

A TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

A.1 EXPERIMENTAL SETUP

All experiments were conducted on NVIDIA A100 GPUs using the official implementation of LLaVA series [Liu et al. \(2024a\)](#), with FlashAttention-2.6.3¹ integrated for efficient attention computation. The context length was set to 4096 tokens, and all generations were performed using greedy decoding with a fixed temperature of 0 to ensure deterministic outputs. Task definitions and groupings follow the standard taxonomy established by MILEBench [Song et al. \(2024\)](#), covering a diverse spectrum of 29 multimodal benchmarks. To accommodate the memory and sequence length variability across datasets, batch sizes were dynamically adjusted: a batch size of 1 was used for long-context datasets such as MMCoQA [Li et al. \(2022\)](#) and GPR1200 [Schall et al. \(2022\)](#), while a batch size of 24 was adopted for the remaining tasks. For tasks with highly imbalanced attention patterns, we applied kernel-based attention merging strategies with top-k region ratios calibrated per dataset. We further incorporated minor task-specific biases—for example, a fixed bias of 0.5 in EgocentricNavigation [Krantz et al. \(2020\)](#) and 1.5 in SlideVQA [Tanaka et al. \(2023\)](#)—while retaining default configurations elsewhere. All preprocessing and evaluation followed the official MILEBench protocol to ensure fair and reproducible comparisons.

Table 7: Detailed Tasks inherited from MILEBench [Song et al. \(2024\)](#).

Category	Task	Dataset	Data Source	Count	Metric
Temporal Multi-image	Action Understanding and Prediction (T-1)	Action Localization Action Prediction Action Sequence	STA Gao et al. (2017) STAR Wu et al. (2024) STAR Wu et al. (2024)	200	Accuracy
	Object and Scene Understanding (T-2)	Object Existence Object Interaction Moving Attribute Object Shuffle	CLEVER Yi et al. (2019) STAR Wu et al. (2024) CLEVER Yi et al. (2019) Perception Test Patraucean et al. (2024)	200	Accuracy
	Visual Navigation and Spatial Localization (T-3)	Egocentric Navigation Moving Direction	VLN-CE Krantz et al. (2020) CLEVER Yi et al. (2019)	200	Accuracy
	Counterfactual Reasoning and State Change (T-4)	Counterfactual Inference State Change Character Order Scene Transition	CLEVER Yi et al. (2019) Perception Test Patraucean et al. (2024) Perception Test Patraucean et al. (2024) MovieNet Huang et al. (2020)	200	Accuracy
Semantic Multi-image	Knowledge Grounded QA (S-1)	Webpage QA Textbook QA Complex Multimodal QA Long Text with Images QA	WebQA Chang et al. (2022) TQA Kembhavi et al. (2017) MultiModalQA Talmor et al. (2021) WikiVQA	200	Accuracy
	Text-Rich Images QA (S-2)	Slide QA Document QA	SlideVQA Tanaka et al. (2023) OCR-VQA Mishra et al. (2019) DocVQA Mithrew et al. (2021)	200	Accuracy
	Visual Relation Inference (S-3)	Visual Change Captioning Visual Relationship Expressing	Spot-the-Diff Jhamiani & Berry-Kirkpatrick (2018) CLEVR-Change Hosseinizadeh & Wang (2021)	200	ROUGE-L
	Dialogue (S-4)	Multimodal Dialogue Conversational Embodied Dialogue	MMCoQA Li et al. (2022) ALFRED Shridhar et al. (2020)	200	Accuracy
	Space Understanding (S-5)	nuScenes	nuScenes Caesar et al. (2020)	200	Accuracy
Diagnostic Evaluation	Needle In A Haystack (N-1)	Text Needle In A Haystack	TextNeedleInAHaystack	320	Accuracy
	Needle In A Haystack (N-2)	Image Needle In A Haystack	ImageNeedleInAHaystack	320	Accuracy
	Image Retrieval (I-1)	Image Retrieval	GPR1200 Schall et al. (2022)	600	Accuracy

A.2 RELATED WORK

With the success of decoder-only large language models (LLMs) [Achiam et al. \(2023\)](#); [Bai et al. \(2023\)](#); [Li et al. \(2025\)](#); [Touvron et al. \(2023\)](#); [Abdin et al. \(2024\)](#), recent advances have extended their capabilities to the multimodal domain, giving rise to Vision-Language Models (VLMs). Early frameworks such as LLaVA [Liu et al. \(2023\)](#), InternVL [Chen et al. \(2024b\)](#), and Qwen-VL [Bai et al. \(2025\)](#) demonstrate that instruction tuning can be adapted to handle flattened textual and visual inputs, enabling strong performance across tasks such as visual reasoning, captioning, and instruction following. Additionally, most multi-modality pre-trained work [Grattafiori et al. \(2024\)](#); [Li et al. \(2023; 2024\)](#); [Zou et al. \(2024\)](#); [Yang et al. \(2025\)](#) also inherit the causal masking design from LLMs, which, while crucial for token generation, may unnecessarily constrain specific modality token (e.g. In our paper it refers to visual tokens). These pre-trained models typically flatten visual and textual tokens into a single sequence and feed them into an decoder-only, which may overlooking modality-specific attention patterns. To mitigate these limitations, recent studies have explored resolution-aware vision encoders [Chai et al. \(2022\)](#); [Chen et al. \(2024a\)](#), multi-modal alignment modules [Singh et al. \(2022\)](#), and fine-grained token interaction strategies [Yao et al. \(2021\)](#), aiming to better adapt LLM-based decoder-only architecture to the visual perception reasoning. And the potential usage of the future tokens has been shown in LLMs architecture [Yin et al. \(2024\)](#). However, the impact of LLM-inherited causal masking on visual token processing remains underexplored and

