, please add them to the description.** If these details cannot be inferred from the image, do not add them. Do not mention the source of the information in the improved description, and do not include phrases like "image title".
 Step 3: Explain the changes made in the description from Step 2.
 Please return the result in the following JSON format:

```
{
  "caption": "the caption",
  "improved_caption": "the improved caption",
  "explanation": "the reason to make the change"
}
```


 Image Title: **a man fishing at sunset on a lake with his guide**

Figure 2: **Prompt Design for Generating informative Captions with GPT-4o.** Our prompt is structured as a three-step procedure designed to elicit informative captions. The instructions highlighted in red are specifically engineered to guide the model in injecting relevant world knowledge.

By constructing SFT data with captions rich in descriptive details, we enable the model, through fine-tuning, to learn to produce detailed and specific descriptions. Therefore, we designed a detailed prompt and used GPT-4o to generate the initial training data. While GPT-4o understands the prompt requirements, the output data still contain a significant number of hallucinations. To address this, we conducted a manual review of the generated captions. By integrating these data processing steps, we constructed the knowledge-enhanced supervised fine-tuning data suitable for training models capable of generating rich image captions, such as the "arctolamia gestro" illustrated in Figure 1.

3.2 ITERATIVE DPO

Reinforcement Learning from Human Feedback (RLHF) has proven to be effective in mitigating hallucinations in large language models (LLM) (Zhang et al., 2023; Huang et al., 2025). As a specific and efficient RLHF paradigm, Direct Preference Optimization (DPO) (Rafailov et al., 2023) offers a compelling method for fine-tuning models on preference data. The DPO loss function, shown in Equations 1 and 2, is designed to optimize for this preference directly. As noted in Rafailov et al. (2023), the gradient of the loss function, $\mathcal{L}_{\text{DPO}}$, systematically increases the likelihood of preferred

completions (y_w) while decreasing that of dispreferred completions (y_l). This mechanism provides a natural and logical framework for combating hallucinations in image captioning. Intuitively, by defining hallucination-free captions as the preferred data (y_w) and captions containing hallucinations as the dispreferred data (y_l), the DPO process directly trains the model to favor factually grounded descriptions. Maximizing the probability of accurate samples while minimizing the likelihood of hallucinatory ones steers our recaption model toward generating captions with the highest possible fidelity.

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_{\theta}(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_{\theta}(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right] \quad (1)$$

$$\begin{aligned} \nabla_{\theta} \mathcal{L}_{\text{DPO}}(\pi_{\theta}; \pi_{\text{ref}}) = & -\beta \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\underbrace{\sigma(\hat{r}_{\theta}(x, y_l) - \hat{r}_{\theta}(x, y_w))}_{\text{higher weight when reward estimate is wrong}} \left[\underbrace{\nabla_{\theta} \log \pi(y_w | x)}_{\text{increase likelihood of } y_w} - \underbrace{\nabla_{\theta} \log \pi(y_l | x)}_{\text{decrease likelihood of } y_l} \right] \right], \end{aligned} \quad (2)$$

where $\hat{r}_{\theta}(x, y) = \beta \log \frac{\pi_{\theta}(y|x)}{\pi_{\text{ref}}(y|x)}$ is the reward implicitly defined by the language model π_{θ} and the reference model π_{ref} .

We observe that standard DPO fails to yield continuous performance improvements in our caption synthesis task, quickly reaching a performance plateau. We attribute this to the static implicit reward signal generated by the fixed reference model in the DPO objective function (Equation 2), which causes the policy to prematurely converge to a local optimum. To overcome this, our iterative DPO framework periodically updates the reference model with the current best policy whenever performance saturates. The updated model then generates a new, more challenging set of preference data, providing a refreshed reward signal that enables the policy to resume optimization and surpass its previous performance ceiling.

Furthermore, to ensure that the model learns from the semantic content of the captions rather than superficial features, we address a potential training bias related to sequence length. Since longer completions may have a higher probability of containing hallucinations, we apply a length-balancing operation to the preference pairs during DPO training. This ensures that the average sequence lengths of the preferred (y_w) and dispreferred (y_l) completions are closely matched, preventing the model from learning a spurious correlation between caption length and quality.

