

# Visual Funnel: Resolving Contextual Blindness in Multimodal Large Language Models

Woojun Jung<sup>1</sup> Jaehoon Go<sup>1</sup> Mingyu Jeon<sup>1</sup> Sunjae Yoon<sup>2</sup> Junyeong Kim<sup>1</sup>  
<sup>1</sup>Chung-Ang University <sup>2</sup>KAIST

{svvma91, junyeongkim}@cau.ac.kr

## Abstract

Multimodal Large Language Models (MLLMs) demonstrate impressive reasoning capabilities, but often fail to perceive fine-grained visual details, limiting their applicability in precision-demanding tasks. While methods that crop salient regions of an image offer a partial solution, we identify a critical limitation they introduce: “Contextual Blindness.” This failure occurs due to structural disconnect between high-fidelity details (from the crop) and the broader global context (from the original image), even when all necessary visual information is present. We argue that this limitation stems not from a lack of information ‘Quantity,’ but from a lack of ‘Structural Diversity’ in the model’s input. To resolve this, we propose Visual Funnel, a training-free, two-step approach. Visual Funnel first performs Contextual Anchoring to identify the region of interest in a single forward pass. It then constructs an Entropy-Scaled Portfolio that preserves the hierarchical context—ranging from focal detail to broader surroundings—by dynamically determining crop sizes based on attention entropy and refining crop centers. Through extensive experiments, we demonstrate that Visual Funnel significantly outperforms naive single-crop and unstructured multi-crop baselines. Our results further validate that simply adding more unstructured crops provides limited or even detrimental benefits, confirming that the hierarchical structure of our portfolio is key to resolving Contextual Blindness.

## 1. Introduction

Multimodal Large Language Models (MLLMs) have recently demonstrated significant capabilities in visual and language understanding. Despite this progress, a persistent challenge remains in perceiving small visual details, such as fine-grained text, distant object attributes, or subtle state differences. This *small detail problem* serves as a major bottleneck, hindering their deployment in applications that require high precision.

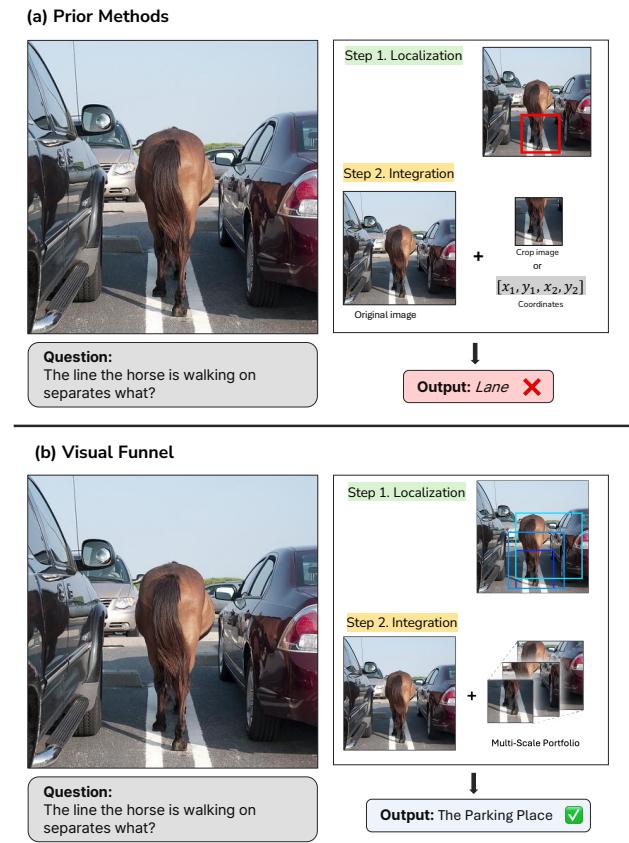
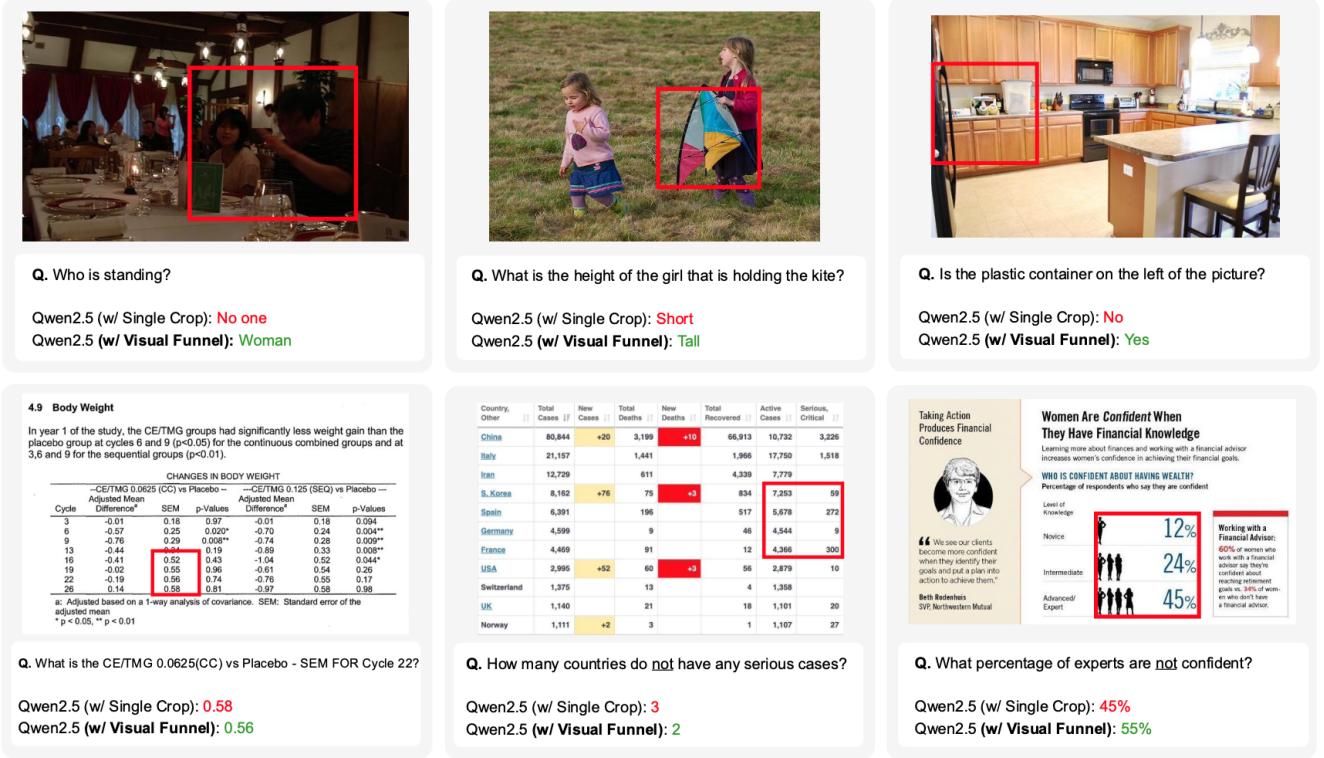


