On the Nature of Attention Sink that Shapes Decoding Strategy in Omni-LLMs

Suho Yoo[1*] **Youngjoon Jang**[1,2*] **Joon Son Chung**[1]
[1]KAIST [2]VGG, University of Oxford
suho.yoo@kaist.ac.kr

https://github.com/kaistmm/Outro

Abstract

The goal of this paper is to strengthen the reasoning of Omnimodal Large Language Models (Omni-LLMs) at inference time, without additional training. These models jointly process video, audio, and text, and given the large number of tokens they consume, how attention is routed across them is central to their behaviour. We focus specifically on *attention sinks*, tokens that absorb a disproportionate share of attention mass regardless of their semantic content, to understand how this routing unfolds. To this end, we conduct a systematic analysis of sink behaviour in Omni-LLMs. Our analysis yields two key findings: (i) high sink attention does not solely indicate head redundancy, suggesting that sink value representations play additional functional roles; (ii) the sink value vector acts as a shared bias added to every token's output, serving as a global signal that organises the representation as a whole. Building on this, we propose ***OutRo***, which correspondingly aligns non-sink token representations with the sink in feature space, and relaxes the causal mask for sink tokens at an early layer to sharpen this bias before the rest of decoding proceeds. This design enhances the reasoning process without requiring additional forward passes or access to attention maps. Based on extensive experiments, OutRo consistently improves performance on seven video QA benchmarks and demonstrates strong generalisation, while incurring only a $1.1\times$ decoding overhead.

1 Introduction

Large language models (LLMs) [1, 2, 5, 12, 37, 60] and their multimodal extensions have demonstrated transformative performance across diverse tasks, driven by the scaling of transformer architectures [3, 4, 10, 22, 31, 65]. Among these, Omnimodal Large Language Models (Omni-LLMs) [24, 48, 50, 58, 59], which jointly process video, audio, and text, present a particularly challenging setting, as modality tokens from video frames and audio segments constitute the vast majority of the input sequence. At the core of these models lies the attention mechanism [52], which allows models to dynamically prioritise relevant information; yet how Omni-LLMs handle the large volume of modality tokens that dominate their inputs remains an open question.

Among the diverse behaviours observed in the attention mechanism [6, 15, 39, 53], the *attention sink* phenomenon has received increasing attention [57]: models tend to allocate a disproportionate share of attention to semantically uninformative tokens, such as special or punctuation tokens in LLMs (*e.g.,* '<BOS>', '.', '\n') [18, 46, 62] and background or weakly relevant patches in vision models [15, 25, 30, 38]. Prior work is divided: some argue these tokens merely stabilise attention distributions and can be pruned or gated for efficiency [18, 21, 40, 43], while others show they carry structured representations beneficial to downstream tasks [15, 25, 38, 46]. In this work, we revisit these two views in the context of Omni-LLMs with complex multimodal inputs.

[*] Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

In this work, we present the first analysis of attention sinks in Omni-LLMs. While prior work in language models has largely focused on attention scores and token norms to characterise sink behaviour, we shift the focus to the representations of sink tokens themselves and their role during decoding. To this end, we formulate three fundamental questions:

Q1. Which criteria are appropriate for identifying sink tokens in Omni-LLMs settings?
Q2. Whether attention sinks solely serve as redundant heads?
Q3. Whether sink token representations act as a global signal across token outputs?

For **Q1**, we examine how sink tokens should be identified in Omni-LLMs by comparing criteria proposed for LLMs [46] and vision-language models (VLMs) [30]. We find that the VLM-based criterion classifies an excessively large fraction of tokens as sinks, including those corresponding to semantically meaningful visual objects. In contrast, the LLM-based criterion identifies only a small number of sink tokens and assigns sink behaviour to structurally defined elements such as the initial token, yielding a more stable and interpretable definition. Building on this criterion, we address **Q2** and **Q3**. For **Q2**, we analyse the functional role of attention heads that attend heavily to sink tokens. While prior work suggests that pruning such heads preserves model performance, implying redundancy [18, 20, 21, 40, 43], our head-by-head ablation experiments reveal no consistent trend: pruning a sink head may either improve or degrade performance depending on the case, indicating that sink heads cannot be characterised solely as inactive components. For **Q3**, we revisit the attention output decomposition in Omni-LLMs. Since most tokens attend strongly to the sink, its value vector acts as a shared bias term added to the attention output across all tokens [46], meaning the direction encoded in the sink value shapes the decoding process throughout the network.

Based on these analyses, we propose ***OutRo***, a novel inference-time decoding method for Omni-LLMs that leverages sink representations to guide token interactions during decoding. OutRo operates through two complementary mechanisms. First, an adaptive geometric transformation rotates the outputs of attention heads at non-sink positions toward the sink value direction, explicitly aligning token representations with the shared bias without requiring additional attention to the sink. Second, it strengthens this global signal by selectively relaxing the causal mask on sink tokens, allowing their representations to better aggregate context from other tokens. Importantly, OutRo operates at the transformer layer level without requiring explicit access to attention maps, avoiding the memory and computational overhead of prior approaches that rely on attention manipulation [23, 30, 55, 62, 68] or multiple forward passes [9, 28, 33, 51, 54]. We evaluate OutRo on seven video QA benchmarks, where it consistently improves performance with only modest decoding overhead, demonstrating that exploiting sink representations is a practical, training-free strategy for improving Omni-LLMs.

2 Related Works

Attention sinks in transformer models refer to a phenomenon where a small set of tokens attract disproportionate attention mass despite carrying limited semantic content [20, 41, 57, 70], commonly attributed to the non-negative normalisation imposed by softmax [20, 46, 57]. Two contrasting views have emerged on their role. The first holds that sink tokens serve a purely structural role. In LLMs, since sink value representations have near-zero norms, heads attending to them produce outputs of small magnitude and are considered inactive [18, 21, 43], motivating pruning and gating strategies that remove sink-dominated heads [40, 43, 45]. In VLMs, visual sink tokens are similarly treated as uninformative patches, and redistributing their attention to more relevant regions has been shown to improve performance [26, 30, 49]. The second view suggests that sink tokens carry meaningful representations. In ViTs, high-norm sink tokens aggregate global information when full attention is available [15, 25]. In vision-language models, whether a visual sink originates from the ViT or emerges within the LLM determines what it encodes, with ViT-propagated sinks capturing coarse-grained global context [38]. In LLMs, sink tokens act as implicit bias terms in the attention output, injecting a shared direction across all token representations [46].

Inference-time methods for multimodal LLMs (MLLMs) focus on enhancing multimodal reasoning and robustness in a training-free manner. One array of work builds upon contrastive decoding paradigms, which improve predictions by contrasting outputs obtained under different conditions, such as perturbed inputs, alternative decoding paths, or representations from different model layers [9, 28, 33, 34, 51, 67]. These methods suppress hallucinations and spurious correlations by comparing biased and unbiased predictions. Another range of work treats attention sinks as explicit intervention targets at inference time. By analysing attention maps, these methods identify

sink tokens and suppress their influence by redistributing attention mass towards more informative tokens [26, 29, 30, 49, 62]. However, these strategies incur overhead: contrastive decoding requires multiple forward passes, while sink-aware methods rely on attention map access or activation manipulation, thus not tailored to optimised attention implementations [13, 14, 44]. Our work addresses this gap by proposing a modality-agnostic, training-free intervention for Omni-LLMs that operates at the level of attention head outputs, without relying on attention maps or repeated inference.

3 On the Nature of Attention Sink

Unless otherwise specified, all experiments in this section are conducted on Qwen2.5-Omni [58].

Preliminaries of Omni-LLMs. Omni-LLMs process a video input $\mathcal{V}$, an audio stream $\mathcal{A}$, and a textual instruction $\mathcal{L}$. Modality encoders transform these into sequences of embeddings: $\mathcal{V}$ and $\mathcal{A}$ are projected into M modality tokens $\mathbf{X}_m \in \mathbb{R}^{M\times D}$ encapsulating visual and auditory features, which are concatenated with L textual tokens $\mathbf{X}_l \in \mathbb{R}^{L\times D}$ from a language model tokenizer to form the input:

$$\mathbf{X} = [\mathbf{X}_m; \mathbf{X}_l] \in \mathbb{R}^{N\times D}, \quad N = M + L, \tag{1}$$

where D denotes the hidden dimension of the transformer.

Multi-head attention. Each transformer layer in Omni-LLMs contains a multi-head attention (MHA) module. For each head $h \in \{1,...,H\}$, the input $\mathbf{X}$ is projected into queries $\mathbf{Q}_h$, keys $\mathbf{K}_h$, and values $\mathbf{V}_h$ via weight matrices $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{D\times D_h}$. A causal mask $\mathcal{M}$ enforces autoregressive generation by setting $\mathcal{M}_{ij} = -\infty$ for $j > i$. The attention distribution for head h is then defined as:

$$A_{h,i,:} = \text{Softmax}\left(\frac{\mathbf{Q}_{h,i}\mathbf{K}_h^\top}{\sqrt{D_h}} + \mathcal{M}_{i,:}\right), \tag{2}$$

and the output representation for each head is computed as:

$$\mathbf{O}_h = \mathbf{A}_h\mathbf{V}_h, \text{where } \mathbf{O}_{h,i} = \sum_{j=1}^{i} A_{h,i,j}\mathbf{v}_{h,j}. \tag{3}$$

The final representation of the layer is obtained by concatenating all head outputs and applying a linear projection $\mathbf{W}_O$: $\mathbf{O}_{final} = [\mathbf{O}_1; \mathbf{O}_2; ...; \mathbf{O}_H]\mathbf{W}_O$.

Q1. Which criteria are appropriate for identifying sink tokens in Omni-LLMs settings?

The criterion originates from analyses on LLMs, where sink phenomena were characterized through *massive activations* [46] in hidden states across layers. Specifically, given the set of token hidden states in a layer $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ with $\mathbf{x}_i \in \mathbb{R}^{1\times D}$, an activation scalar $z = \mathbf{x}_i[d]$ for some $i \in \{1,...,N\}$ and $d \in \{1,...,D\}$ is regarded as massive if

$$\Phi_{\text{LLM}}(\mathbf{x}_i) = \max_{d\in\{1,...,D\}} |\mathbf{x}_i[d]| > \max\big(100, 1000 \times \text{median}_{\mathbf{z}\in\mathbf{X}}(|\mathbf{z}|)\big). \tag{4}$$

More recently, this notion was extended to *visual attention sink* by examining sink dimensions inherited from LLMs [30]. Let D_{sink} denote a fixed set of dimensions known to exhibit sink behaviour in the underlying LLM. For a hidden state $\mathbf{x}_i \in \mathbb{R}^{1\times D}$, the sink dimension value was defined as:

$$\Phi_{\text{VLM}}(\mathbf{x}_i) = \max_{d_s\in D_{\text{sink}}} \left| \frac{\mathbf{x}_i[d_s]}{\sqrt{\frac{1}{D}\sum_{d=1}^{D}\mathbf{x}_i[d]^2}} \right| \geq \tau = 20. \tag{5}$$

Observation. Fig. 1a reveals a striking degeneracy when the visual attention sink criterion Φ_{VLM} is applied to Omni-LLMs. As depth increases, the majority of tokens are classified as sinks, eventually encompassing even semantically meaningful ones. This suggests that Φ_{VLM} fails to discriminatively characterise sink tokens in Omni-LLM settings: rather than isolating a sparse subset of anomalous tokens, the criterion collapses, marking the bulk of representations as sinks (See Sec. A.3 of our appendix for more examples).

