Rethinking Causal Mask Attention for Vision-Language Inference

Xiaohuan Pei[1] **Tao Huang**[2] **Yanxiang Ma**[1] **Chang Xu**[1]*

[1] The University of Sydney, Australia [2] Shanghai Jiao Tong University, China

ICLR 2026

Introduction & Motivation

Autoregressive VLMs inherit causal masking from LLMs, restricting every token to attend only to past positions. The standard causal attention is defined as:

$$A = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}} + M^c\right)V$$
$$M^c_{i,j} = 0 \text{ if } j \le i, \quad -\infty \text{ otherwise}$$

While essential for **text generation**, this constraint is **overly rigid for visual tokens**, which carry holistic spatial information and are inherently non-sequential.

Key Discovery:

- Breaking **visual** causal masks → **improves performance**
  (ROUGE-L: 15.29 → 15.55 on ALFRED)
- Breaking **textual** causal masks → **catastrophic collapse**
  (ROUGE-L: 15.29 → 0.11)

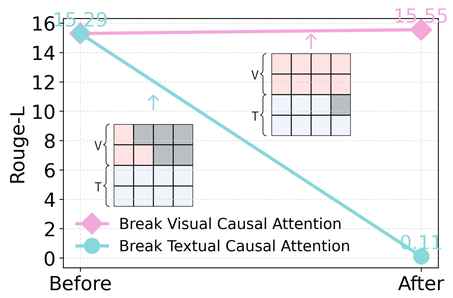

Figure 1: Breaking causal masks in LLaVA-7B (ALFRED). Visual mask relaxation improves scores; textual mask relaxation collapses output.

Research Questions

Q1: *Does causal attention from LLMs fit visual tokens in VLMs?*

Q1: *Does causal attention from LLMs fit visual tokens in VLMs?*

Q2: *How should causal attention be revised for multi-modal settings?*

Q3: *Which future tokens should vision tokens access?*

Q4: *How does pre-seen visual semantics impact vision- vs. text-tasks?*

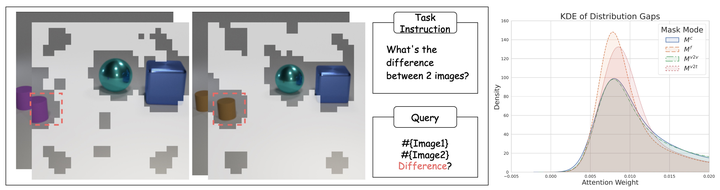

Figure 2: Visual relation task — M^{v2v} enables visual queries to compare across future visual tokens for change captioning and relation inference.

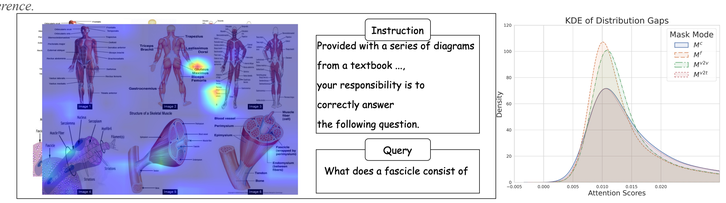

Figure 3: Text-rich VQA task — M^{v2t} lets visual queries preview future textual tokens, boosting OCR-VQA (+0.5) and TextVQA (+6.5).

Future-Aware Causal Masks

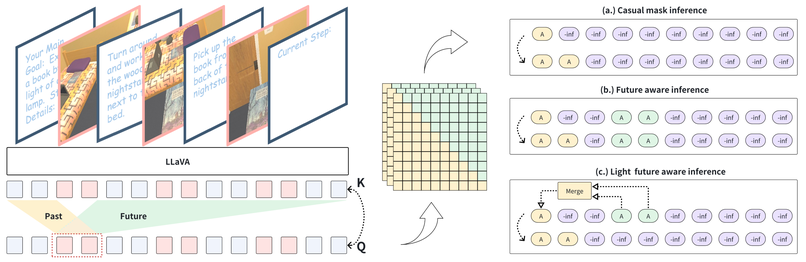

Figure 4: (a) Causal mask blocks future. (b) Future-aware lets vision tokens preview future. (c) Light variant compresses future into prefix.

$$M^f_{i,j}: \quad 0 \text{ if } j \le i \text{ or } (j > i \text{ and } i \in \mathcal{V})$$
$$M^{v2v}_{i,j}: \quad 0 \text{ if } j \le i \text{ or } (j > i \text{ and } i, j \in \mathcal{V})$$
$$M^{v2t}_{i,j}: \quad 0 \text{ if } j \le i \text{ or } (j > i \text{ and } i \in \mathcal{V}, j \in \mathcal{T})$$

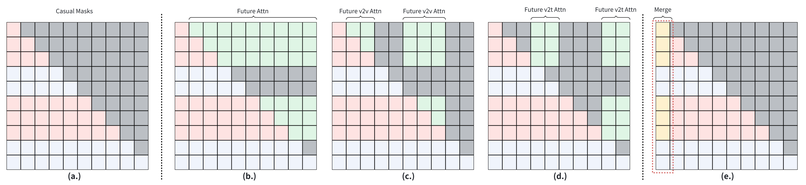

Figure 5: Attention mask designs — (a) Causal, (b) M^f full, (c) M^{v2v}, (d) M^{v2t}, (e) Merged lightweight.

