Task-Aware Resolution Optimization for Visual Large Language Models

Anonymous ACL submission

Abstract

Real-world vision-language applications demand varying levels of perceptual granularity. However, most existing visual large language models (VLLMs), such as LLaVA, pre-assume a fixed resolution for downstream tasks, which leads to subpar performance. To address this problem, we *first* conduct a comprehensive and pioneering investigation into the resolution preferences of different vision-language tasks, revealing a correlation between resolution preferences with ❶ image complexity, and ❷ uncertainty variance of the VLLM at different image input resolutions. Building on this insight, we propose an empirical formula to determine the optimal resolution for a given vision-language task, combining these two factors. *Second*, based on rigorous experiments, we propose a novel parameter-efficient fine-tuning technique to extend the visual input resolution of pre-trained VLLMs to the identified optimal resolution. Extensive experiments on various vision-language tasks validate the effectiveness of our method.

1 Introduction

Visual Large Language Models (VLLMs) represent a powerful class of models capable of handling vision-language tasks (Yin et al., 2023; Liu et al., 2023a, 2024; Alayrac et al., 2022). There is a growing body of research focused on the application of VLLMs in real-world scenarios, where different tasks necessitate varying levels of perceptual granularity. For instance, autonomous driving systems require high resolution to capture multiple objects and intricate details (Zhou et al., 2023; Ding et al., 2023), whereas image classification tasks involving singular, simple objects can be effectively performed at lower resolutions (Li et al., 2024a, 2023d; Zhang et al., 2024). Despite this, most existing VLLMs, *e.g.*, LLaVA, pre-assume a fixed resolution for downstream tasks, which leads to sub-optimal performance (Liu et al., 2023b,a; Li et al., 2023b). A direct "*exhaustive training*" strategy to adapt current VLLMs for diverse vision-language applications by training the models at different resolutions during the pre-training phase to create a series of checkpoints corresponding to various image input resolutions, followed by the selection of the most effective checkpoint for downstream tasks. While this method is viable, it incurs significant training costs. Consequently, we pose the first research question (***RQ1***):

For a given vision-language task, how can we accurately determine the optimal resolution ***without such exhaustive training*** *for VLLMs?*

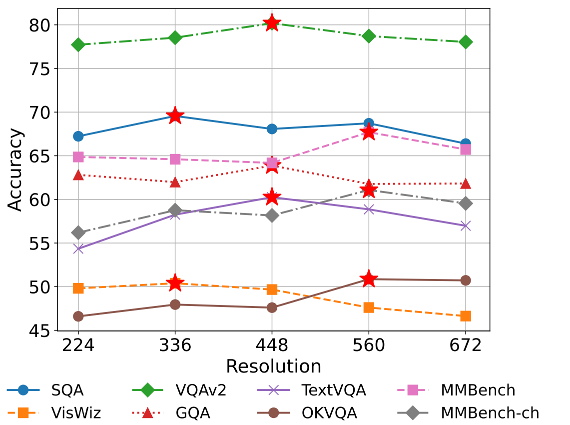


Figure 1: Resolution preference across eight tasks; ★ marks the optimal resolution for each task.

To answer ***RQ1***, we conduct a comprehensive and pioneering investigation into the resolution preferences across eight widely-studied vision-language tasks, utilizing VLLMs with five varying input image resolutions, as shown in Figure 1. Our findings reveal that directly choosing the lowest (224^2) and highest (672^2) resolution leads to sub-par performance across tasks. On the other hand, we observe diverse preferences for the intermediate resolutions, with optimal choices scattered among 336^2, 448^2, and 560^2.

To determine the resolution preference for different tasks, we propose two heuristic methods:

❶ image complexity, which measures the intrinsic complexity of a given image [❖ Secion 3.2.1]. ❷ uncertainty variance, which measures the variance of uncertainty in the model predictions at different image input resolutions [❖ Secion 3.2.2]. Through empirical analysis across eight vision-language tasks, we find that both the complexity scores and model uncertainty variance exhibit a generally positive correlation with the preferred resolution for each task. Building on this insight, we propose an empirical formula integrating both heuristics to determine the optimal resolution for each vision-language task [❖ Secion 3.2.3]. We utilize three reference tasks to optimize a single hyperparameter of this empirical formula, and the fitting results across five additional tasks affirm its generalizability.

