TWILIGHT: THINKING WITH GENERATED IMAGES VIA LIGHTWEIGHT KV CACHE

Anonymous authors
Paper under double-blind review

ABSTRACT

Thinking with Generated Images (TwGI) has emerged as a promising paradigm that enables vision-language models to actively synthesize novel visual content as intermediate reasoning steps, transcending the limitations of static input images. However, due to the inclusion of numerous image tokens and multiple classifier-free guidance (CFG) samples, TWGI's reasoning process incurs significant memory pressure, limiting scalability to long-sequence generation. By analyzing attention patterns, we reveal that TwGI's KV cache contains significant redundancy, as some heads maintain low attention scores on previous tokens. Therefore, we propose Twilight, the first training-free KV cache eviction framework tailored for TwGI. Specifically, we introduce a novel normalized standard deviation metric to identify modality-specific retrieval heads that attend globally to historical tokens. Leveraging the distinctive attention patterns of retrieval heads and streaming heads across text and image generation stages, we design targeted KV cache eviction strategies and sparse attention mechanisms. Our approach selectively retains critical tokens while discarding redundant ones, significantly reducing memory consumption without compromising generation quality. Extensive experiments validate the effectiveness of our approach. Twilight maintains or even improves generation performance through a training-free approach, while reducing KV-cache memory by over 60% on average.

1 INTRODUCTION

Although Vision-Language Models (VLMs) have significantly enhanced multimodal reasoning capabilities by Chain-of-Thought (CoT) mechanisms Bai et al. (2025); Zhan et al. (2026); Wang et al. (2026b), their reasoning processes remain constrained within a text-dominant paradigm where visual inputs serve as passive context rather than active components Su et al. (2025b); Song et al. (2026). The *thinking with images* paradigm emerges to address this limitation by enabling models to actively interleave visual exploration with textual reasoning, transforming images from static inputs into a dynamic and intrinsic component of reasoning process Zheng et al. (2025); Guo et al. (2026b); Tu et al. (2026). This synergistic integration can establish an exhaustive reasoning framework Hong et al. (2025); Li et al. (2026b), not only facilitating fine-grained visual understanding Chen et al. (2025); Li et al. (2026c) but also empowering high-quality image generation Duan et al. (2025); Jiao et al. (2026). However, most existing methods still operate solely on input images without generating novel visual content, limiting their applicability to complex multimodal tasks.

By integrating unified multimodal models Chern et al. (2024); Deng et al. (2025) into the *thinking with images* paradigm, such models can actively synthesize novel visual content as intermediate reasoning steps rather than being constrained to fixed user-provided images. This paradigm transcends input image limitations, eliminates external tool dependency, and shifts models from passive perception to active visual creation, significantly expanding multimodal reasoning capacity Li et al. (2025b;a). While this research direction remains an emerging frontier, *Thinking with Generated Images* (TwGI) represents a pioneering exploration Chern et al. (2025). TwGI trains unified architectures to support native long-multimodal reasoning chain with seamless text-image token interleaving. Through visual subgoal decomposition and iterative refinement, TwGI substantially enhances generation quality and reasoning fidelity. However, TwGI requires numerous tokens to generate a single image and introduces multiple classifier-free guidance (CFG) samples, causing substantial KV cache overhead during interleaved text-image generation. As shown in Figure 1(a),

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

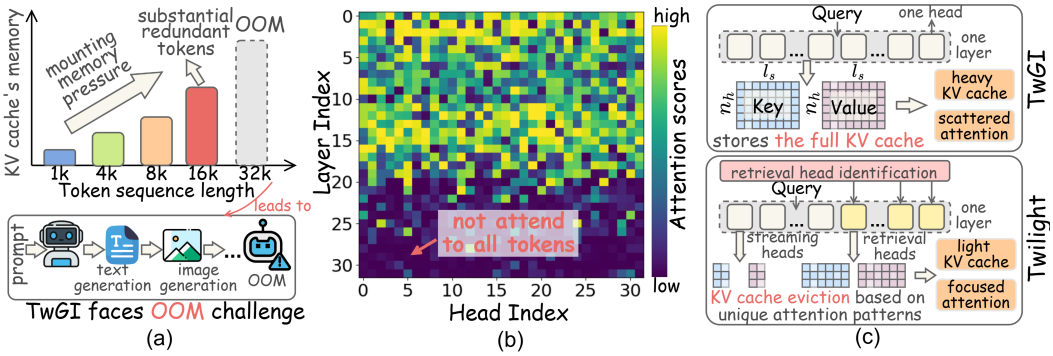


Figure 1: (a) As the token sequence length increases, TwGI suffers from mounting memory pressure and encounters out-of-memory challenges. (b) The attention map for a subset of intermediate tokens across all heads. Many heads do not attend to all tokens, indicating redundancy in the KV cache. (c) Our Twilight can identify retrieval heads in advance and evict redundant KV cache based on unique attention patterns, alleviating memory pressure while achieving focused attention.

memory consumption grows rapidly with progressive token generation and eventually triggers an Out-of-Memory (OOM) result. Furthermore, we analyze the attention map across all heads, which is shown in Figure 1(b). It reveals that not all heads maintain high attention scores for all previous tokens when generating new tokens, indicating that TwGI's original KV cache stores significant redundant tokens with minimal contribution to attention computation.

Therefore, it is feasible to design an eviction scheme that reduces KV cache redundancy while preserving generation quality, which has been empirically validated in long-text LLMs Feng et al. (2026). Due to TwGI's multi-modal generation nature and the substantial divergence in attention patterns across different modalities, most existing eviction strategies designed for text generation cannot be directly applied. Accordingly, we propose **Twilight**, a training-free eviction scheme to compress KV cache tokens while preserving critical tokens. Specifically, to assess the redundancy of each head, we first introduce a normalized standard deviation metric to identify retrieval heads during image and text generation stages through preliminary experiments on a random sample set. We then analyze attention patterns of retrieval heads and streaming heads across different modalities to determine which token positions consistently receive higher attention scores. Finally, leveraging these empirically derived attention patterns, we design a KV cache eviction strategy with incorporating sparse attention to achieve high-quality generation while simultaneously reducing KV cache occupancy and attention computation overhead. The advantages of Twilight are shown in Figure 1(c).

Our contributions can be summarized as follows: (1) Through comprehensive preliminary analysis, we reveal the unique characteristics inherent to the TwGI, including distinctive retrieval head properties and attention patterns, providing valuable insights for KV cache compression. (2) We propose a training-free framework **Twilight**, which is the first work to apply KV cache eviction to *thinking with generated images*. Moreover, it designs a sparse attention mechanism based on TwGI's substantial divergence in attention patterns across different modalities. (3) We introduce a novel retrieval head identification method that introduces the normalized standard deviation of attention distributions, enabling rapid and accurate identification of retrieval heads for different modalities. (4) We conduct extensive experiments demonstrating the effectiveness of our training-free approach, which substantially reduces KV cache usage while preserving generation quality.

2 Related Works

Thinking with images. Thinking with images breaks the limitations of text-centric reasoning Su et al. (2025a); Wu et al. (2026); Wang et al. (2026a). Pixel Reasoner Wang et al. (2025a) and DeepSketcher Zhang et al. (2026) perform dynamic visual operations to enhance reasoning, while Mini o3 Lai et al. (2025) and Simple o3 Wang et al. (2025b) explore visual reasoning with refined datasets. DeepScan Li et al. (2026b) and CVSearch Li et al. (2026a) introduce training-free visual focusing for fine-grained understanding, while Thyme Zhang et al. (2025) and V-Thinker Qiao et al. (2025) enable interaction with images by codes. Recently, several works have explored actively generating novel visual content during reasoning Qin et al. (2026); Gu et al. (2026); Diao et al. (2026). Zebra-CoT Li et al. (2025a) leverages intermediate images to solve scientific problems, while MathCanvas

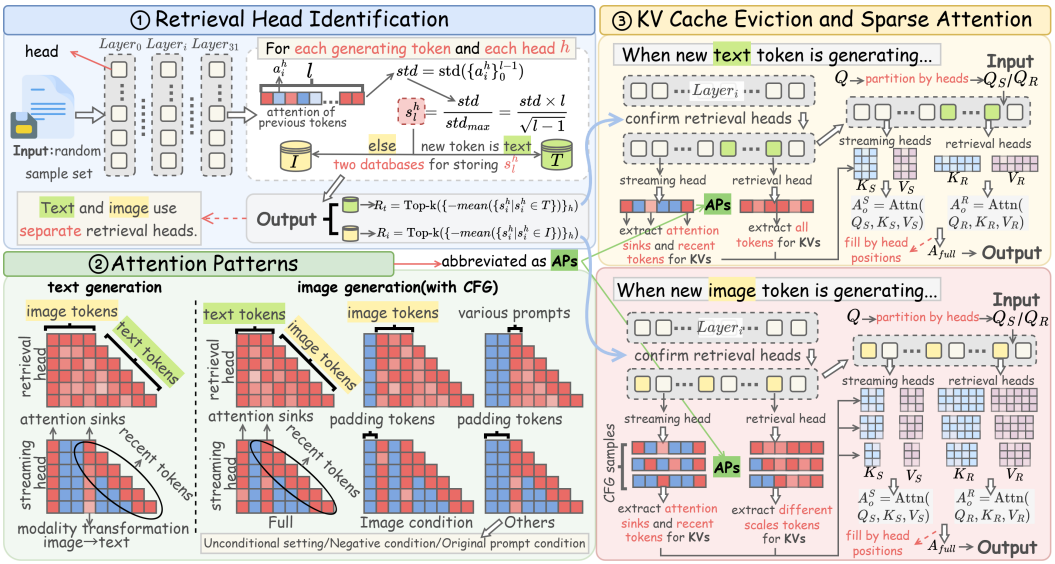


Figure 2: The overview of Twilight. We propose a normalized standard deviation criterion to identify retrieval heads on a random sample set. Leveraging the distinct attention patterns of text and image generation in a unified autoregressive architecture, we design tailored KV-cache eviction and sparse attention mechanisms for text generation and image generation phases. Our training-free approach reduces KV-cache memory consumption while preserving generation performance.