Table 1: **Sink/outlier dimensions.** Sink dimensions strongly overlap with outlier feature patterns.

	Dims
Sink	458, 2570, ...
Outlier	**458**, **2570**, 3206, 3281, 1427, ...

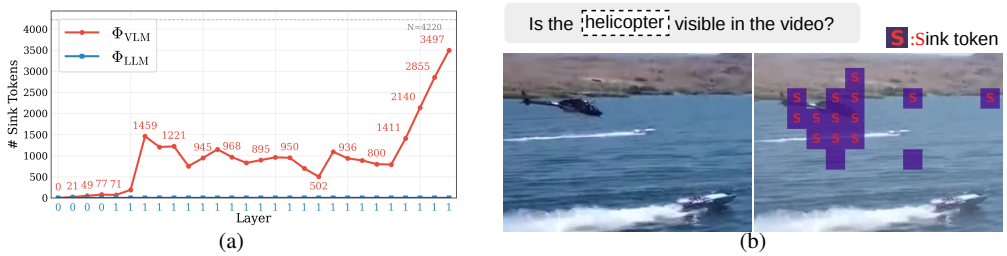


Figure 1: **Φ_{VLM} incorrectly identifies semantic tokens as sinks. (a)** Φ_{VLM} rapidly over-identifies sink tokens in deeper layers, unlike the sparse behaviour of Φ_{LLM}. **(b)** For the query "Is the *helicopter* visible?", highly attended object tokens are still identified as sinks under Φ_{VLM}.

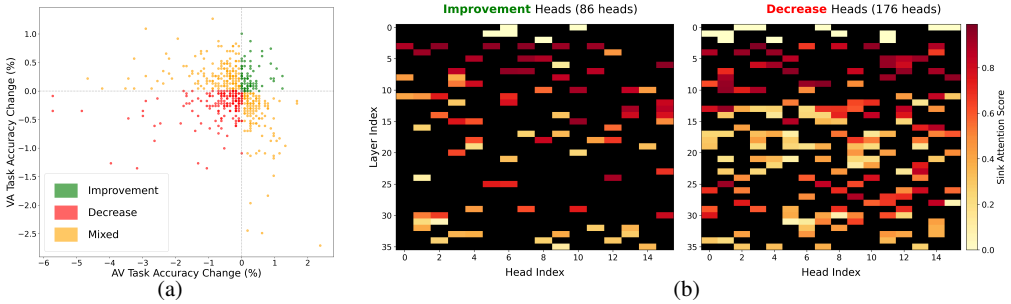


Figure 2: **High sink attention does not imply head redundancy. (a)** Performance change after removing individual heads, distinguishing improvement and decrease heads. **(b)** High sink scores appear in both improvement and decrease heads, indicating sink-heavy heads are not always redundant.

To investigate this discrepancy, Tab. 1 compares dimensions identified via the outlier feature criterion [16, 46] with the sink dimensions D_{sink} in Eq. (5), defined over the 3584-dimensional hidden space. An activation is treated as an outlier if its magnitude exceeds 6.0 in more than 25% of layers and 6% of tokens across at least 90 out of 100 evaluation sequences. All dimensions in D_{sink} are also identified as outliers, suggesting that sink dimensions largely correspond to norm-amplifying patterns. In contrast, Φ_{LLM} from Eq. (4) yields a sparse and stable set of sinks, typically one dominant token per layer, shown to be critical for model performance [57, 62], indicating it provides a more robust criterion for Omni-LLMs. Detailed results are provided in Secs. A.1 and A.2 of our appendix.

Overall, these findings demonstrate that Φ_{VLM} does not generalise reliably to Omni-LLM settings. To obtain a more general identification of sink behaviour, we adopt Φ_{LLM} for subsequent analyses.

Q2. Whether attention sinks solely serve as redundant heads?

A prevailing interpretation of attention sinks is that they are functionally negligible due to the near-zero norms of their value representations [18, 21, 40]. Under this view, even if a head assigns substantial attention to sink tokens, the resulting head output remains relatively small, and the head is therefore considered effectively suppressed. This perspective has motivated pruning strategies that remove sink-dominated heads under the assumption that they are redundant.

Observation. Prior work has explored pruning strategies based on sink attention scores [43, 45], under the assumption that heads attending heavily to sink tokens are largely redundant. However, our head-by-head ablation experiments on Omni-LLMs reveal a different picture. We evaluate on AVHBench [47], which comprises an audio-driven video hallucination subset ($A \rightarrow V$) and a video-driven audio hallucination subset ($V \rightarrow A$). As shown in Fig. 2, removing certain sink heads leads to marked performance improvements, while removing others causes clear degradation, with no consistent trend across cases. Tab. 2 further confirms this: heads with comparably high sink attention scores appear in both groups yet yield opposite performance impacts. These results demonstrate that high sink attention alone is insufficient to determine head importance, suggesting that sink value representations encode functional information beyond mere attention absorption.

Table 2: **Head ablation performance.** Heads with similar sink scores can yield opposite performance changes.

Head	Sink score	$A \rightarrow V$	$V \rightarrow A$	Overall
L15H3	0.7553	+0.88%	+0.48%	+1.36%
L18H12	0.6633	+0.26%	+0.87%	+1.14%
L11H6	0.7003	+0.62%	+0.09%	+0.70%
L24H10	0.7755	-3.43%	-0.48%	-3.91%
L16H14	0.8773	-2.55%	-0.13%	-2.68%
L14H12	0.7497	-1.32%	-0.44%	-1.76%

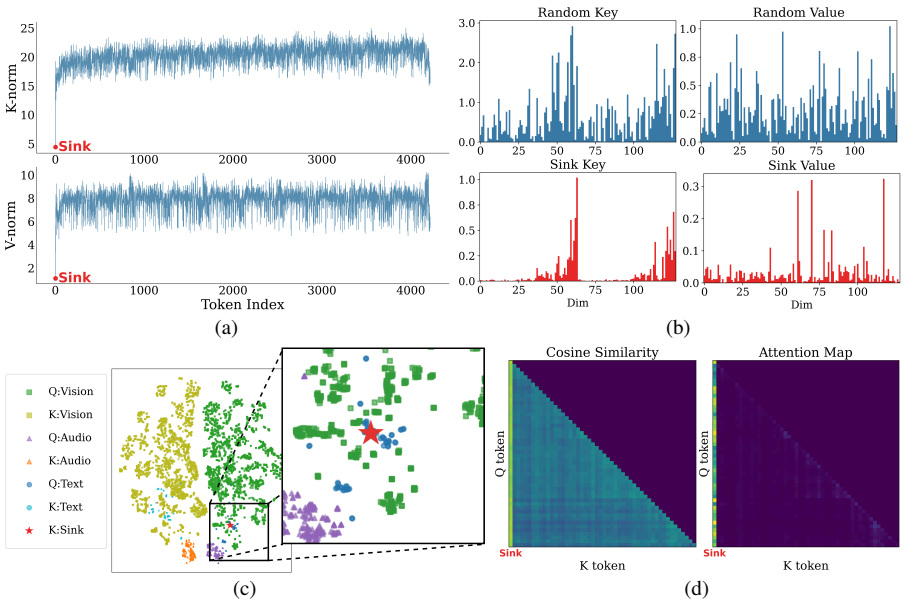


Figure 3: **Geometric analysis of sink token representations.** **(a)** Sink keys and values have substantially smaller norms than other tokens, yet **(b)** their activations are not uniformly small across dimensions. **(c)** In t-SNE, sink keys cluster near the query distribution, and **(d)** their cosine similarity with queries remains high, yielding large attention scores.

Q3. Whether sink token representations act as a global signal across token outputs?

While **Q2** suggests that small-magnitude sink value representations do not imply redundancy, we make this explicit by analysing sink token representations in the feature space. Noting that sink key representations similarly exhibit near-zero norms, we first examine whether such representations can nonetheless carry meaningful structure, before turning to their value counterparts.

Observation. As shown in Fig. 3a, sink key and value representations exhibit smaller norms than other tokens, yet remain structured rather than uniformly suppressed (Fig. 3b). Moreover, sink keys are geometrically aligned with query representations, showing high cosine similarity and attention scores despite their reduced magnitudes (Figs. 3c and 3d). These observations suggest that low norm does not imply irrelevance. To further validate this, we examine the internal structure of sink key activations, which are not uniformly suppressed (Fig. 3b). We hypothesise that these dimensions drive sink behaviour, such that zeroing them should disrupt the characteristic attention pattern. We test this by zeroing out the top-magnitude dimensions in sink key representations, which we term Zero-K. As shown in Figs. 4 and 5, zeroing only the top-1 dimension substantially eliminates the attention sink pattern, with attention to sink positions becoming consistently suppressed across layers.

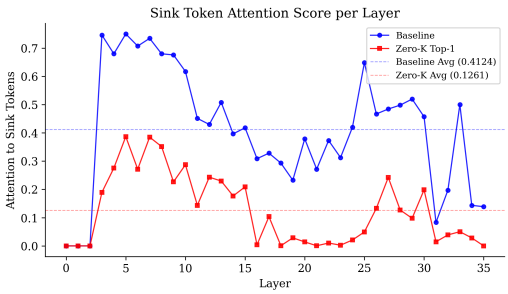


Figure 4: **Sink token attention score.** Compared to the baseline, zeroing the top-1 dimension in the sink key representations substantially reduces attention to sink tokens across layers.

Notably, although Zero-K effectively removes the sink attention pattern, it does not improve performance; as reported in Tab. 3, removing dominant sink dimensions consistently leads to performance degradation. These results demonstrate that sink key representations, despite their low magnitudes, carry meaningful information that shapes the attention mechanism.

