Blink: Dynamic Visual Token Resolution for Enhanced Multimodal Understanding

Supplementary Material

A. Training Details

A.1. Configurations

For each involved transformer layer, we attach a lightweight token super-resolution (TokenSR) module to refine the expanded saliency tokens. Each TokenSR module consists of three sequential convolutional layers. The input and output channels are both set to 4096, matching the hidden dimension of the MLLM backbone, while the intermediate layers use 2048 and 1024 channels, respectively. The convolution kernel sizes are 5, 3, and 1, each applied with symmetric padding to preserve spatial resolution.

A.2. Datasets

To train the TokenSR modules, we use the processed LLaVA-1.5 [32] training set, which is constructed from several widely used vision–language datasets. For each image, we additionally generate four quadrant crops (top-left, top-right, bottom-left, bottom-right) and record their positions relative to the original image. Each processed sample therefore consists of the full image, one cropped image, and the corresponding positional metadata of the cropped quadrant.

During training, both the full image and the cropped image are converted into pixel-value tensors and fed into the frozen MLLM backbone to obtain hidden representations. The cropped image provides the teacher features, while the TokenSR module takes as input the token hidden states corresponding to the same spatial region in the full image after bilinear interpolation, and generates enhanced token hidden states. The optimization objective is to align the enhanced token hidden states with the teacher features of the corresponding spatial region.

A.3. Hyper-parameters

The training hyper-parameters for the TokenSR modules are presented in Tab. 4. All experiments are conducted on 8 H800 (80G) GPUs in total, where each TokenSR module is trained on 2 GPUs with a global batch size of 8, using BFloat16 precision. We set the warmup ratio to 3% and employ a cosine learning rate scheduler, decaying the learning rate from 1e-4 to 0. Our implementation is based on HuggingFace Transformers [44] and DeepSpeed [39].

Training Parameter	Value
# GPUs	2
Sequence length	1024
Data type	Bfloat16
Learning rate	1e-4
Learning rate scheduler	Cosine
Warmup ratio	0.03
Optimizer	AdamW
Global batch size	8
Epoch	1
DeepSpeed	Zero-2

Table 4. Training parameters used in the experiments.

B. Implementation Details

In our main experiments in Sec. 4, the inference settings are configured as follows. For the layer ranges and saliency thresholds, we use layers 12–18 with $\tau_{\text{exp}} = 0.5$ and $\tau_{\text{drop}} = 0.4$ for MME and MM-Vet. For GQA, the same layer range is used with $\tau_{\text{exp}} = 0.6$ and $\tau_{\text{drop}} = 0.4$. For POPE, we apply layer 18 with $\tau_{\text{exp}} = 0.5$ and $\tau_{\text{drop}} = 0.4$. For MM-Bench, MMBench-CN and ScienceQA, we use the same layer with $\tau_{\text{exp}} = 0.25$.

Additionally, in the ablation studies on the MME benchmark, the High τ_{exp} / Low τ_{drop} setting corresponds to values of 0.7 and 0.3, respectively. The High τ_{exp} configuration refers to using $\tau_{\text{exp}} = 0.7$ while keeping τ_{drop} fixed at 0.4, consistent with the main experiment on MME.

C. Evaluation on Other Backbones

C.1. Evaluation on LLaVA-NeXT

We evaluate Blink on the LLaVA-NeXT-7B [33] backbone, a moderate variable-resolution MLLM. As shown in Tab. 5, our method consistently improves performance across a wide range of downstream benchmarks. Blink yields a notable gain of 15.00 on $\text{MME}_{\text{Cognition}}$ compared with the vanilla model and achieves the highest scores on GQA, MMBench-CN, POPE, and MM-Vet. Although its $\text{MME}_{\text{Perception}}$ score is slightly lower than that of Blink-interp, Blink shows stronger advantages on most tasks, suggesting that the fully trained TokenSR module strengthens fine-grained multimodal understanding.

Blink-interp achieves the best performance on $\text{MME}_{\text{Perception}}$ and improves over the backbone model on ScienceQA and MM-Vet, highlighting the complementary benefits of the dynamic inference pipeline. These results collectively demonstrate that Blink generalizes well to stronger backbones and provides robust gains across diverse downstream tasks.