Once the optimal resolution for a given vision-language task is identified, the next step is adapting the current VLLM to the identified resolution. While the training-free method exists for resolution extension, we empirically find it would lead to performance degradation, suggesting that training-based approaches are essential. However, re-training a VLLM with another resolution from scratch incurs significant costs. This prompts our second research question (***RQ2***):

*How can we **efficiently** adapt a pre-trained VLLM to the designated resolution without compromising performance?*

To tackle this problem, we propose a post-training strategy that extends the image input resolution of an existing VLLM checkpoint. We conduct a preliminary experiment to identify which parameters within the VLLM are crucial for performance enhancement. Based on the findings, we propose a parameter-efficient fine-tuning (PEFT) approach, which only requires updating a few parameters in each VLLM component: the positional embedding parameters of the visual encoder, the projector parameters, and the LoRA adapter parameters of the LLM backbone. Empirical studies demonstrate that our method achieves a compelling efficiency-performance trade-off. In summary, this paper has the following contributions:

- **Novel Discovery.** Through a comprehensive and pioneering investigation, we discover that different vision-language tasks prefer distinct resolutions.
- **Empirical Formula.** We find these preferences correlated with image complexity and model uncertainty variance on samples at different input image resolutions. We then propose an empirical formula to adaptively determine the optimal resolution for various downstream vision-language tasks without exhaustively training VLLMs.
- **Efficient Adaptation.** We introduce a PEFT approach to extend the input image resolution of LLaVA through post-training, containing three components, including vision module PEFT, language module PEFT, and the projector tuning.

2 Related Work

VLLMs and Resolution Sensitivity. VLLMs extend the capabilities of LLMs to multimodal tasks by processing both text and visual inputs (Alayrac et al., 2022; Liu et al., 2023a). This work focuses on VLLMs employing an encoder-decoder architecture with a modality connector, a common paradigm represented by models like LLaVA (Liu et al., 2023b). However, a prevalent limitation is their reliance on a fixed input resolution, which can lead to suboptimal performance across diverse downstream tasks. The sensitivity of visual models like CNNs and ViTs to resolution is well-known (Borji, 2021; Dehghani et al., 2023), a challenge VLLMs inherit and which our work addresses by proposing task-aware optimization. Further details on VLLM architectures and the historical context of resolution sensitivity are provided in Appendix A.1.

Strategies for Adapting VLLMs to Varying Resolutions. To address fixed-resolution limitations, various strategies exist. Many recent VLLMs natively support dynamic resolutions via architectural innovations (e.g., 2D RoPE in Qwen2VL (Wang et al., 2024), efficient high-resolution processing in MiniCPM (Yao et al., 2024), or varied aspect ratio handling in LLaVA-UHD (Guo et al., 2025)), but these typically require extensive pre-training. Other techniques focus on processing high-resolution inputs through methods like image patching (Chen et al., 2024; wen Dong et al., 2024), region-aware mechanisms (Wu and Xie, 2023; Zhao et al., 2024; Zhang et al., 2023), or by optimizing computational costs (Li et al., 2024a).

Our approach differs significantly by enabling lightweight, post-training adaptation of *existing* VLLM checkpoints. We first determine an optimal task-level resolution using interpretable heuristics and then efficiently adapt the model using a PEFT

Table 1: Key Distinctions: Our Task-Aware Adaptation vs. Native Dynamic Resolution VLLMs

Comparative Aspect	Our Method (Task-Aware Adaptation)	Native Dynamic Resolution VLLMs
Resolution Handling	✓ Task-Optimized (Post-hoc PEFT)	✓ Inherent (Architectural Design)
Optimal Resolution for Task	✓ Explicitly Selected (Heuristic-driven)	✗ Generally Implicit / Not Primary Focus
Adaptation Approach	✓ Lightweight PEFT (on existing models)	✗ Extensive Pre-training / Full Fine-Tuning
Base Model Architecture	✗ Unchanged (Adapts standard VLLMs)	✓ Often Modified (e.g., RoPE, specialized ViTs)
Resolution Decision Informed by Textual Context	✓ via model uncertainty with text	✗ Typically visual input properties only
Adaptation Cost	✓ Low (Efficient for existing checkpoints)	✗ High (Resource-intensive initial training)