Hypothesis. These findings raise the question of what information is carried by sink value representations. Following the attention output decomposition of [46], since most tokens attend strongly to the sink, the attention output at each token can be decomposed as:

$$\mathbf{O}_{h,i} = \sum_{j \in \mathcal{S}} A_{h,i,j} \mathbf{v}_{h,j} + \sum_{j \notin \mathcal{S}} A_{h,i,j} \mathbf{v}_{h,j}, \tag{6}$$

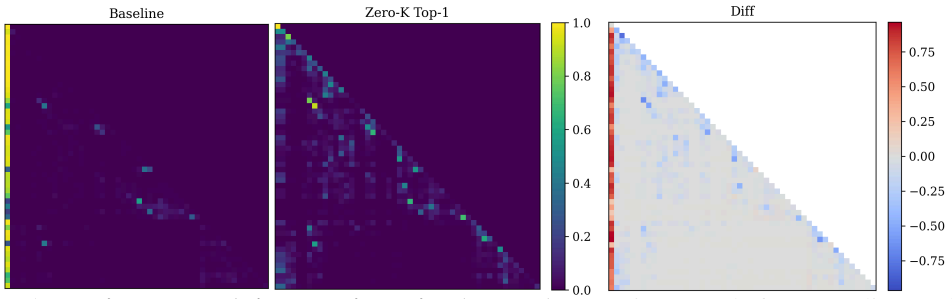
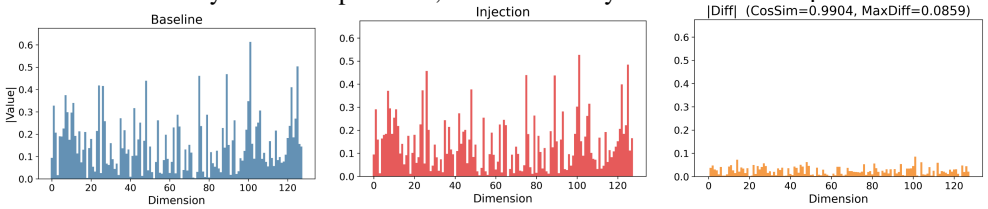


Figure 5: **Zeroing the top-1 sink key dimension (Zero-K).** Attention maps before (Baseline) and after (Zero-K Top-1) the intervention show that suppressing the dominant sink key dimension effectively redirects attention away from sink positions, as confirmed by the difference map.



(a) (b) (c)

Figure 6: **Sink token head outputs before and after injection.** The outputs remain nearly identical, with only minimal differences observed at the last layer.

Table 3: **Zero-K performance.** Removing dominant sink dimensions degrades performance on AVUT [61] and DailyOmni [69] across both Qwen2.5-Omni and SALMONN2+, suggesting that sink key representations carry functional information that shapes the attention mechanism.

Acc.↑	Qwen2.5-Omni [58]		SALMONN2+ [48]	
	AVUT	DailyOmni	AVUT	DailyOmni
Baseline	65.79	54.14	56.73	51.71
Top-1	61.31	48.62	51.07	37.76
Top-5	59.94	47.28	40.16	25.98
Top-10	58.87	45.45	37.82	25.90

Table 4: **Information injection** into sink tokens leads to consistent gains.

Acc.↑	Qwen2.5-Omni		SALMONN2+	
	AVH	DailyOmni	AVH	DailyOmni
Baseline	70.91	55.38	62.07	51.71
Injection	**71.08**	**55.47**	**62.20**	**51.96**

Table 5: **Head output rotation** toward the sink value direction improves performance.

AVHBench	$A \to V$	$V \to A$	Matching
Baseline	80.11	75.41	59.87
Rotation	**80.81**	**75.72**	**60.50**

where the first term, the value update from sink positions, acts as a shared bias term added uniformly across all token outputs. We hypothesise that this bias encodes a structured direction that sink-attending heads align with and propagate across layers, and examine this from two perspectives: the role of sink positions themselves, and the effect of aligning non-sink positions with the sink direction.

For sink positions, we examine whether explicitly enriching sink representations sharpens this bias direction by relaxing causal masking from layers where sink behaviour first emerges through to the second-to-last layer, allowing contextual representations to flow into sink positions. As shown in Figs. 6a to 6c, the resulting representations are nearly identical to the original, indicating that the bias direction is already stably encoded prior to injection. Nevertheless, Tab. 4 shows that enriching sink representations yields consistent performance gains, suggesting that even small refinements to the bias direction are beneficial. **For non-sink positions**, we rotate head outputs toward the sink value direction (i.e., projecting each head output onto the sink value vector, adding the projected component, and rescaling to preserve the original magnitude) and observe consistent performance improvements (Tab. 5). As shown in Fig. 7, rotation reduces attention allocated to sink tokens, with the freed attention redistributing toward modality tokens. Notably, this redistribution is not uniform: attention shifts preferentially toward the modality that the model originally attended to

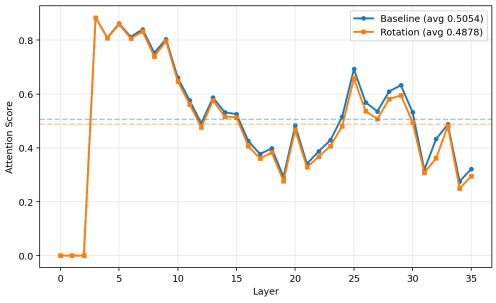


Figure 7: **Attention to sink tokens before and after rotation.** Rotation consistently reduces the total attention allocated to sink tokens.

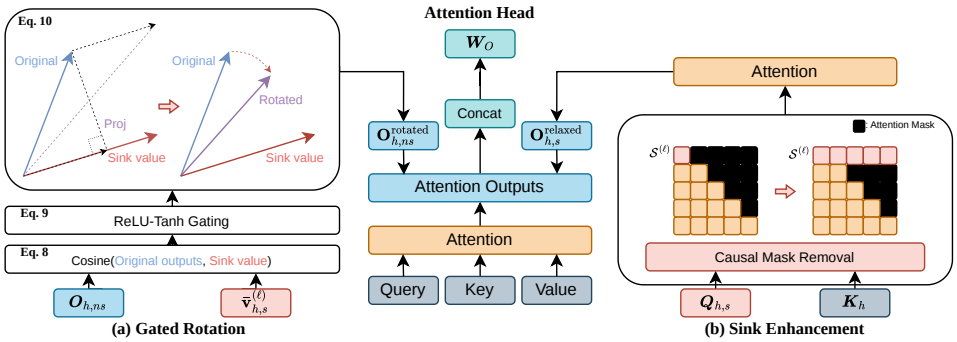


Figure 8: **Overall OutRo pipeline. (a)** Gated head output rotation. Non-sink outputs are rotated toward the sink value direction via adaptive gating. **(b)** Sink enhancement. The causal mask is removed for sink positions, sharpening the shared bias direction.

less (see Sec. B.2 of our appendix for details), suggesting that by explicitly providing the sink bias to non-sink positions, the model no longer needs to attend to the sink to obtain it, and the released attention capacity is used to better balance across modalities.

4 Method

Based on our analysis, we introduce ***OutRo***, a unified inference-time modulation method that combines (i) ReLU-tanh gating to align non-sink representations with the sink bias direction and (ii) sink information enhancement via one-time mask relaxation. The method operates without parameter updates and is implemented via lightweight forward interventions on attention head outputs.

Sink value direction. At each layer ℓ, sink token indices $\mathcal{S}^{(\ell)}$ are identified at inference time using Φ_{LLM} (Eq. (4)). For each head h, the value representations at these positions are averaged as:

$$\bar{\mathbf{v}}^{(\ell)}_{h,\mathcal{S}} = \frac{1}{|\mathcal{S}^{(\ell)}|} \sum_{s \in \mathcal{S}^{(\ell)}} \mathbf{v}^{(\ell)}_{h,s}. \tag{7}$$

ReLU–tanh gated head output rotation. Our finding that aligning head outputs with the sink direction improves performance (Tab. 5) motivates a gated rotation mechanism. For efficiency, we adopt a directional-alignment proxy (see Sec. B.1 of our appendix) for the sink score. Specifically, we compute, for each head h and non-sink position $ns \notin \mathcal{S}^{(\ell)}$, the directional alignment between the head output $\mathbf{O}_{h,ns}$ and the sink value direction:

$$c_{h,ns} = \cos\left(\mathbf{O}_{h,ns}, \bar{\mathbf{v}}^{(\ell)}_{h,\mathcal{S}}\right). \tag{8}$$

To control the rotation strength, we convert this alignment into a soft gate:

$$g_{h,ns} = \tanh\left(\frac{\text{ReLU}(c_{h,ns})}{t}\right), \quad \text{where } t = 0.1. \tag{9}$$

We then rotate the head output by adding a gated projection onto sink direction:

$$\hat{\mathbf{O}}_{h,ns} = \mathbf{O}_{h,ns} + \gamma g_{h,ns} \frac{\mathbf{O}_{h,ns} \cdot \bar{\mathbf{v}}^{(\ell)}_{h,\mathcal{S}}}{\|\bar{\mathbf{v}}^{(\ell)}_{h,\mathcal{S}}\|^2} \bar{\mathbf{v}}^{(\ell)}_{h,\mathcal{S}}, \qquad \mathbf{O}^{\text{rotated}}_{h,ns} = \frac{\|\mathbf{O}_{h,ns}\|}{\|\hat{\mathbf{O}}_{h,ns}\|} \hat{\mathbf{O}}_{h,ns}, \tag{10}$$

where $\gamma > 0$ controls the rotation strength. This gating selectively and adaptively strengthens heads that are strongly aligned with the sink value direction.

Sink information enhancement via mask relaxation. Our analysis in Sec. 3 showed that sink tokens already encode the bias direction stably, and that enriching this via causal mask relaxation yields consistent improvements (Tab. 4). We implement this by relaxing the causal mask at a selected layer ℓ_{enh}. At layer ℓ_{enh}, attention for sink positions $s \in \mathcal{S}^{(\ell_{\text{enh}})}$ is computed without the causal constraint. Concretely, for sink positions, we remove the causal mask and compute

$$\mathbf{A}^{\text{relaxed}}_{h,s,:} = \text{Softmax}\left(\frac{\mathbf{Q}_{h,s}\mathbf{K}_h^\top}{\sqrt{D_h}}\right), \tag{11}$$

Table 6: **Performance comparison of *OutRo*.** OutRo consistently improves performance across models and benchmarks, demonstrating its generality.

Models	Audio-Visual QA				Visual-Only QA		
	OmniBench	AVUT	AVHBench	DailyOmni	Holmes	MME	ActivityNet
Qwen2.5-Omni$_{3B}$	41.33	61.79	70.91	55.38	43.54	58.00	41.82
+OutRo	**42.03**	**62.87**	**71.67**	**55.56**	43.54	**58.11**	**41.86**
Qwen2.5-Omni$_{7B}$	47.37	65.79	71.60	54.14	47.46	62.44	43.59
+OutRo	**48.25**	**66.57**	**73.78**	**54.64**	**47.63**	**62.96**	**44.01**
video-SALMONN2+$_{3B}$	36.69	56.73	62.07	51.71	42.68	73.67	44.31
+OutRo	**36.95**	**57.31**	**62.33**	**52.05**	**42.90**	**73.89**	**44.41**

allowing sink queries to attend to the full sequence. The relaxed attention is then used to compute the corresponding head outputs

$$\mathbf{O}_{h,s}^{\text{relaxed}} = \mathbf{A}_{h,s,:}^{\text{relaxed}} \mathbf{V}_h, \tag{12}$$

which replace the masked outputs at sink positions.