3.3 LARGE-SCALE PREFERENCE DATA

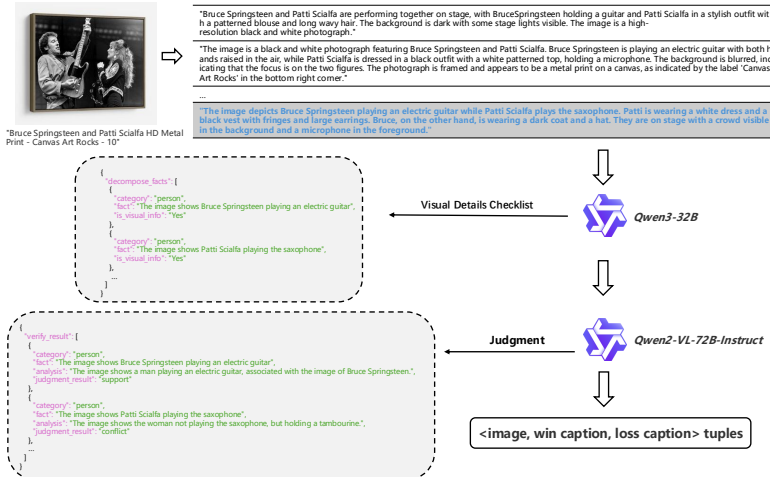


Figure 3: The illustration of the generating preference captions pipeline.

The efficacy of DPO is critically dependent on high-quality preference data, yet quantifying caption hallucination is a challenging task. While large-scale human annotation is infeasible due to cost, direct model-based judgments are often unreliable. We therefore introduce a visual checklist methodology for generating superior preference data. Our approach mimics human reasoning: a Large Language Model (LLM, Qwen3-32B) first deconstructs a caption into a checklist of visual details, and a Vision-Language Model (VLM, Qwen2-VL-72B) then verifies the factuality of each detail individually. This two-step process simplifies the detection task.

To validate this methodology, we conducted a controlled experiment. We randomly sampled 200 images and used our initial SFT model to generate eight candidate captions for each image.¹ We then created preference data using two different scoring methods: 1) a baseline approach where a VLM directly scores the entire caption for hallucinations, and 2) our proposed method, which uses the visual checklist-based scoring. As demonstrated in Table 2, our method not only substantially outperforms direct VLM-based evaluation using the same base model but also matches or exceeds the performance of stronger, closed-source VLMs, confirming the superiority of our structured approach. Specifically, when using the same VLM for judgment, our visual checklist-based method improves upon the direct scoring baseline by 23.3% in precision and 36.3% in recall. This significant improvement indicates that, compared to direct, holistic scoring, our visual checklist approach can more comprehensively and accurately identify individual points of hallucinatory information within a caption.

Model Evaluation	Gemini-2.5-Pro VLM score	GPT4o-latest VLM score	Qwen2-VL-72B VLM score	Qwen2-VL-72B Visual Checklist
Precision	82.8%	70.8%	68.2%	91.5%
Recall	68.7%	45.8%	19.3%	55.6%

Table 2: **Evaluation of Caption Preference Data.** In this table, Precision measures the fraction of retrieved correct preference pairs, while Recall measures the fraction of all existing correct preference pairs that were successfully retrieved.

3.4 TRAINING PIPELINE

This section details our comprehensive training pipeline. The pipeline consists of two primary stages: an initial SFT phase, followed by several rounds of iterative DPO. For our experiments, we employ three distinct VLMs as base models: Qwen2-VL-7B, Gemma3-27B, and Qwen2-VL-72B. The entire training process is implemented using the LLaMAFactory (Zheng et al., 2024) framework. A comparative analysis of the characteristics and performance of each baseline is presented in Section 4.

Supervised Fine-Tuning. We initiated our pipeline by sampling a raw dataset of image-text pairs from several open-source multimodal corpora, including CC3M (Sharma et al., 2018), CC12M (Changpinyo et al., 2021), DataComp (Gadre et al., 2023), and Wikipedia (Srinivasan et al., 2021). Using the prompt detailed in Figure 2, we then employed GPT-4o to generate an initial set of recaptioned data, from which we retained 43,408 high-quality entries. Subsequently, we performed Low-Rank Adaptation (LoRA) fine-tuning on the base model using this curated data. Fine-tuning was performed for 10 epochs with a global batch size of 128 and a peak learning rate of $1e-5$, which produced our initial SFT model.