Shi et al. (2025) learns to generate diagrams for multimodal mathematical reasoning. Visual planning Xu et al. (2025) generates visual steps to address spatial reasoning tasks, whereas TwGI Chern et al. (2025) adopts generated images to enhance generation quality. Freed from the constraints of fixed input images, these approaches are gradually emerging as promising reasoning paradigms.

KV cache eviction. KV cache eviction has been validated to effectively alleviate memory pressure while maintaining output quality Haoyang et al. (2025); Jiang et al. (2026). StreamingLLM Xiao et al. (2023) keeps sink tokens plus sliding window to stabilize generation, while DuoAttention Xiao et al. (2024) introduce different head types to cut KV cost. CAKE Qin et al. (2025) performs layer-preference-guided cascading KV eviction, and SnapKV Li et al. (2024) retains salient KV positions. KeyDiff Park et al. (2026) keeps low-redundancy tokens using key similarity, while Ada-KV Feng et al. (2026) allocates adaptive budgets to realize KV eviction. R-KV Cai et al. (2026) jointly considers token importance and redundancy during decoding, while DefensiveKV Guo et al. (2026a) uses worst-case aggregation to robustify cache eviction. The above methods primarily target text-dominated reasoning models and cannot be directly applied to the thinking with images paradigm.

3 Method

3.1 Overview of Framework

Although TwGI improves image generation quality through an interleaved text-image reasoning process, its growing sequence length incurs substantial KV cache overhead, hindering efficient deployment and scaling. To address this issue, we propose Twilight, a training-free framework for high-quality image generation with a lightweight KV cache, as illustrated in Figure 2. Specifically, we use the normalized standard deviation of attention distributions to pre-identify intrinsic retrieval heads from a small random sample set. Given the substantial modality gap, retrieval heads are identified separately for text and image generation (Section 3.2). We further analyze the distinctive attention patterns induced by cross-modal interactions and the use of CFG in TwGI (Section 3.3). Based on these observations, we design modality-specific KV cache eviction and sparse attention mechanisms, reducing memory usage while preserving generation quality (Section 3.4).

3.2 Retrieval Head Identification

Attention heads in LLMs can be categorized into retrieval heads and streaming heads Wu et al. (2024). Retrieval heads attend broadly to historical tokens for long-context understanding, whereas

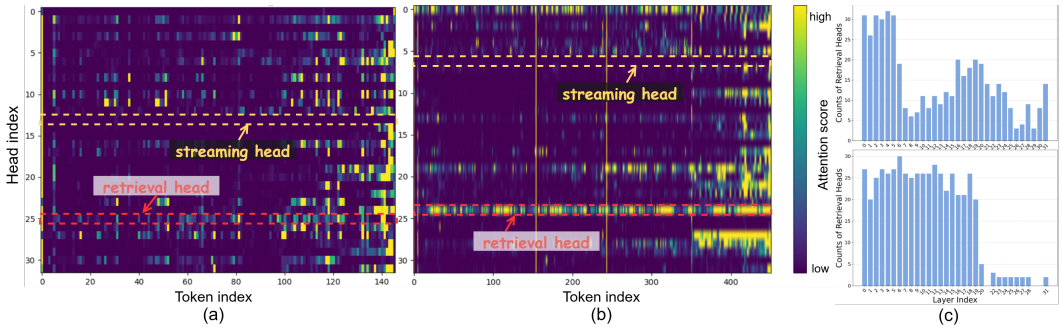


Figure 3: (a) Attention map in the text generation stage. The horizontal and vertical axes denote previous-token indices and head indices, respectively, with each entry (i, j) representing the attention score of head j to token i. (b) Attention map in the image generation stage. (c) Layer-wise distribution of retrieval heads for text generation (top) and image generation (bottom).

Algorithm 1: Retrieval Head Identification

Input: Model with H attention heads in total across all layers; random prompt set $\mathcal{S}$; retrieval budget k; modality mapper $m(\cdot) \in \{\text{text}, \text{image}\}$

Output: Global text and image retrieval-head sets R_t and R_i, where each element is the head index h

Initialize $\mathcal{Z}_\mu[h] \leftarrow \emptyset$ for all $h \in [1, H]$ and $\mu \in \{\text{text}, \text{image}\}$

foreach *prompt* $x \in \mathcal{S}$ **do**

  Run autoregressive generation on x

  foreach *generated token* y *at sequence length* l **do**

    $\mu \leftarrow m(y)$ // $\mu \in \{\text{text}, \text{image}\}$

    for $h = 1$ **to** H **do**

      $\{a_i^h\}_{i=0}^{l-1} \leftarrow \text{Softmax}\left(\frac{Q^h {K^h}^\top}{\sqrt{d_h}}\right)$ // Compute attention scores

      $s_l^h \leftarrow \text{std}\left(\{a_i^h\}_{i=0}^{l-1}\right) \frac{l}{\sqrt{l-1}}$ // Compute normalized std

      Append s_l^h to $\mathcal{Z}_\mu[h]$

for $h = 1$ **to** H **do**

  $s^{h\text{-}t} \leftarrow \text{mean}(\mathcal{Z}_{\text{text}}[h])\,, \quad s^{h\text{-}i} \leftarrow \text{mean}(\mathcal{Z}_{\text{image}}[h])$

$R_t \leftarrow \text{Top-}k\left(\{-(s^{h\text{-}t})\}_{h=1}^{H}\right), \quad R_i \leftarrow \text{Top-}k\left(\{-(s^{h\text{-}i})\}_{h=1}^{H}\right)$

return R_t, R_i

streaming heads mainly focus on recent tokens and attention sinks. To determine whether TwGI also exhibits these head types, we visualize the attention maps of its 21-st layer. As shown in Figure 3(a) and Figure 3(b), certain heads attend to global tokens, whereas others focus on specific positions, confirming the presence of retrieval heads and streaming heads in TwGI. Moreover, retrieval heads differ between text and image generation stages, and identifying them in advance facilitates more targeted KV-cache eviction strategies. Since existing methods are ill-suited to TwGI Wu et al. (2024); Xiao et al. (2024), we propose a training-free and efficient identification algorithm tailored to TwGI.

We first feed a set of randomly sampled inputs into TwGI and compute a retrieval score for each head separately for each modality to identify modality-specific retrieval heads. At each generation step, the attention scores Vaswani et al. (2017) are computed as $\{a_i^h\}_{i=0}^{l-1} = \text{Softmax}\left(Q^h {K^h}^\top / \sqrt{d_h}\right)$, where l is the current token sequence length, d_h is the head dimension, Q^h and K^h are the query and key of head h, and a_i^h denotes the attention score of head h for the i-th token.

When a head attends to global tokens, its attention score distribution $\{a_i^h\}_{i=0}^{l-1}$ tends to be uniform. We therefore introduce the standard deviation to measure its uniformity. However, the standard deviation naturally decreases as the sequence length increases, causing short sequences at early generation steps to disproportionately affect retrieval head assessment. To mitigate this effect, we derive an upper bound on the standard deviation of attention score distribution for a length-l sequence and use it for normalization. Specifically, the maximum standard deviation is $\frac{\sqrt{l-1}}{l}$, with the detailed proof provided in **Appendix A**. The normalized standard deviation is computed as follows:

$$s_l^h = \frac{\text{std}(\{a_i^h\}_{i=0}^{l-1})}{\text{std}_{\max}(\{a_i^h\}_{i=0}^{l-1})} = \text{std}(\{a_i^h\}_{i=0}^{l-1}) \frac{l}{\sqrt{l-1}}, \tag{1}$$

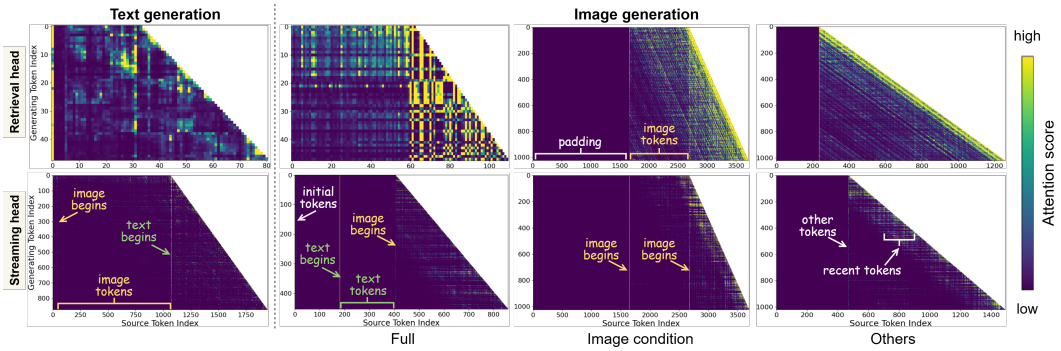


Figure 4: Visualization of attention heatmaps during text and image generation. Retrieval and streaming heads are visualized separately for each modality, along with the attention patterns of different CFG samples during image generation. The 'Others' category includes the unconditional setting, negative condition, and original prompt condition.