Figure 1. **Illustration of Contextual Blindness and our proposed solution, Visual Funnel.** (a) Prior single-crop methods successfully localize the area of interest but perform a naive integration by providing only a tight crop. This isolates the detail from its necessary context, leading to an incorrect answer (e.g., misidentifying a ‘parking space’ as a ‘lane’). (b) Our Visual Funnel performs a more sophisticated integration by generating a multi-scale portfolio. This portfolio preserves the hierarchical context, enabling the MLLM to resolve the ambiguity and provide the correct answer.

Efforts to address this bottleneck generally follow a two-step paradigm: (1) **Localization**, identifying where the



**Figure 2. Examples of Contextual Blindness.** Single-crop methods systematically remove essential context needed for correct reasoning, even when tight crops (red boxes) successfully isolate fine-grained details. **Top row:** (left) cropping only seated diners leads to overlooking the standing person; (center) excluding background reference objects results in incorrect height judgments; (right) different object positions within crop versus full image confuse spatial reasoning about left/right. **Bottom row:** (left) without surrounding column headers and labels, the model cannot identify which specific value corresponds to the requested metric; (center) excluding the “Serious, Critical” column header prevents identifying countries with empty cells in that column; (right) without seeing the question context about “not confident,” the model reports the visible “45%” value instead of computing the inverse as 55%. Even when provided with both the original image and the tight crop, MLLMs struggle to integrate information across these disparate scales, demonstrating the critical need for intermediate-scale representations that preserve hierarchical context. Results shown using Qwen2.5-VL-3B-Instruct.

relevant detail is, and (2) **Integration**, determining how to structure and present that detail to the MLLM. Recent works have made significant strides in the Localization step by leveraging the inherent capabilities of MLLM to pinpoint areas of interest [27, 32]. These approaches either perform a multi-step, guided search to iteratively refine the region of interest, or directly analyze the model’s internal signals, such as attention, in a single forward pass. Both philosophies have been effective in isolating the most salient region.

However, despite their differences in Localization strategies, these methods often rely on a simplistic approach for the crucial Integration step. Typically, a single, tightly-cropped high-resolution region is fed back into the model, sometimes alongside the original image. We identified that such naive integration introduces a critical limitation: while the model gains detail, it loses the intermediate context needed to interpret that detail. This overemphasis on an

isolated region—despite the availability of global context—often fails due to **Contextual Blindness**, which we term as the underlying issue. As illustrated in Figure 2, this issue is not due to missing pixels, but rather to a lack of *structure*. This observation leads to our central premise: “what constrains MLLM’s performance is not the absolute *Quantity* of information, but the lack of *Structural Diversity* in its input”. To address Contextual Blindness, we argue that a more sophisticated Integration strategy is required, one that goes beyond the limitations of single-crop methods.

To this end, we introduce Visual Funnel (VF), a training-free methodology that holistically addresses both Localization and Integration in a holistic manner. Drawing inspiration from the efficiency of internal signal-based methods, Visual Funnel enhances Localization while fundamentally rethinking the Integration process. It operates in two key steps:

1. **Contextual Anchoring:** We first refine the Localization

- step by using a specialized “search” prompt to yield a more precise attention map from a single forward pass.
- Entropy-Scaled Portfolio Generation:** For Integration, we leverage the attention map to construct a *multi-scale information portfolio*. This portfolio dynamically adjusts crop sizes based on attention entropy and hierarchically refines crop centers, preserving the focal detail, its immediate surroundings, and the broader context.

By explicitly addressing both Localization and Integration, Visual Funnel equips the MLLM with the structural diversity necessary to overcome Contextual Blindness. Our experiments validate this, demonstrating that Visual Funnel significantly outperforms baselines that rely on naive, single-crop integration. Additionally, we show that unstructured, repetitive information can be detrimental—introducing a ‘Redundancy Penalty’—further confirming that the hierarchical structure of our portfolio is the key to its success.

## 2. Related Work

### 2.1. Perception Limitations in Multimodal Large Language Models

Recent Multimodal Large Language Models predominantly adopt modular pretrained architectures [1, 3, 10, 13], connecting frozen Vision Transformers [4] with Large Language Models via learnable connectors. This architecture inherently compresses visual information into a fixed number of tokens, creating a structural bottleneck for fine-grained perception, particularly for small visual elements. Prior work has systematically demonstrated MLLMs’ sensitivity to visual concept size [32], establishing a causal relationship between object size and perception accuracy. Related studies document additional perception failures including object hallucination [11] and visual blind spots [30]. Recent surveys further analyze these challenges, highlighting weak spatial reasoning and poor fine-grained visual perception in MLLMs [8]. These limitations primarily address *what* models fail to perceive, leaving *how* visual information should be structured largely unexplored.

### 2.2. High-Resolution Training for Enhanced Visual Perception

Advances in high-resolution MLLM training have improved fine-grained visual understanding. LLaVA-NeXT [14] introduces AnyRes for adaptive grid configurations. Qwen2-VL [24] uses Naive Dynamic Resolution for variable visual tokens. InternVL2 [2] employs pixel shuffle with higher resolutions. Additional methods include patch-based strategies [12], cross-resolution fusion [25], and attention-efficient architectures [23, 28]. Recent work scales vision pre-training to 4K resolution using selective processing of local regions [20]. While these approaches

achieve performance gains, they require significant computational resources and apply static processing uniformly.

