Positional Bias in Multimodal Embedding Models: Do They Favor the Beginning, the Middle, or the End?

Kebin Wu[1], **Fatima Albreiki**[1]

[1]Technology Innovation Institute (TII), Abu Dhabi, UAE
kebin.wu@tii.ae, fatima.albreiki@tii.ae

Abstract

Positional bias—where models overemphasize certain positions regardless of content—has been shown to negatively impact model performance across various tasks. While recent research has extensively examined positional bias in text generation models, its presence and effects in representation models remain underexplored. Even less is known about such biases in multimodal models. In this work, we investigate positional bias in multimodal representation models, specifically in the context of image-text retrieval. We begin by distinguishing between context importance and positional bias, and then assess the presence and extent of positional bias across different models and datasets. Our experiments demonstrate that positional bias is prevalent in multimodal models, but manifests differently across modalities: text encoders tend to exhibit bias toward the beginning of the input, whereas image encoders show bias at both the beginning and end. Furthermore, we find that this bias arises from, or is amplified by, a combination of factors, including the positional encoding scheme, training loss, context importance, and the nature of using image-text pairs in multimodal training.

Code — https://github.com/tiiuae/PosBias/

1 Introduction

Transformer-based models have achieved notable success, especially in natural language processing, leading to the development of numerous language models. However, emerging research indicates that their ability to capture contextual information is influenced by the position of that information within the input sequence—an issue known as positional bias. For example, Liu *et al.* (Liu et al. 2024) showed that models often prioritize content at the beginning or end, while neglecting the middle—a phenomenon termed "lost in the middle." This weakens their reasoning abilities and introduces instability in evaluations. In response, many studies have explored the underlying causes of positional bias and proposed mitigation strategies.

Despite recent progress, research on positional bias has largely focused on text generation. We extend this by analyzing positional bias in multimodal representation learning, specifically within cross-modal retrieval tasks. Our study examines representative multimodal embedding models such as CLIP (Radford et al. 2021) and evaluates positional bias in their text and image encoders separately. Figure 1 shows a text-to-image retrieval example where shifting the position of the ground-truth text within the query significantly alters the retrieved results.

We begin by distinguishing between positional bias and contextual importance. To investigate positional bias, we segment the image or text into multiple parts and iteratively shift a selected segment across different positions, while replacing the remaining segments with either dummy perturbations or modality- and model-specific masking content. This approach reveals substantial positional biases in both the image and text encoders of multimodal models. We extend the analysis to several multimodal models and investigate the prevalence and causes of such bias in multimodal representation learning.

In this paper, we make the following contributions. First, we reveal the presence of positional bias in multimodal models—a phenomenon not previously reported, to our knowledge. Second, we empirically characterize distinct bias patterns: in text encoders, performance is consistently higher when a segment appears at the beginning of the sequence; in image encoders, bias is more variable, favoring either the beginning or both ends. Third, through extensive experiments, we show that this bias persists across a wide range of models, regardless of data distribution, positional encoding, training length, model size, resolution, patch size, loss function, or vision encoder architecture. Finally, we identify several contributing factors that may drive or modulate positional bias.

2 Related Work

2.1 Positional Bias in Text Generation

Positional bias has been extensively studied in text generation models, especially in the context of long-context language modeling. Liu *et al.* (Liu et al. 2024) identified the "lost in the middle" phenomenon, where language models underperform when relevant information appears mid-sequence. Hsieh *et al.* (Hsieh et al. 2024) found that large language models often exhibit a U-shaped attention pattern, emphasizing the beginning and end of sequences over the middle, regardless of semantic importance. Several mitigation strategies have been proposed, including attention calibration (Hsieh et al. 2024), data augmentation (He et al. 2023), differential attention mechanisms (Ye et al.

Figure 1: Top-5 text-to-image retrieval results. The caption is divided into six segments. Each row corresponds to a different position where only Segment 0 (highlighted in dark green) is placed, while the remaining positions are masked. Across the six rows, Segment 0 is shifted through all six possible positions.

2024), and replacing causal attention with bidirectional variants (Wang et al. 2024). While some studies have extended this analysis to vision-language models (Tian et al. 2025), their focus remains on text generation tasks like image captioning.

2.2 Positional Bias in Text Representation

Positional bias has been observed not only in generation tasks but also in text representation models used for retrieval and classification. Coelho *et al.* (Coelho et al. 2024) investigated this bias in web document retrieval and found that such models tend to prioritize the beginning of documents. They attribute this effect to contrastive training and the inverted pyramid structure commonly found in writing. Similar head-biased tendencies were found in Named Entity Recognition (NER), Part-of-Speech (POS) tagging (Ben Amor, Granitzer, and Mitrović 2024), and extractive question answering (Ko et al. 2020). Goel *et al.* (Goel, Lee, and Ramchandran 2024) analyzed this bias through semantic similarity measurement, showing consistent overemphasis on early tokens across diverse embedding models. These biases have been hypothesized to result from characteristics of the training data and preprocessing steps such as input truncation. In this paper, we extend the investigation of positional bias to the multimodal setting by empirically examining its existence, patterns, and the underlying causes or contributing factors.

2.3 Multimodal Representation Learning

Multimodal representation learning models, particularly CLIP (Radford et al. 2021) and its variants, align visual and textual inputs in a shared embedding space using contrastive objectives. While CLIP demonstrates strong performance, it relies on large batch sizes and is limited to short textual inputs. SigLIP (Zhai et al. 2023) mitigates the batch size issue by replacing the contrastive loss with a sigmoid-based alternative. To address text length limitations, Long-CLIP (Zhang et al. 2024) introduces a knowledge-preserving positional embedding stretching strategy, extending the token limit to 248 and fine-tuning the model on a dataset specifically curated for long image-text pairs (Chen et al. 2024). In contrast to such approaches that rely on absolute positional encoding, TULIP (Najdenkoska et al. 2024) adopts relative positional encoding, enabling the model to process text of arbitrary length. This design allows CLIP-like models to make full use of longer captions, thereby enhancing their ability to capture fine-grained and detailed semantic information embedded in extended textual contexts.

Other multimodal models, such as LLaVA (Liu et al. 2023) and BLIP-2 (Li et al. 2023), have gained prominence for enabling complex, instruction-based multimodal generation tasks. However, in this work, we focus on CLIP and its variants, which are designed for multimodal representation learning through contrastive alignment of image and text embeddings. This focus is motivated by the fact that CLIP-style models frequently serve as the key backbone for more advanced multimodal systems, including those used in generative settings. In addition, cross-modal retrieval plays a crucial role in multimodal generation pipelines that incorporate Retrieval-Augmented Generation (RAG) techniques (Wu et al. 2024; Xia et al. 2024). As such, a deeper understanding of CLIP-like models in the context of cross-modal retrieval is critical for advancing the broader landscape of vision-language learning.

A more detailed discussion of related work is provided in the supplementary material.