where s_l^h measures how uniformly head h attends to tokens in the sequence of length l. At each generation step, we compute s_l^h for every head and store it in a modality-specific database based on the generated token's modality. After generating samples in the random sample set, we average the normalized standard deviations of head h in each modality-specific database to obtain the retrieval scores $s^{h\text{-}t}$ and $s^{h\text{-}i}$. Lower scores indicate more uniform global attention and a greater likelihood of being a retrieval head. We identify the k lowest-scoring heads for each modality as follows:

$$\begin{aligned} R_t &= \text{Top-}k(\{-(s^{h\text{-}t})\}_{h=1}^{H}) = \text{Top-}k(\{-\text{mean}(\{s_l^h \mid s_l^h \in \mathcal{Z}_{\text{text}}[h]\})\}_{h=1}^{H}), \\ R_i &= \text{Top-}k(\{-(s^{h\text{-}i})\}_{h=1}^{H}) = \text{Top-}k(\{-\text{mean}(\{s_l^h \mid s_l^h \in \mathcal{Z}_{\text{image}}[h]\})\}_{h=1}^{H}), \end{aligned} \tag{2}$$

where H is the total number of attention heads, and $\mathcal{Z}_{\text{text}}[h]$ and $\mathcal{Z}_{\text{image}}[h]$ are the databases of head h for text and image generation, respectively. We determine Top-k by attention-score coverage and set the default threshold to $\eta = 0.90$. Negating the scores allows Top-k to select the k lowest-scoring heads, yielding the retrieval-head sets R_t and R_i for each modality. Figure 3(c) shows distinct distributions of R_t and R_i, indicating that retrieval heads are not shared across modalities. Algorithm 1 summarizes the identification process. Furthermore, TwGI involves two generative models stemming from training differences: subgoal and critique. We also observe different retrieval-head sets between models, even within the same modality. During image generation, however, the multiple conditional samples introduced by CFG share the same retrieval heads throughout generation. Detailed validation and more analysis are provided in Section 4.4 and **Appendix B**, respectively.

3.3 Attention Patterns

To design the eviction strategy for Twilight, we analyze attention heatmaps throughout token generation to identify specific attention patterns, as shown in Figure 4. TwGI exhibits attention patterns distinct from those of conventional autoregressive LLMs Xiao et al. (2024). During text generation in the left panel, retrieval heads consistently maintain strong attention to all historical tokens, while streaming heads attend strongly to initial tokens, recent tokens, and modality boundary tokens, i.e., the starting and ending tokens of each modality segment. Accordingly, we designate initial tokens and modality boundary tokens as attention sinks Xiao et al. (2023) for streaming heads during text generation to preserve critical contextual anchors during cache eviction.

The right panel shows the attention heatmaps during image generation. Since CFG introduces multiple conditional samples, we analyze each condition separately. **Full** denotes the full condition, preserving the original token sequence such that the reasoning chain preceding image generation ends with text tokens. Retrieval heads attend globally to all historical tokens, while streaming heads treat initial tokens and modality boundary tokens as attention sinks and also focus on recent tokens. **Image condition** retains only image tokens from the preceding reasoning chain and uses padding tokens to maintain the full input length. Retrieval heads attend exclusively to historical and newly generated image tokens while ignoring padding tokens. Streaming heads focus solely on modality boundary tokens and recent tokens, excluding initial tokens. **Others** includes three conditions that replace the original reasoning chain with substitute tokens. Retrieval heads attend only to newly generated image tokens and inserted substitute tokens, while streaming heads focus on inserted tokens and recent tokens, forming compact attention patterns adapted to the modified context. Furthermore, these attention patterns remain stable across generative models, as shown in Figure 5.

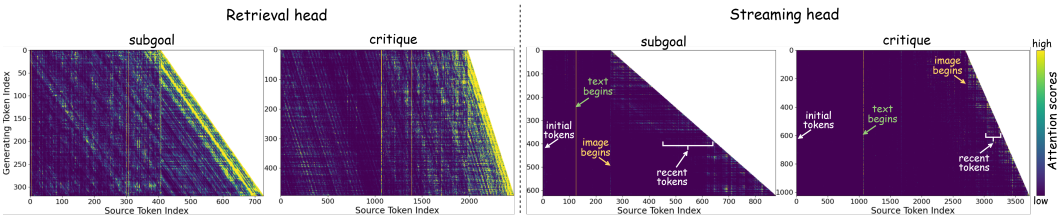


Figure 5: Attention heatmaps across generative models. Regardless of the models, retrieval heads attend to global tokens, whereas streaming heads focus only on specific token positions.

3.4 KV Cache Eviction and Sparse Attention

KV cache eviction retains high-attention tokens for subsequent computation, preserving the principal components of the attention matrix needed to reconstruct the semantic structure of the original attention distribution. As shown in Figure 4, numerous low-attention tokens introduce substantial KV cache redundancy during text-image generation. Based on the analysis in Section 3.2 and 3.3, we propose an eviction strategy and implement sparse attention. During text generation, the retrieval heads and streaming heads in each layer L_i are defined as follows:

$$R_t^{L_i} = \{h_n \in R_t \mid h_n \in L_i\}, \quad S_t^{L_i} = \{h_n \in H \setminus R_t \mid h_n \in L_i\}, \tag{3}$$

where $R_t^{L_i}$ and $S_t^{L_i}$ are retrieval heads and streaming heads in layer L_i. The full KV cache is retained for retrieval heads to preserve global context, while streaming heads keep only initial tokens, modality boundary tokens, and recent tokens. Attention for each head in L_i is computed as

$$\text{Attn}_h(Q,K,V,M) = \begin{cases} \text{softmax}\left(\dfrac{Q_h K_R^{(h)\top}}{\sqrt{d_h}} + M_R^{(h)}\right) V_R^{(h)} & \text{if } h \in R_t^{L_i}, \\ \text{softmax}\left(\dfrac{Q_h K_S^{(h)\top}}{\sqrt{d_h}} + M_S^{(h)}\right) V_S^{(h)} & \text{if } h \in S_t^{L_i}, \end{cases} \tag{4}$$

where Q_h denotes the query for head h, and d_h is the head dimension. $K_R^{(h)}, V_R^{(h)}$ are full-length key-value pairs for retrieval heads, while $K_S^{(h)}, V_S^{(h)}$ are sparse key-value pairs for streaming heads. $M_R^{(h)}$ and $M_S^{(h)}$ are attention masks dynamically sliced to match their respective key sequence lengths. The final output of L_i is assembled via position-wise assignment:

$$\text{Attn_out}_{L_i}[:,h,:,:] = \text{Attn}_h(Q,K,V,M). \tag{5}$$

During image generation, for each layer L_i, we obtain the retrieval-head set $R_i^{L_i}$ and streaming-head set $S_i^{L_i}$ according to Equ. 3. Since Classifier-Free Guidance (CFG) is employed, we configure the KV caches of retrieval and streaming heads separately for each CFG conditional sample based on the attention patterns in Section 3.3, retaining high-attention tokens to preserve critical context. For the c-th CFG sample, attention at each head in L_i is computed as follows:

$$\text{Attn}_h(Q,K,V,M,c) = \begin{cases} \text{softmax}\left(\dfrac{Q_h[c] K_R^{(h)}[c]^\top}{\sqrt{d_h}} + M_R^{(h)}[c]\right) V_R^{(h)}[c] & \text{if } h \in R_i^{L_i}, \\ \text{softmax}\left(\dfrac{Q_h[c] K_S^{(h)}[c]^\top}{\sqrt{d_h}} + M_S^{(h)}[c]\right) V_S^{(h)}[c] & \text{if } h \in S_i^{L_i}, \end{cases} \tag{6}$$

where c indexes the CFG conditional samples. The subgoal generative model uses three CFG samples: full-condition, unconditional, and negative-condition samples, In contrast, the critique generative model employs all CFG samples. The attention output at L_i is then given by

$$\text{Attn_out}_{L_i}[c,h,:,:] = \text{Attn}_h(Q,K,V,M,c). \tag{7}$$

Because retrieval heads vary across generative models, we apply model-specific retrieval-head and streaming-head sets. This heterogeneous attention mechanism preserves comprehensive context awareness in retrieval heads while reducing memory usage in streaming heads without sacrificing critical contextual anchors. The detailed algorithm of sparse attention is provided in **Appendix C**.

Table 1: The results on GenEval benchmark. We reproduce KeyDiff and R-KV on TwGI. (vh.) denotes the initial visual hypothesis, and (f.) indicates the final results.

Model	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall↑
LlamaGen Sun et al. (2024)	0.71	0.34	0.21	0.58	0.07	0.04	0.32
Chameleon-7b Team (2024)	-	-	-	-	-	-	0.39
Emu3 Wang et al. (2024)	0.95	0.56	0.30	0.76	0.09	0.10	0.46
LWM Liu et al. (2025)	0.93	0.41	0.46	0.79	0.09	0.15	0.47
SEED-X Liu et al. (2025)	0.96	0.65	0.31	0.80	0.18	0.14	0.51
janus_flow Ma et al. (2025)	0.94	0.46	0.28	0.75	0.32	0.25	0.50
mmada Yang et al. (2026)	0.89	0.50	0.31	0.74	0.13	0.21	0.46
Anole-7b Chern et al. (2024)	-	0.38	-	-	0.04	0.04	-
TwGI-Subgoal-7b Chern et al. (2025)	-	0.57	-	-	0.11	0.16	-
Twilight-Subgoal	-	0.59	-	-	0.12	0.18	-
TwGI-Critique-7b(vh.) Chern et al. (2025)	0.98	0.51	0.24	0.84	0.06	0.08	0.45
Twilight-Critique(vh.)	0.98	0.53	0.24	0.85	0.08	0.08	0.46
TwGI-Critique-7b(f.) Chern et al. (2025)	0.98	0.59	0.23	0.87	0.12	0.10	0.48
KeyDiff Park et al. (2026)	0.35	0.04	0.03	0.30	0.00	0.00	0.12
R-KV Cai et al. (2026)	0.44	0.06	0.04	0.38	0.03	0.02	0.16
Twilight-Critique(f.)	0.98	0.63	0.24	0.88	0.18	0.12	**0.51**

Table 2: The results on DPG-Bench. ρ_{KV} denotes the average KV-cache budget at the final decoding step over all samples. Twilight achieves superior generation performance with reduced KV cache.