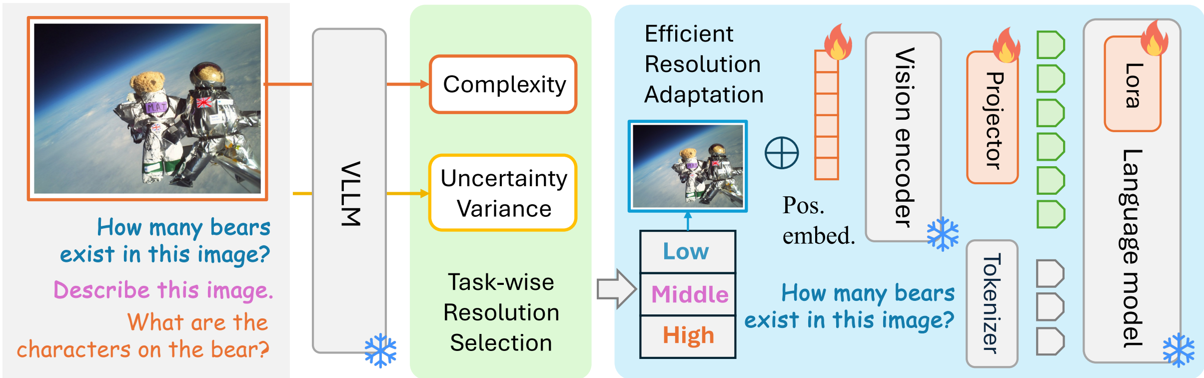


Figure 2: Our method comprises two components: the first component identifies the optimal image input resolution for a given vision-language task (depicted in green), while the second component adapts the VLLM to the selected image input resolution (depicted in blue).

strategy, without architectural changes or retraining from scratch. This offers a practical pathway to enhance existing models. Key differences between our method and native dynamic resolution VLLMs are summarized in Table 1. Further details on these dynamic resolution models and other high-resolution techniques are in Appendix A.2.

3 Methodology

This section elaborates on our proposed methodology. Section 3.1 presents an overview, followed by a detailed explanation of each component in Sections 3.2 and 3.3.

3.1 Method Framework

Figure 2 illustrates our proposed two-stage approach. The first stage, task-specific resolution selection, aims to identify the optimal input resolution for a given vision-language task. This is achieved by first employing two heuristic metrics: image complexity (detailed in Section 3.2.1) and model uncertainty variance across different resolutions (Section 3.2.2). Building on these heuristics, we then introduce an empirical formula (Section 3.2.3) to determine this optimal task-level resolution. Once the optimal resolution is identified for a particular task, the second stage, VLLM adaptation, adjusts the pre-trained VLLM to operate effectively at this new resolution. This adaptation is performed using a PEFT strategy (detailed in Section 3.3), which involves post-training an existing VLLM checkpoint without requiring a full retraining from scratch. Subsequently, this adapted model is deployed to process all samples, including previously unseen ones, for that specific task at the determined optimal resolution.

3.2 Task-wise Optimal Resolution Selection

As highlighted in Section 1, different vision-language tasks have varying requirements for the perceptual capacity of VLLMs. Therefore, it is critical to do task-wise resolution selection. While tuning VLLMs at different image input resolutions and obtaining the best-performing one is feasible, it imposes heavy training costs, which leads to ***RQ1***. In this section, we propose a training-free method for determining the optimal resolution for a specific vision-language task, utilizing two heuristic approaches. We then derive an empirical formula to guide the resolution selection process.

3.2.1 Measuring Image Complexity

The initial stage of VLLM processing involves visual perception. Intuitively, more complex images, requiring finer perceptual granularity, may benefit from higher input resolutions. Consequently, for a given vision-language task, the inherent complexity of its associated images can serve as an indicator of resolution preference.

To quantitatively assess image complexity, we adopt the method by Mahon and Lukasiewicz (2023), which leverages the Minimum Description Length (MDL) principle for hierarchical pixel clustering to identify perceptually meaningful structures. Key steps involve initial MDL-based pixel clustering, followed by constructing and recursively clustering patch signatures to capture multi-scale complexity, with the final score derived from summed entropies. For a comprehensive algorithmic description, we refer the reader to the original publication (Mahon and Lukasiewicz, 2023) and their publicly available implementation[1].