Iterative DPO. As outlined in Section 3.2, our methodology begins with the initial SFT model. To generate preference data, we first randomly sample 200,000 image-text pairs and perform inference using this SFT model. For each pair, we generate eight diverse candidate captions using the following sampling hyperparameters: TOPP=1.0, TOPK=20, and TEMPERATURE=1.0. From this set of eight candidates, we then construct a preference pair (y_w, y_l) by selecting the best (chosen) and worst (rejected) outputs determined by a critic model. We employ Qwen2-VL-72B as our critic model, which operates via a two-step process: it first decomposes each caption into fine-grained visual details and subsequently identifies the number of hallucinations by assessing the factuality of each detail. For each set of candidates, the caption with zero hallucinations is selected as the chosen response (y_w) , while the one with the highest number of hallucinations is designated as the rejected response (y_l) . To mitigate potential length-induced bias in training, we apply a length-balancing sampling strategy to the initial preference data. This curation step ensures a similar length distribution between the chosen (y_w) and rejected (y_l) captions. We then use this dataset to perform DPO

¹Decoding parameters were set to TOPP=1.0, TOPK=20, and TEMPERATURE=1.0.

training on the SFT model, resulting in our first-stage DPO model. This first-stage DPO model is then used to bootstrap subsequent iterations of data generation. We repeat this process—sampling new image-text pairs, generating preference data using the latest DPO model for inference, and applying our critic model and length-balancing filtration for a total of four iterations. This iterative refinement yields a final set of preference pairs, with the total volume varying slightly for each base model: 357k for Qwen2-VL-7B, 340k for Qwen2-VL-72B, and 569k for Gemma3-VL-27B.² The specific hyperparameters for each stage are detailed in Table 3.

	Stage-1-SFT	Stage-2-DPO
Resolution # pixels	[3136, 12845056]	[3136, 12845056]
Dataset # Samples	43K	[357K, 340k, 569k]
Finetuning Type	LoRA	LoRA
Batch Size, Learning Rate, Learning Epoch	128, $1e-5$, 10	64, $5e-6$, 1

Table 3: **Training Configurations for the Recaption Model Stages.** This table details the specifications for each stage of our model’s training process, including the vision parameter settings, dataset characteristics, model details, and training hyperparameters.

4 EXPERIMENTS

This section presents a comprehensive quality assessment of our recaptured data. We begin by outlining the evaluation criteria, followed by an analysis across two key dimensions: 1) the level of hallucination present in the captions, and 2) the data’s empirical impact on the pre-training of large-scale vision-language models.

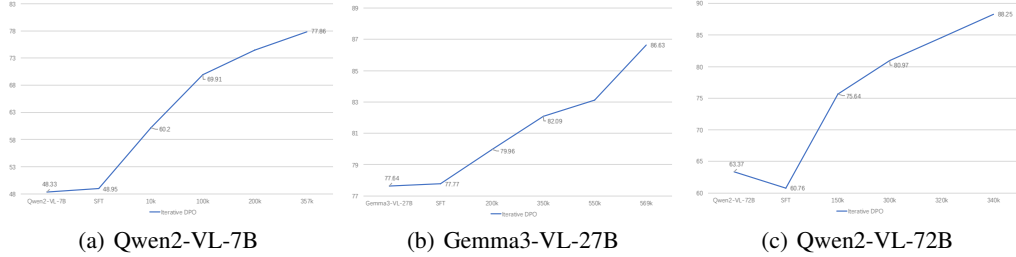


Figure 4: **Performance of Iterative DPO Strategies.** This figure demonstrates the effectiveness of our iterative DPO method in mitigating hallucinations across three different base models. The horizontal axis represents the scale of the training data, while the vertical axis indicates the percentage of hallucination-free captions on the validation set.

4.1 CAPTION EVALUATION METRICS

As established in Section 3.3, our visual checklist method is highly effective for generating preference data. We therefore adopt this framework as our primary methodology for the quantitative assessment of hallucination levels. To ensure maximum precision for this critical evaluation task, we leverage a powerful model, GPT-4o, for both the detail decomposition and factuality judgment steps. The reliability of this approach was confirmed through a manual verification study: on a set of 200 randomly sampled cases, two NLP experts concurred with our GPT-4o-based judgments with an accuracy exceeding 95%. We then conduct a comprehensive evaluation to benchmark our method’s superiority, validating it on two fronts: 1) against prominent open-source synthetic caption datasets, and 2) against various open-source VLM models.