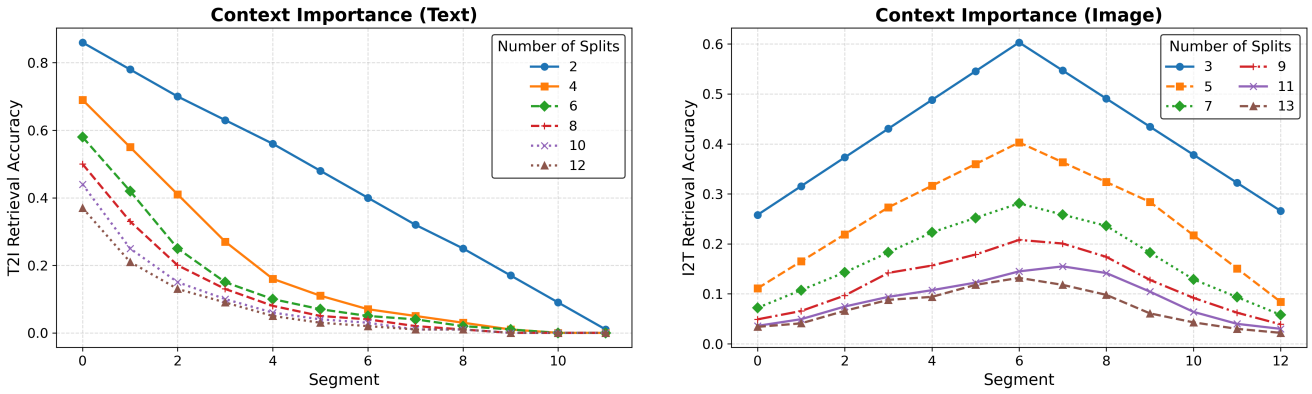


Figure 2: Contextual importance analysis across modalities (image and text). The x-axis shows positions, and colors indicate different number of splits.

3 Methods

3.1 Context Importance

Positional bias is sometimes confounded with contextual importance, as models may attend more to certain positions not solely due to their placement, but because those positions often contain semantically important content. To disentangle these factors, we first identify the regions of contextual importance for the text and image encoders separately.

To evaluate contextual importance in the **text encoder** of a trained CLIP model or its variants, we divide the input tokens into uniform segments. For each trial, we retain only the tokens within one segment and mask the rest using the model's padding token. The masked input is then passed through the text encoder to obtain its representation, with the image input held constant to avoid cross-modal interference. A similar approach is used for the **image encoder**: one segment of the image is kept while the others are masked using CLIP's RGB mean values ([0.481, 0.458, 0.408]), while the text input remains fixed to isolate the contribution of each image segment.

Figure 2 demonstrates the results for Long-CLIP (ViT-B/16) on the Urban1k dataset (Zhang et al. 2024) under various segment split configurations. Since different numbers of splits result in different segment granularities, we apply interpolation to align the segment-wise importance values onto a common scale for visually coherent comparison. For the contextual importance of text (left), retrieval accuracy is highest with the first segment and decreases monotonically with later ones, reflecting the common writing pattern where key information appears early. For the image modality (right), the central region shows the highest importance, with a gradual drop toward the edges—consistent with the typical compositional bias in natural images, where subjects are centered and the top/bottom often contain less informative background like sky or ground. These trends remain consistent across segment granularities, and generalize across datasets and model architectures (see Section C.1 in the supplementary material). Building on these observations, we next investigate positional bias, with a specific focus on whether regions of bias and contextual importance overlap.

3.2 Positional Bias

To investigate positional bias in the **text encoder** of a multimodal model, we adopt two strategies: text perturbation and token masking. In the perturbation approach, the input text is divided into sub-texts, one of which is moved across positions while the others are replaced with Lorem Ipsum dummy placeholders to minimize semantic interference (Goel, Lee, and Ramchandran 2024). To further reduce semantic interference, we apply token masking: the tokenized input is split into segments, and one segment is shifted across positions while the rest are masked with the model's padding token. In both cases, the image input remains fixed, and we evaluate retrieval accuracy for each manipulated position. Similarly, to assess positional bias in the **image encoder**, a single visual segment is isolated and moved across spatial locations, with all other regions masked using CLIP's RGB mean values (see Figure 6 in the supplementary material).

Unlike previous studies on positional bias—where the order of multiple documents is shuffled while preserving the internal content of each—we isolate a single segment of the input sequence and systematically move it across different positions. This design is motivated by two key considerations. First, the models examined in our study operate within a fixed and limited context window; including the full input sequence would occupy all available positions, leaving no room to vary token placement for positional analysis. Second, prior work typically concatenates several candidate inputs into a single sequence, where only one is relevant to the query and the others function as distractors. In contrast, in our setting, all image or text segments are closely aligned with the query, making it inappropriate to treat any segment as an irrelevant perturbation. Consequently, simple shuffling would not isolate positional effects in a meaningful way. Importantly, using a single segment per input does not compromise generalization, as our goal is not to measure absolute retrieval accuracy, but rather to analyze how retrieval perfor-

mance varies with positional changes.

4 Experiments

4.1 Experimental Setting

Dataset. We use three image-text retrieval datasets: Urban1K (Zhang et al. 2024) and DOCCI (Onoe et al. 2024) for long captions, and COCO (Lin et al. 2014) for short captions. Unlike prior work focusing only on long texts, we analyze positional bias across both short and long inputs.

Models. In this work, we empirically assess positional bias across models trained under diverse settings. This enables us to evaluate the (1) generality of the observed bias and to (2) provide evidence that either supports or challenges prevailing assumptions on the causes of positional bias, such as position encoding schemes, loss functions, training stages, data distribution, and preprocessing methods. Our analysis includes the following models:

- **Long-CLIP (ViT-B/16)** (Zhang et al. 2024): Fine-tuned on a long-caption dataset (Chen et al. 2024) with a 248-token context window and absolute positional encoding.
- **TULIP (ViT-L/14)** (Najdenkoska et al. 2024): Uses relative positional encoding for long captions, allowing comparison with Long-CLIP's absolute encoding.
- **Shuffled Long-CLIP (ViT-B/16)**: To examine the role of data structure in positional bias, we train a variant of Long-CLIP on shuffled captions. Each original caption is split into sub-captions at sentence boundaries (e.g., ".", "?", "!"). The sub-captions within the same caption are then randomly reordered (with internal word order preserved), concatenated, and used for training. All other settings are kept identical to Long-CLIP.
- **CLIP** (Radford et al. 2021): Official models (ViT-B/32, B/16, L/14, L/14-336, ResNet-50) are included to assess bias in short-caption settings and to examine the effects of architecture, patch size, resolution, and model size.
- **SigLIP-Base** (Zhai et al. 2023): This model differs from CLIP primarily in its loss function and is included to assess the role of contrastive loss in inducing positional bias (Coelho et al. 2024).

We use official checkpoints for TULIP, CLIP, and SigLIP. Long-CLIP is reproduced using the official recipe with two changes: mixed precision is disabled, and training is run on 4 GPUs with batch size 128. Shuffled Long-CLIP uses the same setup with shuffled sub-captions.

Evaluation Metric. Most experiments in this paper focus on cross-modal retrieval tasks. For the long-caption dataset, we report Recall@1 for both image-to-text (I2T) and text-to-image (T2I) retrieval. For the short-caption dataset COCO, we instead report Recall@10, as Recall@1 is generally too low to support meaningful comparisons. However, T2I (respectively, I2T) is considered when evaluating positional bias in the text (respectively, image) encoder, as the bias analysis involves modifying the text (image) input while keeping the other modality unchanged. In investigating the cause of positional bias, we also conducted experiments on a classification task, where top-1 classification accuracy is utilized as the evaluation metric.

4.2 Experimental Results on Positional Bias

In this section, we first examine positional bias in the multimodal model Long-CLIP (ViT-B/16) on the long-caption dataset Urban1K. We extend the analysis to other datasets and models in a later section, as part of our discussion on the reasons of positional bias.