Model	Global	Entity	Attribute	Relation	Other	Overall↑	ρ_{KV} ↓
Anole-7b Chern et al. (2024)	74.16	70.31	69.42	80.12	43.60	58.32	100%
TwGI-Subgoal-7b Chern et al. (2025)	74.47	78.06	76.17	83.21	42.00	68.44	100%
KeyDiff Park et al. (2026)	58.62	46.50	59.13	71.71	20.83	27.73	54.1%
R-KV Cai et al. (2026)	58.62	48.65	62.24	73.71	29.17	33.95	64.5%
Twilight-Subgoal	74.44	79.45	76.52	82.98	45.00	**68.56**	**29.2%**
TwGI-Critique-7b(vh.) Chern et al. (2025)	77.81	71.45	69.24	78.34	58.00	62.83	100%
Twilight-Critique(vh.)	77.85	72.65	70.29	78.04	60.00	63.10	100%
TwGI-Critique-7b(f.) Chern et al. (2025)	76.90	75.70	73.88	81.86	55.60	67.14	100%
KeyDiff Park et al. (2026)	55.17	33.03	50.83	64.14	12.50	20.34	60.0%
R-KV Cai et al. (2026)	62.07	46.50	59.96	70.92	25.00	28.81	69.7%
Twilight-Critique(f.)	76.92	76.11	75.04	81.59	57.00	**67.39**	**35.3%**

4 Experiment

4.1 Benchmarks and Implementation Details

We evaluate Twilight's image generation performance and efficiency on GenEval Ghosh et al. (2023) and DPG-Bench Hu et al. (2024), and provide an in-depth analysis of each module's impact. GenEval examines images from six aspects, and DPG-Bench evaluates various attributes under complex prompts. We implement Twilight with RoPE Su et al. (2024) on 8 NVIDIA A800 GPUs, and apply chunked prefill by Flashinfer Ye et al. (2025). Further details are provided in **Appendix D**.

4.2 Comparison on Generation Performance

We evaluate the performance of Twilight in two generative models on GenEval and DPG-Bench, with results presented in Tables 1 and 2. To validate Twilight's superiority, we reproduce other advanced eviction methods on TwGI. Subgoal and critique represent distinct reasoning models. Since some GenEval categories require only a single object, we omit them in the subgoal model. On GenEval, Twilight improves most metrics over the baseline, demonstrating enhanced spatial localization and visual compositional reasoning. For the critique model, the 'Overall' score increases from 0.48 to 0.51 with 37.5% of the original KV-cache memory. On DPG-Bench, Twilight achieves superior overall performance while eliminating the greatest amount of KV-cache redundancy. In contrast, other eviction methods struggle to retain critical information, resulting in substantial performance degradation. Twilight maintains performance across models, demonstrating the generalizability of our approach. In summary, our method can maintain visual generation performance while reducing KV cache redundancy. More experimental analysis will be provided in **Appendix E**.

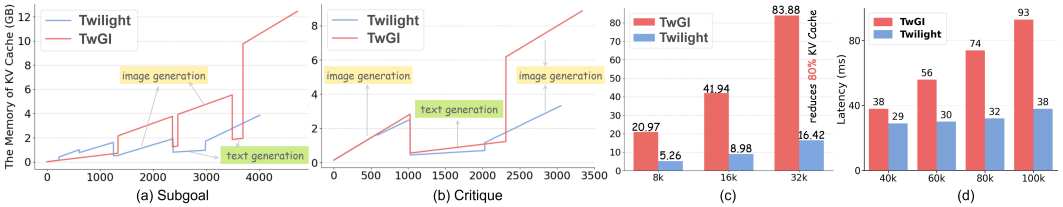


Figure 6: Comparison of KV cache memory and latency between TwGI and Twilight. (a) KV cache memory dynamics in subgoal model. (b) KV cache memory dynamics in critique model. (c) KV cache memory during long-sequence generation. (d) Latency during long-sequence generation.

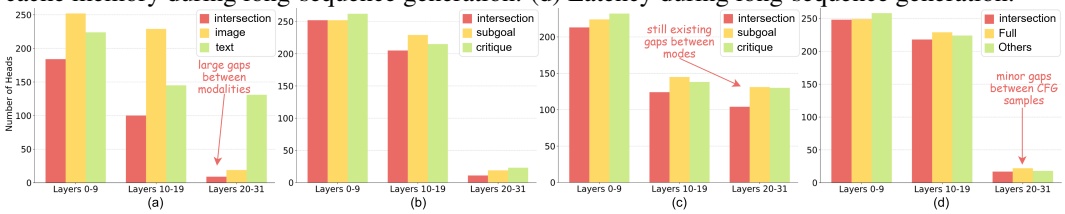


Figure 7: Analysis of retrieval heads distribution. (a) Retrieval-head distributions across modalities, showing significant cross-modal differences. (b) Retrieval head distribution across models during image generation. (c) Retrieval head distribution across models during text generation, with notable inter-model differences (d) Retrieval head distribution across CFG samples shows minor gaps.

4.3 Comparison on KV Cache Memory and Latency

We evaluate the effectiveness of Twilight's KV cache eviction in Figure 6. Across generative models, Twilight significantly reduces KV cache storage. Alternating text-image generation and the introduction of CFG cause a step-like memory pattern, and Twilight exhibits a much smaller amplitude. Moreover, the memory gap between TwGI and Twilight becomes increasingly pronounced as generation progresses, while Twilight's lower growth rate demonstrates its suitability for long-sequence generation. Under longer sequences, as shown in Figure 6(c), Twilight consistently maintains less KV cache memory, achieving an 80% reduction at 32K tokens. Figure 6(d) compares per-token decoding latency during image generation under subgoal model. Twilight exhibits slower latency growth with increasing context length, achieving a 2.46× speedup at 100K tokens. These results show that Twilight's eviction strategy substantially reduces memory usage and delivers stable speedups, highlighting its scalability for long-sequence generation.

4.4 In-depth Analysis of Retrieval Heads

Table 3: Ablation results of retrieval heads in critique model.

Setting	Two Obj.	Position	Color Attri.
Twilight	0.63	0.18	0.12
$R_t = R_i$	0.60	0.11	0.08
$R_{\text{Sub}} \to R_{\text{Crit}}$	0.61	0.15	0.11
$R_i[c_j] \neq R_i[c_k]$	0.63	0.16	0.14

Retrieval heads are critical for long-context understanding, motivating an in-depth analysis. Figure 7(a) shows significant differences across modalities, while the second row of Table 3 confirms that sharing retrieval heads across modalities degrades generation quality. Figures 7(b) and 7(c) reveal that retrieval head distributions also vary across generative models, particularly during text generation. The third row of Table 3 further corroborates the drawback of sharing retrieval heads across models. In contrast, Figure 7(d) shows substantial overlap across CFG samples, and the last row of Table 3 further demonstrates that employing CFG-sample-specific retrieval heads does not improve performance. To validate the configuration of retrieval heads, we conduct ablation experiments in subgoal model, as reported in Table 4, Figure 8, and Figure 9. Thresholding at η yields 476 and 523 heads for text and image, and we set Top-$k = 500$ to balance quality and efficiency. Twilight's performance is sensitive to Top-k below 500, where insufficient retrieval heads provide inadequate coverage, but stabilizes or declines thereafter. More experimental analysis can be found in **Appendix B**.

4.5 In-depth Analysis of Attention Patterns

Due to distinct attention patterns, we conduct ablation studies to investigate the role of each component. As shown in Table 5, removing either modality boundary tokens or initial tokens degrades performance. Figure 10(a) also reveals that removing both tokens leads to poor instruction-following capability. Reducing the recent token length from 512 to 128 degrades image quality, as shown in Figure 10(b), indicating impaired ability to capture local information. Table 6 analyzes the attention

Table 4: Analysis of the scale of retrieval heads.

Top-k	Two Obj.	Position	Color Attri.
k=0	0.24	0.02	0.04
k=100	0.46	0.07	0.10
k=300	0.57	0.10	0.15
k=500	0.59	0.12	0.18
k=700	0.58	0.13	0.17

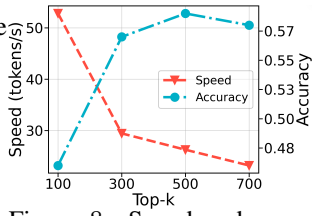
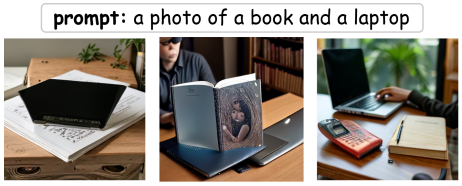


Figure 8: Speed and accuracy across Top-k.



(a) k=0 (b) k=100 (c) k=500

Figure 9: The visual results under different scales of retrieval heads.

Table 5: Ablation of streaming heads' attention pattern. 'mdt.' is modality boundary tokens, and 'it.' means initial tokens. 'rt.' is recent tokens and '→' is length change.

Setting	Two Obj.	Position	Color Attri.
Twilight	0.59	0.12	0.18
w/o mdt.	0.52	0.09	0.13
w/o it.	0.41	0.04	0.07
w/o mdt., it.	0.35	0.04	0.05
rt.(512→128)	0.57	0.10	0.17

Table 6: Ablation of retrieval heads' attention pattern. '↓' is halving retrieval head's original receptive field, and '↑' denotes expanding receptive field to all tokens.