In our framework, this score, averaged across all sampled images for a given task T, is denoted as $C(T)$ and serves as a key heuristic (Section 3.2.3). We chose this recent method for its efficacy in capturing perceptual complexity and its favorable comparisons to alternatives (Khan et al., 2022; Machado et al., 2015; Redies et al., 2012; De Siqueira et al., 2013), as demonstrated in Mahon and Lukasiewicz (2023).

3.2.2 Measuring Uncertainty Variance Across Resolutions

Beyond static image complexity (Section 3.2.1), VLLM prediction uncertainty offers insights into visual-linguistic interplay and sensitivity to resolution variations. We thus introduce a second heuristic based on model uncertainty variance.

Specifically, consider a VLLM pre-trained at a fixed resolution (e.g., 336^2 for LLaVA). We first extend its visual encoder's capacity to handle a different, typically higher, resolution by interpolating its positional embeddings, a technique employed in prior works (Bai et al., 2023; Li et al., 2023b). Let M_1 denote the original model operating at its native resolution, and M_2 denote the same model adapted to operate at the extended resolution (without further fine-tuning at this stage). To assess uncertainty robustness, we apply random augmentations to the input images of a given task T using the RandAugment algorithm (Cubuk et al., 2020). Inference is then performed on these augmented task samples using both M_1 and M_2, from which we extract the softmax probability distributions for each generated token.

[1] https://github.com/Lou1sM/meaningful_image_complexity

Token uncertainty is quantified by information entropy: $H(p) = -\sum_{i=1}^{n} p_i \log p_i$, where p_i is the i^{th} token's softmax probability. Sample-level uncertainty is the average entropy of all generated tokens in an output sequence (computed independently for M_1, M_2). Task-level average uncertainties, $U_1(T)$ and $U_2(T)$, are then derived by averaging these sample-level uncertainties across all selected samples for task T. The uncertainty variance, $V(T)$, for task T is the relative change: $V(T) = \frac{U_2(T)-U_1(T)}{U_1(T)}$. A higher $V(T)$ indicates greater sensitivity of model uncertainty to resolution changes for task T. This $V(T)$ is the second heuristic for our empirical formula (Section 3.2.3).

This uncertainty-based heuristic offers two main advantages to complement the static image complexity: (1) by computing entropy from tokens generated by the VLLM, it inherently accounts for both visual and linguistic features during inference; and (2) it directly quantifies the variance induced by resolution changes, thereby capturing the dynamic effects of such shifts. Notably, calculating this heuristic involves extending VLLM input resolution without parameter tuning, avoiding extra training costs at this stage.

3.2.3 Empirical Formula for Optimal Resolution Estimation

Inspired by the intuition that tasks with more complex imagery or higher resolution sensitivity (in terms of model prediction uncertainty) might benefit from increased input resolutions, we propose an empirical formula to estimate the optimal resolution for a given vision-language task. This intuition, regarding the positive correlation of image complexity and uncertainty variance with preferred resolution, is further explored and validated in Section 4.2. The proposed formula is:

$$Reso(T) = Reso_0 \cdot (1 + k \cdot C(T) \cdot V(T)) \quad (1)$$

Here, $Reso_0$ is the VLLM's baseline input resolution (e.g., 336 for LLaVA), serving as a reference for scaling. $C(T)$ is the average normalized image complexity for task T (Section 3.2.1), and $V(T)$ is its average uncertainty variance. The term k is a user-specified, non-negative hyperparameter modulating the heuristics' combined influence. The expression $(1+k\cdot C(T)\cdot V(T))$ thus acts as a scaling factor, adjusting $Reso_0$ based on task characteristics. The value of k is determined empirically using reference tasks, as discussed in Section 4.3.1.

Resolution	SciQA-IMG	VizWiz	VQAv2	GQA	TextVQA	OKVQA	MMBench	MMBench-CN
224 × 224	67.23	49.81	77.72	62.81	54.35	46.60	64.86	56.19
336 × 336	**69.56**	**50.39**	78.53	61.98	58.25	47.95	64.60	58.76
448 × 448	68.07	49.67	**80.19**	**63.87**	**60.25**	47.60	64.18	58.16
560 × 560	68.72	47.61	78.71	61.77	58.86	**50.86**	**67.70**	**61.08**
672 × 672	66.39	46.63	78.04	61.82	56.98	50.72	65.72	59.54

Table 2: A comprehensive investigation conducted to explore resolution preferences across eight vision-language tasks. For each task, the accuracy scores corresponding to five different resolutions are presented.