First, to ensure a fair comparison against other data generation methods, we apply our recaptioning pipeline to the same input image-text pairs used in previous work, including samples from Laion2B (used in CapsFusion (Yu et al., 2024)) and Datacomp (used in Recap-DataComp-1B (Li et al., 2024b) and BLIP3-KALE (Awadalla et al., 2024)). The results presented in Table 1 demonstrate a substantial

²The dataset for Gemma3-VL-27B is notably larger as it was trained on a multilingual corpus.

reduction in hallucinations. Specifically, our generated captions achieve the following: (1) a 56.9 percentage point increase in the non-hallucination rate compared to Recap-DataComp-1B. (2) an 8.6 and 30.6 percentage point increase in the non-hallucination rate compared to CapsFusion and BLIP3-KALE, respectively.

Second, when compared with other leading open source VLM models on a fixed dataset, our approach again yields a significant decrease in hallucination rates. In both evaluation scenarios, our method consistently produces higher-fidelity captions, proving its superior ability to generate factually grounded training data. Finally, we evaluated the effectiveness of our iterative DPO method for hallucination mitigation. As illustrated in Figure 4, all base models exhibit a significant and consistent increase in the proportion of hallucination-free captions as the DPO dataset is iteratively expanded.

Dataset	MathVista testmini	Scientific QA MMMU val	ALTE test	ScienceQA test	OCRBench test	DocVQA val	TextVQA val	OCR QA InfographicsVQA val	ChartQA test (aug./num.)	MMBench en-dev	Benchmark QA MMStar test	RealworldQA test	MMVet test	Hallucination Tasks HallusionBench	None-Hallu. Rate	Avg.
Alt-text	50.10	45.0	73.18	85.57	59.8	77.34	68.47	43.99	62.08	72.42	47.89	61.57	42.09	33.80	39.27	57.50
Recap-DataComp-1B	53.70	44.44	74.65	86.12	62.1	79.11	66.39	45.38	64.28	65.89	50.44	61.96	39.34	32.77	43.62	58.01
Queen2-VL-7B	52	45.22	77.19	88.60	66.7	83.13	70.77	51.92	66.84	74.31	52.71	65.49	47.29	31.98	66.20	62.69
X-Recap	56.20	44.00	77.91	89.74	65.9	83.06	71.37	50.32	68.96	72.34	54.66	65.49	43.17	36.38	71.86	63.42

Table 4: **Experimental results on 15 vision-language tasks.** Avg. denotes the average score across all evaluated vision-language tasks. The Non-Hallucination Rate is a hallucination assessment metric defined in Section 4.1. For evaluation, we randomly sampled 1000 instances from the DataComp dataset to quantify the non-hallucination rate. In the table, bold text indicates the optimal performance.

4.2 EFFECTIVENESS OF RECAPTION DATA

Our primary validation assesses the effectiveness of X-Recap for vision-language pre-training at two distinct scales. We first conduct experiments in a small-scale, open-source "toy setting" to ensure rapid community reproducibility. We then perform a comprehensive validation on a larger one-billion-pair dataset to verify that the benefits of our high-quality captions are sustained at scale. Furthermore, we demonstrate the broader extensibility of our data through exploratory experiments on a cross-modal text-to-image generation task.

Vision-Language Model Pre-Training.

We first validate the effectiveness of our X-Recap dataset for pre-training Vision-Language Models (VLMs). The objective is to demonstrate that, compared to existing alt-text and other recaption datasets, X-Recap facilitates more effective VLM training, leading to superior performance on both general multimodal benchmarks and our internal knowledge-intensive test sets.

Experimental Setup. To ensure a fair comparison, we standardized our evaluation across two distinct experimental settings designed to isolate the impact of different data sources and scales.