Together, these components allow OutRo to improve model performance at inference time without modifying attention maps or requiring iterative forward passes. Importantly, OutRo is fully compatible with optimised attention implementations such as FlashAttention[2], making it a practical and scalable inference-time strategy across diverse multimodal configurations.

5 Experiments

5.1 Experimental Setup

Baselines. We evaluate on representative Omni-LLMs, including Qwen2.5-Omni [58] and video-SALMONN2+ [48]. All models are evaluated using greedy decoding.

Datasets and metrics. We evaluate OutRo on seven video QA benchmarks across audio-visual and visual-only settings. For audio-visual QA, we report results on OmniBench [36], AVUT [61], AVHBench [47], and DailyOmni [69], covering cross-modal consistency, hallucination detection, and audio-visual reasoning. For visual-only QA, we evaluate on VideoHolmes [7], VideoMME (medium) [17], and ActivityNetQA [63], focusing on temporal reasoning and event understanding. Accuracy is reported on all benchmarks.

Implementation details. For gated rotation, we search the rotation strength $\gamma \in [0.5, 4.0]$ and fix a single value per model: $\gamma = 0.5$ for video-SALMONN2+ and $\gamma = 3.0$ for Qwen2.5-Omni. Rotation is applied to non-sink positions at all layers where sink tokens are present, except the final few layers. Sink information enhancement is applied once at approximately one-seventh of total depth: layer 4 for Qwen2.5-Omni$_{7B}$ (28 layers), and layer 5 for Qwen2.5-Omni$_{3B}$ and video-SALMONN2+ (36 layers). All experiments use a single NVIDIA RTX A6000 GPU.

5.2 Experimental Results

Results. Tab. 6 reports results on seven video QA. Across Qwen2.5-Omni and video-SALMONN2+, OutRo consistently improves performance over base models on both audio-visual and visual-only QA benchmarks. For Qwen2.5-Omni$_{7B}$, OutRo achieves the largest gain on AVHBench (+2.18), which explicitly evaluates audio-visual hallucination. Consistent improvements are observed on OmniBench and AVUT. Similarly, Qwen2.5-Omni$_{3B}$ benefits from gains across all audio-visual benchmarks. For video-SALMONN2+, OutRo improves performance across most settings, including both audio-visual and visual-only benchmarks. Importantly, these consistent gains across video benchmarks demonstrate that OutRo remains effective under complex cross-modal interactions.

5.3 Further Analysis

Here, we conduct all analysis below with the Qwen2.5-Omni model.

Decoding efficiency. To assess the computational overhead of OutRo, we measure decoding latency in seconds per token on 100 examples from AVHBench (Tab. 7). OutRo increases latency from 0.98

[2] `https://github.com/Dao-AILab/flash-attention`

Table 7: **Decoding efficiency.** OutRo introduces only 1.11× overhead with negligible memory overhead compared to Flash-attn.

Decoding	GPU Mem.↓ (GiB)	Latency↓ (sec/token)	Rel.↓
Flash-attn	11.48	0.98	1.00
VCD [34]	13.54	1.98	2.02
OutRo	**11.48**	**1.09**	**1.11**
Eager	17.87	2.00	2.04
VAR [30]	19.02	2.71	2.77
AVCD [28]	27.65	7.88	8.04

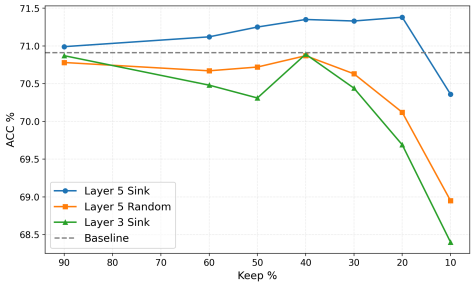


Figure 9: **Token pruning using sink queries.** Pruning tokens by sink-query attention scores improves performance, suggesting sink queries attend to informative modality tokens.

Table 8: **Contrastive decoding.** OutRo is compatible with CD, yielding additional gains.

Qwen2.5-Omni [58]	OmniBench	DailyOmni
Baseline	38.79	54.55
AVCD [28]	39.58	55.56
+OutRo	**40.81**	**55.89**
VideoLLaMA3 [66]	Holmes	MME
Baseline	41.15	61.78
VCD [34]	42.90	62.00
+OutRo	**43.22**	**62.22**

Table 9: **Sink-query token pruning.** Pruning tokens based on sink-query attention reduces latency as fewer tokens are retained, while performance improves and reaches its peak when 20% of the tokens are kept.

Keep %	Latency↓ (sec/token)	Rel.↓	Acc.↑
Baseline	0.996	1.00	70.91
60	0.904	0.91	71.12
50	0.864	0.87	71.25
40	0.824	0.83	71.35
30	0.784	0.79	71.33
20	0.749	0.75	71.38
10	0.714	0.72	70.36

to 1.09 sec/token, a 1.11× slowdown relative to baseline. In contrast, AVCD [28], the Omni-LLM extension of contrastive decoding, is the most expensive, requiring up to four forward passes per step. VCD [34] nearly doubles inference time with two forward passes per step, and VAR [30] incurs overhead from explicit attention map modification that disables optimised attention implementations.

OutRo with contrastive decoding. Tab. 8 reports results when combining OutRo with existing contrastive decoding (CD) methods. For Qwen2.5-Omni$_{3B}$[3] with AVCD, contrastive decoding improves performance, and integrating OutRo yields further gains. Since VCD is not applicable to Omni-LLMs, we additionally evaluate VCD on VideoLLaMA3$_{7B}$ [66] and observe the same trend, showing that OutRo remains effective alongside advanced decoding strategies.

Sink queries encode informative directions. We analyse whether sink queries attend to informative modality tokens by performing token pruning based on attention scores between the sink query and modality tokens on AVHBench. As shown in Fig. 9, pruning tokens by sink-query attention at Layer 5, where OutRo applies sink enhancement, improves performance, suggesting that the sink query at this layer already encodes a structured direction identifying informative tokens. Detailed quantitative results are in Tab. 9. In contrast, random pruning degrades performance, and pruning at Layer 3, where sink tokens first emerge, performs even worse, indicating that this structured direction only emerges at appropriate layers where sufficient context has been aggregated.

Omitted details. Additional ablation studies and analyses are provided in Sec. B of our appendix.

6 Conclusion

In this work, we presented the first analysis of attention sinks in Omni-LLMs and showed that sink tokens are not merely structural artifacts: they function as shared bias terms in the attention output, injecting a consistent direction across all token representations. Building on this insight, we proposed ***OutRo***, a lightweight inference-time method that explicitly aligns non-sink token representations with this bias direction and sharpens it by relaxing causal masking at sink positions. OutRo requires no additional forward passes or attention map access, introduces only a 1.1× decoding overhead, and consistently improves performance across diverse video QA benchmarks. Our findings offer a new perspective on how attention sinks shape decoding in Omni-LLMs, and suggest that leveraging the sink bias is a practical and modality-agnostic strategy for inference-time improvement.

[3] Due to GPU memory constraints when using eager attention, we resize video inputs (256 × 256) for Tab. 8 experiments, which results in slightly lower baseline performance compared to the main results in Tab. 6.

Acknowledgments

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) through a grant funded by the Korean government (MSIT; RS-2024-00457882, National AI Research Lab Project, 50%) and by the IITP–ITRC program funded by the Korean government (MSIT; IITP-2026-RS-2023-00259991, 50%).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv*, 2023.

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv*, 2023.

[3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv*, 2025.

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv*, 2025.

[5] 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. In *Proc. NeurIPS*, 2020.

[6] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proc. CVPR*, 2021.

[7] Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-Holmes: Can mllm think like holmes for complex video reasoning? *arXiv*, 2025.

[8] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv*, 2024.

[9] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. In *Proc. ICLR*, 2024.

[10] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 2024.

[11] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv*, 2021.

[12] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv*, 2025.

[13] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *Proc. ICLR*, 2023.

[14] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with io-awareness. In *Proc. NeurIPS*, 2022.

[15] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *Proc. NeurIPS*, 2024.

[16] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. GPT3.int8(): 8-bit matrix multiplication for transformers at scale. In *Proc. NeurIPS*, 2022.

[17] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proc. CVPR*, 2025.

[18] Zizhuo Fu, Wenxuan Zeng, Runsheng Wang, and Meng Li. Attention is not always needed: Attention sink forges a native moe in attention layers. OpenReview, 2025.

[19] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv*, 2024.

[20] Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *Proc. ICLR*, 2025.

[21] Tianyu Guo, Druv Pai, Yu Bai, Jiantao Jiao, Michael I Jordan, and Song Mei. Active-Dormant Attention Heads: Mechanistically demystifying extreme-token phenomena in llms. *arXiv*, 2024.

[22] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv*, 2020.

[23] Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective Decoding: Alleviating hallucinations for large vision-language models. In *Proc. ICLR*, 2025.

[24] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[25] Nick Jiang, Amil Dravid, Alexei Efros, and Yossi Gandelsman. Vision transformers don't need trained registers. In *Proc. NeurIPS*, 2025.

[26] Pengkun Jiao, Bin Zhu, Jingjing Chen, Chong-Wah Ngo, and Yu-Gang Jiang. Don't Deceive Me: Mitigating gaslighting through attention reallocation in lmms. *arXiv*, 2025.

[27] Chaeyoung Jung, Youngjoon Jang, Jongmin Choi, and Joon Son Chung. Fork-Merge Decoding: Enhancing multimodal understanding in audio-visual large language models. *arXiv*, 2025.

[28] Chaeyoung Jung, Youngjoon Jang, and Joon Son Chung. AVCD: Mitigating hallucinations in audio-visual large language models through contrastive decoding. In *Proc. NeurIPS*, 2025.

[29] Jihoo Jung, Chaeyoung Jung, Ji-Hoon Kim, and Joon Son Chung. Probing cross-modal information hubs in audio-visual llms. In *Proc. ICML*, 2026.

[30] Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See What You Are Told: Visual attention sink in large multimodal models. In *Proc. ICLR*, 2025.

[31] 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*, 2020.

[32] Samir Khaki, Junxian Guo, Jiaming Tang, Shang Yang, Yukang Chen, Konstantinos N Plataniotis, Yao Lu, Song Han, and Zhijian Liu. SparseVILA: Decoupling visual sparsity for efficient vlm inference. In *Proc. ICCV*, 2025.

[33] Junho Kim, Hyunjun Kim, Kim Yeonju, and Yong Man Ro. CODE: Contrasting self-generated description to combat hallucination in large multi-modal models. In *Proc. NeurIPS*, 2024.

[34] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proc. CVPR*, 2024.