3.3 Parameter-efficient Resolution Adaptation

After determining the optimal resolution for a given task, the next step is adapting the VLLM to the selected resolution. To answer ***RQ2***, We propose a parameter-efficient fine-tuning (PEFT) approach that post-train an existing VLLM checkpoint, thus avoiding retraining from scratch.

As depicted in Figure 2, existing VLLMs (e.g., LLaVA) consist of three main components: a visual encoder, a projector mapping visual features to the text embedding space, and an LLM backbone generating language tokens. Increasing input resolution introduces more image patches, causing incompatibility with the original position embeddings. To address this, we interpolate the position embeddings from the initial number of patches (e.g., 24^2) to the extended number (e.g., 32^2), following previous research (Bai et al., 2023; Li et al., 2023b). Although this allows the VLLM to process extended resolutions, performance degrades without further adaptation (as discussed in Secion 3.2). To counter this performance decline, we employ a PEFT method that fine-tunes three key components: (1) position embeddings within the visual encoder, essential for handling additional patches; (2) the lightweight projector parameters; and (3) the parameters of the LoRA adapters integrated into the LLM backbone. By keeping all other parameters frozen, the PEFT approach offers an efficient method for adaptation. Figure 2 provides a visual representation of the components that are fine-tuned versus those that remain frozen.

4 Experiments

This section presents the empirical evaluation of our proposed method. We first introduce the implementation details in Section 4.1, followed by an in-depth analysis of the results, including the investigation into resolution preferences, task-wise resolution selection, and the findings from the ablation study in Section 4.2, 4.3, and 4.4, respectively.

4.1 Implementation Details

VLLM Selection. For our experiments, we select the LLaVA-1.5-7B checkpoint (Liu et al., 2023b) as the representative VLLM for evaluation.

Resolution Configurations. We explore five image resolutions: 224^2, 336^2, 448^2, 560^2, and 672^2. These values cover the resolution spectrum commonly used in previous studies (Liu et al., 2023b,a).

Vision-Language Tasks. Our evaluation encompasses eight vision-language tasks, with details introduced in Appendix B.1.

Baseline Methods. In addition to the original LLaVA model, we compare our method with several state-of-the-art approaches. Besides, we report the performance of position embedding interpolation as a representative of the training-free methods to extend the image input resolution of VLLMs. The details are introduced in Appendix B.2.

Post-training Details. To initialize the position embedding parameters of the visual encoder (Vision Transformer) in LLaVA during resolution adaptation, we employ extended position embeddings derived through positional embedding interpolation, as described in Appendix B.2. Following the instructions provided by the LLaVA authors[2], we concentrate on stage 2 fine-tuning, incorporating the additional parameters for position embeddings in the visual encoder, alongside the LoRA adapter and projector parameters. The fine-tuning process utilizes images from five datasets: COCO (Lin et al., 2014), GQA (Hudson and Manning, 2019), OCR-VQA (Mishra et al., 2019), TextVQA (Singh et al., 2019), and Visual Genome (Krishna et al., 2017). For more details on the construction of the image-text pairs used in training, we refer readers to (Liu et al., 2023a). It is crucial to note that this post-training stage is designed solely to adapt the VLLM to the newly selected input resolution, not to specialize it for a particular task.

Further details regarding the overall method im-

[2]https://github.com/haotian-liu/LLaVA/tree/main?tab=readme-ov-file#train

Table 3: Distributions of image complexity and uncertainty variance across eight tasks.

	vizwiz	SciQA-IMG	TextVQA	GQA	VQAv2	OKVQA	MMBench	MMBench-CN
Resolution Preference	336 × 336		448 × 448			560 × 560		
Complexity (C)	0.2191	0.1437	0.2919	0.3236	0.3017	0.3112	0.2323	0.2329
Average	0.1814		0.3058			0.2588		
Uncertainty Variance (V)	1.83%	6.47%	4.88%	5.34%	5.26%	6.72%	10.79%	10.45%
Average	4.15%		5.16%			9.32%		
C × V	0.0040	0.0093	0.0142	0.0173	0.0159	0.0209	0.0251	0.0243
Average	0.0067		0.0158			0.0234		

Table 4: Comparison between our method and baseline approaches, highlighting the best scores in bold. *indicates that the training images or annotations of the datasets were observed during training.