Textual Positional Bias. On Urban1K, we divide each caption into six segments, based on data analysis indicating that six is the modal number of sub-captions. In Figure 3(a), we present results using a token masking strategy, where each color represents the specific segment (among the six) that is selected and moved across different positions in the input sequence to evaluate positional effects. Here, the step size matches the segment length, resulting in six valid positions. We observe that retrieval accuracy is consistently highest when any given segment is placed at the beginning of the sequence, regardless of its original position in the caption. This indicates a strong positional bias favoring the beginning. When applying the text perturbation method, we observe a similar trend, with the strongest bias again appearing at the beginning of the sequence (see Figure 8 in the supplementary material). Therefore, for brevity, we include only the token masking results in the main paper and provide perturbation-based results in Figure 9 in the supplementary material. In addition to positional effects, Figure 3(a) also reveals context importance. By comparing different segments placed at the same fixed position (i.e., comparing different colors along a vertical slice), we find that the first segment consistently yields the highest retrieval performance.

Visual Positional Bias. Extending the methodology used to study positional bias in text encoder, we examine positional effects on the image side by setting the split number to 7. Choosing an odd number allows the central segment, which often contains the most informative visual content as shown in Section 3.1, to occupy a distinct middle position and enables the detection of potential central bias. Figure 4(a) shows the positional bias in the image encoder of the Long-CLIP (ViT-B/16) model on Urban1K. The results reveal a clear bias toward both the beginning and end positions in most segments (observed in 5 out of 7 cases), with the bias at the beginning being notably stronger. Given that the central segment is the most semantically informative for images as discussed previously, the tendency to favor the beginning and end, particularly the beginning, is unexpected.

4.3 Investigating Reasons for Positional Bias

Previous work has attributed positional bias—mainly in text generation and representation models—to factors such as human writing style (Coelho et al. 2024), training objectives (Coelho et al. 2024), data preprocessing (Goel, Lee, and Ramchandran 2024), positional embeddings, and causal attention (Wang et al. 2024). Here, we conduct experiments to evaluate these assumptions and either support or challenge them in the context of **multimodal** models.

Data Distribution. It is commonly assumed that positional bias in representation models arises from the data distribution: when certain image regions or text segments consistently enhance performance, models learn to favor those po-

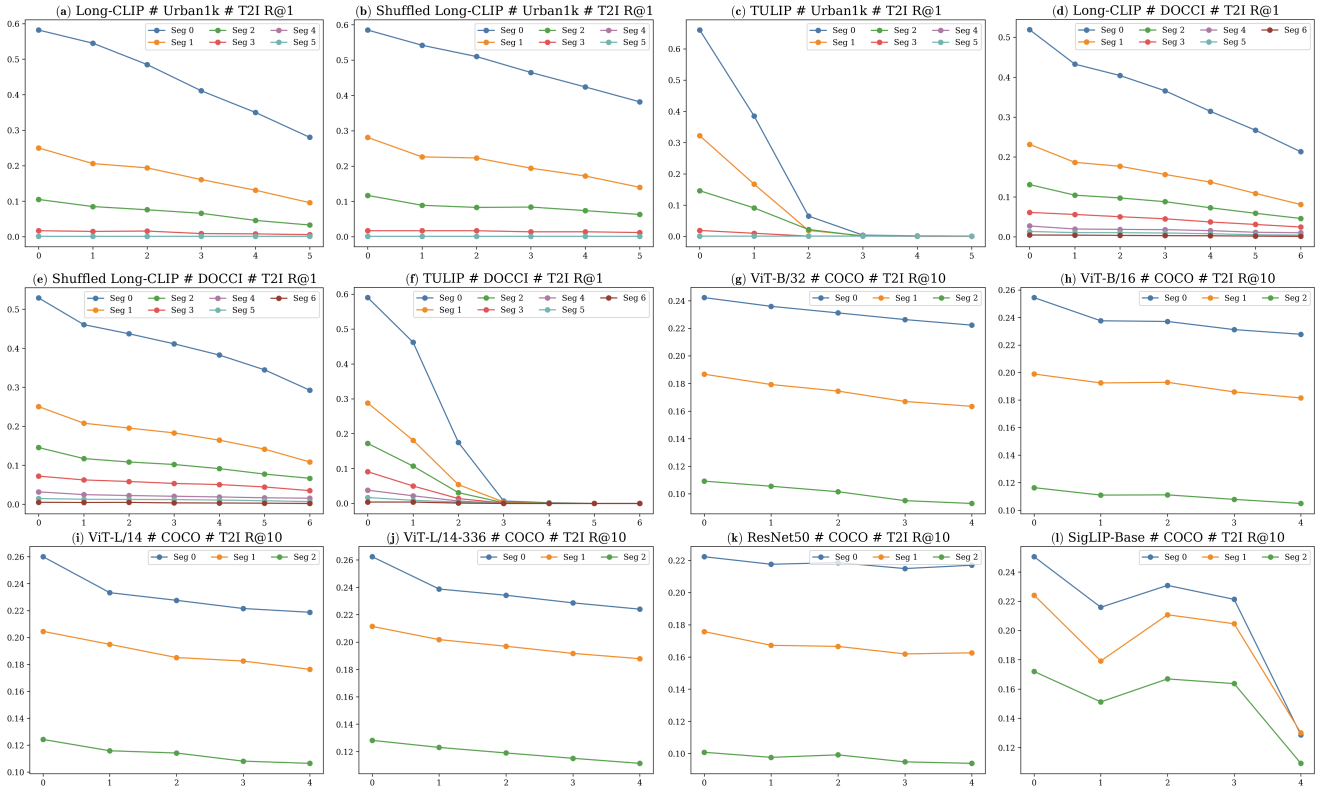


Figure 3: Positional bias analysis of text encoders across multimodal models and datasets. Each figure title follows the format: model name # dataset # metric. The x-axis shows positions, and colors indicate different segments.

sitions, thus developing positional bias towards them. However, based on our findings on context importance (Section C.1) and positional bias (Section 4.2), we argue that regions with high contextual importance do not always overlap with those exhibiting positional bias.

For images, central regions are the most semantically informative, yet the observed bias favors the beginning and the end. *This misalignment does not support the initial hypothesis*. On the text side, moving any segment to the beginning consistently improves retrieval accuracy over its original position. Since the most informative segment also tends to appear at the beginning, aligning with the strongest positional bias, it remains unclear whether importance alone drives the bias. To disentangle these factors, we introduce Shuffled Long-CLIP (ViT-B/16), which randomly reorders sub-captions for training to reduce position-specific cues on the text side. Yet, the model still shows a strong bias toward the beginning, as shown in Figure 3(b). However, compared to the original Long-CLIP, Shuffled Long-CLIP shows a less pronounced drop in accuracy across positions. In particular, shifting the first segment from the beginning to the end causes a drop of 0.199 (from 0.581 to 0.382) in Shuffled Long-CLIP, whereas the original Long-CLIP sees a larger drop of 0.303 (from 0.581 to 0.278). *These findings suggest that contextual importance contributes to positional bias in text representations to some extent, but it does not fully account for the observed bias.*

Notably, these findings are not limited to a single dataset. Similar patterns of positional bias (Figure 3(d) for text and Figure 4(d) for image) and contextual importance (Figure 7 in the supplementary material) are observed on dataset DOCCI, which features more diverse images and human-annotated captions than Urban1K. For DOCCI, we set the number of text splits to seven, matching the average number of sub-sentences per caption.