### 2.3. Training-Free Inference-Time Enhancement

Inference-time interventions leverage MLLMs’ existing capabilities without modifications, categorized by external tool reliance.

**External Tool-Based Approaches** use specialized vision models for visual attention. V\* [27] refines regions iteratively with tools like YOLO [18] and SAM [9]. Visual programming [6, 22] orchestrates tools via code. Set-of-Mark [29] overlays spatial marks on segmented regions. These require multiple passes and external dependencies.

**Internal Signal-Based Approaches** exploit internal representations without tools. Compositional chain-of-thought [17] uses scene graphs for compositional knowledge. ControlMLLM [26] optimizes latent variables to control attention for region description. ViCrop [32] analyzes attention patterns to identify salient regions based on MLLMs’ sensitivity to visual subject size [31]. However, by providing only a single tight crop alongside the original image, it loses intermediate-scale context essential for relational reasoning—a limitation we term *contextual blindness*. We address this by constructing adaptive multi-scale visual portfolios that preserve contextual hierarchies, as detailed in Section 3.

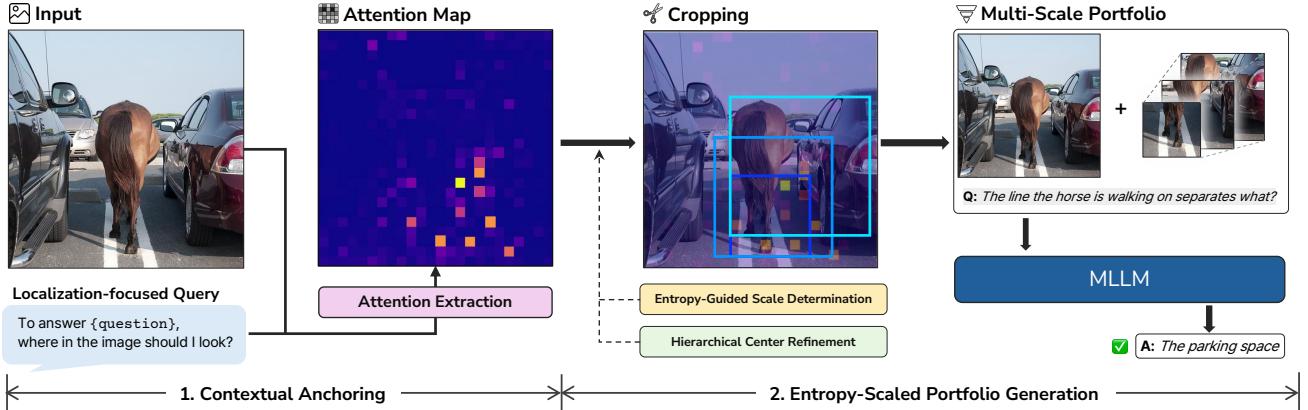
## 3. Method

To address the challenge of perceiving small details in MLLMs, previous works have advanced localization through attention-guided cropping [32] or iterative search [27]. However, these methods rely on a single-scale crop for perception, discarding intermediate context and creating a structural disconnect between global view and focal detail. We formalize this critical limitation as **Contextual Blindness** where coexisting global context and focal detail remain unbridgeable due to the lack of intermediate scales.

We propose Visual Funnel, which is illustrated in Figure 3 as a solution to Contextual Blindness through adaptive multi-scale information portfolios. We begin by analyzing the problem (Section 3.1), then describe context-aware attention extraction (Section 3.2.1) and Entropy-Scaled Portfolio Generation through hierarchical center refinement and entropy-guided scaling (Section 3.2.2).

### 3.1. Contextual Blindness

MLLMs often struggle to perceive small visual details in images. Previous approaches address this through a two-step paradigm: (1) *Localization*, identifying where to focus, and (2) *Integration*, enabling the model to recognize details in that region. Recent attention-based methods [32] have demonstrated strong localization capability, while iterative



**Figure 3. An overview of our proposed Visual Funnel methodology.** Visual Funnel resolves Contextual Blindness through a two-step, training-free process. **(1) Contextual Anchoring:** A localization-focused query guides the MLLM to establish a semantic anchor by generating a precise spatial attention map for the region of interest. **(2) Entropy-Scaled Portfolio Generation:** This attention map then serves as the foundation for generating a multi-scale information portfolio. Crop sizes are dynamically scaled based on attention entropy, and their positions are hierarchically refined to preserve crucial context. This final portfolio, comprising the original image and multiple contextual crops, provides the MLLM with the necessary structural diversity to correctly answer the question (e.g., identifying the context as a ‘parking space’ instead of a ‘lane’), a task where naive single-crop methods often fail.

search approaches [27] achieve precise localization through multiple forward passes. These methods suggest that the primary bottleneck is not localization, but integration.

Current integration strategies, such as providing the best single crop alongside the original image [32], rely on the core assumption that coexisting global context (the original image) and focal detail (the crop) are sufficient for accurate perception. However, we often observe failure cases where contextual understanding is required. Even with both a global view and focal detail, the MLLM cannot answer correctly. The root cause of this failure is the missing intermediate-scale context that bridges the two. We identify and term this underlying issue as “Contextual Blindness”: a failure of reasoning that occurs even when the MLLM possesses all necessary visual information (i.e., in the global image and the focal crop), simply because the structural disconnect between the focal detail and the global context makes them unbridgeable. This “blindness” results in critical failures for robust visual understanding, as illustrated in Figure 2.

For instance, an object’s attribute may be relative rather than absolute, defined by its relation to the surrounding context. In Figure 2, answering the question regarding the kite-holding girl’s height (*Short*) requires comparing her to the other girl in the frame. A single-scale crop focused only on the target girl would sever this relational link, causing the model to lose the necessary reference point and potentially misinterpret her height. Similarly, such failure occurs when an answer requires synthesizing information from spatially distinct pieces of chart or table. To correctly answer the question regarding the percentage of experts who are *not*

*confident* (Figure 2), the MLLM must first locate the ‘Advanced/Expert’ category, associate it with the corresponding ‘45%’ value from the bar chart, and understand that this value represents those who *are* confident, before correctly processing the negation in the query. A single crop focused on either the text or the bar chart in isolation would fragment this information, making the compositional reasoning required for this task impossible. In such cases, the MLLM does not lack information—the necessary pixels are available—but rather lacks the correct information *structure* to process it correctly.