Three Masking Strategies:

- **M^f (Full):** Vision queries attend to ALL future tokens. Best for temporal multi-image tasks.
- **M^{v2v} (V→V):** See only future visual tokens. Best for visual relation tasks.
- **M^{v2t} (V→T):** See only future text tokens. Best for text-rich VQA.

Light Future-Aware Attention

Compress future attention into **prefix tokens** during prefill via kernel pooling:

- 1D kernel pooling aggregates future visual attention scores
- Pooled scores merged into attention sink positions
- Standard causal structure fully preserved during decoding
- Even a single prefix token suffices to absorb future context

$$C(B,\mu)_{i,1} = \sum_{s=1}^{T-k+1} \max_{t=0}^{k-1} (B \odot M^p(\mu))_{i,\, i+s+t}$$

Kernel pooling aggregates future attention into prefix (sink) positions

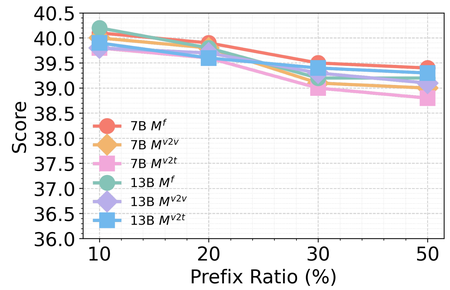

Figure 6: Prefix ratio effect — smaller ratios (fewer sink tokens) yield best results. Future semantics compress efficiently.

Key Results

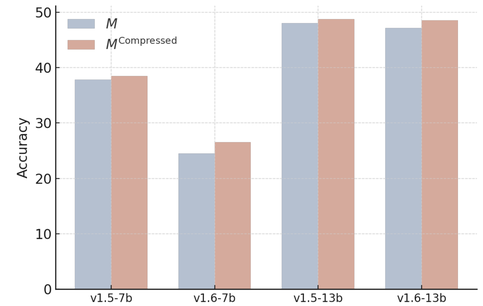

Figure 7: Compressed future-aware masks consistently outperform standard causal masks across all LLaVA model variants and sizes.

Findings (15 tasks, LLaVA-7B & 13B):

- **Temporal multi-image tasks** (navigation, action prediction) consistently benefit from future-aware masks
- **M^{v2v}** improves visual relation tasks; **M^{v2t}** improves text-rich QA
- Merge variants retain most gains with 2–3× decoding speedup
- Text-dominant tasks still require strict autoregressive masking
- Future semantics can be compressed into a single attention sink token

Decoding Latency Reduction:

```
M^f  :   83.18 → 26.54 ms/tok  (+merge)   3.1× faster
M^v2v:   64.13 → 26.40 ms/tok  (+merge)   2.4× faster
M^v2t:   43.04 → 26.11 ms/tok  (+merge)   1.6× faster
```

Conclusions

- **Standard causal masking** from LLMs **misaligns with non-sequential visual inputs** in VLMs
- **Standard causal masking** from LLMs **misaligns with non-sequential visual inputs** in VLMs
- **Selectively relaxing future masking** for vision tokens **improves reasoning** across 15 benchmarks
- **Lightweight merge** compresses future attention into prefix tokens with **2–3× speedup**
- **Optimal mask is task-dependent:** M^f for temporal, **M^{v2v} for visual, M^{v2t} for text**
- **A single sink token** suffices to absorb and propagate **future visual context** effectively

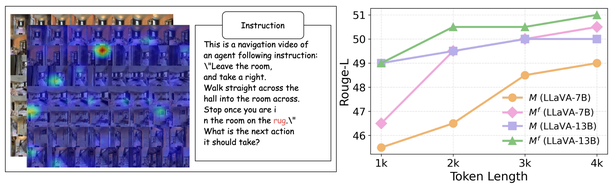

Figure 8: Temporal navigation task — M^f allows visual tokens to preview future frames for improved action prediction.

Key References:

[1] Liu et al. — LLaVA: Visual Instruction Tuning (NeurIPS 2023)
[2] Song et al. — MILEBench (NeurIPS 2024)
[3] Yin et al. — StableMask (2024)
[4] Qi et al. — Beyond Position: Vision Tokens in LLMs (2025)