Positional Encoding. Encoding positional information is essential for contextual understanding. CLIP and its variants typically achieve this using learnable absolute positional encodings for both image and text encoders. In contrast, TULIP adopts rotary positional encoding (RoPE) (Su et al. 2024) in its text encoder to support longer contexts. To examine how positional encoding schemes affect positional bias, we conduct experiments with TULIP on Urban1K and DOCCI. Results for text-based bias appear in Figure 3(c,f), and image-based results are shown in Figure 4(c,f). Compared to Long-CLIP, which uses absolute encoding, TULIP exhibits a stronger positional bias. For example, on Urban1K, shifting the first segment from position 0 to 2 causes TULIP's retrieval accuracy to drop sharply from 0.66 to 0.065, while Long-CLIP shows a more modest decline from 0.582 to 0.485. *These results highlight that the choice of positional encoding significantly influences positional bias: although absolute encodings mitigate the effect, neither strategy eliminates it entirely.* Interestingly, although both the text and image encoders in Long-CLIP use the same absolute positional encoding, the text encoder shows a clear

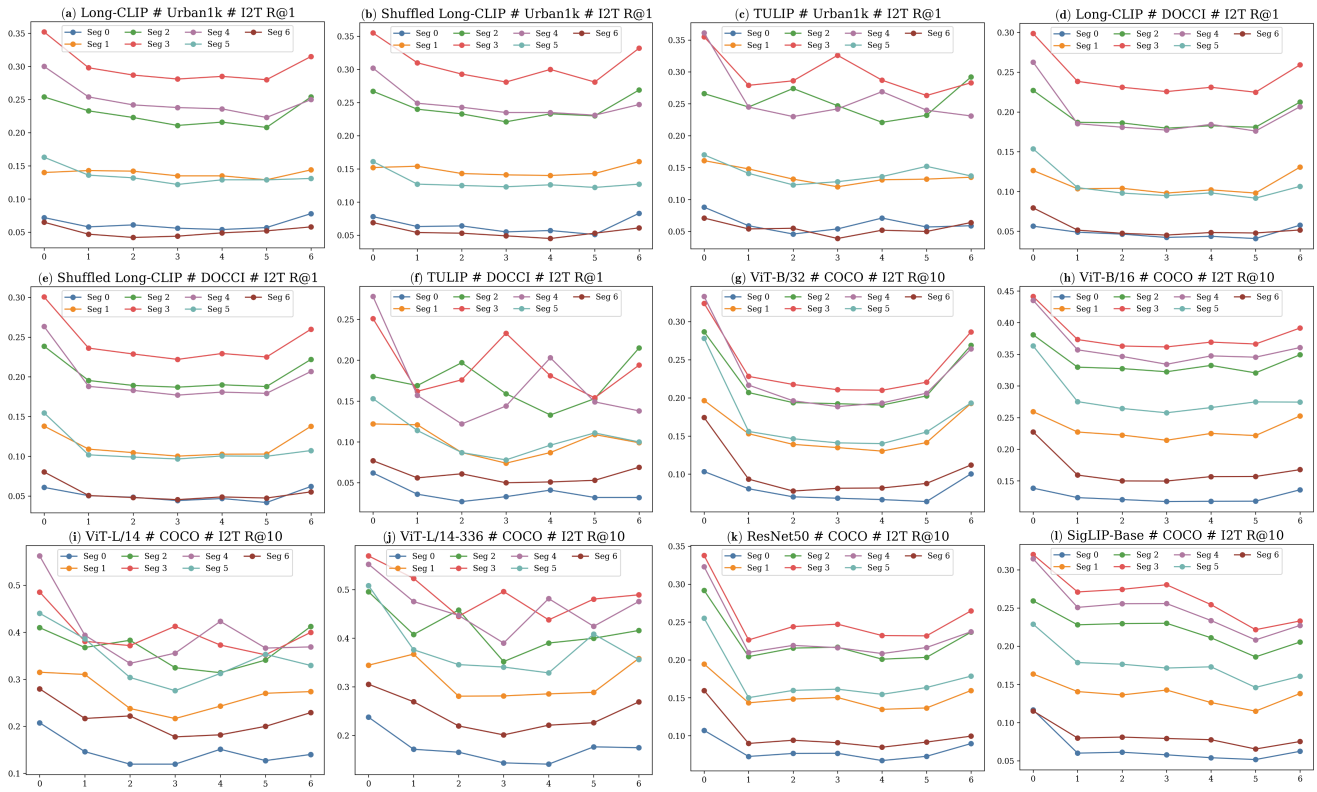


Figure 4: Position bias analysis of image encoders across models and datasets. Each figure title follows the format: model name # dataset # metric. The x-axis shows positions, and colors indicate different segments.

bias toward the beginning, while the image encoder exhibits bias toward both the beginning and end.

Text Length. While our earlier experiments focus on multimodal models trained with long captions, it remains unclear whether similar positional bias arises when training uses only short-caption image-text pairs. To explore this, we evaluate CLIP-ViT-B/16 on the COCO dataset, where captions average 11.53 valid tokens. Due to the short captions, we analyze the first 12 tokens—split into three segments—and shift them across five positions spanning the full context window. As shown in Figure 3(h), CLIP-ViT-B/16 still shows positional bias in the text encoder, favoring early positions, though the effect is weaker. A similar pattern appears in the image encoder, with higher accuracy at both the beginning and end (Figure 4(h)). *These findings suggest that positional bias persists regardless of caption length.*

Model Size. To further validate the prevalence of such bias, we repeat the experiments across models of different parameter sizes, including CLIP-ViT-B/16 and CLIP-ViT-L/14. From Figure 3(i) and Figure 4(i), *it is obvious that such bias is prevailing irrespective of model size*. In fact, CLIP-ViT-L/14 exhibits even stronger positional bias in both the text and image encoders, as evidenced by higher coefficient of variation across different positions compared to CLIP-ViT-B/16 (see Table 1 in the supplementary material). Beyond overall bias strength, we also observe differences in how positional information is retained across models. In the image encoder bias analysis, CLIP-ViT-L/14 maintains high retrieval accuracy when a segment is shifted back to its original position—typically lower than at the beginning and end, but still higher than at other positions. This phenomenon is either absent or substantially less pronounced in CLIP-ViT-B/16, suggesting that CLIP-ViT-L/14 retains more positional information, which may contribute to its superior retrieval performance. Notably, a similar pattern can also be observed in the visual positional bias analysis of TULIP (Figure 4(c,f)), which is also based on the ViT-L/14 architecture.

Image Resolution and Patch Size. Within the CLIP series, we further investigate the effects of image resolution and patch size on positional bias. A comparison between CLIP-ViT-B/16 and CLIP-ViT-B/32 (Figure 3(g)) and Figure 4(g)) reveals that positional bias remains evident and follows a similar distribution pattern. Notably, CLIP-ViT-B/32 shows a stronger bias in the image encoder, as reflected by the higher coefficient of variation reported in Table 1 in the supplementary material. Furthermore, comparing CLIP-ViT-L/14 with CLIP-ViT-L/14@336 (Figure 3(j)) and Figure 4(j)) suggests that increasing image resolution helps reduce bias for the image encoder. Although the highest accuracy consistently occurs when segments are moved to the beginning (positions 0 or 1) or end (position 6), indicating persistent bias, CLIP-ViT-L/14@336 exhibits a lower coefficient of variation across visual segment positions in 5 out

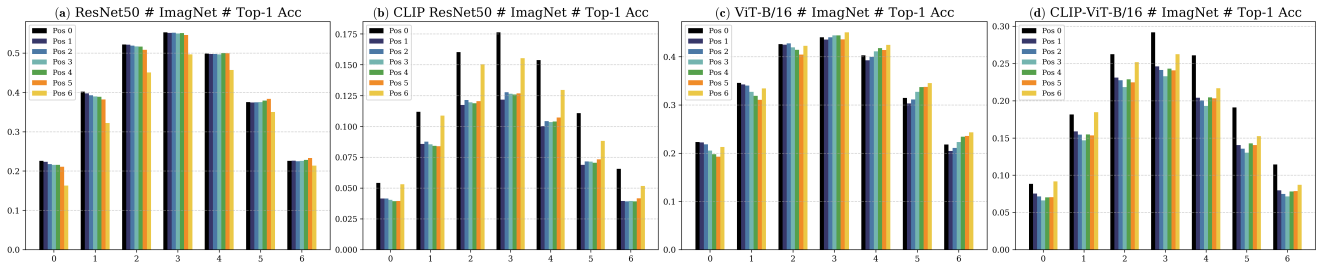


Figure 5: Classification performance on ImageNet for vision-only and CLIP-based models. We report Top-1 accuracy for ResNet-50 and ViT-B/16, both as standalone vision models and as vision backbones within CLIP.

of 7 cases (see Table 1). In summary, reducing the patch size or increasing the image resolution is beneficial for lowering positional bias in the vision encoder.

