Prioritizing Image-Related Tokens Enhances Vision-Language Pre-Training

Yangyi Chen, Hao Peng, Tong Zhang, Heng Ji
University of Illinois Urbana-Champaign
yangyic3@illinois.edu

Reviewed on OpenReview: https://openreview.net/forum?id=jDcnL1hB1Z

Abstract

In standard large vision-language models (LVLMs) pre-training, the model typically maximizes the joint probability of the caption conditioned on the image via next-token prediction (NTP); however, since only a small subset of caption tokens directly relates to the visual content, this naive NTP unintentionally fits the model to noise and increases the risk of hallucination. We present PRIOR, a simple vision-language pre-training approach that addresses this issue by **prior**itizing image-related tokens through differential weighting in the NTP loss, drawing from the importance sampling framework. PRIOR introduces a reference model—a text-only large language model (LLM) trained on the captions without image inputs, to weight each token based on its probability for LVLMs training. Intuitively, tokens that are directly related to the visual inputs are harder to predict without the image and thus receive lower probabilities from the text-only reference LLM. During training, we implement a token-specific re-weighting term based on the importance scores to adjust each token's loss. We implement PRIOR in two distinct settings: LVLMs with visual encoders and LVLMs without visual encoders. We observe 19% and 8% average relative improvement, respectively, on several vision-language benchmarks compared to NTP. In addition, PRIOR exhibits superior scaling properties, as demonstrated by significantly higher scaling coefficients, indicating greater potential for performance gains compared to NTP given increasing compute and data. The code is available at https://github.com/Yangyi-Chen/PRIOR.

1 Introduction

Vision-language pre-training enhances both visual perception and visual-textual association capabilities in large vision-language models (LVLMs) (Zhang et al., 2021; Wu et al., 2024). However, our preliminary human annotations on 100 examples from Capsfusion (Yu et al., 2024) reveal that only 31.3% of words in web-scale image-caption pairs directly relate to the associated images, with the remainder containing irrelevant information, stylistic elements, or website-specific content. For example, considering the second example in Fig. 1 (Top), only "house", "front yard", and "lawn" are image-related, while the remaining tokens lack visual correspondence, such as the location and price information of the house. The standard next-token prediction (NTP) objective treats all tokens equally, regardless of their relevance to visual content. Thus, besides learning visual perception and vision-language alignment, LVLMs inevitably model the entire caption distribution, potentially overfitting to noise and contributing to hallucination (Sharma et al., 2018; Changpinyo et al., 2021; Liu et al., 2023b).

In this work, we present PRIOR, a simple approach that enhances vision-language pre-training by **prior**itizing optimization on image-related tokens. PRIOR addresses the above challenge by effectively distinguishing between image-related tokens and other tokens in the training corpus. We train a text-only reference model to capture the caption distribution without visual context in the vision-language corpus. This allows us to identify which tokens are likely to be image-related and establish an importance distribution over the token space. The key insight is that tokens easily predicted by the reference model likely contain minimal visual

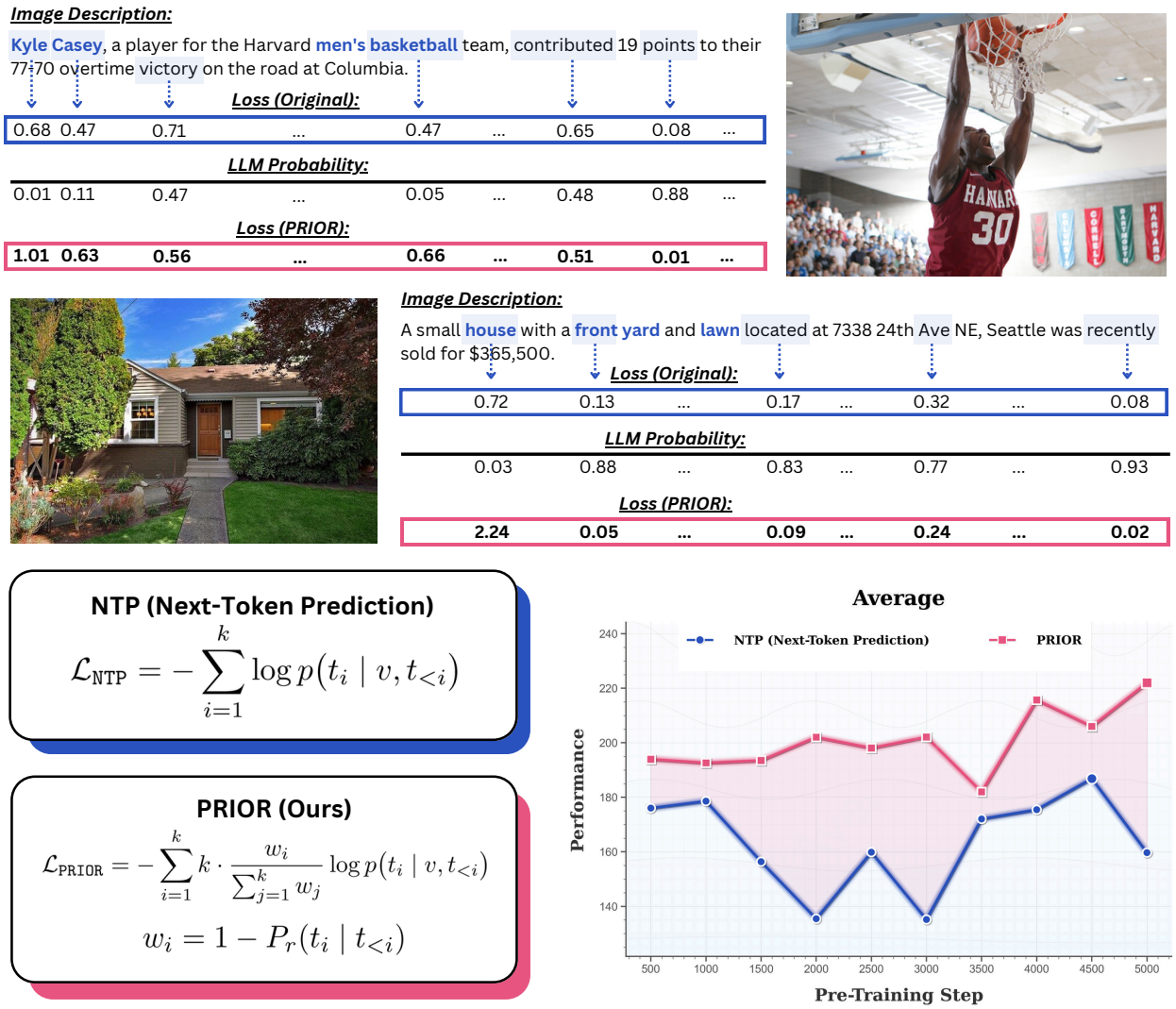

Figure 1: **(Top) Synthetic examples to highlight the motivation of PRIOR.** Only a few tokens in the captions (highlighted in blue, word-level for better visualization) are related to the associated images. PRIOR utilizes probability scores from a text-only LLM to recalibrate the original loss function at the token level, prioritizing image-related tokens that receive lower probability scores from the LLM. **(Bottom Left) PRIOR formulation.** Given an image v paired with a caption '$t_1, t_2, ..., t_k$', PRIOR enhances vision-language pre-training by assigning a normalized weight to each token loss, which is computed based on the LLM probability $P_r(t_i|t_{<i})$. **(Bottom Right) Performance of PRIOR.** PRIOR demonstrates consistent performance improvement (average over several vision-language benchmarks) and better training stability compared to the widely used next-token prediction objective in vision-language pre-training. In addition, PRIOR shows superior scaling behaviors in both performance predictability and potential improvement with increased compute and data (§4).

information, while those difficult for text-only models to predict more likely convey image-specific content. We formalize this intuition by using the text-only model's predictive probability to reweight the NTP loss for each token, as shown in Fig. 1 (Bottom left). This training algorithm, inspired by importance sampling principles, prioritizes the optimization on tokens containing image-related information while preserving general language generation capabilities.

PRIOR addresses several crucial limitations in previous vision-language pre-training approaches. Existing methods, as detailed in §6, primarily fall into two categories, each with their own challenges. The first

category involves supplementing NTP with additional training objectives such as contrastive loss (Li et al., 2019) or distillation loss (Liao et al., 2025), but these approaches sacrifice NTP's simplicity and create challenges for large-scale, efficient implementation with existing frameworks (Rasley et al., 2020). In contrast, PRIOR can be easily integrated with existing pre-training frameworks through minimal code changes for the loss computation and offline-computed token importance scores. The second category focuses on optimizing training data through distillation (Chen et al., 2024a), filtering pipelines (Guo et al., 2024), and related approaches. While conceptually promising, these methods face inherent scalability constraints due to their dependence on teacher models or human-designed heuristics. In contrast, PRIOR's text-only reference model offers a scalable alternative that dynamically extracts useful knowledge from noisy web-scale datasets, ensuring broad compatibility with advancements across vision-language datasets.

To evaluate PRIOR's broad applicability across different LVLMs architectures, we conduct experiments with both vision encoder-based LVLMs (Zhu et al., 2023; Li et al., 2023b) and end-to-end unified LVLMs with a simple linear projector for image processing (Chen et al., 2024f; Diao et al., 2024; Tao et al., 2024). Our results demonstrate that PRIOR consistently outperforms the naive NTP pre-training baseline, achieving 19% and 8% average relative improvement across these two architectures respectively, while also enhancing training stability. The average performance comparisons are visualized in Fig. 1 (Bottom Right) and Fig. 4. In §3.4, we compare PRIOR with additional baselines and alternative token selection methods, and show the unique advantages of PRIOR. Further analysis in §4 reveals that PRIOR exhibits superior scaling behaviors in terms of performance predictability (Fig. 6) and potential improvement with increased computational resources (Fig. 7). Moreover, as detailed in §5.2, PRIOR accelerates pre-training by more effectively optimizing loss on both image-related and image-unrelated tokens.

