Visual Funnel: Resolving Contextual Blindness in Multimodal Large Language Models

Supplementary Material

A. Hyperparameter Sensitivity Analysis

In Section 3.2.2 of the main paper, we introduced the *Entropy-Guided Scale Determination* mechanism. The crop expansion factors, α_1 (for immediate context) and α_2 (for broader context), are computed as linear functions of the normalized attention entropy H_{norm}:

$$\alpha_k(I, q) = \beta_k + \gamma_k \cdot H_{\text{norm}}(I, q) \tag{7}$$

where β_k represents the base expansion factor (minimum context size) and γ_k represents the sensitivity coefficient to the model's uncertainty. Our default configuration uses $\mathcal{C}_{\text{default}} = \{\beta_1 = 1.2, \gamma_1 = 0.6, \beta_2 = 1.6, \gamma_2 = 1.2\}$.

To demonstrate that our method is robust and not overfitted to specific "magic numbers," we conduct a comprehensive sensitivity analysis using the Qwen2.5-VL-3B-Instruct backbone on the DocVQA dataset. Importantly, these default parameters were determined using a small held-out validation set from GQA and kept fixed across all benchmarks reported in the main paper.

A.1. Impact of Entropy Sensitivity (γ)

First, we investigate the necessity of the adaptive scaling mechanism. We vary the sensitivity coefficients γ_1 and γ_2 while keeping the base factors β fixed. Setting $\gamma = 0$ represents a *Static* baseline where crop sizes are fixed regardless of attention uncertainty.

As shown in Table 3, the adaptive configuration ($\gamma > 0$) consistently outperforms the static approach. The performance peaks at our default setting but remains stable within a reasonable range ($\gamma_1 \in [0.4, 0.8]$), confirming that allocating broader context to uncertain regions is crucial for resolving Contextual Blindness.

Configuration	γ_1	γ_2	DocVQA Acc. (%)	Δ
Static (Fixed Size)	0.0	0.0	59.5	-1.6
Weak Adaptation	0.3	0.6	60.4	-0.7
Default (Ours)	**0.6**	**1.2**	**61.1**	**–**
Strong Adaptation	0.9	1.8	60.8	-0.3

Table 3. **Ablation on Entropy Sensitivity.** We analyze the impact of the sensitivity coefficient γ. The results validate that adaptive scaling based on attention entropy yields better performance than static cropping ($\gamma = 0$).

A.2. Robustness of Base Expansion Factors (β)

Next, we analyze the stability of the base crop sizes. We shift the intercept values β_1 and β_2 by ± 0.2 from the default settings to simulate tighter or looser base crops.

Table 4 illustrates the robustness of Visual Funnel. Tighter crops ($\beta - 0.2$) lead to a slight performance drop due to the severance of immediate local context. However, the performance variance across the tested range is minimal ($< 0.6\%$), indicating that our method does not require precise hyperparameter tuning to achieve significant gains.

Base Scale Shift	β_1	β_2	DocVQA Acc. (%)	Δ
Tighter Crops (-0.2)	1.0	1.4	60.5	-0.6
Default	**1.2**	**1.6**	**61.1**	**–**
Wider Crops ($+0.2$)	1.4	1.8	60.9	-0.2

Table 4. **Robustness of Base Expansion Factors.** Shifting the base crop size β shows minimal impact on performance, demonstrating the method's stability.

B. Ablation on Portfolio Size

In Visual Funnel, we construct a hierarchical portfolio consisting of three specific crops: *Focal* (μ_0), *Immediate Context* (μ_1), and *Broader Context* (μ_2), in addition to the original image. A critical question arises: *Is the performance gain simply due to the increased quantity of visual tokens, or is the three-layer hierarchical structure optimal?*

To answer this, we evaluate the impact of the number of portfolio crops (K) on the DocVQA dataset using Qwen2.5-VL-3B-Instruct. We incrementally add crops following our hierarchical expansion strategy:

- $K = 1$: Focal crop only (similar to standard ViCrop).
- $K = 2$: Focal + Immediate Context.
- $K = 3$: Focal + Immediate + Broader Context (**Ours**).
- $K = 4$: Focal + Immediate + Broader + Global Context (an even wider crop).

As presented in Table 5, the results support our structural design:

1. **Significant Gain from Hierarchy ($K = 1 \rightarrow 3$):** Moving from a single focal crop ($K = 1$) to our three-layer portfolio ($K = 3$) yields a substantial improvement (+6.0%). This confirms that resolving Contextual Blindness requires not just the high-resolution detail of the target, but also the intermediate scales that bridge the detail to the global view.