[35] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proc. EMNLP*, 2023.

[36] Yizhi Li, Yinghao Ma, Ge Zhang, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, et al. OmniBench: Towards the future of universal omni-language models. *arXiv*, 2024.

[37] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv*, 2024.

[38] Jiayun Luo, Wan-Cyuan Fan, Lyuyang Wang, Xiangteng He, Tanzila Rahman, Purang Abolmaesumi, and Leonid Sigal. To Sink or Not to Sink: Visual information pathways in large vision-language models. In *Proc. ICLR*, 2025.

[39] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv*, 2022.

[40] Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, et al. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *Proc. NeurIPS*, 2025.

[41] Enrique Queipo-de Llano, Álvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. Attention sinks and compression valleys in llms are two sides of the same coin. In *Proc. ICLR*, 2026.

[42] Qwen, :, 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, Tianyi Tang, 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, 2025.

[43] Pedro Sandoval-Segura, Xijun Wang, Ashwinee Panda, Micah Goldblum, Ronen Basri, Tom Goldstein, and David Jacobs. Identifying and evaluating inactive heads in pretrained llms. In *Proc. ICLR*, 2026.

[44] Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. FlashAttention-3: Fast and accurate attention with asynchrony and low-precision. In *Proc. NeurIPS*, 2024.

[45] Jaewon Sok, Jewon Yeom, Seonghyeon Park, Jeongjae Park, and Taesup Kim. Garbage Attention in Large Language Models: Bos sink heads and sink-aware pruning. *arXiv*, 2026.

[46] Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *Proc. COLM*, 2024.

[47] Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. AVHBench: A cross-modal hallucination benchmark for audio-visual large language models. In *Proc. ICLR*, 2025.

[48] Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-SALMONN 2: Caption-enhanced audio-visual large language models. *arXiv*, 2025.

[49] Lexiang Tang, Xianwei Zhuang, Bang Yang, Zhiyuan Hu, Hongxiang Li, Lu Ma, Jinghan Ru, and Yuexian Zou. Not All Tokens and Heads Are Equally Important: Dual-level attention intervention for hallucination mitigation. *arXiv*, 2025.

[50] Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

[51] Bingkui Tong, Jiaer Xia, and Kaiyang Zhou. Mitigating hallucination in multimodal llms with layer contrastive decoding. In *Proc. NAACL*, 2025.

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. NeurIPS*, 2017.

[53] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing Multi-Head Self-Attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proc. ACL*, 2019.

[54] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of ACL*, 2024.

[55] Yu Wang, Kamalika Das, Xiang Gao, Wendi Cui, Peng Li, and Jiaxin Zhang. Gradient-guided Attention Map Editing: Towards efficient contextual hallucination mitigation. In *Findings of NAACL*, 2025.

[56] Houjing Wei, Yuting Shi, and Naoya Inoue. Phase Diagram of Vision Large Language Models Inference: A perspective from interaction across image and instruction. *arXiv*, 2024.

[57] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *Proc. ICLR*, 2024.

[58] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *arXiv*, 2025.

[59] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

[60] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[61] Yudong Yang, Jimin Zhuang, Guangzhi Sun, Changli Tang, Yixuan Li, Peihan Li, Yifan Jiang, Wei Li, Zejun Ma, and Chao Zhang. Audio-centric video understanding benchmark without text shortcut. In *Proc. EMNLP*, 2025.

[62] Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. In *Proc. ICML*, 2024.

[63] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. ActivityNet-QA: A dataset for understanding complex web videos via question answering. In *Proc. AAAI*, 2019.

[64] Zhuoran Yu and Yong Jae Lee. How Multimodal LLMs Solve Image Tasks: A lens on visual grounding, task reasoning, and answer decoding. In *Proc. COLM*, 2025.

[65] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proc. CVPR*, 2022.

[66] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding. *arXiv*, 2025.

[67] Hongxiang Zhang, Hao Chen, Muhao Chen, and Tianyi Zhang. Active layer-contrastive decoding reduces hallucination in large language model generation. In *Proc. EMNLP*, 2025.

[68] Xiaofeng Zhang, Yihao Quan, Chen Shen, Chaochen Gu, Xiaosong Yuan, Shaotian Yan, Jiawei Cao, Hao Cheng, Kaijie Wu, and Jieping Ye. Shallow Focus, Deep Fixes: Enhancing shallow layers vision attention sinks to alleviate hallucination in LVLMs. In *Proc. EMNLP*, 2025.

[69] Ziwei Zhou, Rui Wang, and Zuxuan Wu. Daily-Omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv*, 2025.

[70] Zayd MK Zuhri, Erland Hilman Fuadi, and Alham Fikri Aji. Softpick: No attention sink, no massive activations with rectified softmax. *arXiv*, 2025.

On the Nature of Attention Sink that Shapes Decoding Strategy in Omni-LLMs

– Appendix –

Contents

A Analysis for Sink Identification **15**

A.1 Sink and Outlier Statistics 15

A.2 Empirical Results 15

A.3 Additional Qualitative Examples 16

A.4 Sink Identification on video-SALMONN2+ 16

A.5 Sensitivity to the Sink Threshold 16

B Analysis of OutRo **17**

B.1 Directional Proxy for Sink Score 18

B.2 Attention Redistribution After Rotation 18

B.3 Sink Enhancements Sources 19

B.4 Results on VLMs 19

B.5 Effect of Rotation Strength γ 19

B.6 Rotation with Different Sink Identification Criteria 20

B.7 Generalisation to Image–Text and Text-Only Settings 21

B.8 Rotation toward Alternative Directions 21

C Discussion **21**

C.1 Broader Impacts 21

C.2 Limitations 22

D Algorithm of OutRo **22**

E Qualitative Results **22**

Table A.1: **Dimensions identified as outliers or sinks.** Under the Φ_{VLM} criterion, sink dimensions (*i.e.,* 458 and 2570) cause many tokens to be classified as sinks.

		Outlier			**Sink Φ_{VLM}**			**Sink Φ_{LLM}**		
Dim	Activation	Layer(%)	Token(%)	Outlier	Φ_{VLM}	Token(%)	Sink	Φ_{LLM}	Index	Sink
458	8768.00	100.0	86.89	✓	46.00	59.83	✓	19072	2	✓
2570	6272.00	100.0	89.80	✓	45.75	44.00	✓	13376	2	✓
2718	2656.00	92.9	2.68		46.00	59.83	✓	6208	2	✓
2730	2128.00	100.0	2.75		46.00	59.83	✓	4960	2	✓
3206	716.00	100.0	57.34	✓	46.00	72.68	✓	1640	2	✓
3281	612.00	100.0	16.74	✓	46.00	59.83	✓	1448	2	✓
1427	604.00	100.0	9.95	✓	46.00	59.83	✓	1400	2	✓
1803	556.00	100.0	3.15		46.00	72.68	✓	1264	2	✓
1692	556.00	100.0	11.10	✓	46.00	59.83	✓	1320	2	✓
3110	544.00	100.0	25.90	✓	46.00	59.83	✓	1256	2	✓
2107	544.00	100.0	42.05	✓	45.75	44.00	✓	1192	2	✓
1923	512.00	100.0	42.13	✓	46.00	59.83	✓	1240	2	✓
32	486.00	100.0	64.11	✓	46.00	59.83	✓	1168	2	✓
1451	400.00	96.4	25.95	✓	46.00	59.83	✓	936	2	
1111	362.00	89.3	8.00	✓	45.75	44.00	✓	800	2	
3197	350.00	100.0	92.90	✓	45.75	44.00	✓	764	2	
392	247.00	96.4	25.92	✓	45.75	9.77	✓	592	2	
3461	229.00	100.0	10.43	✓	45.75	9.77	✓	548	2	
662	212.00	100.0	24.51	✓	46.00	59.83	✓	484	2	
1790	190.00	100.0	14.45	✓	46.00	59.83	✓	456	2	
637	184.00	89.3	6.60	✓	45.75	9.77	✓	440	2	
1627	173.00	100.0	19.78	✓	45.75	44.00	✓	382	2	
608	169.00	100.0	32.64	✓	46.00	72.68	✓	396	2	
162	169.00	89.3	6.09	✓	1.54	0.12		80	120	
1431	160.00	89.3	28.71	✓	45.75	44.00	✓	350	2	
143	150.00	100.0	7.67	✓	45.75	9.77	✓	358	2	
2069	148.00	89.3	6.39	✓	45.75	9.77	✓	354	2	
2591	142.00	100.0	8.54	✓	4.50	0.12		68	15	
68	137.00	89.3	6.30	✓	46.00	59.83	✓	326	2	
882	135.00	100.0	12.82	✓	45.75	9.77	✓	322	2	

A Analysis for Sink Identification

This section provides additional analyses of the sink identification criterion discussed in Sec. 3 (Q1), including statistics of sink and outlier tokens, empirical results, and qualitative examples. All analyses are conducted on Qwen2.5-Omni.

A.1 Sink and Outlier Statistics

We report the top 30 activation dimensions in Tab. A.1. A clear overlap is observed between the dimensions identified as outliers and those detected as sink dimensions. In particular, prominent sink dimensions such as 458 and 2570 exhibit large activations across a wide range of tokens.

Under the Φ_{VLM} criterion, these large activations cause many tokens to be classified as sinks, resulting in a substantial fraction of tokens being labelled as sink tokens, including semantically meaningful visual tokens. This suggests that Φ_{VLM} tends to over-identify sinks in multimodal settings.

In contrast, the Φ_{LLM} criterion identifies only a single token as the sink and consistently assigns the sink index to a structurally defined token, leading to a more selective and stable identification of the sink token.

A.2 Empirical Results

To further validate the sink identification criteria, we conduct a controlled intervention by deactivating the sink dimensions D_{sink} (Tab. A.2).

First, when the dimensions D_{sink} are disabled for all tokens, the model performance collapses across all tasks. Similarly, when applying the LLM-based criterion Φ_{LLM}, disabling D_{sink} only at the token identified as a sink also leads to a severe performance drop, suggesting that the token detected by Φ_{LLM} functions as a stabilising anchor in the model's computation. In contrast, when a token identified by the VLM-based criterion Φ_{VLM} is randomly selected and its D_{sink} dimensions are disabled, the model performance remains close to the baseline. Furthermore, even when D_{sink} are disabled for

Table A.2: **Effect of removing sink dimensions under Φ_{LLM} and Φ_{VLM} on AVHBench.** Removing the Φ_{LLM} sink token collapses performance, while removing tokens identified by Φ_{VLM} has little effect.