2 PRIOR

2.1 Problem Formulation: Vision-Language Pre-training with Image-Caption Pairs

We review the general vision-language pre-training that leverages a web-scale dataset comprising image-caption pairs (v, c) to develop models capable of generating relevant textual descriptions based on visual inputs. v represents the image, and $c = t_1, t_2, ..., t_k$ represents the corresponding caption consisting of k tokens. The widely used vision-language pre-training objective, next-token prediction ($\mathcal{L}_{\texttt{NTP}}$), trains the model to predict each token in the caption based on the image and all previous tokens (Liu et al., 2023d; Dai et al., 2023; Lu et al., 2024), formally expressed as:

$$\mathcal{L}_{\texttt{NTP}} = -\sum_{i=1}^{k} \log p\big(t_i \mid v, t_{<i}\big), \tag{1}$$

where t_i is the current token to predict, $t_{<i}$ represents the prefix.

2.2 PRIOR: Prioritizing Image-Related Tokens for Vision-Language Pre-Training

Eq. 1 outlines a model that conditions token prediction on both images and text but lacks a mechanism to verify whether visual information is actually being effectively utilized. The model can optimize this objective by relying exclusively on textual context, potentially resulting in LVLMs that overfit to supplementary text that doesn't correspond to visible image content. To address this, we introduce PRIOR, a simple method to advance the original vision-language pre-training by prioritizing image-related tokens, which are automatically identified by a text-only reference LLM. Drawing from the importance sampling framework (§2.3), we use the reference model to construct a target distribution that assigns higher probability to tokens likely requiring visual information. This approach effectively concentrates optimization on image-related content while reducing emphasis on tokens predictable from text alone.

Text-Only LLM for Modeling Text Distribution PRIOR introduces a text-only LLM as a reference to model the caption distribution without the image inputs. Specifically, we apply NTP loss exclusively on the

text tokens, and the training objective is:

$$\mathcal{L}_{\texttt{TEXT}} = -\sum_{i=1}^{k} \log p_r\big(t_i \mid t_{<i}\big). \tag{2}$$

Vision-Language Pre-Training with Reweighted Tokens Loss PRIOR utilizes the reference model's token probability to calculate the **importance score** w_i for each token:

$$w_i = \big(1 - p_r(t_i \mid t_{<i})\big)^{\alpha}, \tag{3}$$

where α (empirically set to 1) modulates the impact of w_i during pre-training. Tokens with higher importance scores are those that the text-only LLM finds difficult to predict, suggesting they are more likely to be image-related. We implement this scoring process offline—computing and storing the importance score for each token beforehand—which eliminates the need for reference model inference during vision-language pre-training.

PRIOR applies these token-specific importance scores to reweight the NTP loss. The training objective is:

$$\mathcal{L}_{\texttt{PRIOR}} = -\sum_{i=1}^{k} k \cdot \frac{w_i}{\sum_{j=1}^{k} w_j} \log p\big(t_i \mid v, t_{<i}\big) \tag{4}$$

The normalization term $w_i / \sum_{j=1}^{k} w_j$ ensures that the importance scores form a proper distribution across all k tokens, while the multiplication by k preserves the overall scale of the loss. This algorithm strategically upweights tokens that the text-only model struggles to predict, which typically correspond to image-specific information, while maintaining sufficient weight distribution across contextual tokens to ensure coherent language generation.

2.3 PRIOR as Importance Sampling

We present the theoretical framework based on importance sampling that underpins and motivates PRIOR. In general, PRIOR can be interpreted as a form of importance sampling where we draw more samples from regions where the reference model is uncertain. In classical importance sampling, we estimate an expectation under a target distribution $p(x)$ using samples from a proposal distribution $q(x)$:

$$\mathbb{E}_{p(x)}[f(x)] = \int f(x)p(x)dx = \int f(x)\frac{p(x)}{q(x)}q(x)dx = \mathbb{E}_{q(x)}\left[f(x)\frac{p(x)}{q(x)}\right], \tag{5}$$

where the term $\frac{p(x)}{q(x)}$ represents the importance weight that corrects for the mismatch between the distributions. For PRIOR, we can formulate our problem as follows:

- **Function** $f(x)$: The NTP loss (*a.k.a,* negative log-likelihood loss) $-\log p_{\texttt{model}}(t_i|v, t_{<i})$.
- **Sampling Distribution** $q(x)$: The empirical distribution in data $p_{\texttt{data}}(t_i|v, t_{<i})$.
- **Target Distribution** $p(x)$: We aim to define a new target distribution $p_{\texttt{target}}(t_i|v, t_{<i})$ that assigns higher probability to tokens that are difficult for the reference model to predict:

$$p_{\texttt{target}}(t_i|v, t_{<i}) \propto p_{\texttt{data}}(t_i|v, t_{<i}) \cdot \big(1 - p_{\text{r}}(t_i|t_{<i})\big), \tag{6}$$

  where $p_{\text{r}}(t_i|t_{<i})$ is the reference model's probability estimate. Intuitively, we upweight tokens that have low reference probability (surprising tokens).

Importance Sampling Formulation: Since we only have samples from $p_{\texttt{data}}(t_i|v, t_{<i})$, we use importance sampling to estimate the expectation under $p_{\texttt{target}}(t_i|v, t_{<i})$:

$$\begin{aligned}\mathbb{E}_{p_{\texttt{target}}(t_i|v,t_{<i})}[-\log p_{\texttt{model}}(t_i|v, t_{<i})] = \\ \mathbb{E}_{p_{\texttt{data}}(t_i|i,t_{<i})}\left[-\log p_{\texttt{model}}(t_i|v, t_{<i}) \cdot \frac{p_{\texttt{target}}(t_i|v, t_{<i})}{p_{\texttt{data}}(t_i|v, t_{<i})}\right]\end{aligned} \tag{7}$$

Thus, the importance weight for each token loss (*i.e.,* $-\log p_{\texttt{model}}(t_i|v, t_{<i})$ in Eq. 7), with the target distribution defined in Eq. 6, is:

$$w(t_i|v, t_{<i}) = \frac{p_{\texttt{target}}(t_i|v, t_{<i})}{p_{\texttt{data}}(t_i|v, t_{<i})} \propto \big(1 - p_{\text{r}}(t_i|t_{<i})\big) \tag{8}$$

To make this a proper probability distribution over the sequence, we normalize:

$$\tilde{w}(t_i|v, t_{<i}) = \frac{w(t_i|v, t_{<i})}{\sum_{j=1}^{k} w(t_j|v, t_{<j})} \tag{9}$$

While this self-normalization introduces bias into our estimation of the expectation over the target distribution, it substantially reduces variance, yielding improved training stability—a well-established tradeoff in previous work (Metelli et al., 2018; Korbak et al., 2022). Our importance-sampled loss thus becomes:

$$L_{\text{IS}} = -\sum_{i=1}^{k} \tilde{w}_t(t_i|v, t_{<i}) \cdot \log p_{\texttt{model}}(t_i|v, t_{<i}) \tag{10}$$

This weighted average represents the expected loss under our target distribution that emphasizes difficult tokens. Note that this formulation is consistent with Eq. 4, with the difference in the multiplication by k, which preserves the expected loss magnitude when transitioning from NTP to importance-weighted loss. Since the normalized weights sum to 1, omitting k reduces the loss to a weighted average rather than a weighted sum, effectively shrinking the gradients by a factor of approximately k. Multiplying by k restores the original scale. This design choice is important from two perspectives. First, it ensures **gradient consistency**: gradient magnitudes are preserved, enabling direct reuse of existing learning rates and optimizer configurations without retuning. Second, it provides **theoretical grounding** by aligning with self-normalized importance sampling, where normalization reduces variance at the cost of introducing a small bias (Metelli et al., 2018). In addition, this importance sampling framework can be easily extended to consider the α term. We also provide a theoretical justification for the key intuition of `PRIOR` through mutual information analysis, as detailed in §A.

3 Experiments

3.1 Implementation Details of `PRIOR`

We adopt CapsFusion (Yu et al., 2024), which comprises 120M image-text pairs for vision-language pre-training experiments. We first sample a subset (∼5M) and use only the caption to pre-train a text-only reference LLM (initialized with Llama-3-8B (Dubey et al., 2024)). Since the captions in vision-language datasets are typically shorter, we pack multiple samples within the context length (8192) for efficient pre-training.

For vision-language pre-training, we sample another subset (∼3M) and compute token-level reference probabilities $p_r(t_i|t_{<i})$ offline using the reference LLM. We store this data as (image, caption, token-level reference probability list) tuples. To measure the extensive applicability of `PRIOR` across diverse LVLMs architectures, we investigate two types of LVLMs:

- **H-LVLMs**: We implement LVLMs with pre-trained visual encoders (*i.e.,***H**eterogeneous architectures). Following the typical LLaVA-Style LVLMs design (Liu et al., 2023d), we adopt a straightforward ViT-MLP-LLM architecture, employing Llama-3.2-3B as the LLM backbone for efficient experimentation and CLIP-ViT-Large-336 (Radford et al., 2021) as the vision encoder. Following Liu et al. (2023c), we only train the MLP component during pre-training to learn the vision-language alignment and maintain consistent language capabilities across all LVLMs for our controlled study.
- **U-LVLMs**: We implement LVLMs with **U**nified architectures, which drop the pre-trained visual encoders and adopt a single end-to-end Transformer architecture for vision-language modeling (Chen et al., 2024f; Lei et al., 2025; Zhang et al., 2025). Following Chen et al. (2024f), we initialize with Llama-8B and pre-train it on ImageNet for 500 steps before further pre-training on image-caption data.