# Crops (K)	Configuration	Token Usage	DocVQA Acc.	Δ
0	Original Image Only	$1\times$	51.5	-9.6
1	Focal Only	$\sim 1.3\times$	55.1	-6.0
2	Focal + Imm.	$\sim 1.6\times$	58.0	-3.1
3	**Focal + Imm. + Broader**	$\sim$ **1.9**$\times$	**61.1**	**–**
4	+ Global Context	$\sim 2.2\times$	60.7	-0.4

Table 5. **Impact of Portfolio Size (K).** Increasing crops saturates at $K = 3$. Adding more leads to a "Redundancy Penalty."

Model Configuration	Avg. Tokens	Latency (ms)	Relative Time	DocVQA Acc.	**Gain/Time**
Base (No Crop)	$\sim 1,200$	450	$1.00\times$	51.5	–
w/ ViCrop	$\sim 1,800$	780	$1.73\times$	54.2	Low
w/ ViCrop (Top-3)	$\sim 2,400$	920	$2.04\times$	55.3	Low
w/ Visual Funnel (Ours)	$\sim 2,300$	**890**	**1.98**$\times$	**61.1**	**High**

Table 6. **Efficiency vs. Performance Trade-off.** Compared to the Base model, Visual Funnel requires approximately $2\times$ the inference time but yields a massive performance gain (+9.6%). Notably, it is more efficient than the naive multi-crop baseline (ViCrop Top-3) in terms of accuracy per computational unit.

2. **The Redundancy Penalty ($K = 4$):** Interestingly, adding a fourth crop ($K = 4$) does not further improve performance; in fact, it leads to a slight degradation (61.1% $\rightarrow$ 60.7%). We attribute this to the *Redundancy Penalty*: providing too much overlapping visual information can overwhelm the MLLM's attention mechanism, causing it to lose focus on the critical details.
3. **Efficiency Trade-off:** Furthermore, $K = 4$ increases the input token count and inference latency without functional benefit. Therefore, we identify $K = 3$ as the optimal configuration that maximizes structural diversity while maintaining computational efficiency.

C. Computational Efficiency Analysis

While Visual Funnel significantly enhances fine-grained perception, it inevitably introduces computational overhead due to the two-step inference process and the processing of additional visual tokens. In this section, we provide a detailed analysis of inference latency and token usage to demonstrate the cost-effectiveness of our approach.

Experimental Setup. We measured the average wall-clock time per query on the DocVQA validation set. All experiments were conducted on four NVIDIA RTX PRO 6000 (96GB) GPU with PyTorch 2.8. The latency includes image preprocessing, visual encoding, and language generation.

Analysis. As shown in Table 6:

- **Latency Overhead:** Visual Funnel increases the inference latency by approximately $1.98\times$ compared to the base model. This is primarily due to the additional forward pass required for *Contextual Anchoring* (Step 1) and the encoding of the multi-scale portfolio (Step 2).
- **Comparison with Baselines:** Compared to `w/ ViCrop (Top-3)`, which processes a similar number of visual tokens, our method is slightly faster (890ms vs. 920ms) and significantly more accurate (61.1% vs. 55.3%). This indicates that the *structure* of the visual input is more important than raw pixel quantity.
- **Parallelization:** It is worth noting that the multiple crops in Step 2 are encoded in a single batch, allowing us to leverage GPU parallelism. This ensures that the latency does not scale linearly with the number of crops.
- **Practicality:** Given the complexity of fine-grained tasks (e.g., reading small text in documents), we argue that a $2\times$ latency increase is a justifiable trade-off for a $\sim 10\%$ accuracy improvement. For real-time applications, Visual Funnel can be selectively applied only when the base model's confidence is low.

D. Qualitative Visualizations

We present further qualitative success and failure cases of `Qwen2.5-VL-3B-Instruct` in Figure 4.