1. **Open-Source Comparative Setting.** In this setting, we aim to directly compare the efficacy of different caption datasets. We adopted the LLaVA (Liu et al., 2023) architecture, utilizing SigLIP (Zhai et al., 2023a) as the visual encoder and Hunyuan-7B (Hunyuan, 2025) as the base large language model. For each dataset under comparison (i.e., various alt-text and our recaption datasets), we sampled a consistent volume of 20M data points for pre-training. Following this, a light, unified supervised fine-tuning (SFT) (Wang et al., 2023; Li et al., 2024a) was applied to all models. It is important to note that our goal here is to isolate the impact of the caption data itself. Therefore, these experiments were conducted at a fixed scale, using only pure caption data without annealing techniques, advanced fine-tuning, or reinforcement learning. As such, the results are intended for comparative analysis and not for direct comparison with state-of-the-art leaderboards.

2. **Large-Scale Scaling Analysis.** To investigate the scaling properties of our recaptioned data, we conducted a series of large-scale experiments, comparing its effectiveness against traditional alt-text. For this analysis, we used the LLaVA architecture with a smaller Hunyuan-2B language model to balance computational resources and efficiency. The pre-training data for each experimental run was a blend: 70% was either our synthetic captions or the alt-text baseline, mixed with a fixed 30% proportion of a diverse in-house corpus (primarily OCR, STEM, and text-only data). We performed experiments at four distinct scales, corresponding to total token counts of 20B, 50B, 200B, and 1T. Following each pre-training run, a common 6B-token in-house dataset was used for a unified SFT stage to consistently align and elicit the model’s final capabilities.

Results. 1. Open-Source Comparative Setting. As presented in Table 4, the Vision-Language Model (VLM) pre-trained on X-Recap significantly outperforms all baselines trained on existing alt-text and synthetic caption datasets across 15 multimodal tasks. This superiority is particularly pronounced on hallucination-related benchmarks, where our dataset provides a substantial performance advantage. For context within the table, "Qwen2-VL-7B" denotes a baseline trained on 20M data generated by an unoptimized model with a higher hallucination rate. Furthermore, the benefits of X-Recap extend beyond final performance metrics; as detailed in Appendix A, models pre-trained with our data also exhibit markedly improved convergence properties during the subsequent SFT phase.

2. Large-Scale Scaling Analysis. As presented in Table 5, our recaptured data demonstrates superior performance compared to the original alt-text data. This advantage is consistent across various data scales and on nearly all evaluation tasks. Crucially, as the data volume increases, X-Recap exhibits a favorable scaling law, indicating that its benefits are sustained and amplified at scale. This superiority is particularly pronounced on hallucination-related benchmarks, a finding that aligns with our results from the Open-Source Comparative Setting and further underscores the significant advantages of low-hallucination synthetic captions. Furthermore, we include the performance of a 2B-scale leading model, Qwen2-VL-2B, for reference. It is important to note that our training methodology was intentionally simplified to isolate the effect of the caption data, without the extensive fine-tuning and optimization typically applied to benchmark models. Despite using less training data and a non-optimized strategy, our model trained on X-Recap achieves better performance than this strong baseline, underscoring the high quality and efficacy of our generated captions.

Dataset	Data size	Scientific QA				OCR QA				Benchmark QA				Hallucination Tasks			
		MathVista testmini	MMMU val	AI2D test	ScienceQA test	OCRBench test	DocVQA val	TextVQA val	InfographicsVQA val	ChartQA test (aug./hum.)	MMBench en-dev	MMStar test	RealWorldQA test	MMVet test	HallusionBench	None-Hallu. Rate	Avg.
Qwen2-VL-2B	1.4T	43	41.1	74.7	80.4	80.9	90.1	79.7	65.5	73.5	74.9	48.0	62.9	49.5	41.7	54.33	64.02
ALT-Text	20B	43.4	34.56	53.34	70.02	64.7	72.55	66.84	39.73	66.32	57.73	37.88	47.84	39.03	29.92	25.05	49.88
X-Recap	20B	44.5	36.56	55.57	71.49	72.2	80.28	70.9	49.99	70.92	60.65	41.1	49.67	44.63	32.9	41.73	54.87
ALT-Text	50B	41.3	34.89	55.28	71.05	71.5	81.24	72.05	49.56	70.88	62.42	40.09	53.86	42.84	35.06	36.37	56.90
X-Recap	50B	46.6	36.11	59.42	74.42	73.8	82.67	73.43	52.91	72.72	62.97	42.5	55.69	43.39	33.37	49.90	57.33
ALT-Text	200B	44.6	35.33	58.1	72.88	79.1	84.33	77.14	53.66	71.6	61.77	41.3	51.37	40.41	34.63	43.29	56.63
X-Recap	200B	50.7	36.22	58.42	74.91	81.4	87.08	77.43	58.48	74.56	66.07	45.58	52.94	43.48	36.37	61.72	60.36
ALT-Text	1T	52	36.78	61.01	74.96	82.5	87.61	78.74	57.99	76.32	62.2	45.72	52.68	48.66	40.38	48.40	60.42
X-Recap	1T	54.1	37.33	68.62	76.3	85	91.38	79.57	66.91	79.2	67.01	51.74	56.47	50.0	44.29	67.63	65.04