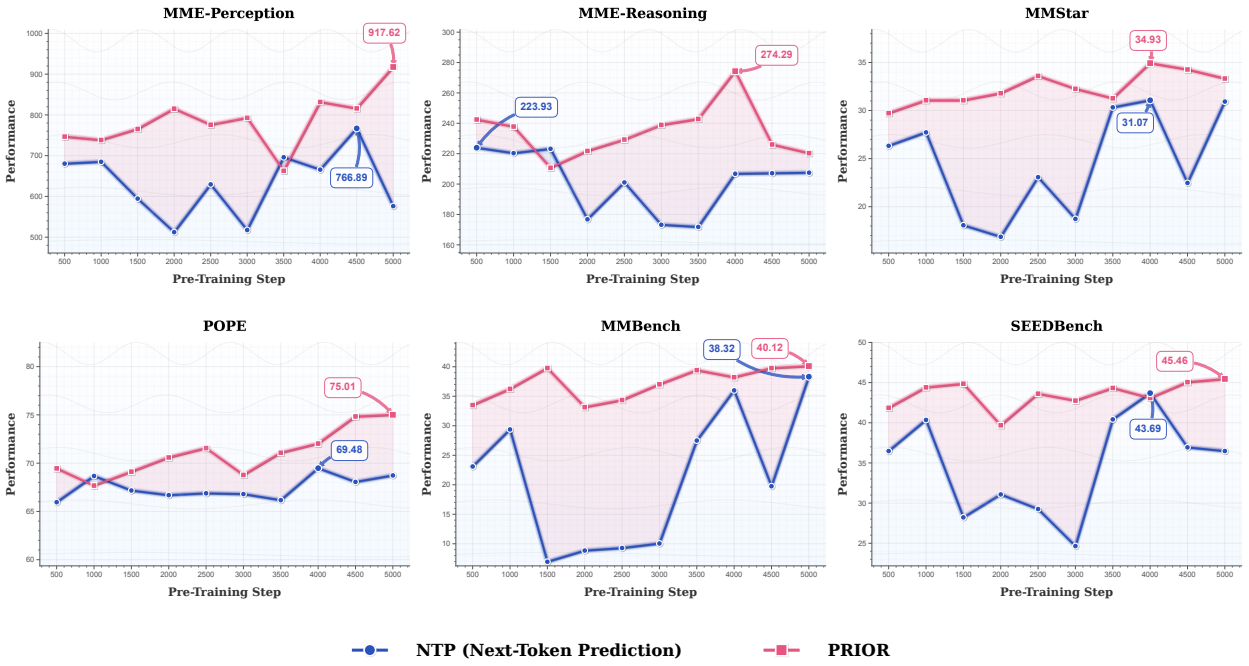

Figure 2: **Main experimental results of LVLMs with pre-trained visual encoders.** We compare PRIOR with the NTP vision-language pre-training across various training steps on LVLMs with pre-trained visual encoders, and we annotate the highest performance for each method, respectively. PRIOR demonstrates both superior performance and greater stability throughout the entire training.

For the two types of LVLMs, we set the total training steps as 5,000, record every 500 steps, with the batch size as 512. During U-LVLM pre-training, we also incorporate an equal proportion of DCLM language dataset (Li et al., 2024b) to maintain general language capabilities.

3.2 Experimental Setting

Evaluation Benchmarks We select the following benchmarks for evaluation: (1) MME (Fu et al., 2024), which measures both the perception and reasoning capabilities. Thus, we report MME-Perception and MME-Reasoning respectively. (2) MMStar (Chen et al., 2024b), which measures advanced vision-language skills. (3) POPE (Li et al., 2023d), a benchmark for object hallucination evaluation. (4) MMBench (Liu et al., 2024c), a benchmark designed for general vision-language capacities evaluation. (5) SEEDBench (Li et al., 2024a), which covers 12 evaluation dimensions including diverse aspects of LVLMs.

Evaluation Setting To elicit meaningful responses beyond image captions during evaluation, we conduct 20 steps for instruction fine-tuning on all pre-trained checkpoints using the same data subset from the post-training set in Chen et al. (2024f). We include a controlled sensitivity study over the number of fine-tuning Steps in §B. We employ VLMEvalKit (Duan et al., 2024) for unified evaluation and comparison across models.

3.3 Experimental Results

The main experimental results for each dataset are shown in Fig. 2 (H-LVLMs) and Fig. 3 (U-LVLMs). Average performance across all datasets is summarized in Fig. 1, bottom right (H-LVLMs) and Fig. 4 (U-LVLMs). We have the following findings:

- **PRIOR generally outperforms the naive NTP pre-training objective by a large margin.** We observe that PRIOR enhances performance across the entire training trajectory for both H-LVLMs and U-LVLMs, demonstrating improvements compared to the NTP objective at most intermediate training

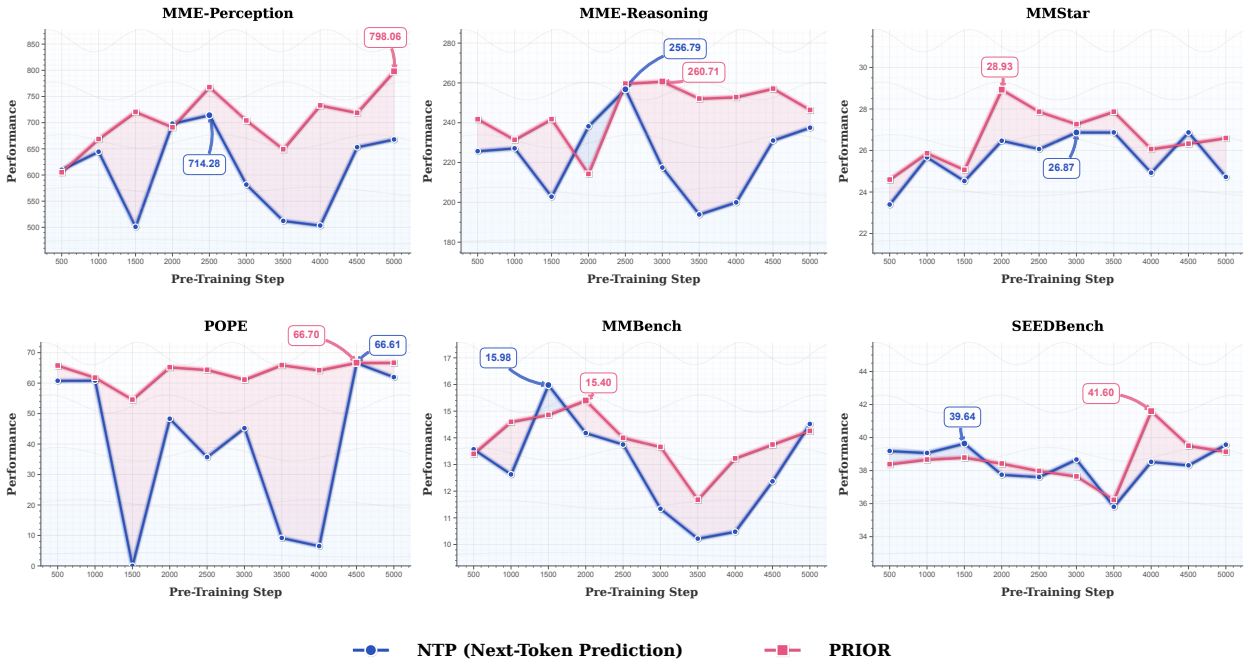

Figure 3: **Main experimental results of LVLMs with unified architectures.** We compare PRIOR with the NTP vision-language pre-training on LVLMs with unified architectures.

Table 1: **A comparison of PRIOR with additional baselines and alternative token-selection approaches implemented on H-LVLMs.** The evaluation is the average of two runs. PRIOR yields consistently better results throughout the evaluation.

Category	Method	MME-P	MME-R	MMStar	POPE	MMBench	SEEDBench
Naive	Next-Token Prediction	576.2	207.5	30.9	68.7	38.3	36.5
Baseline	Image-Text Matching	580.3	215.0	30.8	68.9	38.1	42.5
	Image-Text Contrastive	635.6	218.6	26.3	67.5	38.2	41.2
	Reconstructive Tuning	648.3	202.1	27.3	69.3	38.0	41.5
Token Selection	Attention Weighting	655.4	215.8	31.2	73.4	36.9	39.2
	Rare Word Weighting	588.1	203.2	30.1	70.4	39.2	38.5
Selected Reference LLMs	PRIOR (8B off-the-shelf LLM)	703.4	203.6	29.3	68.6	31.4	41.6
	PRIOR (3B fine-tuned LLM)	824.2	222.9	32.1	71.0	38.0	44.6
	PRIOR (On-the-fly inference)	856.0	210.7	31.7	67.0	38.3	42.5
	PRIOR (8B fine-tuned LLM)	**917.6**	**220.4**	**33.3**	**75.0**	**40.1**	**45.5**

steps. For fully converged models (*i.e.,* at 5,000 steps), we observe average relative improvements across all datasets of 18.61% for H-LVLMs and 7.93% for U-LVLMs.

- **PRIOR demonstrates better training stability.** We observe performance fluctuations during training across both H-LVLMs and U-LVLMs. The NTP objective causes significant performance drops in MMBench and SEEDBench for H-LVLMs, while producing sharp performance fluctuations in POPE and MMBench for U-LVLMs. PRIOR significantly reduces these fluctuations, demonstrating superior training stability compared to the naive NTP pre-training objective.
- **PRIOR exhibits higher potential in performance.** The highest performance scores are highlighted in Fig. 2 and Fig. 3 for each dataset. Our results show PRIOR outperforms the NTP objective in peak performance for both LVLM types. For H-LVLMs specifically, PRIOR achieves optimal performance in final checkpoints, while NTP peaks earlier in training. This suggests PRIOR benefits from extended training, likely due to its comprehensive optimization approach that continues refining multimodal representations in later stages. PRIOR's sustained improvement trajectory indicates greater performance potential.