### 3.2. Visual Funnel

To overcome Contextual Blindness, we propose Visual Funnel, a training-free approach that constructs adaptive multi-scale information portfolios. Visual Funnel dynamically determines both crop locations and the extent of context required, using hierarchical refinement and entropy-guided scaling, respectively. It provides complementary views across focal, immediate, and broader context scales, bridging fine-grained details with the global structure.

#### 3.2.1. Step 1: Contextual Anchoring

To identify focal regions requiring detailed examination, we extract attention maps from the MLLM’s internal representations. Unlike direct answering, which may produce hallucinations when visual information is insufficient, we prompt the MLLM with a localization-focused query: ‘*To answer ‘{question}’, where in the image should I look?*’. This query encourages the model to identify the *region* containing relevant information, rather than prematurely com-

mitting to an answer when details are unclear.

**Attention Extraction.** Following the approach in ViCrop [32], we extract cross-attention weights from the backbone LLM’s final layer during a single forward pass. Given an image-question pair  $(I, q)$ , modern MLLMs process the image through a Vision Transformer (ViT) encoder into  $N \times N$  patch tokens, which are then projected into  $T$  image tokens for the LLM. We then extract the softmax cross-attention of the first response token to all image tokens, yielding  $\mathbf{A}(I, q) \in \mathbb{R}^{H \times 1 \times T}$  where  $H$  is the number of attention heads. Averaging over the heads gives the attention map:

$$\hat{\mathbf{A}}(I, q) = \frac{1}{H} \sum_{h=1}^H \mathbf{A}^h(I, q) \in \mathbb{R}^{1 \times T}. \quad (1)$$

For MLLMs using Transformer-based connectors (*e.g.*, InstructBLIP), we combine the LLM-to-token attention with the connector’s token-to-patch cross-attention to establish spatial correspondence. For MLLMs with direct projection (*e.g.*, LLaVA-1.5), image tokens directly correspond to spatial patches, producing a spatial attention map  $\mathbf{A} \in \mathbb{R}^{B_h \times B_w}$  over ViT output patches. We normalize  $\mathbf{A}$  to form a probability distribution:

$$\mathbf{A}_{\text{norm}}[i, j] = \frac{\mathbf{A}[i, j]}{\sum_{i', j'} \mathbf{A}[i', j']}, \quad (2)$$

where  $\mathbf{A}_{\text{norm}} \in \mathbb{R}^{B_h \times B_w}$  represents the probability that spatial block  $(i, j)$  contains information relevant to answering the question. This attention map serves as the foundation for our Entropy-Scaled Portfolio Generation (Section 3.2.2), requiring only a single forward pass without architectural modifications.

### 3.2.2. Step 2: Entropy-Scaled Portfolio Generation

Given the spatial attention map  $\mathbf{A}_{\text{norm}}$  from Section 3.2.1, we construct an adaptive multi-scale portfolio through two interleaved mechanisms: *entropy-guided scaling* determines *how much* context each crop requires based on attention uncertainty, while *hierarchical center refinement* determines *where* to center crops to handle asymmetric attention distributions.

**Entropy-Guided Scale Determination.** We observe that attention entropy directly correlates with the contextual requirements of a region. Low entropy ( $H \approx 0$ ) indicates highly confident, localized attention, requiring minimal additional context. High entropy ( $H \approx \log |\mathbf{A}|$ ) suggests diffuse attention, indicating ambiguity or relationships between multiple elements that demand broader context to resolve uncertainty.

We compute the normalized Shannon entropy over the

spatial attention distribution:

$$H_{\text{norm}}(I, q) = -\frac{1}{\log(B_h \cdot B_w)} \sum_{i,j} \mathbf{A}_{\text{norm}}[i, j] \log \mathbf{A}_{\text{norm}}[i, j], \quad (3)$$

where  $H_{\text{norm}} \in [0, 1]$  quantifies attention uncertainty. We define adaptive expansion factors as linear functions of  $H_{\text{norm}}$ :

$$\alpha_1(I, q) = 1.2 + 0.6 \cdot H_{\text{norm}}(I, q) \in [1.2, 1.8], \quad (4)$$

$$\alpha_2(I, q) = 1.6 + 1.2 \cdot H_{\text{norm}}(I, q) \in [1.6, 2.8], \quad (5)$$

These hyperparameters were empirically determined. Our analysis shows that the model’s performance remains stable across a reasonable range of these values, indicating that our method is not sensitive to specific choices. A full analysis is provided in the Appendix.

These mappings ensure minimal context expansion ( $1.2\times$ ,  $1.6\times$ ) even for confident attention—preventing Contextual Blindness by always providing intermediate context—while allowing aggressive expansion ( $1.8\times$ ,  $2.8\times$ ) for uncertain cases requiring broader relationships.

**Hierarchical Center Refinement.** In standard multi-scale approaches, attention is assumed to be centered within crops, but attention distributions are often asymmetric. In document images, target cells may lie near table edges, and in outdoor scenes, salient objects may occupy corners. We address this issue through hierarchical refinement where each level’s center is computed based on attention within its parent crop region.

Starting from the global attention center  $\mu_0$  over the entire image, we iteratively refine the center at each scale level  $\ell$  within the region  $\mathcal{R}_\ell$  defined by the previous level’s crop:

$$\mu_\ell(I, q) = \frac{\sum_{(i,j) \in \mathcal{R}_\ell} \mathbf{c}_{ij} \cdot \mathbf{A}_{\text{norm}}[i, j]}{\sum_{(i,j) \in \mathcal{R}_\ell} \mathbf{A}_{\text{norm}}[i, j]} \quad (6)$$

where  $\mathbf{c}_{ij} \in \mathbb{R}^2$  denotes the center coordinate of spatial block  $(i, j)$  in image space. This refinement automatically corrects for asymmetry: if attention within a crop is skewed toward one edge,  $\mu_\ell$  shifts in that direction relative to  $\mu_{\ell-1}$ , ensuring the next scale is optimally positioned to capture relevant context.