Method	$\text{MME}_{\text{Perception}}$										
	Exist.	Count	Pos.	Color	Poster	Celeb.	Scene	Landm.	Artw.	OCR	**Total**
Vanilla	195.00	**158.33**	**135.00**	175.00	150.68	146.47	**164.25**	**167.50**	**134.75**	**102.50**	1529.48
Blink-interp (Ours)	195.00	**158.33**	**135.00**	**180.00**	150.68	**147.35**	**164.25**	**167.50**	134.00	**102.50**	**1534.62**
Blink (Ours)	**200.00**	**158.33**	130.00	**180.00**	**153.74**	142.94	163.50	166.00	134.00	**102.50**	1531.02

Method	$\text{MME}_{\text{Cognition}}$					GQA	MMBench	$\text{MMBench}_{\text{CN}}$	POPE	SQA_{Img}	MM-Vet
	CS	Num	Text	Code	**Total**						
Vanilla	**120.00**	**57.50**	102.50	**35.00**	315.00	64.26	**69.07**	60.82	86.37	70.40	39.80
Blink-interp (Ours)	**120.00**	55.00	110.00	**35.00**	320.00	64.22	**69.07**	60.82	86.23	**70.50**	40.40
Blink (Ours)	**120.00**	**57.50**	**117.50**	**35.00**	**330.00**	**64.29**	**69.07**	**60.91**	**86.40**	70.45	**40.50**

Table 5. Downstream task performance across multiple benchmarks on LLaVA-NeXT. Vanilla denotes the base model, and our methods correspond to two configurations of Blink, where -interp indicates the variant that replaces the amplifier with training-free interpolation while retaining the Blink inference pipeline. The best scores are in **bold**.

Method	$\text{MME}_{\text{Perp.}}$	$\text{MME}_{\text{Cogn.}}$	GQA	POPE	SQA_{Img}
Vanilla	1638.16	598.57	58.14	87.63	80.81
Blink-interp (Ours)	1638.16	606.07	**58.37**	**87.82**	81.06
Blink (Ours)	**1645.65**	**608.57**	58.28	87.75	**81.21**

Table 6. Downstream task performance on Qwen2.5-VL-7B. Vanilla denotes the base model, and our methods correspond to two configurations of Blink, where -interp indicates the variant that replaces the amplifier with training-free interpolation while retaining the Blink inference pipeline. The best scores are in **bold**, and the second-best results are underlined.

C.2. Evaluation on Qwen2.5-VL

To further validate the effectiveness of Blink on modern multimodal backbones, we conduct additional experiments on Qwen2.5-VL-7B [4], a recent and strong vision-language model. We evaluate our method across four benchmarks, including MME [17], GQA [20], POPE [29], and ScienceQA [35]. We consider both Blink and Blink-interp, where the trainable modules are replaced with bilinear interpolation while retaining the same inference pipeline.

As shown in Tab. 6, both Blink and Blink-interp consistently outperform the vanilla model across multiple benchmarks, demonstrating the effectiveness of our design. Specifically, Blink achieves the best performance on MME, improving $\text{MME}_{\text{Perp.}}$ from 1638.16 to 1645.65 and $\text{MME}_{\text{Cogn.}}$ from 598.57 to 608.57. It also attains the highest score on ScienceQA, indicating stronger multimodal reasoning ability. Meanwhile, Blink-interp achieves the best results on GQA of 58.37 and POPE of 87.82, showing that even a training-free variant can effectively enhance visual grounding and perception. Notably, Blink-interp also improves $\text{MME}_{\text{Cogn.}}$ to 606.07 and ScienceQA to 81.06, outperforming the vanilla baseline. These results demonstrate consistent improvements across backbones, highlighting the robustness and general applicability of our method.

Method	$\text{MME}_{\text{Total}}$	GQA	POPE	FLOPs	#Fwd Passes
ViCrop	1804.65	60.98	**87.25**	21.5T	3
Blink (Ours)	**1881.53**	**61.98**	85.23	**9.73T**	**1**

Table 7. Downstream task performance and efficiency of different methods on LLaVA-1.5-7B. FLOPs denote the theoretical maximum computation per image. #Fwd Passes indicates the number of full model inferences required for each input. The best scores are in **bold**.

D. Comparison with ViCrop

Direct comparison with dynamic perception methods is non-trivial, as they typically require multiple forward passes and incur higher inference costs. To ensure a fair evaluation, we reproduce ViCrop [51] on LLaVA-1.5-7B and report results on three benchmarks. As shown in Tab. 7, our method achieves better performance on MME and GQA, while maintaining comparable results on POPE. These results demonstrate that Blink consistently improves multimodal perception over prior methods.

In terms of efficiency, Blink significantly reduces computational overhead by lowering FLOPs from 21.5T to 9.73T (a 54.7% reduction) and requiring only a single forward pass instead of three, thereby reducing inference latency. Meanwhile, the additional memory overhead compared to vanilla LLaVA-1.5-7B is minimal, with GPU usage increasing from 19.7 GB to 20.8 GB (a 1.1 GB increase) under FP16. Overall, Blink achieves a better balance between performance and efficiency, making it more suitable

Method	# Patches	Perc.	Cogn.	Total
Blink-interp	2×2	1514.08	353.21	1867.29
	3×3	1507.44	350.36	1857.80
	4×4	1499.38	335.36	1834.74
Blink	2×2	1519.74	361.79	1881.53
	3×3	1498.16	374.64	1872.80
	4×4	1500.24	358.21	1858.45

Table 8. Performance of Blink and its interpolation variant on the MME benchmark with different numbers of patches. The 2×2 setting is used in our main configuration, and the other configurations correspond to alternative patch partition settings.

for practical deployment.