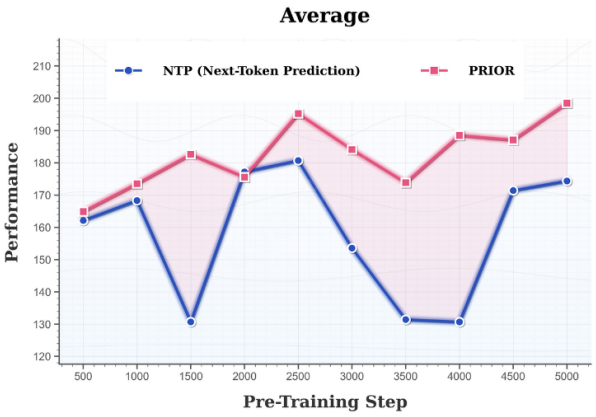

Figure 4: **The average performance comparison on LVLMs with unified architectures.** PRIOR demonstrates better performance and stability across the entire training process.

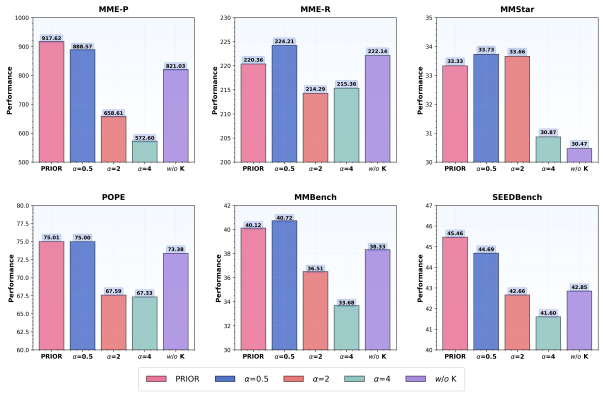

Figure 5: **The ablation study of PRIOR regarding α and k on H-LVLMs.** Results show optimal performance with α=1 and scaling factor k, justifying the design choices in PRIOR.

3.4 Comparison with Additional Baselines, Token Selection Methods, and Selected Reference LLMs

We compare PRIOR with the following approaches specifically designed for H-LVLMs: (1) **Image-text matching**: The training objective that pre-trains the MLP connector to map the [CLS] token embedding close to the caption embedding (Li et al., 2023b; Chen et al., 2024g). (2) **Image-text contrastive learning**: The training objective that pre-trains the MLP connector to align projected image representations with text representations (Chen et al., 2024g; Radford et al., 2021). (3) **Reconstructive tuning**: The training objective that supervises LVLMs to reconstruct images, focusing on the inherent richness and detail within input images (Wang et al., 2024). All these methods are combined with NTP for 5,000 training steps.

The results presented in Tab. 1 demonstrate that PRIOR consistently outperforms these baselines across all benchmarks, with particularly significant gains observed on complex reasoning tasks (*i.e.,* MME-R, MMStar). Additionally, PRIOR integrates more seamlessly with existing pre-training frameworks, requiring only modifications to the loss computation when data is stored offline. This enables scalable pre-training of LVLMs on large-scale datasets across compute clusters.

To further validate the effectiveness of our approach, we conduct another study comparing PRIOR against alternative token-weighting strategies. Since PRIOR operates by prioritizing optimization on *important* tokens within LVLMs, we examine two additional heuristic methods for computing token importance scores: (1) **Attention weighting**, which leverages cross-attention scores from a pre-trained CLIP model (Radford et al., 2021) to assign weights to text tokens based on their visual relevance. (2) **Rare word weighting**, which computes importance scores using inverse token frequency statistics from large-scale text corpora (Das et al., 2023), under the assumption that rare tokens carry more semantic significance.

As shown in Tab. 1, PRIOR substantially outperforms these two token-weighting methods across all evaluation benchmarks, validating the superiority of our principled approach to token importance estimation, with gains of +262.2 and +329.5 on MME-P, a critical benchmark that examines the visual perception capabilities of LVLMs, respectively. This substantial margin highlights that our importance sampling formulation of token importance, grounded in the PRIOR objective, more effectively identifies semantically critical tokens than attention-based or frequency-based heuristics.

We further conduct ablation studies to examine the effect of reference LLM configuration (*i.e.,* fine-tuned vs. off-the-shelf, 3B vs. 8B) as well as an on-the-fly inference alternative, with results on H-LVLMs after 5,000 pretraining steps reported in Tab. 1. We observe that using an off-the-shelf 8B LLM without fine-tuning leads to substantial degradation across all benchmarks (*e.g.,* MME-Perception drops from 917.6 to 703.4), demonstrating that task-specific adaptation is critical for generating reliable importance scores. Among fine-tuned variants, scaling the reference model from 3B to 8B yields consistent improvements, with notable

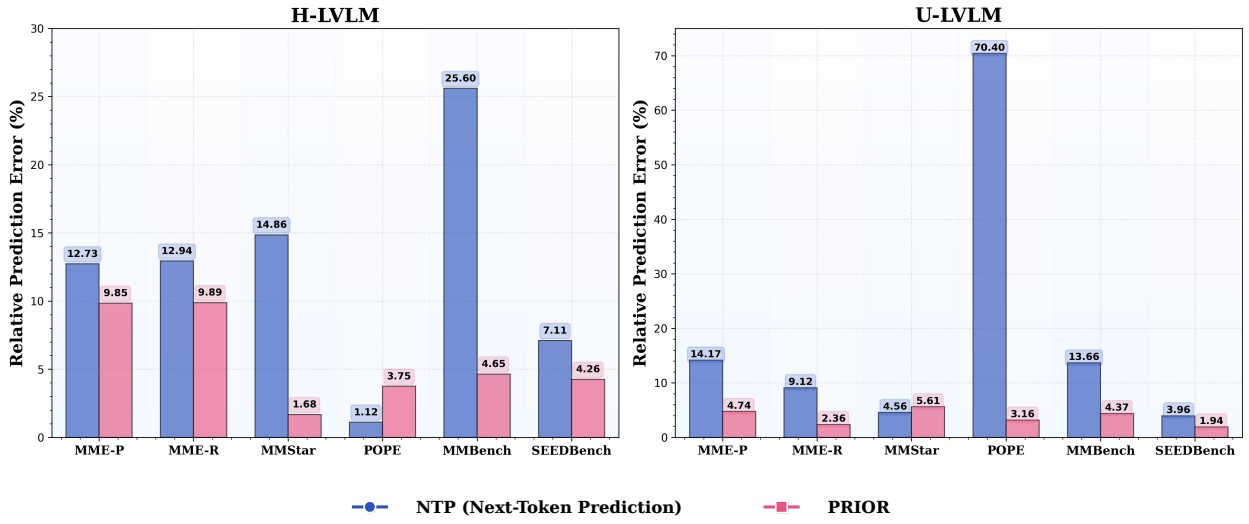

Figure 6: **The relative prediction error of NTP and PRIOR.** We observe that the performance of LVLMs trained via PRIOR is more predictable at scale.

gains on MME-Perception (+93.4) and POPE (+4.0). Nonetheless, even the smaller 3B fine-tuned model provides meaningful guidance for larger LVLMs, indicating that the proposed framework is robust to scale mismatch between the reference and target models, though a scale-matched configuration remains preferable. We also compare against an on-the-fly baseline that estimates text-only predictability during training by running a caption-only forward pass (with images masked) to compute importance weights, followed by an image+caption forward pass optimizing the reweighted loss. PRIOR with offline precomputed scores from the 8B fine-tuned LLM substantially outperforms this approach across all benchmarks, with notable improvements on MME-Perception (+61.6), POPE (+8.0), and SEEDBench (+3.0). The reason is that the caption-domain fine-tuned reference LLM is specifically optimized for modeling caption distributions, yielding more reliable estimates of text-only predictability than the LVLM's masked forward pass.

3.5 Ablation Study

We conduct an ablation study to understand the influence of α and k in Eq (4). While PRIOR sets α to 1 by default, we experiment with $\alpha = 0.5, 2, 4$ to test sensitivity. Higher α increases focus on optimizing loss for image-related tokens identified by the reference model. Additionally, we evaluate the necessity of k, which maintains the token loss scale, by comparing performance with and without k while fixing $\alpha = 1$. All experiments are conducted on H-LVLMs. Results in Fig. 5 demonstrate that performance degrades as α increases beyond 1 (to 2 or 4), yet remains stable within the moderate range of 0.5-1. Furthermore, removing k negatively impacts performance on several benchmarks, validating our design choice to preserve the original loss scale. These findings suggest that a balanced approach to token weighting is crucial, as excessive emphasis on image-related tokens can impede the model's overall learning dynamics. Our empirical selection of $\alpha = 1$ and inclusion of scaling factor k represents an optimal trade-off between focusing on visual content and maintaining stable training.