Table 5: Experimental results on 15 vision-language tasks for the Large-Scale Scaling Analysis. Results in bold indicate the superior performance between the two methods under the same data scale setting. An underline denotes the best overall result across all methods, including the comparison with a leading, similarly-sized open-source model, Qwen2-VL-2B.

Text-to-Image Model Training. We also validated X-Recap on the text-to-image generation task by fine-tuning the Hunyuan-DiT model (Li et al., 2024d). The fine-tuned model achieved a significantly lower FID and a higher CLIP score than the public baseline (Table 6), confirming that our data enhances the generation of authentic, high-fidelity images. Examples are presented in Appendix B.

Dataset	FID↓			CLIP Score↑		
	Row Caption	Our Recaption	Fine-tuning with Our Recaption	Row Caption	Our Recaption	Fine-tuning with Our Recaption
MSCOCO Test set	28.79	21.09	15.46	0.307	0.305	0.313

Table 6: **Text-to-Image Evaluation** of the Hunyuan-DiT model on MSCOCO-30K, comparing performance with and without fine-tuning using our recaption data. It should be noted that, due to computational resource limitations, we only sampled 2M data points from the X-Recap dataset for LoRA fine-tuning.

5 CONCLUSION

This work tackles the critical challenge of hallucination in large-scale synthetic image captioning. Our core finding is that Direct Preference Optimization (DPO) is exceptionally effective at reducing caption hallucinations. We further show that by applying DPO iteratively, we can progressively increase the ratio of hallucination-free captions in our synthetic data, a process that demonstrates a distinct and predictable scaling law. Leveraging this methodology, we construct X-Recap, a new large-scale dataset of high-fidelity, low-hallucination captions. Our extensive experiments demonstrate that pre-training Vision-Language Models (VLMs) on X-Recap yields substantial performance gains over models trained on both raw alt-text and existing synthetic datasets. To foster future research in multimodal learning, we will release the X-Recap dataset to the open-source community. For a detailed discussion of limitations and potential risks, refer to Sections C and D.

REFERENCES

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.
- Anas Awadalla, Le Xue, Manli Shu, An Yan, Jun Wang, Senthil Purushwalkam, Sheng Shen, Hannah Lee, Oscar Lo, Jae Sung Park, Etash Guha, Silvio Savarese, Ludwig Schmidt, Yejin Choi, Caiming Xiong, and Ran Xu. Blip3-kale: Knowledge augmented large-scale dense captions, 2024. URL <https://arxiv.org/abs/2411.07461>.
- Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.
- James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3):8, 2023.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>.
- Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. <https://github.com/kakaobrain/coyo-dataset>, 2022.
- Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for lite vision-language models. *arXiv preprint arXiv:2402.11684*, 2024a.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *ArXiv preprint*, abs/2204.02311, 2022. URL <https://arxiv.org/abs/2204.02311>.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

- 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. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL <https://openreview.net/forum?id=YicbFdNTTy>.
- Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36: 27092–27112, 2023.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, 35:26418–26431, 2022.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Kashmir Hill and Jeremy White. Designed to deceive: Do these people look real to you. *The New York Times*, 11, 2020.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.
- Hunyuan. Tencent-hunyuan-7b, 2025. URL <https://github.com/Tencent/Tencent-Hunyuan-7B>.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.
- Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, et al. What if we recaption billions of web images with llama-3? *arXiv preprint arXiv:2406.08478*, 2024b.
- Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and LINGYU DUAN. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. *Advances in Neural Information Processing Systems*, 37:18535–18556, 2024c.
- Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xincheng Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li, Jihong Zhang, Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yangyu Tao, Jianchen Zhu, Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Di Wang, Yong Yang, Jie Jiang, and Qinglin Lu. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding, 2024d.

- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.
- Yulong Liu, Guibo Zhu, Bin Zhu, Qi Song, Guojing Ge, Haoran Chen, GuanHui Qiao, Ru Peng, Lingxiang Wu, and Jinqiao Wang. Taisu: A 166m large-scale high-quality dataset for chinese vision-language pre-training. *Advances in Neural Information Processing Systems*, 35:16705–16717, 2022.
- Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruvi Shah, Xianzhi Du, Futang Peng, Anton Belyi, et al. Mml: methods, analysis and insights from multimodal llm pre-training. In *European Conference on Computer Vision*, pp. 304–323. Springer, 2024.
- OpenAI. Gpt-4 technical report, 2023.
- Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pp. 1143–1151, 2011. URL <https://proceedings.neurips.cc/paper/2011/hash/5dd9db5e033da9c6fb5ba83c7a7e9bea9-Abstract.html>.
- 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. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL <http://proceedings.mlr.press/v139/radford21a.html>.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.
- Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. URL <https://aclanthology.org/P18-1238>.
- Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 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*, pp. 2443–2449, 2021.

- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.
- Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Will we run out of data? limits of llm scaling based on human-generated data. *arXiv preprint arXiv:2211.04325*, 2022.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2023.
- Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- Qiyang Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wang, and Jingjing Liu. Capsfusion: Rethinking image-text data at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14022–14032, 2024.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 11941–11952. IEEE, 2023a. doi: 10.1109/ICCV51070.2023.01100. URL <https://doi.org/10.1109/ICCV51070.2023.01100>.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023b.
- Yifan Zhang, Yifan Luo, Yang Yuan, and Andrew Chi-Chih Yao. Autonomous data selection with language models for mathematical texts. *arXiv preprint arXiv:2402.07625*, 2024.
- Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren’s song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyao Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL <http://arxiv.org/abs/2403.13372>.
- Qihao Zhu, Daya Guo, Zhihong Shao, Dejian Yang, Peiyi Wang, Runxin Xu, Y Wu, Yukun Li, Huazuo Gao, Shirong Ma, et al. Deepseek-coder-v2: Breaking the barrier of closed-source models in code intelligence. *arXiv preprint arXiv:2406.11931*, 2024.

A LOSS OF VISION-LANGUAGE MODEL TRAINING

To address the issue of loss value incomparability arising from diverse pre-training data distributions, we examine the convergence behavior of pre-trained models when fine-tuned on a common supervised dataset (Wang et al., 2023; Li et al., 2024a), as illustrated in Figure 5. Figure 5 shows that models trained with X-Recap data exhibit superior convergence. Quantitatively, these models converge to a lower final loss of 0.4796 compared to models pre-trained on other datasets.



Figure 5: Loss curve of vision-language model pre-trained with alt-text, Recap-DataComp-1B, Qwen2-VL-Zeroshot or X-Recap. To ensure a fair comparison, all pre-trained models analyzed here utilized the same amount of data (20M) for pre-training. Furthermore, the four convergence curves presented are obtained by fine-tuning these models on an identical supervised dataset.

B EXAMPLES OF TEXT-TO-IMAGE GENERATIONS

This section presents examples of text-to-image generation in Figure 6; the corresponding text prompts are provided in Table 7. Observation of these examples reveals that the original Hunyuan-DiT model struggles to accurately render certain concepts (e.g. abacus, Yueqin) and produces less accurate depictions for others (e.g. jellyfish, acerola). However, after fine-tuning with the X-Recap data, the model’s understanding of these concepts becomes more precise, resulting in generated images with a more realistic style. These qualitative improvements are consistent with the quantitative results shown in Table 6, such as the observed reduction in FID. This consistency highlights how the comprehensive concept coverage in our data enhances the model’s perceptual fidelity, which is further corroborated by the significant performance improvement of the VLM trained with X-Recap on numerous vision-language tasks.

C LIMITATIONS

The pre-training of large vision-language models is an inherently resource-intensive endeavor. In our work, for instance, pre-training the 2B-scale VLM on 1B tokens alone required 123,776 A800 GPU hours. Due to these substantial computational demands, a key limitation of this study is that we were unable to explore the full scaling potential of our recaptioned data with even larger VLM architectures. We leave this promising direction for future investigation.