¹<https://github.com/Dao-AILab/flash-attention>

despite its potential misalignment with the non-sequential nature of many visual reasoning tasks, which motivating the core investigation in our work.

A.3 DISTRIBUTION ANALYSIS FOR FUTURE-AWARE ATTENTION.

To better understand the limitations of causal masking in vision-language models (VLMs), we analyze the predictive uncertainty from an information-theoretic perspective, following the previous work [Yin et al. \(2024\)](#). In particular, we examine the mutual information between the model output and the observed context under different masking strategies. Let $\mathbf{x}^v = \{x_1^v, \dots, x_m^v\}$ and $\mathbf{x}^t = \{x_1^t, \dots, x_n^t\}$ denote the visual and textual tokens respectively, and let $\mathbf{X} = \mathbf{x}^v \oplus \mathbf{x}^t$ be the unified input sequence of total length $L = m + n$. The autoregressive model predicts the output token x_o based on a masked prefix $\mathbf{X}_{\leq i}$. The mutual information between the output and its visible prefix context is $I(\mathbf{X}_{\leq i}; x_o) = H(x_o) - H(x_o | \mathbf{X}_{\leq i})$, where $H(\cdot)$ denotes the entropy. Depending on the specific causal mask, the prefix $\mathbf{X}_{\leq i}$ may include different subsets of visual and textual tokens. For example,

$$I(x_o; \mathbf{x}_{1:i}^v \cup \mathbf{x}_{1:j}^t) = H(x_o) - H(x_o | \mathbf{x}_{1:i}^v, \mathbf{x}_{1:j}^t), \quad (15)$$

which isolates the contributions of each modality. As shown in our empirical study in Section [3](#) and [4](#), breaking the visual-based causal inference procedure by exposing future tokens leads to a distribution shift because of the rich semantic information in the masked future region.

Theoretical Properties. Based on the preceding information-theoretic derivations in [Yin et al. \(2024\)](#); [Yun et al. \(2019\)](#), and assuming the vision language models induce causally isotropic intermediate representations, we further derive the following properties of mutual information:

Property A.1. For any $i \leq L$, the mutual information between the output token x_o and the l -th layer intermediate representation $\omega_{\leq i}^{(l)}$ is upper-bounded by the mutual information from the raw prefix input $\mathbf{X}_{\leq i}$:

$$I(x_o; \omega_{\leq i}^{(l)}) \leq I(x_o; \mathbf{X}_{\leq i}).$$

This follows from the data processing inequality and reflects that internal representations cannot increase information about the target beyond what is available from the input.

Property A.2. If the VLM decoder is contextual, then its intermediate representation preserves all information in the input:

$$I(x_o; \omega_{\leq L}^{(l)}) = I(x_o; \mathbf{X}_{\leq L}).$$

This implies that the decoder faithfully encodes the entire causal context without losing predictive power.

Property A.3. If the input distribution is causally isotropic and $\omega_{\leq i}^{(l)}$ is uniquely determined by $\mathbf{X}_{\leq i}$, then the representation retains no more information than the original prefix:

$$I(x_o; \omega_{\leq i}^{(l)}) \leq I(x_o; \mathbf{X}_{\leq i}) \quad \text{for all } i \leq L.$$

This reinforces that isotropic settings do not amplify mutual information through intermediate computation.

Property A.4. If both the decoder is contextual and the data distribution is causally isotropic, then the mutual information is exactly preserved:

$$I(x_o; \omega_{\leq L}^{(l)}) = I(x_o; \mathbf{X}_{\leq L}).$$

This guarantees no loss of information between the raw prefix and the intermediate representation.

Property A.5 (Upper-Triangular Future Visibility in Multimodal Masking). For any visual query position $i \in \mathcal{V}$ and ground-truth output x_o , the ratio of mutual information satisfies:

$$\frac{I(\mathbf{X}_{\leq i}; x_o)}{I(\mathbf{X}_{\leq L}; x_o)} = \frac{H(x_o) - H(x_o | \mathbf{X}_{\leq i})}{H(x_o) - H(x_o | \mathbf{X}_{\leq L})} \geq \frac{I(\Omega_{\leq i}^{(l)}; x_o)}{I(\Omega_{\leq L}^{(l)}; x_o)}, \quad (16)$$

where $\Omega_{\leq i}^{(l)}$ represents the intermediate representation at layer l computed from prefix $\mathbf{X}_{\leq i}$. This inequality quantifies how future-aware visual masking contributes to reducing uncertainty of the output, and suggests that semantically rich upper-triangle access allows earlier layers to preserve more predictive information.