Method	$A \to V$	$V \to A$	Matching	Overall
Baseline	82.31	82.58	51.71	71.60
All tokens	0.00	0.00	0.00	0.00
Sink token (Φ_{LLM})	0.00	0.00	0.00	0.00
Random token (Φ_{VLM})	80.02	78.95	52.67	69.87
Sink tokens (Φ_{VLM})	78.96	82.14	52.77	71.06

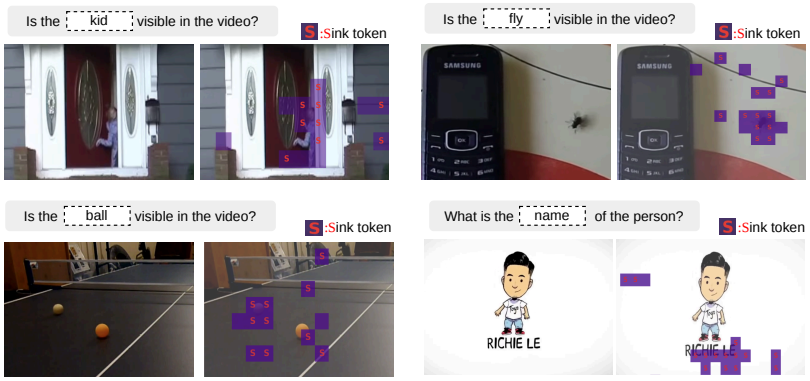


Figure A.1: **Additional qualitative examples.** Each example visualises attention scores for query tokens (*e.g.,* kid, fly, ball, name) on AVHBench videos.

all tokens detected by Φ_{VLM} except the token identified by Φ_{LLM}, the performance remains largely unchanged.

These results suggest that the token identified by Φ_{LLM} plays a more critical role in the model's inference process than those detected by Φ_{VLM}.

A.3 Additional Qualitative Examples

We provide further qualitative examples illustrating the misbehaviour of the visual attention sink criterion Φ_{VLM} when applied to Omni-LLMs. In Fig. A.1, we observe the same pattern where a large fraction of tokens are classified as sinks, including tokens corresponding to semantically meaningful visual content.

A.4 Sink Identification on video-SALMONN2+

We provide additional sink identification results on video-SALMONN2+ to support the findings in Q1. As shown in Fig. A.2, the same pattern observed in Qwen2.5-Omni holds: Φ_{VLM} collapses and over-identifies the majority of tokens as sinks in deeper layers, while Φ_{LLM} maintains a sparse and stable identification throughout. This confirms that Φ_{LLM} generalises reliably across different Omni-LLM architectures.

A.5 Sensitivity to the Sink Threshold

We examine whether the over-identification of sink tokens under Φ_{VLM} can be resolved by adjusting its threshold τ. We vary $\tau \in \{20, 25, 30\}$ and measure the proportion of tokens identified as sinks across four audio-visual benchmarks on Qwen2.5-Omni 3B and 7B. Increasing τ substantially reduces the number of detected sinks, but its effect varies markedly across model scales. On the 3B model, the proportion falls from 10.02–15.33% at $\tau = 20$ to 0.03–0.13% at $\tau = 30$. In contrast, the same threshold of $\tau = 30$ still identifies 5.98–8.27% of tokens as sinks on the 7B model, compared with 24.53–33.32% at $\tau = 20$. These results indicate that Φ_{VLM} is highly sensitive to threshold selection, and that a threshold yielding sparse detections at one model scale does not necessarily transfer to another. In comparison, Φ_{LLM} consistently identifies a single structurally defined sink token without model-specific threshold tuning.

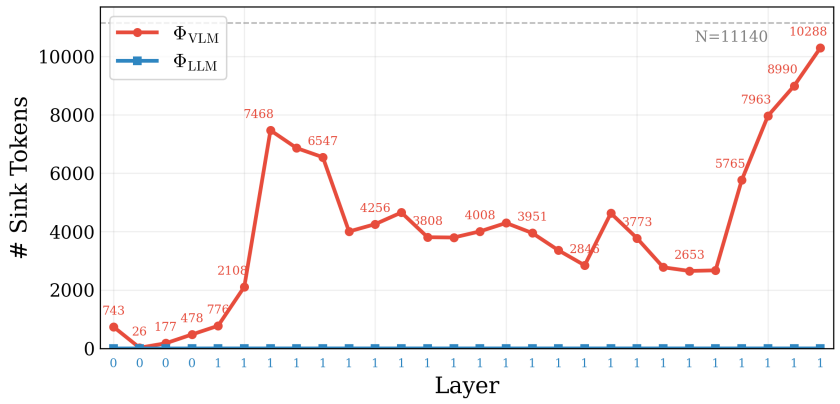


Figure A.2: **Number of sink tokens across layers under Φ_{VLM} and Φ_{LLM} on video-SALMONN2+.** Φ_{VLM} rapidly over-identifies sink tokens as depth increases, eventually classifying nearly all tokens ($N = 11140$) as sinks in the final layers. In contrast, Φ_{LLM} consistently identifies only a sparse and stable set of sink tokens throughout all layers.

Table A.3: **Sensitivity of Φ_{VLM} to the sink threshold.** We report the mean proportion of input tokens identified as sinks. Increasing τ reduces the proportion of detected sinks, but the same threshold yields substantially different proportions across model scales.

Models	τ	**Sink Token Proportion (%)**			
		OmniBench	AVUT	AVHBench	DailyOmni
Qwen2.5-Omni$_{3B}$	20	15.33	13.79	10.02	11.29
	25	2.50	1.91	1.30	1.38
	30	0.13	0.03	0.06	0.03
Qwen2.5-Omni$_{7B}$	20	33.32	26.16	26.46	24.53
	25	16.41	11.83	12.09	11.44
	30	8.27	5.98	6.10	6.11

B Analysis of OutRo

In this section, we further investigate the design choices and the analysis of OutRo. Unless otherwise specified, all analyses are conducted on Qwen2.5-Omni.

Gated rotation. We also investigate using the sink score directly as a rotation criterion. Although tuned thresholds (*e.g.,* 0.7) can yield larger improvements over the baseline (Tab. B.1), this approach requires access to attention maps or pre-computed sink statistics, introducing overhead and hyper-parameter tuning. In contrast, the proxy based on cosine similarity (Eq. (8)) achieves performance gains while preserving FlashAttention compatibility and computational efficiency.

We therefore adopt gated rotation (Eq. (9)). Instead of uniform rotation, we modulate head outputs with a Tanh gate, enabling adaptive control over the rotation strength. This outperforms plain rotation, indicating that controlled modulation is empirically more effective (Tab. B.2).

Sink enhancement. The timing of sink enhancement is critical. If applied too early, sink tokens have not yet aggregated sufficient global information and their queries remain weakly structured, leading to limited or degraded performance (Tab. B.3). This aligns with prior findings that transformer layers exhibit an early stage of modality structuring before deeper integration, typically spanning the first seventh of layers [27, 32, 56, 64]. We also compare applying enhancement once at this depth with applying it at every subsequent layer. A single application performs better, indicating that intervention at the transition depth is sufficient, while repeated enhancement introduces redundancy.

Table B.1: **Sink score vs. cosine similarity.** Cosine gating improves over the baseline.

Criterion	Acc.
Baseline	70.91
0.6	72.46
0.7	**72.82**
0.8	72.72
0.9	71.46
Cosine	71.33

Table B.2: **Gate study.** Tanh gate outperforms uniform and polynomial alternatives on AVHBench.

Gate	Acc.
Baseline	70.91
Uniform	71.33
Quadratic	70.78
Cubic	70.67
Tanh	**71.44**

Table B.3: **Enhancement layer ablation.** A single application at L5 outperforms repeated enhancement on AVHBench.

Layer	Acc.
Baseline	70.91
L3	70.67
L4	70.94
L5	71.08
L5 (Once)	**71.25**

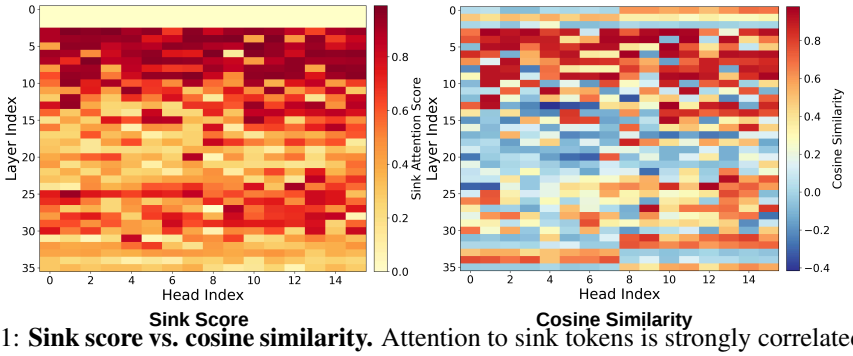


Figure B.1: **Sink score vs. cosine similarity.** Attention to sink tokens is strongly correlated with the cosine similarity between head outputs and the sink value direction.

Table B.4: **Attention redistribution after rotation.** Attention to video/audio values are baseline attention mass normalised by the number of modality tokens. Deltas show the change in total attention mass after rotation. Rotation reduces sink attention and preferentially increases attention toward the modality that was originally attended to less.

Model	Attn. to video / tok	Attn. to audio / tok	Δ Sink	Δ Video	Δ Audio
Qwen2.5-Omni [58]	2.20e-04	**2.87e-04**	-0.0176	**+0.0127**	+0.0022
video-SALMONN2+ [48]	6.45e-05	**3.64e-04**	-0.0095	**+0.0088**	+0.0005
VideoLLaMA2 [8]	**3.22e-04**	2.15e-04	-0.0234	+0.0103	**+0.0123**

B.1 Directional Proxy for Sink Score

Fig. B.1 shows a strong positive correlation between attention to sink tokens and the cosine similarity between each head output vector and the sink value direction, with an average Pearson correlation of approximately 0.7. This indicates that directional alignment serves as an effective proxy for sink score, enabling a more efficient formulation used in our method.

B.2 Attention Redistribution After Rotation

We analyse how rotation affects the distribution of attention across token types. For each model, we measure the change in total attention mass allocated to sink, video, and audio tokens before and after rotation. To account for differences in the number of tokens per modality, baseline attention values are normalised by the number of modality tokens.

As shown in Tab. B.4, rotation consistently reduces attention to sink tokens across all models, with the freed attention redistributing toward modality tokens. Notably, the redistribution is not uniform across modalities: in Qwen2.5-Omni [58] and video-SALMONN2+ [48], where audio tokens originally receive more attention per token than video tokens, the increase is larger for video; in VideoLLaMA2 [8], where video tokens receive more attention per token, the increase is larger for audio. This suggests that rotation normalises the attention imbalance across modalities, directing the released capacity toward the modality that the model originally attended to less.

Table B.5: **Effect of different sink attention sources.** Using internal features achieves the best performance, whereas encoder inputs do not improve performance.

Mode	$A \to V$	$V \to A$	Matching	Overall
Baseline	80.11	75.41	59.86	70.91
Internal (Ours)	**80.37**	**75.81**	60.18	**71.25**
Encoder	79.93	74.93	**60.55**	70.91
Encoder w/o text	80.11	75.33	59.59	70.78

Table B.6: Performance of ***OutRo*** on VLMs.