**Multi-Scale Portfolio.** Let  $S$  denote the MLLM’s input image resolution. Our final portfolio consists of three crops, each centered at its hierarchically refined location and scaled according to entropy:

- **Crop<sub>focal</sub>**:  $S \times S$  pixels centered at  $\mu_0$  (*focal detail*)
- **Crop <sub>$\alpha_1$</sub>** :  $(\alpha_1 \cdot S) \times (\alpha_1 \cdot S)$  pixels centered at  $\mu_1$  (*immediate context*)
- **Crop <sub>$\alpha_2$</sub>** :  $(\alpha_2 \cdot S) \times (\alpha_2 \cdot S)$  pixels centered at  $\mu_2$  (*broader context*)

Each crop is resized to  $S \times S$  pixels, encoded by the MLLM’s vision encoder, and concatenated with the original

image tokens. The MLLM then processes this enriched token sequence—comprising global context (original image) and multi-scale focal information (three crops)—to generate the final answer.

## 4. Experiments

We implement our proposed Visual Funnel methodology in PyTorch. As Visual Funnel is a training-free, inference-time approach, our implementation only requires the standard infrastructure necessary for running inference on the base MLLMs. For all experiments, we use the official, publicly available implementations of the models evaluated, as detailed in Section 4.2.

### 4.1. Datasets

We evaluate our method on eight widely-used VQA benchmarks, categorized into two groups based on evaluation focus.

**Grounded Visual QA.** This category emphasizes fine-grained visual perception requiring models to recognize and reason about small textual or visual details. We evaluate on: (i) **TextVQA** [21], containing 45K questions requiring scene text reading without external OCR tokens; (ii) **DocVQA** [15], featuring 50K questions on document understanding with complex layouts; (iii) **InfoVQA** [16], comprising 5.4K questions on infographics with charts and diagrams.

**Recognition Visual QA.** This category assesses broader visual recognition, spatial reasoning, and knowledge-based understanding. We evaluate on: (i) **GQA** [7], with 22M compositional questions on Visual Genome scene graphs; (ii) **POPE** [11], evaluating object hallucination through binary presence questions; (iii) **A-OKVQA** [19], requiring commonsense and world knowledge integration; (iv) **VQAv2** [5], containing 1M+ questions with balanced answer distributions.

### 4.2. Models

We apply Visual Funnel to three representative MLLMs with different architectural designs, demonstrating generalizability across varying model capacities.

**LLaVA.** The LLaVA series [13] pioneered a simple yet effective approach to vision-language instruction tuning, employing a linear projection to map CLIP visual features into the LLM’s token space. LLaVA-1.5 enhances this design by replacing the linear projection with a two-layer MLP. In our experiments, we evaluate LLaVA-1.5-7B, which uses a CLIP ViT-L/14 visual encoder at a  $336 \times 336$  input resolution, producing 576 visual tokens.

**InstructBLIP.** Built on BLIP-2 [10], InstructBLIP [3] employs a Q-Former architecture that compresses visual information through learnable query tokens. The Q-Former

cross-attends to both visual features and task prompts simultaneously, creating context-dependent visual representations. We evaluate the Vicuna-7B variant, which uses 32 learnable queries and a default input resolution of  $224 \times 224$ .

**Qwen-VL.** The Qwen-VL family [1] introduces versatile vision-language models capable of understanding, localization, and text reading. Qwen2-VL [24] advances this with Naive Dynamic Resolution, which adaptively processes images at their native aspect ratios. We evaluate Qwen2.5-VL-3B-Instruct, a compact instruct-tuned variant suitable for efficient deployment.

### 4.3. Baselines

To evaluate the effectiveness of Visual Funnel, we compare its performance against three key baselines. These are presented in our main results in Table 1 and are designed to isolate the impact of our structured, multi-scale approach.

**Base MLLM (No Cropping).** Our first baseline is the standard zero-shot performance of each base MLLM. The model is applied directly to the benchmark questions without any cropping or other interventions. This baseline establishes the lower-bound performance and serves as the reference point for quantifying the improvements gained from any cropping-based strategy.

**w/ViCrop.** Our primary baseline for comparison is the single-crop enhancement method inspired by ViCrop [32], which we denote as w/ViCrop. This approach leverages the model’s internal attention map from a single forward pass to identify the most salient visual region relevant to the query. A single tight crop of this region is then resized and fed back to the model alongside the original image. This represents the standard, state-of-the-art for single-crop, training-free enhancement methods.

**w/ViCrop (Top-3).** To specifically test our core hypothesis that ‘Structural Diversity > Quantity’, we designed a crucial multi-crop baseline, termed w/ViCrop (Top-3). This baseline uses the same attention map generated for w/ViCrop. However, instead of selecting only the single most salient region, it identifies the top three non-overlapping regions with the highest average attention scores. These three crops are then individually resized and concatenated with the original image tokens. Crucially, this baseline provides the same number of additional crops as our Visual Funnel, allowing us to isolate the effect of our portfolio’s *hierarchical structure* from the simple effect of adding more visual tokens. It directly challenges our method by providing an equal ‘Quantity’ of unstructured information.

Finally, we compare these baselines against our proposed w/Visual Funnel method, which also uses three crops but, unlike w/ViCrop (Top-3), constructs them into an adaptive, hierarchical portfolio as detailed in Section 3.

Model	Grounded Visual QA				Recognition Visual QA		
	TextVQA	GQA	DocVQA	InfoVQA	POPE	AOKVQA	VQAv2
LLaVA-1.5-7B	47.9	60.1	15.9	12.0	85.6	58.7	75.4
w/ViCrop	54.1	60.4	19.4	12.6	87.4	60.4	76.1
w/ViCrop (Top-3)	53.5	60.5	19.2	12.9	87.5	60.6	76.6
<b>w/Visual Funnel</b>	<b>59.1 (+11.2)</b>	<b>61.3 (+1.2)</b>	<b>22.8 (+7.0)</b>	<b>15.1 (+3.1)</b>	<b>88.3 (+2.7)</b>	<b>60.6 (+1.9)</b>	<b>76.7 (+1.3)</b>
InstructBLIP-7B	33.4	49.4	9.2	12.8	84.7	59.9	76.3
w/ViCrop	45.3	49.7	9.9	15.8	86.6	61.3	76.8
w/ViCrop (Top-3)	45.8	49.8	10.1	16.0	87.0	61.5	77.1
<b>w/Visual Funnel</b>	<b>49.8 (+16.4)</b>	<b>50.6 (+1.2)</b>	<b>18.5 (+9.3)</b>	<b>25.1 (+12.3)</b>	<b>87.1 (+0.5)</b>	<b>61.6 (+1.7)</b>	<b>77.2 (+0.9)</b>
Qwen2.5-VL-3B-Instruct	70.1	61.2	51.5	34.2	87.1	57.9	78.9
w/ViCrop	76.0	60.8	54.2	39.4	88.4	59.4	78.2
w/ViCrop (Top-3)	76.7	61.4	55.3	39.9	88.5	60.3	79.4
<b>w/Visual Funnel</b>	<b>79.8 (+9.7)</b>	<b>62.2 (+1.0)</b>	<b>61.1 (+9.6)</b>	<b>49.6 (+15.4)</b>	<b>88.5 (+1.4)</b>	<b>60.4 (+2.5)</b>	<b>79.5 (+0.6)</b>