Property [A.5](#) establishes that, under a future-aware masking strategy, visual queries that access upper-triangular regions (i.e., future tokens) can retain a higher fraction of mutual information with the output token x_o compared to standard causal masking. This suggests that even partial access to semantically informative future tokens allows intermediate representations to encode more predictive context. The inequality further implies that the proportion of retained information in early layers is lower bounded by the proportion of information retained in their corresponding representations $\Omega^{(l)}$. In practice, this supports the design of selective future access in vision-language inference, where relaxing strict causality on visual queries can effectively enhance downstream prediction without fully compromising autoregressive generation.

A.4 FUTURE-AWARE FLASH-ATTENTION

To efficiently support our future-aware causal masking strategies defined in Section [3](#), we integrate them with the FlashAttention framework for scalable inference. As detailed in Algorithm [1](#), we implement our masking logic by applying the selected future-aware mask $\mu \in \{M^f, M^{v2v}, M^{v2t}\}$ directly into the attention score computation, replacing the standard causal mask. During runtime, both the queries and key-value pairs are processed in blocks to fit within on-chip memory, and attention scores are computed with fused softmax operations to ensure numerical stability and memory efficiency. The masked attention scores are exponentiated and normalized via a log-sum-exp trick, and aggregated token-wise to produce final outputs. This fusion enables our proposed vision-language attention design to retain the efficiency advantages of FlashAttention while supporting flexible, modality-aware causal constraints.

Algorithm 1 Future-Aware Mask equipped with FlashAttention

Require: Matrices $\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{M} \in \mathbb{R}^{L \times d}$, future-aware mask $\mu \in \{M^f, M^{v2v}, M^{v2t}\}$, block sizes B_r, B_c

- 1: Divide $\mathbf{Q}$ into $T_r = \lceil \frac{L}{B_r} \rceil$ blocks $\mathbf{Q}_1, \dots, \mathbf{Q}_{T_r}$.
- 2: Divide $\mathbf{K}, \mathbf{V}, \mu$ into $T_c = \lceil \frac{L}{B_c} \rceil$ blocks $\mathbf{K}_j, \mathbf{V}_j, \mu_j$ of size B_c each
- 3: Initialize output $\mathbf{O} \in \mathbb{R}^{L \times d}$ and logsumexp $\mathbf{L} \in \mathbb{R}^L$
- 4: **for** $i = 1$ to T_r **do**
- 5: Load $\mathbf{Q}_i$ into on-chip SRAM
- 6: Initialize $\mathbf{O}_i^{(0)} \leftarrow 0, \ell_i^{(0)} \leftarrow 0, \mathbf{m}_i^{(0)} \leftarrow -\infty$
- 7: **for** $j = 1$ to T_c **do**
- 8: Load $\mathbf{K}_j, \mathbf{V}_j, \mu_j$ into on-chip SRAM
- 9: Compute masked attention score:

$$\mathbf{S}_i^{(j)} = \mathbf{Q}_i \mathbf{K}_j^\top / \sqrt{d} + \mu_{i,j}$$

- 10: Normalize: $\tilde{\mathbf{S}}_i^{(j)} = \exp(\mathbf{S}_i^{(j)} - \mathbf{m}_i^{(j)})$
- 11: Update max: $\mathbf{m}_i^{(j)} = \max(\mathbf{m}_i^{(j-1)}, \max(\mathbf{S}_i^{(j)}, \dim = 1))$
- 12: Update sum: $\ell_i^{(j)} = \exp(\mathbf{m}_i^{(j-1)} - \mathbf{m}_i^{(j)}) \odot \ell_i^{(j-1)} + \sum \tilde{\mathbf{S}}_i^{(j)}$
- 13: Output partial result:

$$\mathbf{O}_i^{(j)} = \exp(\mathbf{m}_i^{(j-1)} - \mathbf{m}_i^{(j)}) \cdot \mathbf{O}_i^{(j-1)} + \tilde{\mathbf{S}}_i^{(j)} \cdot \mathbf{V}_j$$

- 14: Final output:

$$\mathbf{O}_i = \mathbf{O}_i^{(T_c)} / \ell_i^{(T_c)}$$

- 15: Logsumexp: $\mathbf{L}_i = \mathbf{m}_i^{(T_c)} + \log(\ell_i^{(T_c)})$
 - 16: Store $\mathbf{O}_i, \mathbf{L}_i$ to global memory
 - 17: **return** $\mathbf{O}, \mathbf{L}$
-

B THE USE OF LARGE LANGUAGE MODELS

In preparing this manuscript, we employed a Large Language Model (LLM) as a writing assistant to refine the clarity and presentation of the text. Its use was strictly confined to linguistic enhancement rather than substantive content generation. Specifically, the LLM was used to:

- Rephrase sentences and paragraphs for greater readability, conciseness, and academic formality.
- Correct grammar, spelling, and punctuation errors.
- Improve logical flow and transitions between sentences.