Models	Holmes	MME	ActivityNet
Qwen2.5-VL$_{3B}$ [4]	41.26	57.89	**47.02**
+OutRo	**42.41**	**58.33**	46.87
Qwen3-VL$_{8B}$ [3]	46.65	67.44	46.65
+OutRo	**46.76**	**68.33**	**46.85**
VideoLLaMA3$_{7B}$ [66]	41.15	61.78	49.32
+OutRo	**42.62**	**62.56**	**49.92**

Table B.7: **Rotation with different sink identification criteria.** We compare rotation toward sinks identified by Φ_{LLM} with rotation toward the top 1% or top 1 token under Φ_{VLM}. All variants use rotation only, without mask relaxation. We report mean accuracy and standard deviation over five seeds. Restricting Φ_{VLM} to fewer tokens does not recover an effective rotation direction.

Models	**Method**	**Audio-Visual QA**			
		OmniBench	AVUT	AVHBench	DailyOmni
Qwen2.5-Omni$_{3B}$	Baseline	40.19±0.57	60.23±0.56	64.39±0.66	55.03±0.70
	OutRo (Φ_{LLM})	**40.90±0.68**	**61.11±0.47**	**64.78±0.74**	**55.77±0.64**
	Φ_{VLM} top 1%	34.20±0.58	48.69±0.50	60.97±0.59	33.02±0.49
	Φ_{VLM} top 1	38.87±0.84	56.38±1.01	61.90±0.52	43.97±0.54
Qwen2.5-Omni$_{7B}$	Baseline	48.13±0.88	**64.53±0.53**	71.13±0.32	**52.89±1.07**
	OutRo (Φ_{LLM})	**48.93±0.82**	64.44±0.57	**72.03±0.27**	52.70±1.16
	Φ_{VLM} top 1%	44.89±1.62	63.32±1.16	70.28±0.43	49.40±1.24
	Φ_{VLM} top 1	47.73±0.96	63.70±0.47	70.28±0.51	51.06±1.37

B.3 Sink Enhancements Sources

We analyse different sources for applying attention between the sink query and modality keys on AVHBench. In addition to the internal features used in OutRo, we also validate encoder inputs by directly applying cross-attention to the visual and audio modalities. As shown in Tab. B.5, using encoder inputs is less effective than using internal features. Removing the text modality and applying cross-attention only to the visual and audio inputs also does not improve performance.

B.4 Results on VLMs

Tab. B.6 reports results on visual-only LLMs. Across Qwen2.5-VL, Qwen3-VL, and VideoLLaMA3, OutRo improves performance on visual-only QA benchmarks. For example, Qwen3-VL achieves a gain of +0.89 on VideoMME, and VideoLLaMA3 improves by +1.47 on VideoHolmes. These results demonstrate that OutRo generalises robustly across diverse VLMs regardless of modality composition, improving visual reasoning.

B.5 Effect of Rotation Strength γ

We analyse the sensitivity of OutRo to the rotation strength parameter γ, which controls the magnitude of modulation applied to attention head outputs. Fig. B.2 reports results across representative Omni-LLMs and VLMs on the benchmarks. We vary γ in the range $[0.5,4.0]$ and observe stable performance trends within each model, with moderate variation across architectures. Based on these results, we select the optimal γ for each model in our experiments.

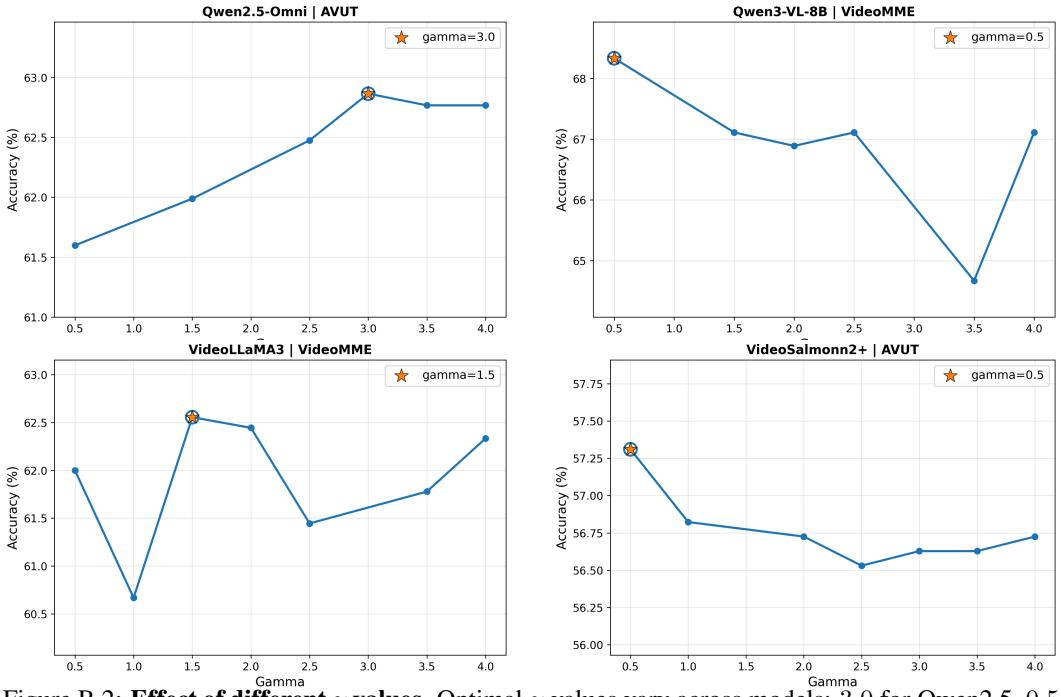


Figure B.2: **Effect of different γ values.** Optimal γ values vary across models: 3.0 for Qwen2.5, 0.5 for Qwen3-VL and video-SALMONN2+, and 1.5 for VideoLLaMA3.

B.6 Rotation with Different Sink Identification Criteria

We examine whether a stricter selection under Φ_{VLM} recovers sink tokens that provide an effective direction for rotation. We retain either the top 1% of tokens or the single highest-scoring token under Φ_{VLM}, and use their value representations to construct the rotation direction. We isolate the rotation component of OutRo and keep the remaining intervention unchanged, allowing us to assess the effect of the sink identification criterion. Experiments are conducted on Qwen2.5-Omni 3B and 7B across four audio-visual benchmarks, using stochastic decoding with temperature 0.7 and top-p 0.9. We report the mean and standard deviation over five seeds.

As shown in Tab. B.7, both selections under Φ_{VLM} degrade performance relative to the baseline across all eight model–benchmark settings. Selecting only the highest-scoring token generally reduces the degradation compared with the top 1% selection, but does not recover baseline performance. In contrast, rotation using Φ_{LLM} outperforms both alternatives in every setting and improves over the baseline in six of the eight settings. These results suggest that the distinction between the criteria is not explained solely by the number of selected tokens: restricting Φ_{VLM} to a small subset does not recover a comparably effective rotation direction.

Table B.8: **Generalisation to image–text and text-only settings.** OutRo improves performance on POPE with image–text inputs and GSM8K with text-only inputs. Results on Qwen2.5-Omni use greedy decoding over the full evaluation sets. For text-only LLMs, we report mean accuracy and standard deviation over five seeds.

Models	**Image–Text**	**Text-Only**
	POPE	GSM8K
Qwen2.5-Omni$_{3B}$	87.67	73.92
+OutRo	**88.06**	**74.45**
Qwen2.5-7B-Instruct	-	59.03±0.73
+OutRo	-	**61.41±0.85**
Llama-3.1-8B-Instruct	-	76.00±0.46
+OutRo	-	**77.60±0.46**

Table B.9: **Rotation toward alternative directions.** We compare the sink value direction with random video, audio, and text token directions and a head-mean direction on Qwen2.5-Omni 3B. All variants use gated rotation without mask relaxation and preserve the original head-output norm. We report mean accuracy and standard deviation over five seeds.

Method	Audio-Visual QA			
	OmniBench	AVUT	AVHBench	DailyOmni
Baseline	40.19 ± 0.57	60.23 ± 0.56	64.39 ± 0.66	55.03 ± 0.70
OutRo	$\mathbf{40.90 \pm 0.68}$	$\mathbf{61.11 \pm 0.47}$	$\mathbf{64.78 \pm 0.74}$	$\mathbf{55.77 \pm 0.64}$
Random video	36.38 ± 1.28	56.52 ± 0.95	61.04 ± 0.39	46.59 ± 1.34
Random audio	37.14 ± 0.70	56.44 ± 0.90	60.57 ± 0.92	45.32 ± 0.91
Random text	36.59 ± 0.73	53.42 ± 0.99	61.75 ± 0.42	44.49 ± 1.55
Head mean	22.50 ± 0.65	26.18 ± 0.70	55.22 ± 0.65	25.98 ± 0.75

B.7 Generalisation to Image–Text and Text-Only Settings

We examine whether the benefits of OutRo extend beyond audio-visual inputs. We first evaluate Qwen2.5-Omni 3B under image–text inputs on POPE [35] and text-only inputs on GSM8K [11]. Both experiments use the full evaluation sets with greedy decoding. As shown in Tab. B.8, OutRo improves performance in both settings, indicating that its effectiveness does not depend on the simultaneous presence of video and audio.

We further evaluate OutRo on two text-only LLMs, Qwen2.5-7B-Instruct [42] and Llama-3.1-8B-Instruct [19], using GSM8K. We report mean accuracy and standard deviation over five seeds, with correctness determined by matching the final numeric answer. The rotation strength γ is selected on 100 validation samples, yielding $\gamma = 3$ for Qwen and $\gamma = 1$ for Llama. OutRo improves performance on both models, suggesting that the intervention also applies beyond Omni-LLM backbones. Together, these results support the applicability of OutRo across different input configurations and model families.

B.8 Rotation toward Alternative Directions

We examine whether the benefit of rotation depends on the sink value direction or can be obtained using alternative directions. We replace the sink value direction with the value representation of a randomly selected video, audio, or text token, and additionally evaluate a head-mean direction. All variants retain the same gated rotation and norm-preserving normalisation. We use the rotation component alone, without mask relaxation, to isolate the effect of the chosen direction. Experiments are conducted on Qwen2.5-Omni 3B across four audio-visual benchmarks, using stochastic decoding with temperature 0.7 and top-p 0.9. We report the mean and standard deviation over five seeds.

As shown in Tab. B.9, all alternative directions degrade performance relative to the baseline across the four benchmarks. Random token directions reduce accuracy regardless of the selected modality, while the head-mean direction produces the largest degradation in every setting. In contrast, rotation toward the sink value direction consistently improves performance. Since all variants preserve the original head-output norm, these differences cannot be explained by changes in output magnitude alone. These results support the importance of the sink value direction, rather than rotation toward an arbitrary token or the head-mean direction.