Training Loss. Previous studies have suggested that positional bias arises and worsens due to contrastive pretraining and fine-tuning (Coelho et al. 2024). In this paper, we examine the role of contrastive loss in multimodal models. Unlike CLIP and its variants, which use a softmax-based contrastive loss, SigLIP employs a sigmoid cross-entropy loss as its training objective (Zhai et al. 2023). After performing similar image-side and text-side bias analyses, we observe that positional bias still persists, as shown in Figure 3(l) and Figure 4(l). Interestingly, the bias pattern on the image side differs from that observed in CLIP-based models. SigLIP exhibits a prominent bias toward the beginning of the sequence. Although the accuracies at the final positions are generally higher than at the penultimate position, they are typically lower than those at most other positions. *These findings suggest that the training loss plays an important role in positional bias, as contrastive learning and sigmoid loss lead to different bias patterns*.

Model Structure. Given the experiments above, a natural question arises: are these biases specific to transformer-based models? To address this, we conducted an additional experiment using CLIP-ResNet-50, where the vision encoder is a ResNet-50 — a convolutional neural network (CNN) rather than a transformer. As presented in Figure 3(k) and Figure 4(k), positional bias still persists, and the observed patterns resemble those seen with ViT-based image encoders. *This suggests that positional bias is not exclusive to transformer-based models*. Future work could explore whether this holds when both image and text encoders are non-transformer-based.

Uni-Modality vs. Multi-Modality. Coelho *et al.* (Coelho et al. 2024) reported a "dwelling at the beginning" bias in certain text representation models. Interestingly, this bias pattern in unimodal models trained solely on text data is consistent with the pattern we observe in the text encoder of multimodal models trained on image-text pairs. Here, we shift focus to the image side, investigating whether the existence and pattern of positional bias remain consistent when the vision encoder is trained with image-only data versus image-text pairs. To this end, we adopt image classification on ImageNet (Deng et al. 2009) as the evaluation task and utilize Top-1 accuracy as the metric. For multimodal models, we use zero-shot classification, where class categories are formatted as natural language prompts. As shown in Figure 5(a), the accuracy of moving a segment across different positions using ResNet classification model remains relatively stable, indicating no clear positional bias. This finding is expected as CNN models usually possesses the property of translation invariance. In contrast, CLIP-ResNet-50 zero-shot classification results in Figure 5(b) exhibit clear positional bias at the beginning and end. To further validate this observation, we conducted the same comparison using ViT-B/16 as the vision encoder and found that the positional bias is significantly more pronounced under the CLIP framework (see the comparison between Figure 5(c) and Figure 5(d)). *These findings suggest that the positional bias on the image side in multimodal models originates from, or is amplified by, training with image–text pairs*.

5 Conclusion and Future Work

This paper presents an empirical analysis of positional bias in multimodal representation models, mainly focusing on image-text retrieval. We find that positional bias is widespread: text encoders tend to emphasize early positions, while image encoders exhibit bias at the beginning or both ends. This bias appears to arise from—or be shaped by—multiple factors, including positional encoding schemes, the contextual importance of training data, training objectives, and the intrinsic use of image-text pairs in multimodal learning.

For future work, we aim to further investigate the underlying causes of positional bias and develop strategies to mitigate it. For instance, the Differential Transformer (Ye et al. 2024), which introduces a differential attention mechanism to reduce attention noise in causal models, may offer promising directions if adapted to bidirectional multimodal representations.

References

Ben Amor, M.; Granitzer, M.; and Mitrović, J. 2024. Impact of Position Bias on Language Models in Token Classification. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, 741–745.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; He, C.; Wang, J.; Zhao, F.; and Lin, D. 2024. Sharegpt4v: Improving large multimodal models with better captions. In *European Conference on Computer Vision*, 370–387. Springer.

Coelho, J.; Martins, B.; Magalhães, J.; Callan, J.; and Xiong, C. 2024. Dwell in the Beginning: How Language Models

Embed Long Documents for Dense Retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 370–377.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. IEEE.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Goel, S.; Lee, R. J.; and Ramchandran, K. 2024. Quantifying Positional Biases in Text Embedding Models. *arXiv preprint arXiv:2412.15241*.

He, J.; Pan, K.; Dong, X.; Song, Z.; Liu, Y.; Sun, Q.; Liang, Y.; Wang, H.; Zhang, E.; and Zhang, J. 2023. Never Lost in the Middle: Mastering Long-Context Question Answering with Position-Agnostic Decompositional Training. *arXiv preprint arXiv:2311.09198*.

Hsieh, C.-Y.; Chuang, Y.-S.; Li, C.-L.; Wang, Z.; Le, L.; Kumar, A.; Glass, J.; Ratner, A.; Lee, C.-Y.; Krishna, R.; et al. 2024. Found in the middle: Calibrating Positional Attention Bias Improves Long Context Utilization. In *Findings of the Association for Computational Linguistics ACL 2024*, 14982–14995.

Ko, M.; Lee, J.; Kim, H.; Kim, G.; and Kang, J. 2020. Look at the first sentence: Position bias in question answering. *arXiv preprint arXiv:2004.14602*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, 740–755. Springer.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12: 157–173.

Najdenkoska, I.; Derakhshani, M. M.; Asano, Y. M.; van Noord, N.; Worring, M.; and Snoek, C. G. 2024. Tulip: Token-length upgraded clip. *arXiv preprint arXiv:2410.10034*.

Onoe, Y.; Rane, S.; Berger, Z.; Bitton, Y.; Cho, J.; Garg, R.; Ku, A.; Parekh, Z.; Pont-Tuset, J.; Tanzer, G.; et al. 2024. Docci: Descriptions of connected and contrasting images. In *European Conference on Computer Vision*, 291–309. Springer.

Peysakhovich, A.; and Lerer, A. 2023. Attention sorting combats recency bias in long context language models. *arXiv preprint arXiv:2310.01427*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; and Liu, Y. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063.

Tian, X.; Zou, S.; Yang, Z.; and Zhang, J. 2025. Identifying and Mitigating Position Bias of Multi-image Vision-Language Models. *arXiv preprint arXiv:2503.13792*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, Z.; Zhang, H.; Li, X.; Huang, K.-H.; Han, C.; Ji, S.; Kakade, S. M.; Peng, H.; and Ji, H. 2024. Eliminating position bias of language models: A mechanistic approach. *arXiv preprint arXiv:2407.01100*.

Wu, T.-H.; Biamby, G.; Quenum, J.; Gupta, R.; Gonzalez, J. E.; Darrell, T.; and Chan, D. M. 2024. Visual Haystacks: Answering Harder Questions About Sets of Images. *arXiv e-prints*, arXiv–2407.