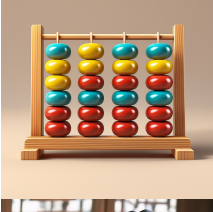
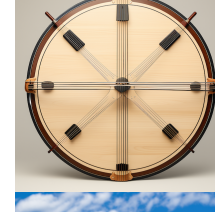
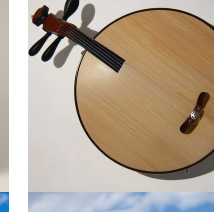
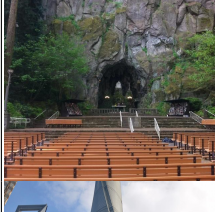
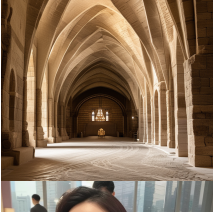
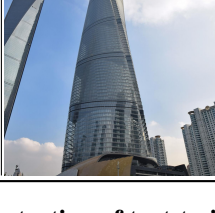
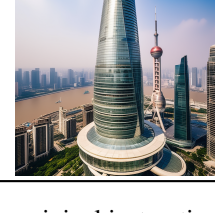
Text	Image	Hunyuan-DiT	Hunyuan-DiT wRecap	Hunyuan-DiT wFT
Jellyfish				
Acerola				
Abacus				
Yueqin				
Alpaca				
The Grotto				
Shanghai Tower				

Figure 6: **Demonstration of text-to-image cases.** Here, the original instruction for image generation is denoted as "Text", the original picture as "Image", the generation effect of the original open-source model as "Hunyuan-DiT", the generation effect of the original open-source model with our Synthetic texts as "Hunyuan-DiT wRecap", and the effect of the model fine-tuned with our recaption data as "Hunyuan-DiT wFT".

Text	Our Recaption
Jellyfish	The picture shows two jellyfishes. Their bodies are orange or light red with distinct stripes or lines. Their tentacles are long, slender, transparent or semi-transparent. The background is black, highlighting the colors of the jellyfish and the details of their tentacles. The jellyfish are swimming or floating in the water, and their tentacles are spread out in the water, forming a distinct linear effect.
Brazilian Acerola	The image shows the branches of a tree laden with red and green fruits. The red fruits look ripe, while the green ones appear unripe. There are also green leaves on the branches. In the background, more green leaves and a few blurred flowers or small buds can be seen. The overall environment seems to be an outdoor natural setting.
Abacus	There is a group of people sitting in front of a wooden table in the picture, using abacuses for learning or operation. There are several abacuses on the table. Wooden doors and windows can be seen in the background, indicating that this is an indoor environment, perhaps a classroom or a study place. People are wearing casual or semi-formal clothes, looking relaxed.
Yueqin	This is a Yueqin, which has a circular resonator and four strings. There are multiple positions for pressing the strings on the body of the instrument, and there are four tuning pegs on the headstock. Overall, it exhibits the characteristics of a traditional musical instrument.
Alpaca	There are two alpacas standing on the grass in the image, with blue sky and white clouds in the background. The alpaca on the left has lighter fur, while the one on the right has darker fur. The grass is green, and the postures of the alpacas suggest that they are either walking or standing.
The Grotto	The Grotto is a church located in a natural environment. The facade of the church is a huge rock cave, covered with green vegetation above the cave, and surrounded by dense trees. Inside the church, there is an altar, above which a cross is hung, surrounded by lighting. In front of the church, there is a row of wooden benches for believers to sit and pray.
Shanghai Tower	Shanghai Tower is a skyscraper with a spiral glass curtain wall design on its exterior. At the top, there is a circular observation deck, surrounded by other high-rise buildings. The overall architectural style is modern and futuristic.

Table 7: The original text of the image in Figure 6 and the synthetic caption after being processed by our recaption model.

D POTENTIAL RISKS

The original image-text pairs are primarily derived from open-source datasets. While we have implemented substantial measures to filter out undesirable content, potential risks remain. These risks are particularly salient in the field of image generation, exemplified by issues such as the creation of fake portraits for social media (Hill & White, 2020), which is a recognized challenge in this research area.

E DECLARATION ON THE USE OF AI TOOLS

During the preparation of this manuscript, the authors utilized a large language model to assist with language editing and refinement. The tool was employed to improve grammar, clarity, and readability. The authors take full responsibility for the final content and all scientific claims presented herein.