Table 1. **Main results of our proposed Visual Funnel method.** The benchmarks are categorized into *Grounded Visual QA*, which is highly sensitive to Contextual Blindness, and standard *Recognition Visual QA*. We report zero-shot accuracy (%). Numbers in parentheses denote the absolute performance gain of Visual Funnel over the Base MLLM (No Cropping) baseline. The best results for each model are shown in **bold**. Our method consistently and significantly outperforms all baselines, with the most substantial gains observed on the detail-oriented Grounded Visual QA tasks.

## 4.4. Results

We evaluate Visual Funnel on three representative MLLMs—LLaVA-1.5-7B, InstructBLIP-7B, and Qwen2.5-VL-3B—across seven benchmarks, categorized into Grounded Visual QA and Recognition Visual QA. The full results are presented in Table 1.

### 4.4.1. Analysis on Grounded Visual QA

Grounded Visual QA (TextVQA, DocVQA, InfoVQA, GQA) is the primary testbed for *Contextual Blindness*, which requires reasoning about the relationship between fine-grained details and their surrounding context. As shown in Table 1, Visual Funnel demonstrates significant performance gains over all baselines on these datasets, with improvements ranging from +9.3 to +16.4 points on DocVQA and InfoVQA across different models.

Crucially, this analysis allows us to validate our core hypothesis: *Structural Diversity > Quantity*. We designed the w/ViCrop (Top-3) baseline to isolate the impact of ‘Quantity’ (adding more detail tokens) from ‘Structural Diversity’ (adding multi-scale hierarchical context). Both w/ViCrop (Top-3) and w/Visual Funnel use identical token budgets and extract crops from the same attention map, differing only in the structural organization of information.

The results for w/ViCrop (Top-3) are striking. For Qwen2.5-VL, adding more crops provides only negligible improvements over the standard w/ViCrop (e.g., TextVQA: 76.0 → 76.7, +0.7). More notably, for LLaVA-

1.5, w/ViCrop (Top-3) consistently performs *worse* than the single-crop w/ViCrop (TextVQA: 54.1 → 53.5, -0.6; DocVQA: 19.4 → 19.2, -0.2). This “Redundancy Penalty” strongly indicates that adding unstructured, repetitive information can be actively detrimental to the MLLM’s reasoning. In sharp contrast, our Visual Funnel consistently and significantly outperforms both baselines (e.g., LLaVA/TextVQA: 59.1 vs. 53.5, +5.6; InstructBLIP/DocVQA: 18.5 vs. 10.1, +8.4). This confirms that the hierarchical multi-scale structure (focal → immediate → broader) is the key factor in resolving these complex reasoning tasks.

We note that GQA shows more modest gains (+1.0 to +1.2). This is consistent with our hypothesis, as GQA questions often feature ‘Larger Visual Concepts’ in natural scenes, which rely less on the intermediate-scale context that defines *Contextual Blindness*.

### 4.4.2. Analysis on Recognition Visual QA

On standard recognition benchmarks (POPE, AOKVQA, VQAv2), where *Contextual Blindness* is not the primary bottleneck, w/Visual Funnel shows modest improvements over w/ViCrop (averaging +0.5 to +1.0 points). These gains are substantially smaller than those observed on Grounded Visual QA (averaging +7.1 to +12.7 points).

This finding reinforces our core claim. In the absence of complex requirements for perceiving small visual details with surrounding context, the unique advantage of VF’s ‘Structural Diversity’ is less pronounced. This demonstrates that VF’s significant advantage is precisely targeted

at solving the small-concept *Contextual Blindness* it was designed to address, while maintaining robust performance elsewhere. The differential effectiveness across benchmark categories directly validates our problem formulation: Visual Funnel is not a generic performance enhancer but a targeted solution for *Contextual Blindness* in small detail perception.

#### 4.5. Ablations

We conduct a comprehensive ablation study using the Qwen2.5-VL-3B-Instruct model to systematically dissect the contributions of the core components within Visual Funnel. Our analysis is designed to isolate the distinct effects of our two main steps: (1) *Step 1: Contextual Anchoring*, and (2) *Step 2: Entropy-Scaled Portfolio Generation*. The results are summarized in Table 2.

**The Limitation of Contextual Anchoring Alone.** When we ablate our portfolio construction (Visual Funnel w/o Step 2), applying only our specialized localization prompt (Step 1) to a single-crop baseline, we observe only a marginal performance increase (+0.9 on DocVQA) over the standard ViCrop baseline. This finding is critical: it demonstrates that even a more precise attention map is insufficient to overcome the core issue. This strongly supports our claim that Contextual Blindness is a fundamental problem of information *structure*, which is addressed by our Step 2, not merely localization accuracy from Step 1.

**The Decisive Impact of Entropy-Scaled Portfolio Generation.** Conversely, when we ablate our specialized prompt (Visual Funnel w/o Step 1) and apply our portfolio construction method (Step 2) to a standard, less-precise attention map, the model dramatically outperforms the ViCrop baseline (+5.6 on DocVQA). This result provides direct evidence for our central hypothesis that providing ‘Structural Diversity’ is the key to resolving complex visual reasoning tasks. The performance gains are primarily driven by the adaptive and hierarchical structure of the portfolio generated in Step 2.