Method	LLM	Resolution	Post-training	VQAv2	GQA	TextVQA	OKVQA	MMBench	MMBench-CN
BLIP-2	Vicuna-13B	224 × 224	-	65.00	41.00	42.50	-	-	-
InstructBLIP	Vicuna-7B	224 × 224	-	-	49.20	50.10	-	36.00	23.70
InstructBLIP	Vicuna-13B	224 × 224	-	-	49.50	50.70	-	-	-
Shikra	Vicuna-13B	224 × 224	-	77.40*	-	-	-	58.80	-
IDEFICS-9B	LLaMA-7B	224 × 224	-	50.90	38.40	25.90	-	48.20	25.20
IDEFICS-80B	LLaMA-65B	224 × 224	-	60.00	45.20	30.90	-	54.50	38.10
Qwen-VL	Qwen-7B	448 × 448	-	78.80*	59.30*	**63.80***	-	38.20	7.40
Qwen-VL-Chat	Qwen-7B	448 × 448	-	78.20*	57.50*	61.50*	-	60.60	56.70
LLaVA-1.5	Vicuna-7B	336 × 336	-	78.53*	61.98*	58.25	47.95	64.60	58.76
LLaVA-1.5	Vicuna-7B	448 × 448	✗	77.82*	61.29*	56.61	47.38	63.32	57.73
LLaVA-1.5	Vicuna-7B	448 × 448	✓	**80.19***	**63.87***	60.25	47.60	64.18	58.16
LLaVA-1.5	Vicuna-7B	560 × 560	✓	78.71*	61.77*	58.86	**50.86**	**67.70**	**61.08**
LLaVA-1.5	Vicuna-7B	Adaptive	✓	**80.19***	**63.87***	60.25	**50.86**	**67.70**	**61.08**
LLaVA-1.5	Vicuna-13B	336 × 336	-	80.00*	63.30*	61.30	-	67.70	63.60

† Shikra, primarily a referential dialogue model, is evaluated here in a VQA instruction-following setting for broader comparison.

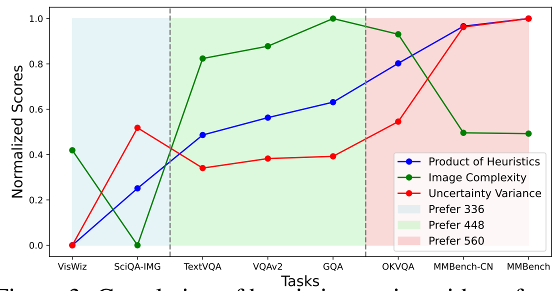


Figure 3: Correlation of heuristic metrics with preferred task resolution. The product of $C(T)$ and $V(T)$ exhibits a more consistent correlation compared to individual heuristics. All metrics are normalized for visualization.

plementation and our PEFT setup are provided in Appendix B.3 and B.4, respectively.

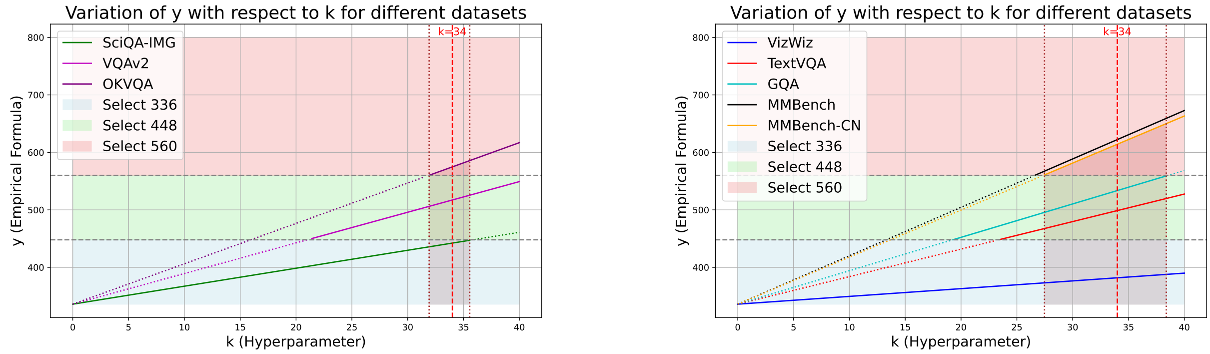
4.2 Analyzing Resolution Preferences Across Vision-Language Tasks

We systematically analyze resolution preferences across vision-language tasks (Table 2), revealing two key findings: ❶ Performance is suboptimal at very low (224^2) or very high (672^2) resolutions—low resolution limits visual detail capture, while high resolution disrupts adaptation and introduces irrelevant tokens. ❷ Optimal resolutions lie in the mid-range (336^2, 448^2, 560^2), varying by task, which underscores the need for task-specific selection.