4 Scaling Laws: PRIOR Scales Predictably and Reliably with Increasing Compute

A fundamental distinction exists between LLMs and LVLMs regarding pre-training objectives. While NTP has proven to be an effective proxy for downstream task performance in LLMs (Chen et al., 2024e; Huang et al., 2024), this correlation does not extend to LVLMs (Chen et al., 2024f). Specifically, optimizing LVLMs to better predict caption given images through NTP does not reliably improve performance on downstream benchmarks. In this section, we conduct a critical analysis of the scaling behaviors of NTP and PRIOR. We train both H-LVLMs and U-LVLMs using different amounts of data (ranging from 7M to 70M training tokens, and including 8 sampling models), and use the following analytical form to estimate the scaling laws of NTP

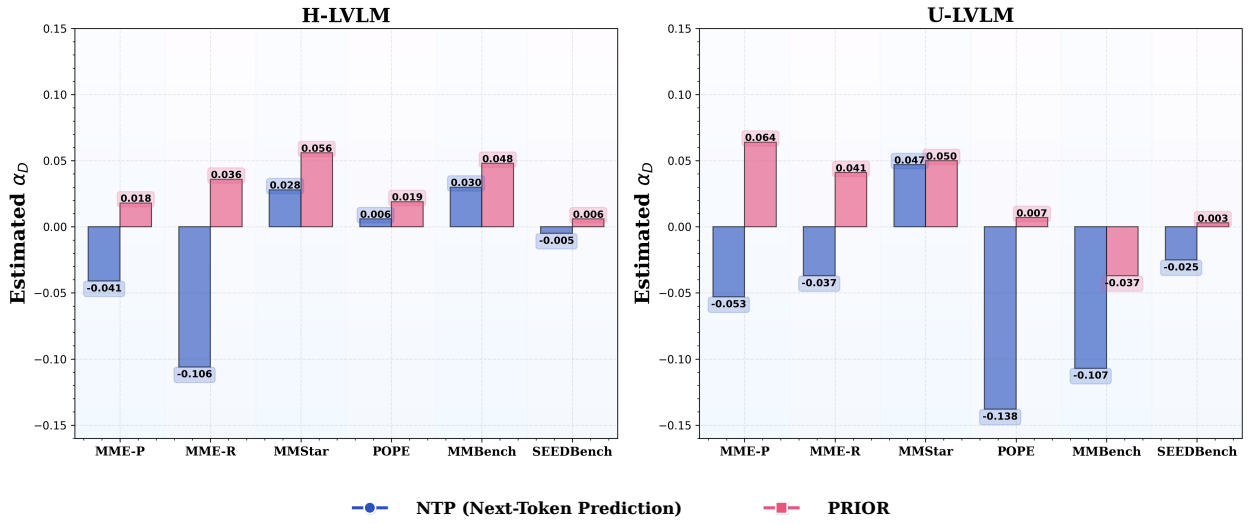

Figure 7: **The scaling behavior comparison of NTP and PRIOR.** PRIOR shows better scaling coefficients, indicating higher efficiency in translating increased resources into performance gains.

and PRIOR (Kaplan et al., 2020):

$$L(D) = \left(\frac{D}{D_c}\right)^{\alpha_D}, \tag{11}$$

where α_D and D_c are constants to be estimated, D is the amounts of token used in training, and L is the predictive performance. This analytical form differs slightly from Kaplan et al. (2020) since we are targeting the benchmark performance rather than the pre-training loss. A larger α_D indicates superior scaling behavior, reflecting greater performance improvement for an equivalent amount of training tokens. With the fitted analytical function, we investigate two problems:

- **Predictability of model performance:** We use the fitted function to estimate the performance of LVLMs trained on 100M tokens, and measure the relative prediction error:

$$\text{Relative Prediction Error} = \frac{|\text{Predictive Performance} - \text{Actual Performance}|}{\text{Actual Performance}} \tag{12}$$

  The results are presented in Fig. 6. The downstream performance of both H-LVLMs and U-LVLMs exhibits significantly higher predictability when trained using PRIOR. This property facilitates more reliable performance estimation for larger-scale deployments, addressing a critical challenge in production-level LVLM implementation

- **Scaling behavior:** Fig. 7 illustrates the scaling factor (*i.e.,* α_D) comparison for both methods. PRIOR demonstrates better scaling properties across two LVLMs architectures, consistently achieving higher α_D than NTP. This indicates that PRIOR more efficiently converts additional training data and compute into improved downstream performance.

5 Further Analysis

5.1 Importance Score Distribution

PRIOR adjusts the loss for each token based on the importance score w_i assigned to that token. We conduct a quantitative analysis to understand the distribution of the importance score. The results are presented in Fig. 8. We examine the relation between text length (*i.e.,* number of tokens) and the average importance score across all tokens within the caption, finding that the average importance score decreases as text length increases, with a linear correlation of r=-0.243. This suggests longer caption introduces more information

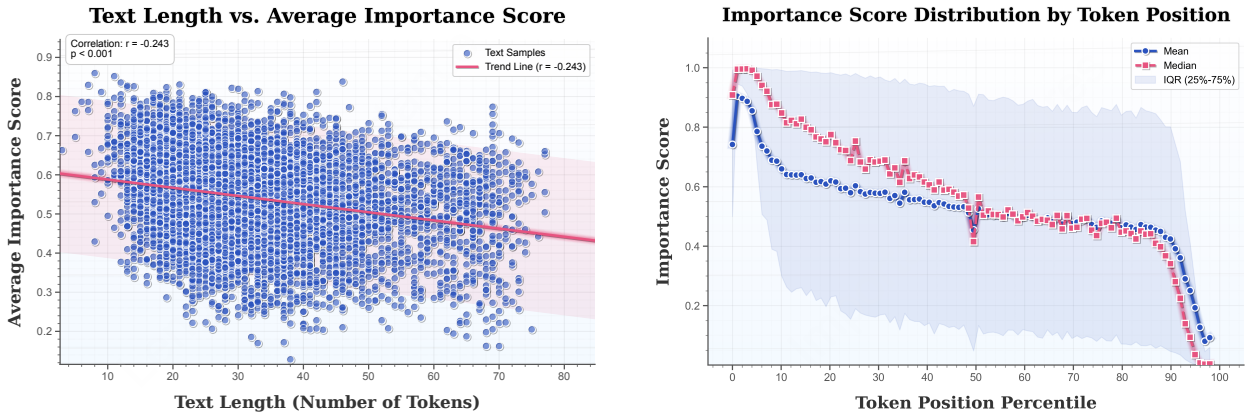

Figure 8: **Quantitative analysis of the importance score distribution.** We find that the average importance score for each caption decreases (linearly) with the text length. Within each caption, the importance score decreases in later positions.

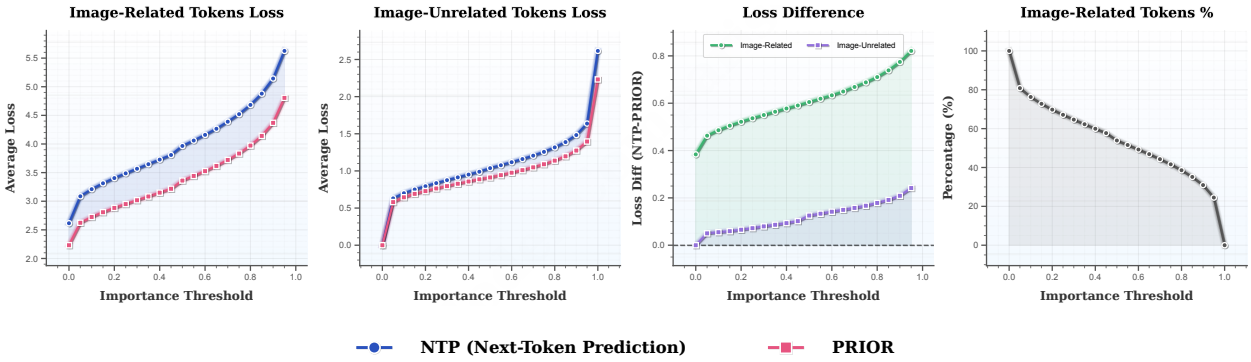

Figure 9: **The comparison of NTP and PRIOR regarding the achieved loss on image-related and image-unrelated tokens.** These two token groups are dynamically categorized based on varying the importance threshold w_i (Eq. 3). We find that PRIOR accelerates the LVLMs training, consistently achieving lower loss on both image-related and image-unrelated tokens.

absent from images, potentially including background content. Analysis of importance scores by token position reveals that initial tokens receive high importance scores, with scores progressively declining for later tokens. This indicates that later tokens can be more readily predicted from previous tokens without image reference.

5.2 Comparative Loss Analysis of NTP and PRIOR

We compare NTP and PRIOR from the loss perspective. By applying various importance thresholds, we categorize tokens into image-related and image-unrelated subsets based on their assigned importance scores w_i. We measure the average NTP loss of H-LVLMs trained via two methods on these two distinct subsets. The results in Fig. 9 reveal a significant pattern: PRIOR deliberately guides models to prioritize image-related tokens, achieving consistently lower loss on these information-rich elements. Interestingly, PRIOR also optimizes learning on image-unrelated tokens, though this performance difference is less pronounced than for image-related tokens. This suggests that PRIOR introduces general model improvement and effectively accelerates LVLMs training overall. Furthermore, our analysis reveals a compelling trend: as we progressively increase the importance threshold, the performance gap between methods widens substantially. This monotonic relationship confirms that PRIOR delivers increasingly significant improvements for tokens with higher image relevance, validating its fundamental design principle of prioritizing tokens that carry the most visually related information.

6 Related Work

6.1 Large Vision-Language Models

Existing research advances the development of LVLMs capable of addressing diverse tasks via a unified interface that can directly generate natural language, thus avoiding task-specific modifications (Wang et al., 2021; 2022; Li et al., 2023b; Agrawal et al., 2024; Luo et al., 2025). Utilizing advanced pre-trained LLMs (Brown et al., 2020; Bubeck et al., 2023; Dubey et al., 2024; Yang et al., 2024) as the language component (Liu et al., 2023d; Zhu et al., 2023), the instruction-following and complex reasoning abilities of LVLMs are significantly improved (Du et al., 2022; Ghosh et al., 2024). Typically, LVLMs leverage extensive image-caption pair datasets (Lin et al., 2014; Schuhmann et al., 2021; 2022) to train a projector that maps image features into the embedding space of LLMs, thereby aligning the two modalities (Liu et al., 2023d; Zhu et al., 2023; Alayrac et al., 2022; Li et al., 2023b; Yu et al., 2023; Awadalla et al., 2024). Furthermore, large-scale vision-language instruction tuning datasets (Su et al., 2023; Wei et al., 2023; Liu et al., 2023a; Gong et al., 2023; Gao et al., 2023; Li et al., 2023a; 2025) and feedback datasets (Chen et al., 2023a; Li et al., 2023c; Sun et al., 2023; Chou et al., 2024; Zhao et al., 2025) are utilized to align LVLMs with human preferences, ensuring their ability to comprehend instructions and generate responses that are user-friendly. In this work, we present a simple vision-language pre-training algorithm applicable to existing datasets that enhances visual-related training outcomes.