Xia, P.; Zhu, K.; Li, H.; Wang, T.; Shi, W.; Wang, S.; Zhang, L.; Zou, J.; and Yao, H. 2024. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*.

Ye, T.; Dong, L.; Xia, Y.; Sun, Y.; Zhu, Y.; Huang, G.; and Wei, F. 2024. Differential transformer. *arXiv preprint arXiv:2410.05258*.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11975–11986.

Zhang, B.; Zhang, P.; Dong, X.; Zang, Y.; and Wang, J. 2024. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, 310–325. Springer.

Supplementary Material

A Related Works

Positional Bias in Text Generation. Most studies on positional bias have focused on text generation models. Several works have explored different aspects of this issue, including the identification of the phenomenon, analysis of its underlying causes, and investigation of potential mitigation strategies. In (Liu et al. 2024), the 'lost in the middle' phenomenon was identified for the first time by analyzing performance of long-context language models on tasks of multi-document question answering and key-value retrieval. This finding also inspired the authors to propose new model evaluation protocols. In (Hsieh et al. 2024), it is found that LLMs usually exhibit an U-shaped attention bias, with the tokens at the beginning and at the end given higher attention, irrespective of their contextual importance. To alleviate such bias, a calibration mechanism called found-in-the-middle was proposed so that the allocated attention can be faithful to the actual relevance. Ye *et al.* (Ye et al. 2024) noticed that Transformer usually allocates a small portion of attention to relevant regions, while putting a majority of its attention on irrelevant context. They proposed to handle this challenge by introducing a modified architecture, namely DIFF Transformer, which adopts a differential attention mechanism to cancel attention allocated to irrelevant context, but amplify these for the relevant context. In (Wang et al. 2024), it was claimed that position bias is the results of two common modules in LMs: causal attention and position embedding. The resulted solution was to replace the causal attention with bidirectional attention, and also re-order the documents based on model attention values. In (He et al. 2023), the approach to overcoming the bias is augmenting the training documents so that the correct answers are located at arbitrary positions in contexts among noisy documents. In (Peysakhovich and Lerer 2023), it was discovered that most models show bias towards context that are close in position to the generated response. And they put forward to sort documents based on attention scores before running model generation.

Most existing studies on positional bias have concentrated on text generation models, investigating different facets of this issue, including its identification, underlying causes, and possible mitigation strategies. The "lost in the middle" phenomenon, first identified by Liu *et al.* (Liu et al. 2024), highlights how long-context language models often underperform when relevant information appears in the middle of an input sequence. Their analysis, conducted on multi-document question answering and key-value retrieval tasks, also motivated the development of new evaluation protocols for long-context understanding. Hsieh *et al.* (Hsieh et al. 2024) further revealed that large language models (LLMs) tend to exhibit a U-shaped attention pattern, prioritizing tokens at the beginning and end of sequences regardless of their contextual relevance. To address this, they proposed a calibration mechanism called *found-in-the-middle*, which adjusts attention to more accurately reflect the true importance of tokens.

In a related line of work, Ye *et al.* (Ye et al. 2024) observed that standard Transformers often allocate disproportionate attention to irrelevant context, while neglecting relevant regions. They introduced the DIFF Transformer, a modified architecture that employs a differential attention mechanism to suppress attention on unimportant tokens while amplifying it for relevant ones. Similarly, Wang *et al.* (Wang et al. 2024) attributed positional bias to two architectural components: causal attention and position embeddings, and proposed a solution that replaces causal attention with bidirectional attention, combined with document reordering based on attention distributions. He *et al.* (He et al. 2023) approached the problem from a data-centric perspective by augmenting training data such that correct answers appear at arbitrary positions within noisy contexts, encouraging models to attend to content irrespective of position. Similarly, Peysakhovich *et al.* (Peysakhovich and Lerer 2023) recommended sorting documents by attention relevance prior to response generation to reduce this bias.

Tian *et al.* (Tian et al. 2025) extend the bias analysis from LLMs to large vision-language models (LVLMs), with a particular emphasis on models capable of reasoning over multiple images. Their findings reveal a positional bias favoring image placement toward the end of the input sequence in open-source models, which they attribute to the use of causal attention. While the study is conducted in the multimodal domain, the underlying architecture remains aligned with text generation models, treating visual inputs as special text tokens and focusing primarily on question answering tasks, which are inherently text generation-based.

Prior work has consistently demonstrated the pervasive nature of positional bias in text generation models, offering a range of perspectives, including architectural modifications, training strategies, and inference-time techniques, to understand and mitigate its effects. In contrast, our work investigates positional bias in multimodal representation learning, where we observe distinct bias patterns that differ notably from those found in text generation models.

Positional Bias in Text Representation. Coelho *et al.* (Coelho et al. 2024) investigated positional biases in text representation modeling, focusing on web document retrieval tasks. Unlike text generation models, their study revealed that text representation models tend to emphasize the beginning. Interestingly, this bias was not found to be a result of language modeling pretraining but emerged after contrastive pretraining. The phenomenon, referred to as "dwell in the beginning," was attributed to the inverted pyramid writing style, where the most important information is typically presented at the start of a document. In (Ben Amor, Granitzer, and Mitrović 2024), the positional bias of language representation models in token classification tasks, such as Named Entity Recognition (NER) and Part-of-Speech (POS) tagging, is investigated. A similar tendency was observed, where models exhibit a bias toward the beginning of sequences. The authors hypothesize that this bias arises due to properties of the training data: the sequences are generally short and the most informative tokens often appear early. Focusing on the application of extractive question answering with text embedding models, the study in (Ko et al. 2020) also reveals a positional bias to-