**Synergistic Effect of Both Steps.** Finally, our full Visual Funnel (Ours) model, which integrates both steps, achieves the best results. It consistently outperforms all other variants, indicating a clear synergistic effect. The high-quality attention map from Step 1 provides a more robust anchor for the portfolio construction in Step 2, and the portfolio in turn provides the necessary structural context that a single crop cannot. This comprehensive analysis validates our two-step design, confirming that both components are essential and work in concert to effectively resolve Contextual Blindness.

Additionally, we conducted further analyses on the optimal number of portfolio crops and a detailed efficiency comparison (e.g., input tokens, latency) against the ViCrop (Top-3) baseline. Due to space constraints,

Configuration	DocVQA	InfoVQA
ViCrop (Baseline)	54.2	39.4
Visual Funnel w/o Step 2	55.1	40.3
Visual Funnel w/o Step 1	59.8	47.9
<b>Visual Funnel (Ours)</b>	<b>61.1</b>	<b>49.6</b>

Table 2. Ablation study on the core components of Visual Funnel, conducted with the Qwen2.5-VL-3B-Instruct model. We report accuracy on DocVQA and InfoVQA.

these results and their detailed discussions are provided in Appendix.

## 5. Conclusion

In this work, we identified and addressed “Contextual Blindness,” a critical failure mode in Multimodal Large Language Models where the structural disconnect between global context and focal detail impedes fine-grained visual reasoning. We argued that resolving this issue requires moving beyond naive, single-crop integration methods. Our approach is rooted in the central premise that for complex perception, ‘Structural Diversity’ is more critical than the mere ‘Quantity’ of visual information. To this end, we proposed Visual Funnel, a training-free, two-step methodology that operationalizes this principle. By first employing a specialized prompt for robust attention-guided localization and then, crucially, constructing an adaptive multi-scale information portfolio, Visual Funnel provides MLLMs with the necessary hierarchical context to bridge the gap between detail and context. Through extensive experiments on several MLLMs and challenging benchmarks, we demonstrated that Visual Funnel significantly outperforms existing methods. Our ablation studies further validated our design, revealing a ‘Redundancy Penalty’ for unstructured multi-crop inputs and confirming that the hierarchical structure of our portfolio is the key to its success.

## 6. Limitations

While Visual Funnel demonstrates strong performance, we acknowledge several limitations.

First, the effectiveness of our method is predicated on a reasonably accurate initial attention map from Step 1. In rare cases where the MLLM completely fails to localize the region of interest, the quality of the generated portfolio may be compromised.

Second, our current approach is designed to resolve questions centered around a single region of interest. Consequently, it may not be suitable for complex queries that require synthesizing information from multiple, spatially distinct focal points simultaneously.

Finally, while our method is training-free, it introduces a computational overhead at inference time due to the processing of multiple image crops. Although our experiments suggest this is a favorable trade-off for the substantial accuracy gains on detail-oriented tasks, this overhead could be a concern in latency-sensitive scenarios.

## References

- [1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. [3](#) [6](#)
- [2] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jia-peng Luo, Zheng Ma, et al. InternVL2: Better than the best—redefining open-source multimodal models with InternVL2.5. *arXiv preprint arXiv:2412.05271*, 2024. [3](#)
- [3] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. [3](#) [6](#)
- [4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021. [3](#)
- [5] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. [6](#)
- [6] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. *arXiv preprint arXiv:2211.11559*, 2023. [3](#)
- [7] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019. [6](#)
- [8] Anubhooti Jain, Mayank Vatsa, and Richa Singh. Words over pixels? rethinking vision in multimodal large language models. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25) Survey Track*, pages 10481–10489, 2025. [3](#)
- [9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. [3](#)
- [10] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. [3](#) [6](#)
- [11] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. [3](#) [6](#)
- [12] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv preprint arXiv:2311.06607*, 2024. [3](#)
- [13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. [3](#) [6](#)
- [14] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. <https:////llava-vl.github.io/blog/2024-01-30-llava-next/>, 2024. [3](#)
- [15] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2200–2209, 2021. [6](#)
- [16] Minesh Mathew, Viraj Bagal, Rubén Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. InfoGraphicVQA. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. [6](#)
- [17] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. [3](#)
- [18] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016. [3](#)
- [19] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162, 2022. [6](#)
- [20] Baifeng Shi, Boyi Li, Han Cai, Yao Lu, Sifei Liu, Marco Pavone, Jan Kautz, Song Han, Trevor Darrell, Pavlo Molchanov, and Hongxu Yin. Scaling vision pre-training to 4k resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. [3](#)
- [21] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. [6](#)
- [22] Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*, 2023. [3](#)
- [23] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akkiraju, Jihan Xiao, Silvio Savarese, Yann LeCun, Bart Ooguz, et al. Cambrian-1: A

- fully open, vision-centric exploration of multimodal LLMs. *arXiv preprint arXiv:2406.16860*, 2024. 3
- [24] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3, 6
- [25] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language models. *arXiv preprint arXiv:2312.06109*, 2023. 3
- [26] Mingrui Wu, Xinyue Cai, Jiayi Ji, Jiale Li, Oucheng Huang, Gen Luo, Hao Fei, Guannan Jiang, Xiaoshuai Sun, and Rongrong Ji. Controlmllm: Training-free visual prompt learning for multimodal large language models. In *Advances in Neural Information Processing Systems*, 2024. 3
- [27] Penghao Wu and Saining Xie. V\*: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024. 2, 3, 4
- [28] Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, Maosong Sun, and Gao Huang. LLava-UHD: An LMM perceiving any aspect ratio and high-resolution images. *arXiv preprint arXiv:2403.11703*, 2024. 3
- [29] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. *arXiv preprint arXiv:2310.11441*, 2023. 3
- [30] Jiarui Zhang, Jinyi Hu, Mahyar Khayatkhoei, Filip Ilievski, and Maosong Sun. Exploring perceptual limitation of multimodal large language models, 2024. 3
- [31] Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Towards perceiving small visual details in zero-shot visual question answering with multimodal LLMs, 2024. 3
- [32] Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 3, 4, 5, 6

## 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,  $\alpha_1$  (for immediate context) and  $\alpha_2$  (for broader context), are computed as linear functions of the normalized attention entropy  $H_{\text{norm}}$ :