Setting	Two Obj.	Position	Color Attri.
Twilight	0.63	0.18	0.12
Full ↓	0.05	0.01	0.01
Image condition↓	0.59	0.15	0.12
Others ↓	0.45	0.08	0.06
all CFG samples↑	0.62	0.17	0.13



Figure 10: Visualization of different settings.

patterns of retrieval head under critique model. Reducing the receptive field for full-condition samples sharply degrades generation quality and instruction following capability, which is also demonstrated in Figure 10(c). Similar reductions for image-condition and other-condition samples also lead to performance degradation, and Figure 10(d) shows the poorer image quality with shrinking receptive field of other-condition samples. These results indicate that the retrieval heads' receptive field is critical for long-context understanding and high-quality generation. Furthermore, expanding the receptive field of retrieval heads across all CFG samples does not improve performance, suggesting that retrieval heads need not attend to additional padding tokens.

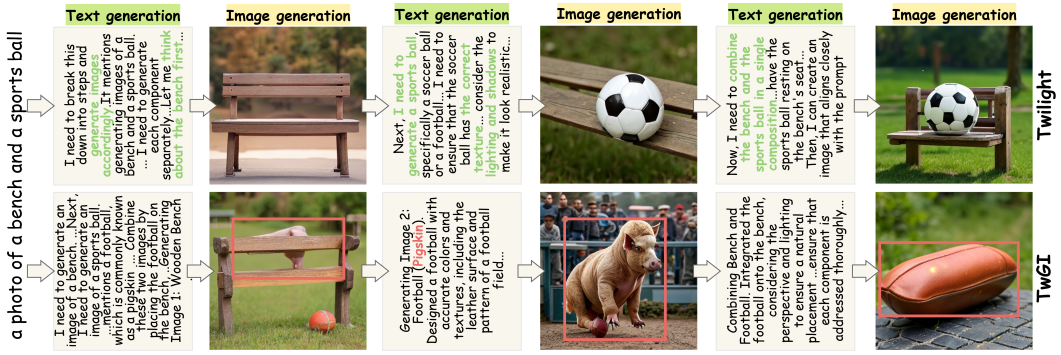


Figure 11: Visualizations of reasoning processes in Twilight and TwGI.

4.6 Visual Analysis

We visualize the reasoning processes of Twilight and TwGI under subgoal model, with results presented in Figure 11. Specifically, Twilight decomposes prompts into multiple generation subgoals during reasoning and combines these generated subgoals as the final output to enhance generation quality. Due to retaining critical tokens, Twilight maintains the original reasoning logic and generation quality even after eliminating substantial KV cache. In contrast, TwGI's generation may deviate from the intended target due to redundant tokens that disrupt attention focus and induce dispersion, with results shown in Figure 11(b). More visualization and analysis are provided in **Appendix F**.

5 Conclusion

In this paper, we propose Twilight, a training-free eviction method tailored for TwGI to compress KV cache memory while maintaining generation quality. By introducing normalized standard deviation, we construct a novel retrieval head identification algorithm. Meanwhile, we analyze the unique attention patterns during alternating text-image generation and design a KV cache eviction algorithm to eliminate redundant tokens. Extensive experiments validate that Twilight can significantly reduce KV cache memory while maintaining or even surpassing TwGI's generation performance.

REFERENCES

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 preprint arXiv:2511.21631*, 2025.

Zefan Cai, Wen Xiao, Hanshi Sun, Yikai Zhang, Ke Wan, Yucheng Li, Yeyang Zhou, Li-Wen Chang, Jiuxiang Gu, Zhen Dong, et al. R-kv: Redundancy-aware kv cache compression for reasoning models. *Advances in neural information processing systems*, 38:60980–61005, 2026.

Xinyan Chen, Renrui Zhang, Dongzhi Jiang, Aojun Zhou, Shilin Yan, Weifeng Lin, and Hongsheng Li. Mint-cot: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning. *arXiv preprint arXiv:2506.05331*, 2025.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1290–1299, 2022.

Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*, 2024.

Ethan Chern, Zhulin Hu, Steffi Chern, Siqi Kou, Jiadi Su, Yan Ma, Zhijie Deng, and Pengfei Liu. Thinking with generated images. *arXiv preprint arXiv:2505.22525*, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, et al. Sensenova-u1: Unifying multimodal understanding and generation with neo-unify architecture. *arXiv preprint arXiv:2605.12500*, 2026.

Chengqi Duan, Rongyao Fang, Yuqing Wang, Kun Wang, Linjiang Huang, Xingyu Zeng, Hongsheng Li, and Xihui Liu. Got-r1: Unleashing reasoning capability of mllm for visual generation with reinforcement learning. *arXiv preprint arXiv:2505.17022*, 2025.

Yuan Feng, Junlin Lv, Yukun Cao, Xike Xie, and S Kevin Zhou. Ada-kv: Optimizing kv cache eviction by adaptive budget allocation for efficient llm inference. *Advances in Neural Information Processing Systems*, 38:113152–113188, 2026.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Jiawei Gu, Yunzhuo Hao, Huichen Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. Thinkmorph: Emergent properties in multimodal interleaved chain-of-thought reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 141405–141447, 2026.

Haoyu Guo, Junlin Lv, S Kevin Zhou, Xike Xie, et al. Defensivekv: Taming the fragility of kv cache eviction in llm inference. In *International Conference on Learning Representations*, volume 2026, pp. 43407–43432, 2026a.

Zirun Guo, Minjie Hong, Feng Zhang, Kai Jia, and Tao Jin. Thinking with programming vision: Towards a unified view for thinking with images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33467–33476, 2026b.

LI Haoyang, Yiming Li, Anxin Tian, Tianhao Tang, Zhanchao Xu, Xuejia Chen, HU Nicole, Wei Dong, Li Qing, and Lei Chen. A survey on large language model acceleration based on kv cache management. *Transactions on Machine Learning Research*, 2025.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. Deepeyesv2: Toward agentic multimodal model. *arXiv preprint arXiv:2511.05271*, 2025.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Jiantong Jiang, Peiyu Yang, Rui Zhang, and Feng Liu. Towards efficient large language model serving: A survey on system-aware kv cache optimization. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 38450–38476, 2026.

Siyu Jiao, Yiheng Lin, Yujie Zhong, Qi She, Wei Zhou, Xiaohan Lan, Zilong Huang, Fei Yu, Yingchen Yu, Yunqing Zhao, et al. Thinkgen: Generalized thinking for visual generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14713–14723, 2026.

Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. *arXiv preprint arXiv:2509.07969*, 2025.

Ang Li, Charles Wang, Deqing Fu, Kaiyu Yue, Zikui Cai, Wang Bill Zhu, Ollie Liu, Peng Guo, Willie Neiswanger, Furong Huang, et al. Zebra-cot: A dataset for interleaved vision language reasoning. *arXiv preprint arXiv:2507.16746*, 2025a.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025b.

Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, He Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 7241–7259, 2022.

Liupeng Li, Haoqian Kang, Zhenyu Lu, Jinpeng Wang, Bin Chen, Ke Chen, and Yaowei Wang. Cvsearch: Empowering multimodal llms with cognitive visual search for high-resolution image perception. *arXiv preprint arXiv:2605.23655*, 2026a.

Yangfu Li, Hongjian Zhan, Jiawei Chen, Yuning Gong, Qi Liu, and Yue Lu. Deepscan: A training-free framework for visually grounded reasoning in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19155–19164, 2026b.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation. *Advances in Neural Information Processing Systems*, 37:22947–22970, 2024.

Zejun Li, Yingxiu Zhao, Jiwen Zhang, Siyuan Wang, Yang Yao, Runzhou Zhao, Jun Song, Bo Zheng, et al. Mixture-of-visual-thoughts: Exploring context-adaptive reasoning mode selection for general visual reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 81981–82016, 2026c.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention. In *International Conference on Learning Representations*, volume 2025, pp. 45953–45977, 2025.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7739–7751. IEEE, 2025.

Junyoung Park, Dalton Jones, Matthew Morse, Raghavv Goel, Mingu Lee, and Christopher Lott. Keydiff: Key similarity-based kv cache eviction for long-context llm inference in resource-constrained environments. *Advances in Neural Information Processing Systems*, 38:5983–6019, 2026.

Runqi Qiao, Qiuna Tan, Minghan Yang, Guanting Dong, Peiqing Yang, Shiqiang Lang, Enhui Wan, Xiaowan Wang, Yida Xu, Lan Yang, et al. V-thinker: Interactive thinking with images. *arXiv preprint arXiv:2511.04460*, 2025.

Luozheng Qin, GONG JIA, Yuqing Sun, Tianjiao Li, Haoyu Pan, Mengping Yang, Xiaomeng Yang, Chao Qu, Zhiyu Tan, and Hao Li. Uni-cot: Towards unified chain-of-thought reasoning across text and vision. In *International Conference on Learning Representations*, volume 2026, pp. 15999–16028, 2026.

Ziran Qin, Yuchen Cao, Mingbao Lin, Wen Hu, Shixuan Fan, Ke Cheng, Weiyao Lin, and Jianguo Li. Cake: Cascading and adaptive kv cache eviction with layer preferences. In *International Conference on Learning Representations*, volume 2025, pp. 89945–89977, 2025.

Weikang Shi, Aldrich Yu, Rongyao Fang, Houxing Ren, Ke Wang, Aojun Zhou, Changyao Tian, Xinyu Fu, Yuxuan Hu, Zimu Lu, et al. Mathcanvas: Intrinsic visual chain-of-thought for multimodal mathematical reasoning. *arXiv preprint arXiv:2510.14958*, 2025.

Zijia Song, Yelin Wang, Zhengyi Ma, Zitong Yu, Tianheng Wang, Jiahuan Zhang, Taorui Wang, and Kaicheng Yu. Oracleanalyser: Analysing implicit semantics of oracle bone scripts through mllms with post-training. *arXiv preprint arXiv:2606.25906*, 2026.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, et al. Openthinkimg: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*, 2025a.