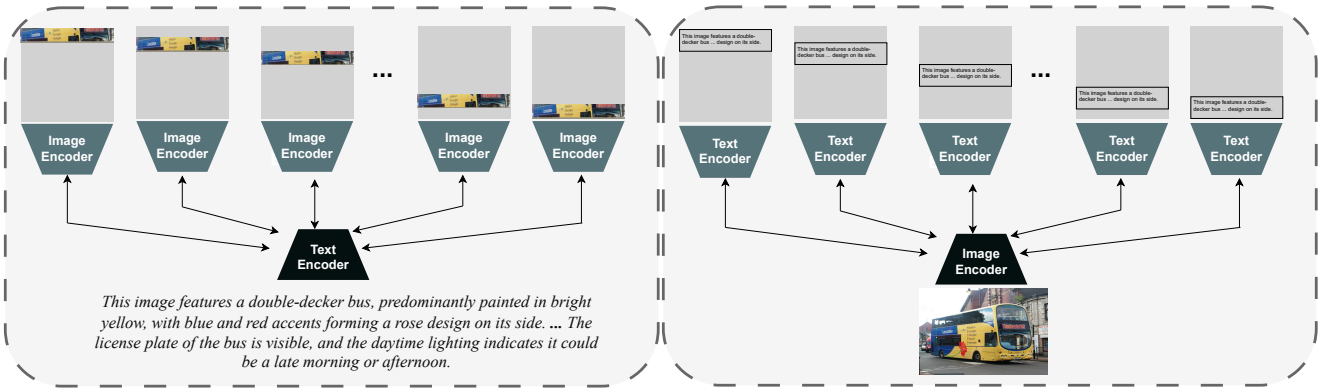
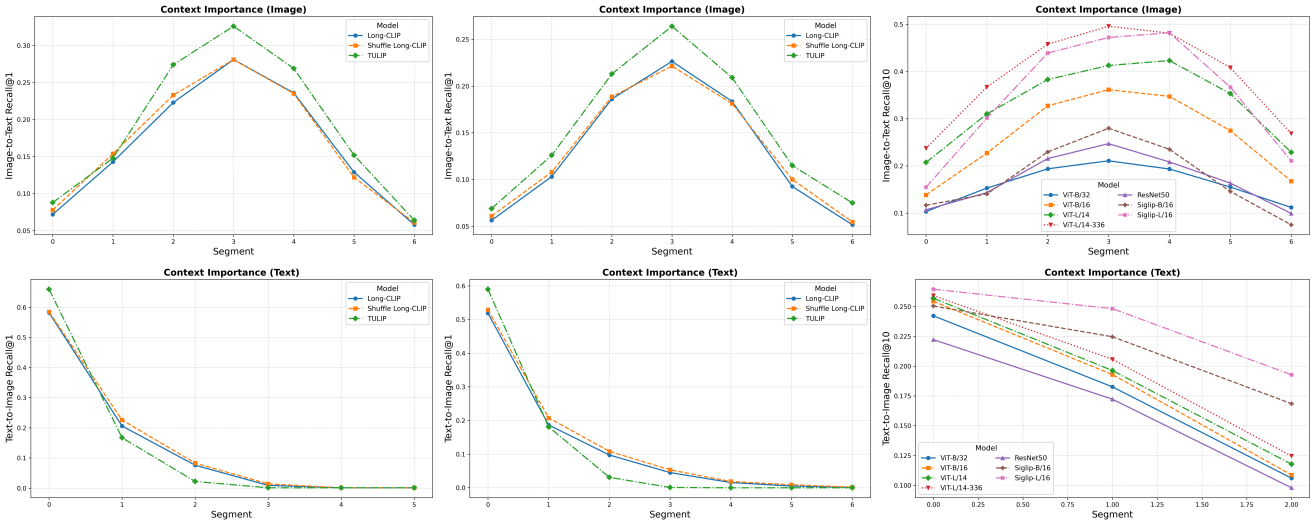


Figure 6: Demonstration of the experimental setup for analyzing positional bias in image and text encoders. For image bias (left), a fixed visual segment is shifted across spatial positions while the caption remains constant. For text bias (right), a fixed text segment is moved across positions in the input sequence with the image held constant. Retrieval scores are computed at each position in both cases.



Urban1k DOCCI COCO

Figure 7: Context importance of different modalities in retrieval tasks.

ward the beginning of the input. The authors attribute this bias to the characteristics of the training data distribution. The study in (Goel, Lee, and Ramchandran 2024) further demonstrates the presence of positional bias in text embedding models, specifically from the perspective of semantic similarity measurement. Across eight different models, regardless of their positional encoding mechanisms, it was consistently observed that the initial portions of input texts are disproportionately emphasized. This bias was attributed to input truncation during preprocessing, a common practice used to constrain inputs within the model's context window. In this paper, we extend the investigation of positional bias to the multimodal setting by empirically examining its existence, patterns, and the underlying causes or contributing factors.

Multimodal Representation Learning. Unlike language models such as BERT (Devlin et al. 2019) and LLaMA (Touvron et al. 2023) that are limited to processing textual input, multimodal models are capable of handling inputs from multiple modalities by projecting them into a shared embedding space, thereby enabling the capture of cross-modal relationships. A prominent and pioneering example is CLIP (Radford et al. 2021), which learns to align image and text representations through contrastive learning. Trained on 400 million image-text pairs, CLIP demonstrates strong performance across various downstream tasks, including zero-shot image classification and image-text retrieval. Despite its success, CLIP also has notable limitations. One major challenge lies in its reliance on contrastive loss, which necessitates large batch sizes to provide sufficient negative samples during training—resulting in significant computational overhead. To mitigate this issue, SigLIP (Zhai et al. 2023) was proposed, introducing a sigmoid-based binary classification loss that eliminates the need for large batch sizes. Ex-

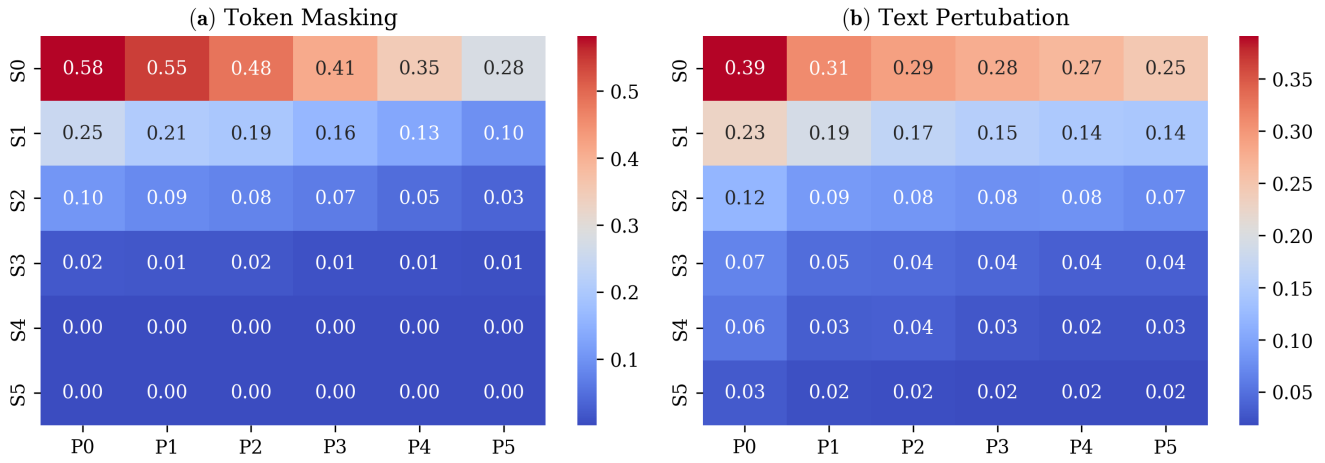


Figure 8: Comparison of text-to-image retrieval using token masking (left) vs. text perturbation (right). Each row shows a different segment shifted across positions; different rows correspond to different segments.

Model	**seg0**	**seg1**	**seg2**	**seg3**	**seg4**	**seg5**	**seg6**
Long-CLIP†	0.146	0.040	0.084	0.087	0.099	0.098	0.159
Shuffled Long-CLIP†	0.185	0.054	0.077	0.089	0.098	0.106	0.144
TULIP†	0.220	0.098	0.098	0.107	0.179	0.112	0.186
Long-CLIP‡	0.140	0.125	0.095	0.110	0.158	0.199	0.223
Shuffled Long-CLIP‡	0.155	0.148	0.102	0.116	0.157	0.190	0.226
TULIP‡	0.308	0.184	0.161	0.188	0.316	0.231	0.170
ViT-B/32	0.163	0.158	0.175	0.178	0.212	0.243	0.274
ViT-B/16	0.060	0.084	0.096	0.094	0.100	0.116	0.136
ViT-L/14	0.162	0.118	0.107	0.104	0.160	0.133	0.121
ViT-L/14–336	0.142	0.125	0.108	0.101	0.104	0.134	0.116
ResNet-50	0.152	0.100	0.110	0.117	0.131	0.157	0.211
SigLIP-Base	0.289	0.126	0.099	0.109	0.118	0.148	0.232

Table 1: Segment-wise variance coefficient across models. We report results for three models (Long-CLIP, Shuffled Long-CLIP, and TULIP) evaluated on two long-caption datasets (Urban1k† and DOCCI‡), as well as six short-caption models evaluated on COCO.

perimental results show that SigLIP achieves comparable or superior performance to CLIP on a range of vision-language tasks while simplifying the training process.