6.2 Vision-Language Training Objective

NTP loss on language tokens serves as the predominant pre-training objective for LVLMs (Wang et al., 2022; Dai et al., 2023; Bai et al., 2023; Lu et al., 2024; Dai et al., 2024; Chen et al., 2024f). Researchers have explored complementary approaches, including but not limited to: (1) Contrastive or matching loss to align the vision-language modalities (Li et al., 2019; Lu et al., 2019; Li et al., 2022). (2) Distillation loss to expedite the training (Diao et al., 2024; Liao et al., 2025). (3) Grounding objective to more effectively align the text tokens to the corresponding image regions (Koh et al., 2023; Rasheed et al., 2024). (4) Reconstructing the image based on the pre-defined codebook or external decoders (Sun et al., 2024; Zou et al., 2023; Ge et al., 2024). (5) External constrains to improve the visual perception in LVLMs (Wang et al., 2024; Luo et al., 2024a; Chen et al., 2023b). These supplementary objectives, though effective, compromise the simplicity of NTP in the original pre-training approach, complicating large-scale, efficient implementation (Shoeybi et al., 2019; Rasley et al., 2020; Liang et al., 2024; Zheng et al., 2023). `PRIOR` prioritizes maintaining NTP simplicity by merely adding a regularization term (*i.e.,* a weighting factor) to each token during training.

6.3 Vision-Language Training Corpus

The quality of training data is a critical determinant of LVLMs' ultimate performance. The established efforts on improving the training data include but not limited to: (1) Distilling knowledge from advanced closed-source LVLMs (Chen et al., 2024a;c; Liu et al., 2024b), like GPT-4o (Hurst et al., 2024). (2) Scaling the size of the dataset with scalable data collection pipelines (Li et al., 2024d; Awadalla et al., 2024; Chen et al., 2024d; Wang et al., 2025). (3) Incorporating the human-written filtering rules or pipelines (Guo et al., 2024; Gohari et al., 2025). (4) Curating domain-specific corpus to facilitate certain abilities in LVLMs (Li et al., 2024c; Yun et al., 2024; Fan et al., 2024; Han et al., 2024; Yang et al., 2025). (5) Relying on the self-evolving ability in LVLMs (Liu et al., 2024a; Luo et al., 2024b). While existing work makes valuable progress toward higher-quality training data, these approaches face fundamental scalability constraints imposed by their data creation components, including the inherent capabilities of teacher models, pre-defined filtering criteria, and established pipelines—all of which create practical upper bounds on quality improvement. `PRIOR` avoids imposing human priors on the training dataset and remains compatible with advancements across all vision-language datasets.

7 Conclusion

This work introduces `PRIOR`, an advanced vision-language pre-training method that prioritizes the loss optimization on image-related tokens that a text-only reference model struggles to predict. Our experiments

demonstrate that PRIOR significantly outperforms NTP, achieving 19% and 8% average relative improvement when implemented on LVLMs with and without pre-trained visual encoders, respectively. Furthermore, PRIOR exhibits more predictable and reliable scaling behaviors given increasing compute, indicating that PRIOR represents a promising pre-training algorithm.

Limitations and Broader Impacts

Limitations PRIOR requires the LVLMs and the reference LLM to share the same tokenizer, preventing the creation of a universal dataset for all LVLMs pre-training. This constraint on the tokenizer necessitates specific implementations and processing, increasing computational overhead when deploying across multiple model architectures. However, this tokenizer constraint is a common practice in knowledge distillation and model training pipelines, and many modern model families (*e.g.,* LLaMA series (Dubey et al., 2024), Qwen series (Yang et al., 2024)) already share tokenizers across their variants of different sizes, making this limitation less restrictive in practice. We also discuss the limitations of PRIOR' token importance heuristic in §C.

Broader impacts PRIOR's token prioritization approach enhances LVLMs' performance across benchmarks, potentially accelerating progress in building advanced general-purpose LVLMs. Additionally, by reducing hallucination risk through better grounding of language in visual content, PRIOR contributes to developing more trustworthy LVLMs models. However, there remains the possibility of misuse, and the differential weighting approach may unintentionally amplify existing biases in training data. We encourage continued research into responsible deployment practices and bias mitigation techniques alongside performance improvements.

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Singh Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Théophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego de Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. Pixtral 12b. CoRR, abs/2410.07073, 2024. doi: 10.48550/ARXIV.2410.07073. URL https://doi.org/10.48550/arXiv.2410.07073.

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, Oscar Lo, Manli Shu, Hannah Lee, Etash Guha, Sheng Shen, Mohamed Awadalla, Silvio Savarese, Caiming Xiong, Ran Xu, Yejin Choi, and Ludwig Schmidt. MINT-1T: scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/40b9196c25fe1d64d87ca3a80a91d0ce-Abstract-Datasets_and_Benchmarks_Track.html.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. CoRR, abs/2308.12966, 2023. doi: 10.48550/ARXIV.2308.12966. URL https://doi.org/10.48550/arXiv.2308.12966.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 2020.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4. CoRR, 2023.

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 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, pages 3558–3568. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00356. URL https://openaccess.thecvf.com/content/CVPR2021/html/Changpinyo_Conceptual_12M_Pushing_Web-Scale_Image-Text_Pre-Training_To_Recognize_Long-Tail_Visual_CVPR_2021_paper.html.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XVII, volume 15075 of Lecture Notes in Computer Science, pages 370–387. Springer, 2024a. doi: 10.1007/978-3-031-72643-9_22. URL https://doi.org/10.1007/978-3-031-72643-9_22.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? arXiv preprint arXiv:2403.20330, 2024b.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Lin Bin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. Sharegpt4video: Improving video understanding and generation with better captions. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024c. URL http://papers.nips.cc/paper_files/paper/2024/hash/22a7476e4fd36818777c47e666f61a41-Abstract-Datasets_and_Benchmarks_Track.html.

Xiaohui Chen, Satya Narayan Shukla, Mahmoud Azab, Aashu Singh, Qifan Wang, David Yang, ShengYun Peng, Hanchao Yu, Shen Yan, Xuewen Zhang, and Baosheng He. Compcap: Improving multimodal large language models with composite captions. CoRR, abs/2412.05243, 2024d. doi: 10.48550/ARXIV.2412.05243. URL https://doi.org/10.48550/arXiv.2412.05243.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. Dress: Instructing large vision-language models to align and interact with humans via natural language feedback. arXiv preprint arXiv:2311.10081, 2023a.

Yangyi Chen, Xingyao Wang, Manling Li, Derek Hoiem, and Heng Ji. Vistruct: Visual structural knowledge extraction via curriculum guided code-vision representation. arXiv preprint arXiv:2311.13258, 2023b.

Yangyi Chen, Binxuan Huang, Yifan Gao, Zhengyang Wang, Jingfeng Yang, and Heng Ji. Scaling laws for predicting downstream performance in llms. arXiv preprint arXiv:2410.08527, 2024e.

Yangyi Chen, Xingyao Wang, Hao Peng, and Heng Ji. A single transformer for scalable vision-language modeling. CoRR, abs/2407.06438, 2024f. doi: 10.48550/ARXIV.2407.06438. URL https://doi.org/10.48550/arXiv.2407.06438.

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, pages 24185–24198, 2024g.

Christopher Chou, Lisa Dunlap, Koki Mashita, Krishna Mandal, Trevor Darrell, Ion Stoica, Joseph E. Gonzalez, and Wei-Lin Chiang. Visionarena: 230k real world user-vlm conversations with preference labels. CoRR, abs/2412.08687, 2024. doi: 10.48550/ARXIV.2412.08687. URL https://doi.org/10.48550/arXiv.2412.08687.

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

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuoling Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NVLM: open frontier-class multimodal llms. CoRR, abs/2409.11402, 2024. doi: 10.48550/ARXIV.2409.11402. URL https://doi.org/10.48550/arXiv.2409.11402.

Mamata Das, PJA Alphonse, et al. A comparative study on tf-idf feature weighting method and its analysis using unstructured dataset. arXiv preprint arXiv:2308.04037, 2023.

Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/5e2217482fa75556f1970be809acd3f8-Abstract-Conference.html.

Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. In Luc De Raedt, editor, Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022, pages 5436–5443. ijcai.org, 2022. doi: 10.24963/IJCAI.2022/762. URL https://doi.org/10.24963/ijcai.2022/762.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. arXiv preprint arXiv:2407.11691, 2024.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of models. CoRR, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

Wan-Cyuan Fan, Yen-Chun Chen, Mengchen Liu, Lu Yuan, and Leonid Sigal. On pre-training of multimodal language models customized for chart understanding. CoRR, abs/2407.14506, 2024. doi: 10.48550/ARXIV.2407.14506. URL https://doi.org/10.48550/arXiv.2407.14506.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. arXiv preprint arXiv:2304.15010, 2023.

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama SEE and draw with SEED tokenizer. In The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net, 2024. URL https://openreview.net/forum?id=0Nui91LBQS.

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. CoRR, abs/2404.07214, 2024. doi: 10.48550/ARXIV.2404.07214. URL https://doi.org/10.48550/arXiv.2404.07214.

Hajar Emami Gohari, Swanand Ravindra Kadhe, Syed Yousaf Shah Adam, Abdulhamid Adebayo, Praneet Adusumilli, Farhan Ahmed, Nathalie Baracaldo Angel, Santosh Borse, Yuan-Chi Chang, Xuan-Hong Dang, et al. Gneissweb: Preparing high quality data for llms at scale. arXiv preprint arXiv:2502.14907, 2025.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. arXiv preprint arXiv:2305.04790, 2023.

Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhu Chen, and Xiang Yue. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. CoRR, abs/2412.05237, 2024. doi: 10.48550/ARXIV.2412.05237. URL https://doi.org/10.48550/arXiv.2412.05237.

Xiaotian Han, Yiren Jian, Xuefeng Hu, Haogeng Liu, Yiqi Wang, Qihang Fan, Yuang Ai, Huaibo Huang, Ran He, Zhenheng Yang, and Quanzeng You. Infimm-webmath-40b: Advancing multimodal pre-training for enhanced mathematical reasoning. CoRR, abs/2409.12568, 2024. doi: 10.48550/ARXIV.2409.12568. URL https://doi.org/10.48550/arXiv.2409.12568.

Yuzhen Huang, Jinghan Zhang, Zifei Shan, and Junxian He. Compression represents intelligence linearly. arXiv preprint arXiv:2404.09937, 2024.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll L. Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, and Dane Sherburn. Gpt-4o system card. CoRR, abs/2410.21276, 2024. doi: 10.48550/ARXIV.2410.21276. URL https://doi.org/10.48550/arXiv.2410.21276.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal inputs and outputs. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, volume 202 of Proceedings of Machine Learning Research, pages 17283–17300. PMLR, 2023. URL https://proceedings.mlr.press/v202/koh23a.html.

Tomasz Korbak, Ethan Perez, and Christopher L Buckley. Rl with kl penalties is better viewed as bayesian inference. arXiv preprint arXiv:2205.11275, 2022.

Weixian Lei, Jiacong Wang, Haochen Wang, Xiangtai Li, Jun Hao Liew, Jiashi Feng, and Zilong Huang. The scalability of simplicity: Empirical analysis of vision-language learning with a single transformer. arXiv preprint arXiv:2504.10462, 2025.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. arXiv preprint arXiv:2305.03726, 2023a.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13299–13308, 2024a.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Scott Keh, Kushal Arora, et al. Datacomp-lm: In search of the next generation of training sets for language models. Advances in Neural Information Processing Systems, 37:14200–14282, 2024b.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA. Pmlr, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. CoRR, 2023b.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silkie: Preference distillation for large visual language models. arXiv preprint arXiv:2312.10665, 2023c.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 14369–14387. Association for Computational Linguistics, 2024c. doi: 10.18653/V1/2024.ACL-LONG.775. URL https://doi.org/10.18653/v1/2024.acl-long.775.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. CoRR, 2019.

Qingyun Li, Zhe Chen, Weiyun Wang, Wenhai Wang, Shenglong Ye, Zhenjiang Jin, Guanzhou Chen, Yinan He, Zhangwei Gao, Erfei Cui, Jiashuo Yu, Hao Tian, Jiasheng Zhou, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Zhenxiang Li, Pei Chu, Yi Wang, Min Dou, Changyao Tian, Xizhou Zhu, Lewei Lu, Yushi Chen, Junjun He, Zhongying Tu, Tong Lu, Yali Wang, Limin Wang, Dahua Lin, Yu Qiao, Botian Shi, Conghui He, and Jifeng Dai. Omnicorpus: A unified multimodal corpus of 10 billion-level images interleaved with text. CoRR, abs/2406.08418, 2024d. doi: 10.48550/ARXIV.2406.08418. URL https://doi.org/10.48550/arXiv.2406.08418.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355, 2023d.

Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, Nadine Chang, Karan Sapra, Amala Sanjay Deshmukh, Tuomas Rintamaki, Matthieu Le, Ilia Karmanov, Lukas Voegtle, Philipp Fischer, De-An Huang, Timo Roman, Tong Lu, José M. Álvarez, Bryan Catanzaro, Jan Kautz, Andrew Tao, Guilin Liu, and Zhiding Yu. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. CoRR, abs/2501.14818, 2025. doi: 10.48550/ARXIV.2501.14818. URL https://doi.org/10.48550/arXiv.2501.14818.

Wanchao Liang, Tianyu Liu, Less Wright, Will Constable, Andrew Gu, Chien-Chin Huang, Iris Zhang, Wei Feng, Howard Huang, Junjie Wang, Sanket Purandare, Gokul Nadathur, and Stratos Idreos. Torchtitan: One-stop pytorch native solution for production ready LLM pre-training. CoRR, abs/2410.06511, 2024. doi: 10.48550/ARXIV.2410.06511. URL https://doi.org/10.48550/arXiv.2410.06511.

Bencheng Liao, Hongyuan Tao, Qian Zhang, Tianheng Cheng, Yingyue Li, Haoran Yin, Wenyu Liu, and Xinggang Wang. Multimodal mamba: Decoder-only multimodal state space model via quadratic to linear distillation. arXiv preprint arXiv:2502.13145, 2025.

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

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. arXiv preprint arXiv:2306.14565, 2023a.

Genglin Liu, Xingyao Wang, Lifan Yuan, Yangyi Chen, and Hao Peng. Examining llms' uncertainty expression towards questions outside parametric knowledge. arXiv preprint arXiv:2311.09731, 2023b.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744, 2023c.

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

Wei Liu, Junlong Li, Xiwen Zhang, Fan Zhou, Yu Cheng, and Junxian He. Diving into self-evolving training for multimodal reasoning. CoRR, abs/2412.17451, 2024a. doi: 10.48550/ARXIV.2412.17451. URL https://doi.org/10.48550/arXiv.2412.17451.

Yangzhou Liu, Yue Cao, Zhangwei Gao, Weiyun Wang, Zhe Chen, Wenhai Wang, Hao Tian, Lewei Lu, Xizhou Zhu, Tong Lu, Yu Qiao, and Jifeng Dai. Mminstruct: A high-quality multi-modal instruction tuning dataset with extensive diversity. CoRR, abs/2407.15838, 2024b. doi: 10.48550/ARXIV.2407.15838. URL https://doi.org/10.48550/arXiv.2407.15838.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In European conference on computer vision, pages 216–233. Springer, 2024c.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding. CoRR, abs/2403.05525, 2024. doi: 10.48550/ARXIV.2403.05525. URL https://doi.org/10.48550/arXiv.2403.05525.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 13–23, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html.

Run Luo, Yunshui Li, Longze Chen, Wanwei He, Ting-En Lin, Ziqiang Liu, Lei Zhang, Zikai Song, Xiaobo Xia, Tongliang Liu, Min Yang, and Binyuan Hui. DEEM: diffusion models serve as the eyes of large language models for image perception. CoRR, abs/2405.15232, 2024a. doi: 10.48550/ARXIV.2405.15232. URL https://doi.org/10.48550/arXiv.2405.15232.

Run Luo, Haonan Zhang, Longze Chen, Ting-En Lin, Xiong Liu, Yuchuan Wu, Min Yang, Minzheng Wang, Pengpeng Zeng, Lianli Gao, Heng Tao Shen, Yunshui Li, Xiaobo Xia, Fei Huang, Jingkuan Song, and Yongbin Li. Mmevol: Empowering multimodal large language models with evol-instruct. CoRR, abs/2409.05840, 2024b. doi: 10.48550/ARXIV.2409.05840. URL https://doi.org/10.48550/arXiv.2409.05840.

Run Luo, Ting-En Lin, Haonan Zhang, Yuchuan Wu, Xiong Liu, Min Yang, Yongbin Li, Longze Chen, Jiaming Li, Lei Zhang, et al. Openomni: Large language models pivot zero-shot omnimodal alignment across language with real-time self-aware emotional speech synthesis. arXiv preprint arXiv:2501.04561, 2025.

Alberto Maria Metelli, Matteo Papini, Francesco Faccio, and Marcello Restelli. Policy optimization via importance sampling. Advances in Neural Information Processing Systems, 31, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, 2021.

Hanoona Abdul Rasheed, Muhammad Maaz, Sahal Shaji Mullappilly, Abdelrahman M. Shaker, Salman H. Khan, Hisham Cholakkal, Rao Muhammad Anwer, Eric P. Xing, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Glamm: Pixel grounding large multimodal model. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024, pages 13009–13018. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01236. URL https://doi.org/10.1109/CVPR52733.2024.01236.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash, editors, KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020, pages 3505–3506. ACM, 2020. doi: 10.1145/3394486.3406703. URL https://doi.org/10.1145/3394486.3406703.

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. Advances in Neural Information Processing Systems, 35:25278–25294, 2022.

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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers. Association for Computational Linguistics, 2018.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. CoRR, abs/1909.08053, 2019. URL http://arxiv.org/abs/1909.08053.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. arXiv preprint arXiv:2305.16355, 2023.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net, 2024. URL https://openreview.net/forum?id=mL8Q9OOamV.

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

Chenxin Tao, Shiqian Su, Xizhou Zhu, Chenyu Zhang, Zhe Chen, Jiawen Liu, Wenhai Wang, Lewei Lu, Gao Huang, Yu Qiao, et al. Hovle: Unleashing the power of monolithic vision-language models with holistic vision-language embedding. arXiv preprint arXiv:2412.16158, 2024.

Haochen Wang, Anlin Zheng, Yucheng Zhao, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Zhaoxiang Zhang. Reconstructive visual instruction tuning. CoRR, abs/2410.09575, 2024. doi: 10.48550/ARXIV.2410.09575. URL https://doi.org/10.48550/arXiv.2410.09575.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA. Pmlr, 2022.

Xiao Wang, Ibrahim Alabdulmohsin, Daniel Salz, Zhe Li, Keran Rong, and Xiaohua Zhai. Scaling pre-training to one hundred billion data for vision language models. arXiv preprint arXiv:2502.07617, 2025.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. arXiv preprint arXiv:2108.10904, 2021.