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025b.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Chongjun Tu, Peng Ye, Dongzhan Zhou, Tao Chen, and Wanli Ouyang. Mitigating low-quality reasoning in mllms: Self-driven refined multimodal cot with selective thinking and step-wise visual enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9576–9584, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025a.

Pengjie Wang, Linger Deng, Zujia Zhang, Shaojie Zhang, Zhenbo Luo, Pei Fu, Jian Luan, Xiang Bai, and Yuliang Liu. Deltav: Thinking with visual state updates in unified large multimodal models. *arXiv preprint arXiv:2607.08434*, 2026a.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

Ye Wang, Qianglong Chen, Zejun Li, Siyuan Wang, Shijie Guo, Zhirui Zhang, and Zhongyu Wei. Simple o3: Towards interleaved vision-language reasoning. *arXiv preprint arXiv:2508.12109*, 2025b.

Yelin Wang, Zijia Song, Shuo Ye, Chuanguang Yang, Miaoyu Wang, Yong Xu, Zhulin An, Yongjun Xu, and Zitong Yu. Rsicclm: A multimodal large language model for remote sensing image change captioning. *arXiv preprint arXiv:2606.28266*, 2026b.

Junfei Wu, Jian Guan, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tienie Tan. Chatting with images for introspective visual thinking. *arXiv preprint arXiv:2602.11073*, 2026.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. *arXiv preprint arXiv:2404.15574*, 2024.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. Duoattention: Efficient long-context llm inference with retrieval and streaming heads. *arXiv preprint arXiv:2410.10819*, 2024.

Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual planning: Let's think only with images. *arXiv preprint arXiv:2505.11409*, 2025.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *Advances in Neural Information Processing Systems*, 38:138867–138907, 2026.

Zihao Ye, Lequn Chen, Ruihang Lai, Wuwei Lin, Yineng Zhang, Stephanie Wang, Tianqi Chen, Baris Kasikci, Vinod Grover, Arvind Krishnamurthy, et al. Flashinfer: Efficient and customizable attention engine for llm inference serving. *Proceedings of Machine Learning and Systems*, 7, 2025.

Yufei Zhan, Yousong Zhu, Hongyin Zhao, Fan Yang, Shurong Zheng, Ming Tang, and Jinqiao Wang. Vision-r1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5807–5817, 2026.

Chi Zhang, Haibo Qiu, Qiming Zhang, Zhixiong Zeng, Lin Ma, and Jing Zhang. Deepsketcher: Internalizing visual manipulation for multimodal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9389–9400, 2026.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

A Upper Bound on the Standard Deviation of Attention Scores

To obtain a normalized standard deviation for retrieval head identification, we derive the following upper bound on the standard deviation of the attention score distribution.

Lemma 1. *Let* $\mathbf{a} = (a_1, \ldots, a_l)$ *denote an attention score vector of length* l*, where*

$$a_i > 0, \qquad \sum_{i=1}^{l} a_i = 1.$$

Define the standard deviation of $\mathbf{a}$ *as*

$$\sigma(\mathbf{a}) \triangleq \sqrt{\frac{1}{l}\sum_{i=1}^{l}\left(a_i - \frac{1}{l}\right)^2}.$$

Then,

$$\sigma(\mathbf{a}) < \frac{\sqrt{l-1}}{l}.$$

Moreover,

$$\sup_{\mathbf{a}} \sigma(\mathbf{a}) = \frac{\sqrt{l-1}}{l}.$$

If the constraint is relaxed from $a_i > 0$ *to* $a_i \geq 0$*, then the bound is attainable, i.e.,*

$$\max_{\mathbf{a}} \sigma(\mathbf{a}) = \frac{\sqrt{l-1}}{l}.$$

Proof. Since $\sum_{i=1}^{l} a_i = 1$, the mean of the attention scores is

$$\mu = \frac{1}{l}.$$

Therefore, the variance can be written as

$$\sigma^2(\mathbf{a}) = \frac{1}{l}\sum_{i=1}^{l}\left(a_i - \frac{1}{l}\right)^2.$$

Expanding the square yields

$$\sigma^2(\mathbf{a}) = \frac{1}{l}\left(\sum_{i=1}^{l} a_i^2 - \frac{2}{l}\sum_{i=1}^{l} a_i + \sum_{i=1}^{l}\frac{1}{l^2}\right).$$

Using $\sum_{i=1}^{l} a_i = 1$, we obtain

$$\sigma^2(\mathbf{a}) = \frac{1}{l}\left(\sum_{i=1}^{l} a_i^2 - \frac{1}{l}\right).$$

Hence, maximizing $\sigma^2(\mathbf{a})$ is equivalent to maximizing $\sum_{i=1}^{l} a_i^2$ under the simplex constraint.

Now, for any nonnegative $\{a_i\}_{i=1}^{l}$ satisfying $\sum_{i=1}^{l} a_i = 1$, we have

$$\left(\sum_{i=1}^{l} a_i\right)^2 = \sum_{i=1}^{l} a_i^2 + 2\sum_{1\leq i<j\leq l} a_i a_j \geq \sum_{i=1}^{l} a_i^2.$$

Since $\sum_{i=1}^{l} a_i = 1$, it follows that

$$\sum_{i=1}^{l} a_i^2 \leq 1.$$

Substituting this into the variance expression gives

$$\sigma^2(\mathbf{a}) \le \frac{1}{l}\left(1-\frac{1}{l}\right) = \frac{l-1}{l^2}.$$

Therefore,

$$\sigma(\mathbf{a}) \le \frac{\sqrt{l-1}}{l}.$$

If zero entries are allowed, the equality is attained by the one-hot distribution

$$\mathbf{a} = (1, 0, \ldots, 0),$$

for which

$$\sigma(\mathbf{a}) = \sqrt{\frac{1}{l}\left[\left(1-\frac{1}{l}\right)^2 + (l-1)\left(\frac{1}{l}\right)^2\right]} = \frac{\sqrt{l-1}}{l}.$$

Under the strict positivity constraint $a_i > 0$, the one-hot distribution is infeasible, so the equality cannot be attained. However, for any sufficiently small $\varepsilon > 0$, consider

$$a_1 = 1 - (l-1)\varepsilon, \qquad a_2 = \cdots = a_l = \varepsilon.$$

Then $a_i > 0$ and $\sum_{i=1}^{l} a_i = 1$. As $\varepsilon \to 0^+$, this sequence converges to the one-hot distribution, and thus

$$\sigma(\mathbf{a}) \to \frac{\sqrt{l-1}}{l}.$$

Hence,

$$\sup_{\mathbf{a}} \sigma(\mathbf{a}) = \frac{\sqrt{l-1}}{l}.$$

This completes the proof. □

B More Analysis of Retrieval Heads

B.1 Why Existing Identification Methods Are Ill-Suited to Thinking with Generated Images

Existing retrieval-head identification methods are primarily designed for text-only long-context models and cannot be directly transferred to Thinking with Generated Images paradigm.

The Needle-in-a-Haystack approach Wu et al. (2024) identifies retrieval heads according to their token-level copying behavior. Specifically, a head receives a high retrieval score when its most-attended input token belongs to the inserted needle and exactly matches the token currently generated in the answer. This criterion relies on the discrete and lexically interpretable correspondence between input and output text tokens. Such a criterion is difficult to extend to image generation because image tokens represent distributed visual latents rather than independent semantic units. Semantically equivalent images may correspond to different token sequences, making exact token matching an unreliable indicator of visual retrieval. Moreover, inserting arbitrary visual tokens may create invalid image contexts, while text-based needle tasks primarily identify heads activated during text decoding and overlook image-specific retrieval heads.

Optimization-based identification methods, such as DuoAttention Xiao et al. (2024), assign a trainable gate to each KV head and optimize these gates on synthetic passkey data to preserve the hidden states of the full-attention model while encouraging more heads to use streaming attention. However, this procedure relies on a specially constructed passkey-retrieval dataset and requires additional optimization with both full and streaming attention. Extending it to TwGI would require multimodal synthetic tasks that explicitly depend on distant visual evidence, together with suitable supervision for image-token generation. A single set of optimized gates may also be insufficient because TwGI exhibits different head behaviors across text and image generation stages, and classifier-free guidance conditions. Performing separate optimization for these settings would further multiply the computational and memory overhead, particularly because image generation contains substantially

more tokens and multiple CFG samples. Therefore, directly applying existing identification methods would require non-trivial multimodal task construction and repeated optimization. In contrast, our normalized attention-statistics criterion identifies modality-specific retrieval heads from their native attention patterns on a small set of randomly sampled prompts, without synthetic visual needles, additional supervision, or gradient-based optimization.