Despite the notable success of CLIP and its successor SigLIP, both models are trained predominantly on image-text pairs with relatively short textual descriptions. In practice, most CLIP variants constrain the input text length to 77 tokens. More notably, recent work has shown that only a small subset of these tokens are effectively utilized during training, with the number of well-trained tokens being fewer than 20 (Zhang et al. 2024). To address this limitation, Long-CLIP (Zhang et al. 2024) introduces a knowledge-preserving positional embedding stretching strategy, extending the token limit to 248 and fine-tuning the model on a dataset specifically curated for long image-text pairs (Chen et al. 2024). In contrast to such approaches that rely on absolute positional encoding, TULIP (Najdenkoska et al. 2024) adopts relative positional encoding, enabling the model to process text of arbitrary length. This design allows CLIP-like models to make full use of longer captions, thereby enhancing their ability to capture fine-grained and detailed semantic information embedded in extended textual contexts.

Other multimodal models, such as LLaVA (Liu et al. 2023) and BLIP-2 (Li et al. 2023), have gained prominence for enabling complex, instruction-based multimodal generation tasks like visual question answering and image captioning. However, in this work, we focus on CLIP and its variants, which are designed for multimodal representation learning through contrastive alignment of image and text embeddings. This focus is motivated by the fact that CLIP-style models frequently serve as the key backbone for more advanced multimodal systems, including those used in generative settings. In addition, cross-modal retrieval plays a crucial role in multimodal generation pipelines that incorporate Retrieval-Augmented Generation (RAG) techniques (Wu et al. 2024; Xia et al. 2024), where retrieval mechanisms are employed to select relevant visual or textual content as input to the generation module. As such, a deeper understanding of CLIP-like models in the context of cross-modal retrieval is critical for advancing the broader landscape of vision-language learning.

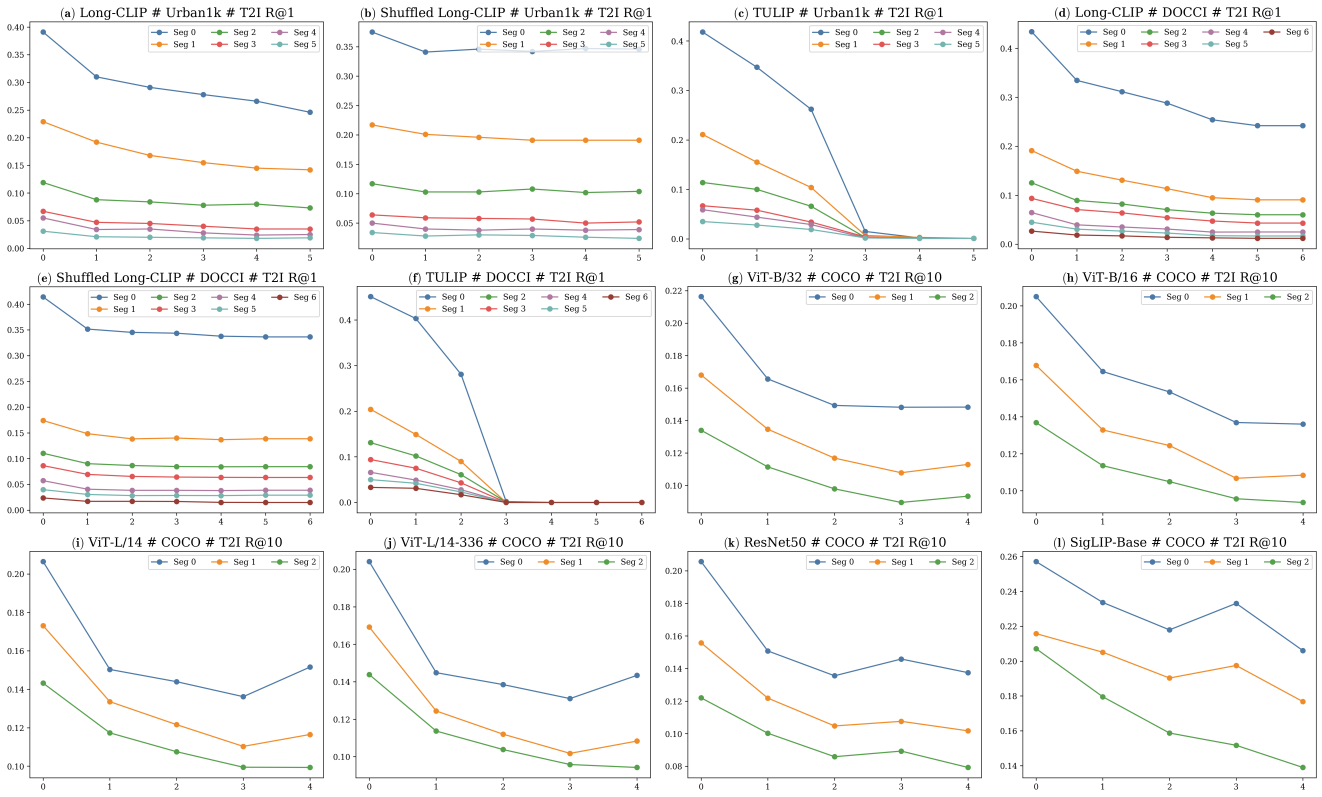


Figure 9: Positional bias analysis of text encoders across multimodal models and datasets using text perturbation strategy. Each figure title follows the format: model name # dataset # metric. The x-axis shows positions, and colors indicate different segments.

B Flowchart of Positional Bias Analysis

Figure 6 illustrates our approach for analyzing positional bias in multimodal models across both image and text modalities. For clarity, we show one example each for an image segment and a text split. In the image-side analysis (left), a single visual segment is isolated and shifted across spatial positions, with the caption fixed and all other image regions masked; retrieval scores are computed at each position. On the text side (right), a text split is selected and moved across different positions in the input text, while the image remains fixed and the rest of the text are masked. This process is repeated for all segments.

C More Experiments

C.1 Additional Experiments on Context Importance

We present additional context importance experiments across image and text modalities (Figure 7). For long-caption models, we include Long-CLIP, Shuffled Long-CLIP, and TULIP; for short-caption models, we evaluate CLIP with various backbones (ViT-B/32, ViT-B/16, ViT-L/14, ViT-L/14-336, ResNet-50) and SigLIP-Base. Experiments are conducted on the DOCCI and Urban1K datasets for long captions, and on COCO for short captions.

Across all settings, context importance trends remain consistent across datasets and architectures. For images, the central region is most important, with declining relevance toward the edges. For text, the first segment yields the highest retrieval accuracy, decreasing steadily in later segments.

C.2 Token Masking vs. Text Perturbation

We explore two strategies for analyzing positional bias in the text encoder: token masking and text perturbation. Figure 8 compares text-to-image retrieval performance under both methods on Urban1K using Long-CLIP. Both approaches yield similar trends across positions (x-axis), consistently revealing strong positional bias toward the beginning.

Furthermore, Figure 9 shows positional bias trends across different multimodal models and datasets using the text perturbation strategy, revealing patterns consistent with those observed in the token masking experiments.

C.3 Measurement of Positional Bias

To quantitatively assess positional bias, we use the coefficient of variation to capture how retrieval accuracy changes when the same segment is shifted across positions, with all other positions masked. Table 1 reports this metric for various models and datasets in our image-side analysis. While the values offer useful insights, our primary goal is to highlight the presence and patterns of positional bias. A more precise quantification and mitigation of this bias is left for future work.