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

where  $\beta_k$  represents the base expansion factor (minimum context size) and  $\gamma_k$  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 ( $\gamma$ )

First, we investigate the necessity of the adaptive scaling mechanism. We vary the sensitivity coefficients  $\gamma_1$  and  $\gamma_2$  while keeping the base factors  $\beta$  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	$\gamma_1$	$\gamma_2$	DocVQA Acc. (%)	$\Delta$
Static (Fixed Size)	0.0	0.0	59.5	-1.6
Weak Adaptation	0.3	0.6	60.4	-0.7
<b>Default (Ours)</b>	<b>0.6</b>	<b>1.2</b>	<b>61.1</b>	–
Strong Adaptation	0.9	1.8	60.8	-0.3

Table 3. **Ablation on Entropy Sensitivity.** We analyze the impact of the sensitivity coefficient  $\gamma$ . 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 ( $\beta$ )

Next, we analyze the stability of the base crop sizes. We shift the intercept values  $\beta_1$  and  $\beta_2$  by  $\pm 0.2$  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	$\beta_1$	$\beta_2$	DocVQA Acc. (%)	$\Delta$
Tighter Crops (-0.2)	1.0	1.4	60.5	-0.6
<b>Default</b>	<b>1.2</b>	<b>1.6</b>	<b>61.1</b>	–
Wider Crops (+0.2)	1.4	1.8	60.9	-0.2

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

# Crops ( $K$ )	Configuration	Token Usage	DocVQA Acc.	$\Delta$
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
<b>3</b>	<b>Focal + Imm. + Broader</b>	$\sim 1.9 \times$	<b>61.1</b>	–
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
<b>w/ Visual Funnel (Ours)</b>	$\sim 2,300$	<b>890</b>	<b><math>1.98 \times</math></b>	<b>61.1</b>	<b>High</b>

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.

## B. Ablation on Portfolio Size

In Visual Funnel, we construct a hierarchical portfolio consisting of three specific crops: *Focal* ( $\mu_0$ ), *Immediate Context* ( $\mu_1$ ), and *Broader Context* ( $\mu_2$ ), 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.
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 degra-

dation (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.

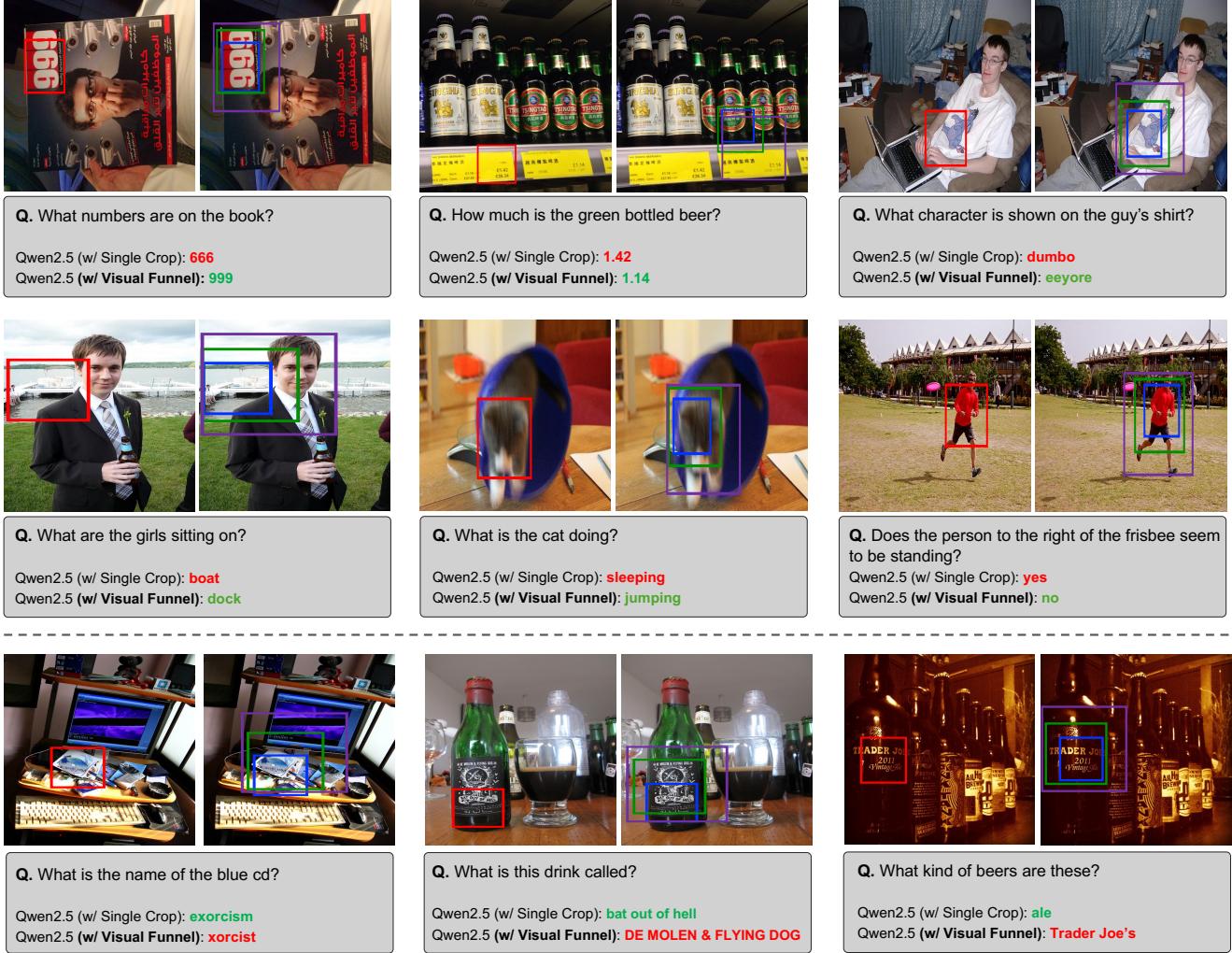


Figure 4. **Qualitative comparison between the Single-Crop baseline and Visual Funnel.** We visualize the inputs and predictions using Qwen2.5-VL-3B-Instruct. The **Red box** represents the input for the standard Single-Crop baseline (w/ ViCrop), while the **Blue, Green, and Purple** boxes represent 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.