Algorithm 2: Automatic Top-k Selection by Attention-Score Coverage

Input: Model with H attention heads; random sample set $\mathcal{S}$; retrieval scores $\{r_h^\mu\}_{h=1}^H$ for $\mu \in \{\text{text}, \text{image}\}$; search range $[n_{\min}, n_{\max}]$; coverage threshold η

Output: Modality-specific retrieval-head numbers k^*_{text} and k^*_{image}

Sort $\{r_h^\mu\}_{h=1}^H$ in descending order to obtain π_μ for each modality μ;
Initialize $Z_\mu[n] \leftarrow 0$ and $N_\mu \leftarrow 0$ for all $n \in [n_{\min}, n_{\max}]$ and $\mu \in \{\text{text}, \text{image}\}$;
foreach *prompt* $x \in \mathcal{S}$ **do**
  Run full-cache autoregressive generation on x;
  foreach *generated token* y_t **do**
    $\mu \leftarrow m(y_t)$;
    Obtain the original full-attention scores $\{a_{t,j}^h\}_{h=1,j=0}^{H,l_t-1}$;
    Construct $P_t^{\text{stream}} \leftarrow P^{\text{init}} \cup P_t^{\text{boundary}} \cup P_t^{\text{recent}}$;
    for $n = n_{\min}$ **to** $n_{\max}$ **do**
      $R_\mu(n) \leftarrow \{\pi_\mu(1), \ldots, \pi_\mu(n)\}$;
      $C_t^\mu(n) \leftarrow \frac{1}{H}\left[n + \sum_{h \notin R_\mu(n)} \sum_{j \in P_t^{\text{stream}}} a_{t,j}^h\right]$;
      $Z_\mu[n] \leftarrow Z_\mu[n] + C_t^\mu(n)$;
    $N_\mu \leftarrow N_\mu + 1$;
foreach $\mu \in \{\text{text}, \text{image}\}$ **do**
  $\overline{C}^\mu(n) \leftarrow Z_\mu[n]/N_\mu$ for all $n \in [n_{\min}, n_{\max}]$;
  $k_\mu^* \leftarrow \min\{n \mid \overline{C}^\mu(n) \geq \eta\}$;
return $k^*_{\text{text}}, k^*_{\text{image}}$;

B.2 Determining Top-k via Attention-Score Coverage

We determine the number of retrieval heads according to the attention mass preserved by the resulting sparse attention pattern. For each modality $\mu \in \{\text{text}, \text{image}\}$, we sort all heads in descending order according to their retrieval scores $r_h^\mu = -s^{h\text{-}\mu}$. Given a candidate number n, the top-n heads form the retrieval-head set $R_\mu(n)$ and retain their complete KV cache, whereas the remaining heads retain only initial, modality-boundary, and recent tokens. At generation step t, the corresponding attention-score coverage is

$$C_t^\mu(n) = \frac{1}{H}\left[n + \sum_{h \notin R_\mu(n)} \sum_{j \in P_t^{\text{stream}}} a_{t,j}^h\right], \tag{8}$$

where P_t^{stream} denotes the token positions retained by streaming heads, and $a_{t,j}^h$ is the original full-attention score of head h on token j before sparse re-normalization. The first term follows from the fact that every retrieval head retains an attention mass of one. During image generation, $C_t^{\text{image}}(n)$ is additionally averaged over all CFG conditional samples.

Table 7: Jaccard similarity of multiple retrieval heads.

Modality	Jaccard similarity
Text	97.11%
Image	95.44%

We average Eq. 8 over the random sample set and select the smallest n whose average coverage reaches a predefined threshold η, as summarized in Algorithm 2. The coverage criterion yields $k^*_{\text{text}} = 476$ and $k^*_{\text{image}} = 523$, identifying a narrow operating range around 500 heads. We therefore set $k = 500$ for both modalities to balance generation quality and efficiency. Meanwhile, to assess robustness of retrieval-head identification to sample choice, we randomly generate multiple sample sets independent of the evaluation set for retrieval-head identification, and report the average pairwise Jaccard similarity to measure differences. As shown in Table 7, the high similarity for both modalities indicates

Table 8: Ablation results on different settings of retrieval heads.

Setting	Single Obj.	Two obj.	Counting	Colors	Position	Color Attri.	Overall
Twilight	0.98	0.63	0.24	0.88	0.18	0.12	0.51
$R_t = R_i$	0.96	0.60	0.21	0.83	0.11	0.08	0.46
$R_{Subgoal} \rightarrow R_{Critique}$	0.98	0.61	0.23	0.86	0.15	0.11	0.49
$R_i[c_j] \neq R_i[c_k]$	0.99	0.63	0.24	0.87	0.16	0.14	0.51

stable retrieval-head selection under distribution shifts. Moreover, Table 9 shows that the same strategy robustly identifies meaningful retrieval heads across domains, suggesting its ability to capture general properties of TwGI-style models. Table 8 provides a more comprehensive presentation of the results reported in Table 3, further substantiating the conclusions drawn in Section 4.4.

Algorithm 3: KV Cache Eviction and Sparse Attention

Input: Layer L_i; query Q; surviving KV cache (K, V); mask M; retrieval-head sets R_t, R_i; current modality $\mu \in \{\text{text}, \text{image}\}$; CFG sample index c

Output: Attention output AttnOut_{L_i}; updated surviving KV cache (K', V')

foreach *head* $h \in L_i$ **do**
  if $\mu = \text{text}$ **then**
    if $h \in R_t^{L_i}$ **then**
      $(K_R^{(h)}, V_R^{(h)}, M_R^{(h)}) \leftarrow$ all surviving entries;
      $\text{AttnOut}_{L_i}[:, h, :, :] \leftarrow \text{softmax}\left(\frac{Q_h K_R^{(h)\top}}{\sqrt{d_h}} + M_R^{(h)}\right) V_R^{(h)}$;
    else
      $(K_S^{(h)}, V_S^{(h)}, M_S^{(h)}) \leftarrow$ surviving initial, modality-boundary, and recent tokens;
      $\text{AttnOut}_{L_i}[:, h, :, :] \leftarrow \text{softmax}\left(\frac{Q_h K_S^{(h)\top}}{\sqrt{d_h}} + M_S^{(h)}\right) V_S^{(h)}$;
  else
    if $c = \texttt{full}$ **then**
      if $h \in R_i^{L_i}$ **then**
        $(K_R^{(h)}[c], V_R^{(h)}[c], M_R^{(h)}[c]) \leftarrow$ all surviving entries;
      else
        $(K_S^{(h)}[c], V_S^{(h)}[c], M_S^{(h)}[c]) \leftarrow$ surviving initial, modality-boundary, and recent tokens;
    else if $c = \texttt{img-condition}$ **then**
      if $h \in R_i^{L_i}$ **then**
        $(K_R^{(h)}[c], V_R^{(h)}[c], M_R^{(h)}[c]) \leftarrow$ surviving image tokens;
      else
        $(K_S^{(h)}[c], V_S^{(h)}[c], M_S^{(h)}[c]) \leftarrow$ surviving modality-boundary and recent tokens;
    else
      if $h \in R_i^{L_i}$ **then**
        $(K_R^{(h)}[c], V_R^{(h)}[c], M_R^{(h)}[c]) \leftarrow$ surviving newly generated image tokens and inserted substitute tokens;
      else
        $(K_S^{(h)}[c], V_S^{(h)}[c], M_S^{(h)}[c]) \leftarrow$ inserted and recent tokens;
    if $h \in R_i^{L_i}$ **then**
      $\text{AttnOut}_{L_i}[c, h, :, :] \leftarrow \text{softmax}\left(\frac{Q_h[c] K_R^{(h)}[c]^\top}{\sqrt{d_h}} + M_R^{(h)}[c]\right) V_R^{(h)}[c]$;
    else
      $\text{AttnOut}_{L_i}[c, h, :, :] \leftarrow \text{softmax}\left(\frac{Q_h[c] K_S^{(h)}[c]^\top}{\sqrt{d_h}} + M_S^{(h)}[c]\right) V_S^{(h)}[c]$;
  Update (K', V') by appending current KV and evicting outdated entries;
return $\text{AttnOut}_{L_i}, (K', V')$;

C Algorithm of Sparse Attention

To mitigate the increasing memory consumption and attention computation overhead caused by the continuously growing KV cache during long-sequence and multimodal generation, we design a KV Cache Eviction and Sparse Attention mechanism, as illustrated in Algorithm 3. The core idea of this algorithm is to maintain only a surviving KV cache during decoding, which stores key-value

Table 9: Results on visual understanding benchmarks.

Method	MathVision	MathVista	VisuLogic	EMMA	MMVP	Blink	Vstar
Anole Chern et al. (2024)	13.80	22.80	8.50	12.80	10.00	26.46	23.60
Zebra-CoT Li et al. (2025a)	16.45	25.30	21.80	**15.02**	15.33	**31.25**	27.20
Twilight	**16.84**	**25.90**	**22.30**	14.92	**16.00**	31.04	**28.80**

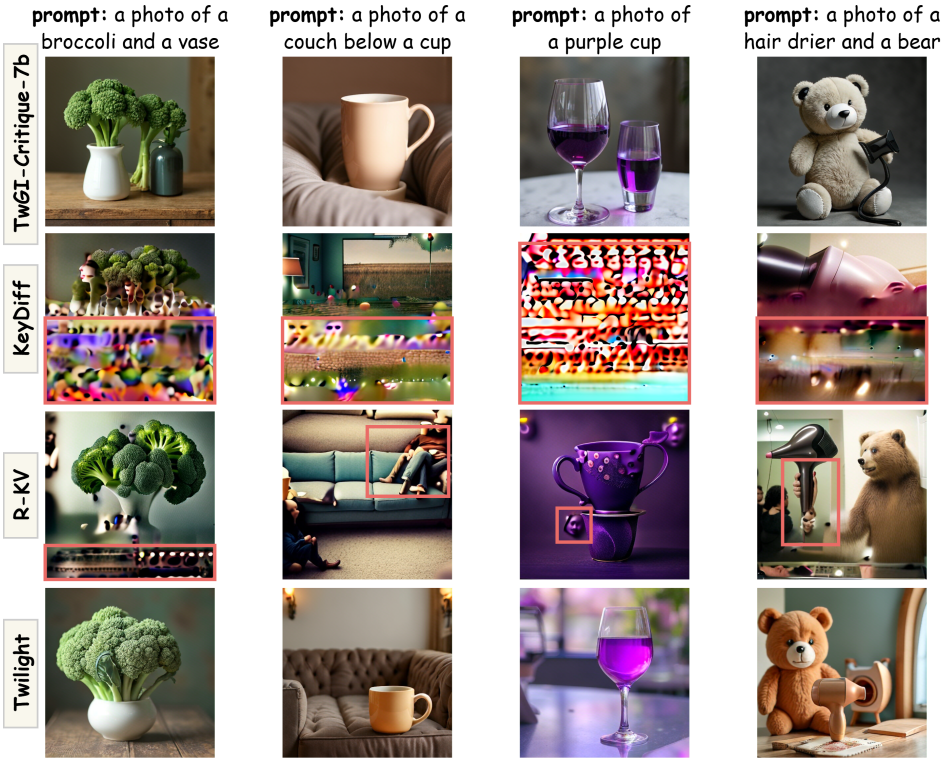


Figure 12: Visualization of the final generated images using different approaches on Geneval.

pairs carrying useful contextual information, while evicted entries are permanently removed and no longer participate in subsequent attention computation.