E. Cases of Attention Redistribution

Following the analysis in Sec. 4.4, we conduct additional visualization experiments to examine how the tokens generated by the TokenSR module influence attention redistribution across transformer layers after expansion. As in the main experiments, expansion is applied at layer 12, and the newly generated tokens are preserved in all subsequent layers without being dropped.

Fig. 8 presents attention maps across different layers to illustrate this spatial redistribution. We analyze the two examples from Sec. 2.2, along with four additional cases. As expected, in all examples, attention on the newly introduced tokens remains more evenly distributed compared to the original visual tokens. Furthermore, in the first three cases, we observe clear layer-wise shifts in attention, consistent with our first key insight that the model progressively adjusts its attention allocation across layers after the expanded tokens are introduced.

F. Ablation Study on Patch Numbers

In our main experimental setup, the reshaped $H \times W$ attention grid used in saliency-guided scanning is uniformly partitioned into 2×2 patches. In this section, we evaluate the effect of varying the partition granularity by dividing the grid into $p \times p$ non-overlapping patches of equal size, and report results on the MME benchmark.

Since changing the number of patches alters the minimum proportion of the image that a salient region can occupy, we scale the saliency ratio thresholds accordingly. For the expansion threshold, the patch-adjusted value is computed as $\tau_{\text{exp}}^{(p)} = 0.5 \times \frac{1/p^2}{1/2^2}$, and similarly for the drop threshold, $\tau_{\text{drop}}^{(p)} = 0.4 \times \frac{1/p^2}{1/2^2}$.

As shown in Tab. 8, neither the 3×3 nor 4×4 partitioning yields improvements over the default 2×2 setting. For Blink-interp, increasing the number of patches consistently degrades both perception and cognition scores, with the 4×4 configuration showing the largest decline. For Blink, the 3×3 variant offers a small gain in cognition but still reduces perception and overall MME performance, while the 4×4 setup leads to drops across all metrics. This may be because the current thresholds are derived using a simple proportional scaling rule, and more fine-grained tuning could further optimize performance for different patch granularities. Nevertheless, across all configurations, Blink equipped with the trained TokenSR module consistently outperforms Blink-interp, confirming that the learned amplifier is more effective than naive interpolation under all patch partitions.

Layers	Perc.	Cogn.	Total
Blink-interp (Ours)	1514.08	353.21	1867.29
w/o interp	1510.58$_{-3.50}$	357.86$_{+4.65}$	1868.44$_{+1.15}$
Blink (Ours)	1519.74	361.79	1881.53
w/o interp	1515.08$_{-4.66}$	355.00$_{-6.79}$	1870.08$_{-11.45}$

Table 9. Performance of Blink and its training-free varianton the MME benchmark with and without the interpolation step. In the w/o interp setting, the selected saliency tokens are inserted directly without being upsampled to match the original image resolution.

G. Ablation Study on Interpolation

In our method, prior to feeding saliency tokens into the TokenSR module, we first perform interpolation to upsample these tokens. The purpose of this interpolation is to align the length of the salient token sequence with that of the original image tokens, allowing both training and inference to proceed directly with standard two-image inputs at the original spatial resolution. This design ensures that the model operates on inputs that preserve the original spatial structure, which is more consistent with conventional visual encoding practices and facilitates effective learning of fine-grained multimodal representations.

Tab. 9 presents the performance on the MME benchmark with and without the interpolation step. In our experiments, the w/o interp condition refers to bypassing this initial upsampling. Specifically, after selecting the saliency tokens, they are directly extracted and inserted between the original visual and text tokens without resizing. For Blink-interp, w/o interp corresponds to inserting the saliency tokens directly into the dynamic inference pipeline without any training. For Blink, w/o interp means feeding the tokens into the trained TokenSR module without prior interpolation. From the experimental results, skipping interpolation for Blink-interp slightly reduces the perception score by 3.50, indicating that aligning the length of the expanded saliency se-

quence with the original image helps preserve spatial information. For Blink, removing interpolation leads to drops of 4.66 in perception, 6.79 in cognition, and 11.45 overall, further demonstrating that the TokenSR module benefits from receiving saliency tokens of the same length as the original image tokens, which allows the amplifier to more effectively refine multimodal representations. Overall, these results highlight that interpolation is a crucial step for fully leveraging spatial context and ensuring stable and accurate performance.