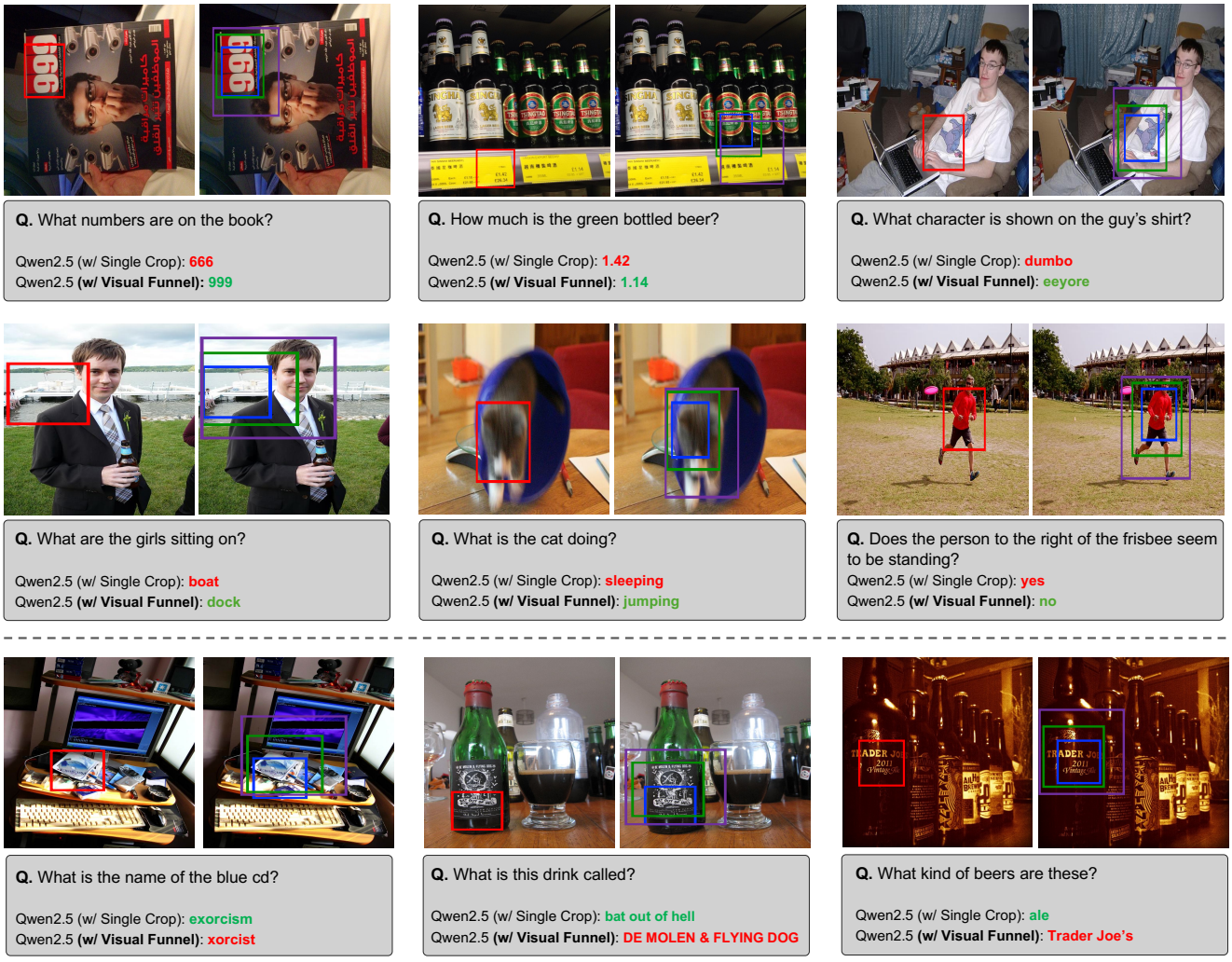

Figure 4. **Qualitative comparison between the Single-Crop baseline and Visual Funnel.** We visualize the inputs and predictions using `Qwen2.5-VL-3B-Instruct`. The Red box represents the input for the standard Single-Crop baseline (w/ ViCrop), while the Blue, Green, and Purple boxes represent the hierarchical portfolio (Focal, Immediate, Broader context) used in Visual Funnel. **(Top two rows) Success Cases:** Visual Funnel successfully resolves *Contextual Blindness* across various tasks, including fine-grained OCR (e.g., identifying "999" instead of inverted "666"), small object recognition ("Eeyore"), and action/state reasoning ("jumping" vs. "sleeping", "dock" vs. "boat"). **(Bottom row) Failure & Ambiguous Cases:** Examples below the dashed line illustrate limitations where the model still struggles despite improved context. These include partial OCR errors ("xorcist"), ambiguity in label hierarchy (Brewery name vs. Drink name), or distinct object attributes (Brand vs. Type), suggesting directions for future work.