Based on this surviving cache, the algorithm further constructs sparse attention patterns by leveraging the functional differences among attention heads. Specifically, retrieval heads, which are primarily responsible for accessing global contextual information, retain all surviving KV entries during text generation. During image generation, they instead preserve context according to the CFG conditional sample index. In contrast, streaming heads mainly capture local sequential dependencies and therefore operate on a compact subset of the surviving cache. Each attention head then performs scaled dot-product attention on its corresponding KV subset, thereby reducing computational cost while preserving essential contextual modeling capability. After computing attention for the current step, the KV cache is updated by appending the key-value pair of the newly generated token and evicting outdated entries, thus dynamically maintaining the surviving KV cache for the next decoding step. Overall, this algorithm integrates cache eviction with head-aware sparse attention, significantly reducing the storage pressure of the KV cache and the complexity of attention computation, while preserving generation quality in both text and image generation scenarios.

D More Implementation Details

For retrieval head identification, we randomly generate 50 samples independent of GenEval and DPG-Bench to form the calibration set, with Top-k set to 500. Regarding the attention patterns of streaming head, the sink tokens comprise the first 128 tokens and modality boundary tokens with length of 20, while the number of recent tokens is set to 512. GenEval introduces object detection models Cheng et al. (2022) and other discriminative models to evaluate generated images across six aspects, with their average as the overall score. DPG-Bench evaluates fine-grained entities, attributes, and relationships in complex prompts using mPLUG Li et al. (2022) and predefined VQA questions. We use the same random seed for all eviction methods.

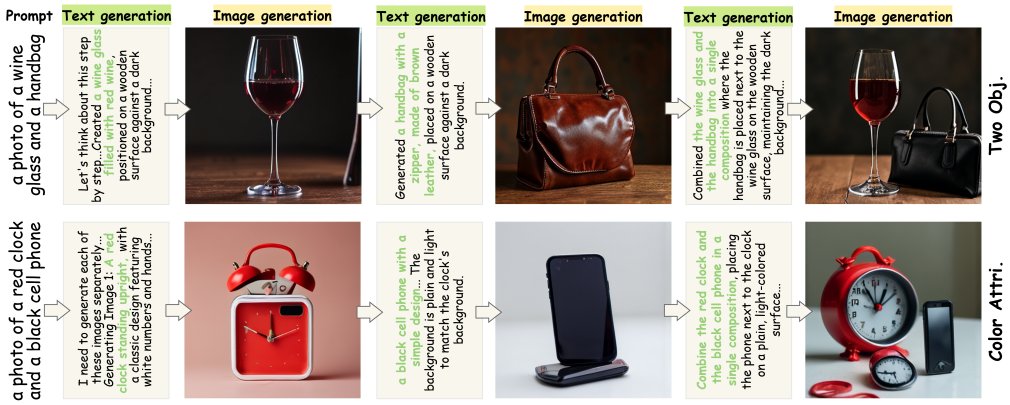


Figure 13: Visualization of the reasoning process using subgoal models for different tasks.

E More Comparative Results

We first define the KV-cache budget ratio in Table 2 as follows:

$$\rho_{\text{KV}} = \frac{M_{\text{retained}}}{M_{\text{full}}} \times 100\%, \tag{9}$$

where M_{retained} and M_{full} denote the KV-cache memory retained by each method and full attention, respectively. Lower ρ_{KV} indicates greater memory reduction.

Table 10: Comparison with other eviction methods on GenEval.

Method	Two Obj.	Position	Color Attri.
StreamingLLM-style	0.24	0.02	0.04
DuoAttention-style	0.52	0.09	0.13
Twilight	**0.59**	**0.12**	**0.18**

Figure 12 shows the final images generated by different approaches on GenEval. The red boxes highlight defects in the generated images. Twilight maintains higher generation quality, while both KeyDiff and RKV exhibit varying degrees of image generation collapse, indicating that their eviction strategies fail to preserve image generation capability. In addition to evaluating generation performance, we further assess Twilight on visual understanding tasks. Zebra-CoT Li et al. (2025a) enhances the multimodal reasoning capabilities through thinking with generated images, and we build Twilight upon Zebra-CoT. The results across multiple visual understanding benchmarks are presented in Table 9. Twilight maintains comparable visual understanding performance with original baseline while reducing KV cache memory usage, further demonstrating its generalizability across diverse tasks. To isolate the effects of different retention patterns, we additionally construct StreamingLLM-style and DuoAttention-style variants corresponding to the first row of Table 4 and the second row of Table 5, respectively. For DuoAttention-style methods, since optimization-based identification is difficult to apply to the thinking-with- generated-images paradigm, we implement the corresponding attention patterns using the retrieval-head sets identified by our criterion. As shown in Table 10, Twilight outperforms other eviction methods, demonstrating that the attention pattern described in Section 3.3 better preserves critical tokens according to the model's characteristics, thereby reducing KV storage while maintaining performance.

F More Visual Analysis

In this section, we present additional visualization results, including Twilight's reasoning processes for different tasks under different models. In Figure 13, Twilight decomposes the generation target into multiple subgoals, generates them one by one, and then assembles them into the final output. In Figure 14, Twilight analyzes the initially generated image, identifies its deficiencies, and then guides the generation of a higher-quality image. These results show that, even after removing a large amount of KV cache, Twilight is still able to preserve its reasoning logic while maintaining image generation quality. Furthermore, on DPG-Bench, we visualize and compare the generation results of state-of-the-art KV cache eviction methods and Twilight in critique model, as shown in Figure 15. Twilight clearly preserves both the image generation capability and the textual reasoning logic more effectively. KeyDiff fails to generate coherent images and exhibits poor instruction following during textual reasoning, ultimately producing corrupted outputs. Although R-KV generates content that generally follows the instructions, its image quality degrades substantially. These findings show that existing eviction methods struggle to accommodate the characteristics of thinking with generated images. Their failure to retain critical information leads to accumulated errors, disrupts the original reasoning structure, and substantially degrades generation quality.

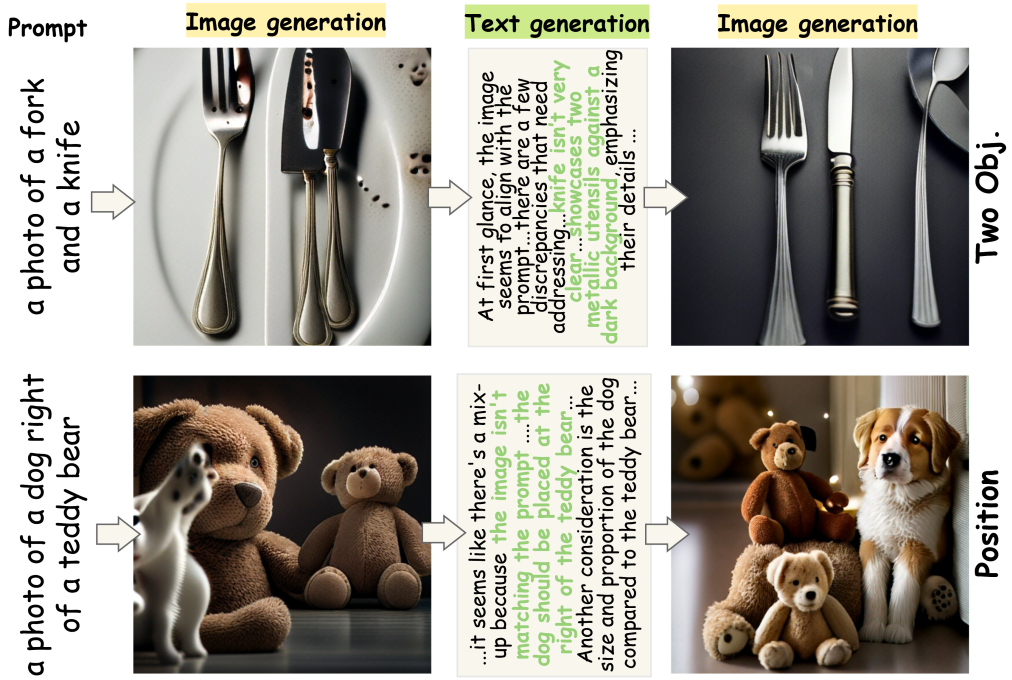


Figure 14: Visualization of the reasoning process under critique model for different tasks.



Figure 15: Visualization of the reasoning process in different approaches on DPG-Bench.

G AI USE STATEMENT

In this work, we used generative AI tools to improve the clarity and fluency of the manuscript.

H FUTURE WORK

Future work could extend Twilight to a broader range of unified multimodal architectures and application domains, examining whether its modality-specific attention patterns remain effective under different tokenization schemes and generation processes. It would also be valuable to evaluate the method on more diverse tasks that require extended, interleaved text-image reasoning. We also plan to explore more aggressive eviction strategies that preserve information needed across successive reasoning steps. These directions may further improve the balance between memory use, decoding speed, and generation quality, particularly for longer and more varied text-image reasoning chains.