H. Limitations and Future Work

Although Blink demonstrates strong generalization across different tasks and backbones, its current design relies primarily on convolution-based upsampling. Exploring alternative expansion modules may further improve flexibility and performance. Moreover, due to computational constraints, all experiments are conducted on models no larger than 7B parameters, leaving the scalability of Blink to larger MLLMs unverified.

Future work will explore alternative TokenSR mechanisms, such as multi-layer perceptrons, as complements or replacements for the convolution-based module, and investigate improvements that enhance the overall efficiency and practicality of Blink.

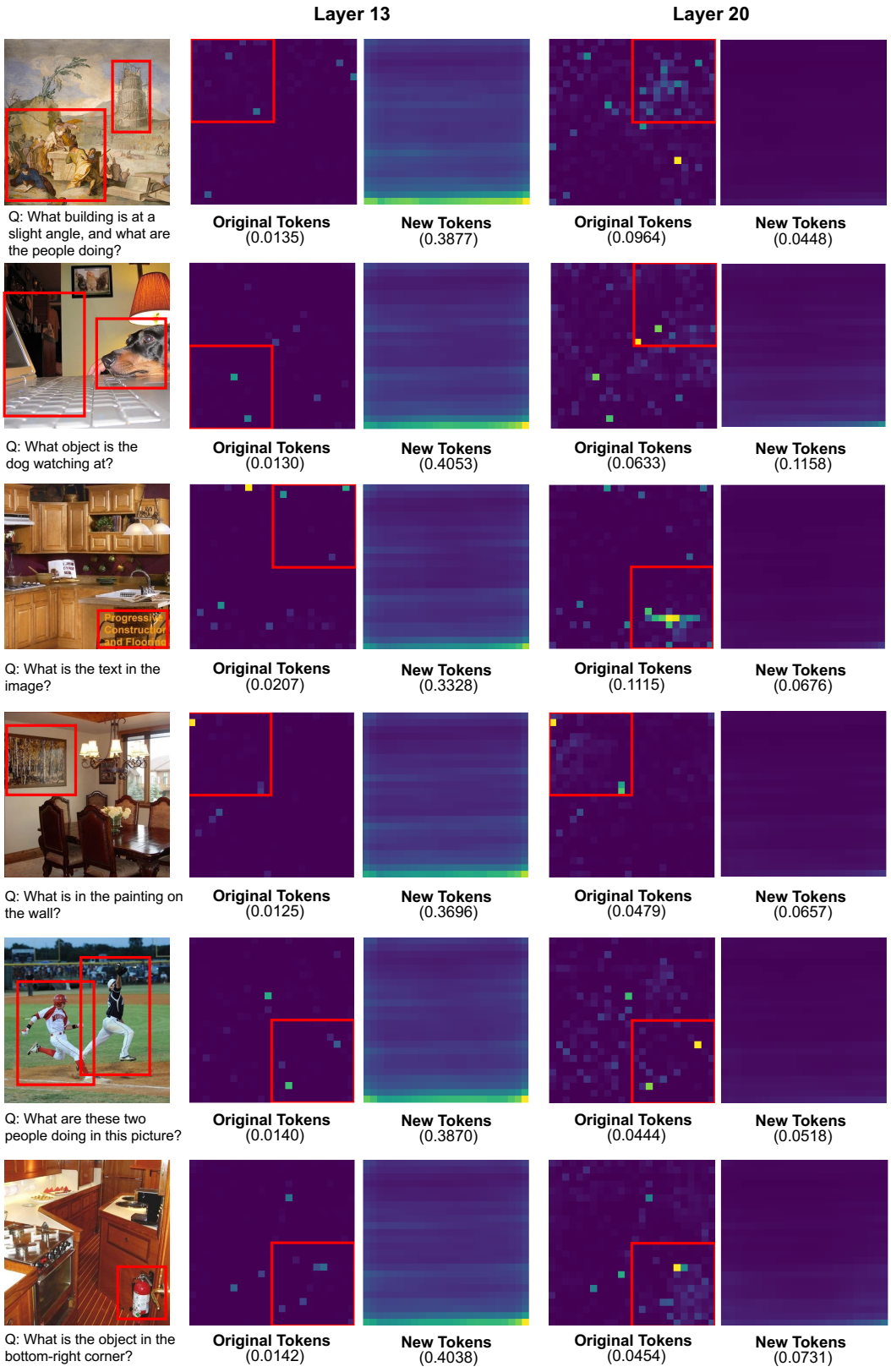

Figure 8. Visualization of attention redistribution after token expansion. Red boxes in the original image indicate the ground-truth important regions. The right panels show attention distributions on the original and expanded visual tokens at layers 13 and 20.