C Discussion

C.1 Broader Impacts

This work provides a better understanding of how attention sinks operate in Omni-LLMs, offering insights into the internal mechanisms of multimodal transformers. Building on this analysis, OutRo improves audio-visual reasoning in a training-free manner, which could benefit real-world applications such as assistive technologies, video summarisation, and multimodal dialogue systems.

As a training-free method that does not modify model weights or training data, OutRo does not introduce new risks beyond those already present in the underlying models. The potential for misuse is inherited from the base Omni-LLMs rather than introduced by our approach.

C.2 Limitations

The rotation strength parameter γ requires manual specification and its optimal value may vary across model architectures. In addition, although OutRo improves performance across tasks, the magnitude of improvement is modest. Developing strategies to reduce the calibration effort for γ and further enhance its effectiveness remains an important direction for future work.

D Algorithm of OutRo

Alg. 1 summarises the OutRo procedure applied at each decoder layer. Given the hidden states, we first identify sink tokens using the sink criterion Φ_{LLM}. We then compute the attention outputs and estimate the sink value direction by averaging the value vectors of the detected sink tokens.

For non-sink positions, head outputs are modulated through gated rotation according to their directional alignment with the sink value direction, controlled by the rotation strength γ. The updated outputs are rescaled to preserve their original magnitude. For sink positions, we apply mask relaxation at a selected layer, allowing sink queries to attend beyond the causal mask and thereby aggregate contextual information. The resulting outputs are then combined across heads and projected to produce the hidden states that are fed into the subsequent FFN layer.

E Qualitative Results

We present qualitative examples illustrating the behaviour of OutRo across different multimodal models. Examples on video-QA tasks are shown in Fig. E.1, while video captioning results are presented in Fig. E.2. These qualitative observations are consistent with the quantitative improvements, where captioning performance on AVHBench increases from 3.01 to 3.11.

Algorithm 1 OutRo in a Decoder Layer

Require: Hidden states $\mathbf{X}^{(\ell)}$, sink criterion Φ_{LLM}, rotation strength γ
Ensure: Updated hidden states $\mathbf{X}_{\mathrm{out}}^{(\ell)}$

Identify sink indices:

$$\mathcal{S}^{(\ell)} \leftarrow \Phi_{\mathrm{LLM}}(\mathbf{X}^{(\ell)})$$

Compute attention outputs

$$\mathbf{O}^{(\ell)} = \mathrm{Attn}(\mathbf{Q}^{(\ell)}, \mathbf{K}^{(\ell)}, \mathbf{V}^{(\ell)})$$

Compute sink value direction

$$\bar{\mathbf{v}}_{h,s}^{(\ell)} = \frac{1}{|\mathcal{S}^{(\ell)}|} \sum_{s \in \mathcal{S}^{(\ell)}} \mathbf{v}_{h,s}^{(\ell)}$$

▷ **Non-sink positions: gated head rotation**

for each $ns \notin \mathcal{S}^{(\ell)}$, head h **do**

Compute directional alignment

$$c_{h,ns} = \cos(\mathbf{O}_{h,ns}, \bar{\mathbf{v}}_{h,s}^{(\ell)})$$

Compute gate

$$g_{h,ns} = \tanh\left(\frac{\mathrm{ReLU}(c_{h,ns})}{0.1}\right)$$

Rotate head output

$$\hat{\mathbf{O}}_{h,ns} = \mathbf{O}_{h,ns} + \gamma g_{h,ns} \frac{\mathbf{O}_{h,ns} \cdot \bar{\mathbf{v}}_{h,s}^{(\ell)}}{\|\bar{\mathbf{v}}_{h,s}^{(\ell)}\|^2} \bar{\mathbf{v}}_{h,s}^{(\ell)}$$

Rescale magnitude

$$\mathbf{O}_{h,ns} \leftarrow \frac{\|\mathbf{O}_{h,ns}\|}{\|\hat{\mathbf{O}}_{h,ns}\|} \hat{\mathbf{O}}_{h,ns}$$

end for

▷ **Sink positions: mask relaxation**

if $\ell = \ell_{\mathrm{enh}}$ **then**

for each $s \in \mathcal{S}^{(\ell)}$, head h **do**

Relax causal mask for sink queries

$$\mathbf{A}_{h,s,:}^{\mathrm{relaxed}} = \mathrm{Softmax}\left(\frac{\mathbf{Q}_{h,s}\mathbf{K}_h^\top}{\sqrt{D_h}}\right)$$

Replace sink-position outputs

$$\mathbf{O}_{h,s}^{(\ell)} \leftarrow \mathbf{A}_{h,s,:}^{\mathrm{relaxed}} \mathbf{V}_h$$

end for

end if

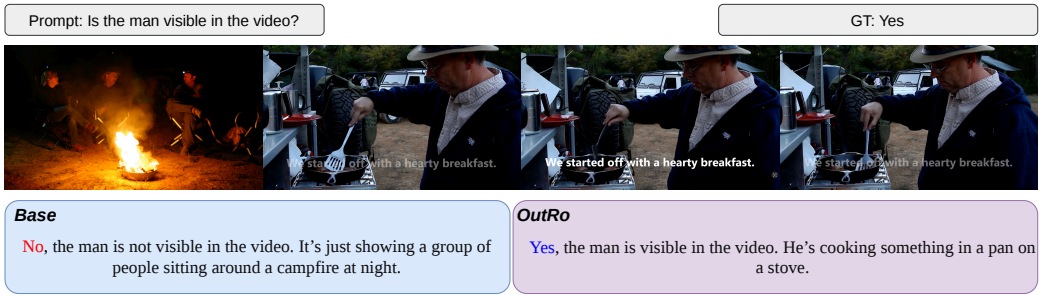
Aggregate heads and apply output projection

$$\mathbf{X}_{\mathrm{out}}^{(\ell)} \leftarrow \mathbf{X}^{(\ell)} + W_O \mathbf{O}^{(\ell)}$$

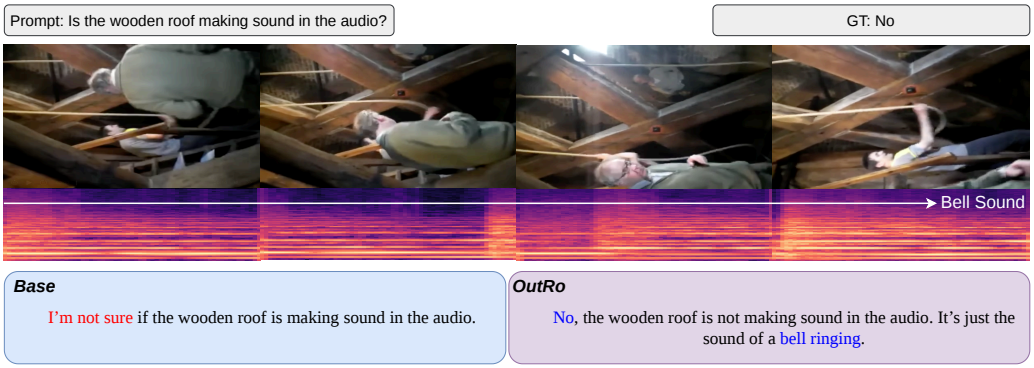
return $\mathbf{X}_{\mathrm{out}}^{(\ell)}$



(a)



(b)



(c)



(d)

Figure E.1: **Qualitative results on video-QA on AVHBench.** Compared with the baseline, OutRo produces answers that better align with visual and audio-visual evidence in the video, leading to more accurate multimodal reasoning.

Prompt: Describe what you see and hear in a single sentence.



GT

A young man whooshes through a rushing stream, holding a fish he caught while fishing.

Base

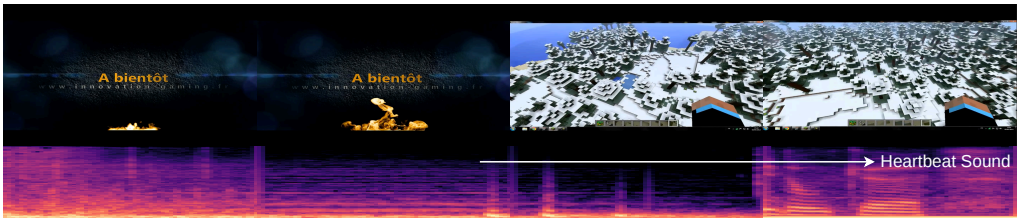
I see a man in a hat and vest with a backpack and net on his back, standing in a shallow river surrounded by rocks and trees, and I hear the sound of flowing water.

OutRo

I see a man in a hat and backpack standing in a shallow river, holding a fishing rod, and I hear the sound of flowing water.

(a)

Prompt: Describe what you see and hear in a single sentence.



GT

In a game, a large fire emerging from a dark background, while a man speaks with a heartbeat in the background.

Base

Well, it seems like there's a video game being played, and the player is in a snowy Minecraft world. There's also a sound effect that could be from the game or something else. And the speech subtitle is in French.

OutRo

I see a dark background with a fire at the bottom and the words "A bientôt" and a website address. There's also a Minecraft game scene with snowy terrain and a player's hand holding a sword. I hear a sound effect and a heartbeat sound.

(b)

Figure E.2: **Qualitative results for video captioning on AVHBench.** OutRo produces captions grounded in both visual and audio-visual cues, whereas the baseline often fails to capture such multimodal details.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the three main contributions: (i) the first analysis of attention sinks in Omni-LLMs, (ii) the shared bias interpretation of sink value representations, and (iii) the OutRo method with its two components. These are fully supported by the experimental results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in the Appendix, noting that the rotation strength γ requires manual tuning per model architecture and that performance gains are modest.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [No]

Justification: The shared bias interpretation of the sink value vector is theoretically motivated through the attention output decomposition, but no formal proof is provided. The claim is instead supported empirically through injection, and rotation experiments.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All hyperparameters, model names, benchmarks, and hardware specifications are provided and the appendix. OutRo is a training-free method applied to publicly available models, making results reproducible through the greedy decoding.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is not yet publicly available but will be released upon acceptance. All datasets and models used are publicly available and cited in the paper.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Hyperparameters, model configurations, benchmark protocols, and hardware details are provided. Additional details are in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not reported. OutRo is a deterministic inference-time method with no stochastic components, so results are fully reproducible given the same model and inputs.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments are conducted on a single NVIDIA RTX A6000 GPU and implementation details are provided.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: Broader impacts are discussed in our appendix, covering both potential positive applications and the inherited risks from the underlying Omni-LLMs.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release new models or datasets. OutRo is an inference-time method applied to existing publicly available models, which poses no additional risks beyond those of the base models.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All models and datasets used in this paper are publicly available and properly cited. No proprietary assets are used.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce new datasets or model weights.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects research and therefore does not require IRB approval.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are the subject of study in this paper, not a component of the core methodology. No LLMs were used as tools in the research process.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.