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

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. CoRR, abs/2412.10302, 2024. doi: 10.48550/ARXIV.2412.10302. URL https://doi.org/10.48550/arXiv.2412.10302.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. CoRR, abs/2412.15115, 2024. doi: 10.48550/ARXIV.2412.15115. URL https://doi.org/10.48550/arXiv.2412.15115.

Yue Yang, Ajay Patel, Matt Deitke, Tanmay Gupta, Luca Weihs, Andrew Head, Mark Yatskar, Chris Callison-Burch, Ranjay Krishna, Aniruddha Kembhavi, et al. Scaling text-rich image understanding via code-guided synthetic multimodal data generation. arXiv preprint arXiv:2502.14846, 2025.

Qiying 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, pages 14022–14032, 2024.

Tianyu Yu, Yangning Li, Jiaoyan Chen, Yinghui Li, Hai-Tao Zheng, Xi Chen, Qingbin Liu, Wenqiang Liu, Dongxiao Huang, Bei Wu, and Yexin Wang. Knowledge-augmented few-shot visual relation detection. CoRR, 2023.

Sukmin Yun, Haokun Lin, Rusiru Thushara, Mohammad Qazim Bhat, Yongxin Wang, Zutao Jiang, Mingkai Deng, Jinhong Wang, Tianhua Tao, Junbo Li, Haonan Li, Preslav Nakov, Timothy Baldwin, Zhengzhong Liu, Eric P. Xing, Xiaodan Liang, and Zhiqiang Shen. Web2code: A large-scale webpage-to-code dataset and evaluation framework for multimodal llms. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/cb66be286795d71f89367d596bf78ea7-Abstract-Datasets_and_Benchmarks_Track.html.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, pages 5579–5588. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.

00553. URL https://openaccess.thecvf.com/content/CVPR2021/html/Zhang_VinVL_Revisiting_Visual_Representations_in_Vision-Language_Models_CVPR_2021_paper.html.

Tao Zhang, Xiangtai Li, Zilong Huang, Yanwei Li, Weixian Lei, Xueqing Deng, Shihao Chen, Shunping Ji, and Jiashi Feng. Pixel-sail: Single transformer for pixel-grounded understanding. arXiv preprint arXiv:2504.10465, 2025.

Xiangyu Zhao, Shengyuan Ding, Zicheng Zhang, Haian Huang, Maosong Cao, Weiyun Wang, Jiaqi Wang, Xinyu Fang, Wenhai Wang, Guangtao Zhai, et al. Omnialign-v: Towards enhanced alignment of mllms with human preference. arXiv preprint arXiv:2502.18411, 2025.

Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens. CoRR, abs/2310.02239, 2023. doi: 10.48550/ARXIV.2310.02239. URL https://doi.org/10.48550/arXiv.2310.02239.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. CoRR, 2023.

Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee, and Jianfeng Gao. Generalized decoding for pixel, image, and language. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, pages 15116–15127. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01451. URL https://doi.org/10.1109/CVPR52729.2023.01451.

Table 2: Sensitivity study over the number of SFT steps on H-LVLMs. We report performance of NTP and PRIOR along with their gap (Δ) across six benchmarks. PRIOR consistently outperforms NTP regardless of the SFT duration, confirming that its gains stem from improved pre-training representations.

SFT Steps	Method	MME-P	MME-R	MMStar	POPE	MMBench	SEEDBench
Step-20	NTP	576.2	207.5	30.9	68.7	38.3	36.5
	PRIOR	**917.6**	**220.4**	**33.3**	**75.0**	**40.1**	**45.5**
	Δ	+341.4	+12.9	+2.4	+6.3	+1.8	+9.0
Step-50	NTP	871.0	224.7	32.7	68.9	42.7	42.8
	PRIOR	**995.8**	**235.2**	**37.0**	**75.5**	**48.1**	**47.6**
	Δ	+124.8	+10.5	+4.3	+6.6	+5.4	+4.8
Step-500	NTP	992.9	237.8	34.5	69.3	52.6	44.3
	PRIOR	**1024.7**	**243.9**	**37.3**	**76.3**	**59.7**	**50.4**
	Δ	+31.8	+6.1	+2.8	+7.0	+7.1	+6.1
Step-1000	NTP	947.0	241.6	37.3	70.3	54.1	47.0
	PRIOR	**1037.9**	**254.3**	**38.6**	**77.1**	**60.0**	**54.0**
	Δ	+90.9	+12.7	+1.3	+6.8	+5.9	+7.0

A Theoretical Justification via Mutual Information

The efficacy of PRIOR can be further understood through the lens of mutual information theory. When optimizing vision-language models, we aim to maximize the predictive power of visual information v with respect to text t. This can be formalized by maximizing the mutual information between v and t, quantifying the uncertainty reduction about t when v is observed. For data sampled from the joint distribution (v, t), the mutual information is expressed as:

$$I(v;t) = \mathbb{E}_{(v,t)}\left[\ln \frac{p(v,t)}{p(v)p(t)}\right] = \mathbb{E}_{(v,t)}\left[\ln \frac{p(t|v)}{p(t)}\right] \tag{13}$$

Decomposing this expression at the token level, we obtain:

$$I(v;t) = \mathbb{E}_{(v,t)}\left[\sum_i \ln p(t_i|v, t_{<i}) - \ln p(t_i|t_{<i})\right] \tag{14}$$

This formulation shows that mutual information is maximized when there's a large discrepancy between token probability given both visual and textual context $p(t_i|v, t_{<i})$ versus textual context alone $p(t_i|t_{<i})$. Tokens exhibiting this difference benefit most from visual inputs.

The weighting mechanism in PRIOR, defined as $w_i = (1 - p_r(t_i|t_{<i}))^\alpha$, implicitly aligns with this mutual information objective. By assigning higher importance scores to tokens that are difficult to predict from text alone, PRIOR effectively prioritizes tokens where visual information potentially provides the greatest reduction in uncertainty. This approach creates a natural emphasis on tokens where $\ln p(t_i|v, t_{<i}) - \ln p(t_i|t_{<i})$ is likely to be large, thus indirectly promoting higher mutual information between vision and language representations.

This intuitively corresponds to focusing loss optimization on the tokens where the LVLMs have the greatest opportunity to outperform text-only predictions by leveraging visual information. In essence, PRIOR adaptively modulates the learning signal based on the potential information gain from incorporating visual context, leading to more efficient and effective vision-language pre-training.

B Controlled Sensitivity Study over the Number of SFT Steps

In the main experiments (§3), all pre-trained checkpoints are evaluated after a lightweight 20-step instruction fine-tuning (SFT) stage. A natural question arises: does PRIOR's advantage originate from the pre-training

objective itself, or is it amplified by this particular fine-tuning regime? To answer this, we conduct a controlled sensitivity study by varying the number of SFT steps $\in \{20, 50, 500, 1000\}$ while keeping the SFT data, optimizer, and evaluation protocol fixed. All SFT runs start from the fully converged 5,000-step pre-trained H-LVLMs checkpoints. The results are presented in Tab. 2. We have the following observations: (1) The PRIOR–NTP gap persists across all SFT regimes. Across all six benchmarks and all four SFT configurations, PRIOR consistently outperforms NTP. This demonstrates that PRIOR's advantage is not an artifact of a particular fine-tuning duration. (2) PRIOR maintains advantages even with extended SFT. At Step-1000, PRIOR still outperforms NTP on all benchmarks (*e.g.,* +90.9 on MME-Perception, +6.8 on MMBench, +7.0 on SEEDBench), confirming that the gains are not merely amplified by limited fine-tuning but reflect genuinely improved representations learned during pre-training.

C Limitations of the Token Importance Heuristic

While PRIOR's text-only probability heuristic effectively identifies image-related tokens in the majority of cases, we acknowledge scenarios where low text-only probability does not correspond to visual relevance. We provide a qualitative analysis of such failure cases to better characterize the limitations of our approach.

Rare Proper Nouns and Specific Identifiers Consider the following caption from our curated study dataset: "A hand-painted ceramic vase crafted by artisan Mikhail Vorontsov, displayed at the Hermitage Museum." In this case, visually verifiable tokens like "hand-painted" and "ceramic" receive moderate importance scores (*e.g.,* $w_i = 0.65$ and 0.55, respectively), appropriately reflecting their image relevance. However, the artisan's first name "Mikhail" and the specific museum "Hermitage" receive very high importance scores ($w_i = 0.98$ and 0.92) because they are rare and difficult to predict from textual context alone. These tokens represent metadata or provenance information that cannot be verified from the image, yet they receive disproportionately high weights in the PRIOR loss.

Invisible Attributes and Non-Visual Metadata A related failure mode arises with attributes that are fundamentally non-visual. For example, given the caption "An authentic antique mahogany chair, appraised at $47,500 and previously owned by aristocrats," the token "aristocrats" receives a high importance score ($w_i = 0.91$) due to its unpredictability from context. Similarly, "authentic" ($w_i = 0.62$) describes a property that requires expert verification rather than visual inspection. While visually informative tokens like "mahogany" ($w_i = 0.85$) are also appropriately upweighted, the heuristic cannot distinguish between tokens that are *hard to predict* because they describe visual content versus those that are hard to predict because they encode non-visual metadata.

Numerical Specificity Exact quantities such as prices, dates, and measurements are inherently unpredictable from preceding text, leading to high importance scores. For instance, a price tag like "$365,500" or a specific address like "7338 24th Ave NE" in a real estate caption will receive elevated importance scores despite being metadata that cannot be inferred from the associated image.

Mitigating Factors Despite these cases, several aspects of PRIOR's design help mitigate their impact. First, the normalization term $w_i / \sum_{j=1}^{k} w_j$ bounds the maximum weight any single token can receive, preventing rare entities from dominating the loss. Second, these failure cases represent a relatively small fraction of all high-importance tokens in practice—the majority of tokens with low text-only probability genuinely correspond to visual content, as evidenced by the consistent performance improvements across all benchmarks.