After identifying task-specific resolution preferences, we explore the correlation between optimal resolutions and our proposed heuristics of image complexity and uncertainty variance, as shown in Table 3. We can draw the following conclusions: ❶ No increasing trend is observed between 448^2 and 560^2 in image complexity, but a noticeable gap exists between 336^2 and 448^2, suggesting that image complexity differentiates tasks favoring 336^2 from those preferring higher resolutions. ❷ There is a positive correlation between preferred resolution and uncertainty variance across tasks, with an upward trend showing that uncertainty variance reliably indicates resolution preference. ❸ Some exceptions exist, e.g., GQA prefers lower resolution than MMbench but has higher image complexity, and SciQA-IMG has higher uncertainty variance but favors a lower resolution than TextVQA. Multiplying the scores of two heuristics provides a more consistent correlation, as shown in Figure 3.

4.3 Evaluating Heuristic-Based Task-Specific Resolution Selection

The investigation presents the correlation between task-specific resolution preferences and two heuristics. This section describes hyperparameter determination for our empirical formula and summarizes the performance of models using this strategy.



(a) Optimization of the hyperparameters in the empirical formula using three reference tasks.

(b) The empirical formula demonstrates effective generalization across five vision-language tasks.

Figure 4: Applying the empirical formula to determine the optimal resolution for vision-language tasks.

4.3.1 Applying the empirical formula to determine the optimal resolution

To optimize the hyperparameter in Equation 1, we select three reference tasks representing different visual perception requirements (Figure 6 in Appendix D shows task images). Tasks with simpler images (e.g., Figure 6a) are considered low resolution, while complex images (e.g., Figure 6c) require higher resolutions. Intermediate tasks (e.g., Figure 6b) represent medium resolution. SciQA-IMG, VQAv2, and OKVQA are separately chosen to reflect low, medium, and high resolution needs.

When tuning the hyperparameter k, we focus on 336^2, 448^2, and 560^2. The constant $Reso_0$ is set to 336 (default LLaVA resolution). The formula selects the resolution based on the value of k. For example, a value of 500 leads to 448^2. Figure 4a visualizes the relationship between hyperparameter values and selected resolutions. For simplicity, we select $k = 34$, which results in optimal resolution selection for the reference tasks. Additionally, as shown in Figure 4b, this value generalizes well to other tasks, achieving the best resolution for each.

While the empirical formula demonstrates good generalization with a fixed k value, its practical application to a new task involves sampling a subset of data from that task to compute $C(T)$ and $V(T)$. Appendix C analyzes the formula's robustness to varying sample sizes, including the relationship between sampling ratio and prediction success, and the influence of heuristic distributions, offering guidance for data-limited applications.

4.3.2 Overall results of Task-wise Adaptive Model and Baselines

Table 4 presents the performance of baseline methods and LLaVA variants across six tasks that demand high visual perception capacity from VLLMs. Among the LLaVA variants, the training-free method to extend the input resolution through PE interpolation shows performance degradation at varying levels. This confirms that the position embeddings in the visual encoder and LLM backbone in LLaVA cannot fully adapt to the increased number of image tokens without post-training. On the other hand, the task-wise adaptive LLaVA variant, which optimally selects the input resolution for each task, achieves the best overall performance compared to fixed-resolution LLaVA variants, regardless of whether the resolution is 336^2, 448^2, or 560^2. Notably, the task-wise adaptive LLaVA variant with a 7B backbone performs comparably to the 13B variant, underscoring the importance of adaptive perception capacity in VLLMs.

When comparing the task-wise adaptive LLaVA variant with other state-of-the-art baselines, it outperforms all but the TextVQA task. In the case of TextVQA, the Qwen-VL and Qwen-VL-Chat methods have observed training images or annotations of the dataset during their training. Importantly, as previous studies (McKinzie et al., 2024a) have highlighted, resolution plays a crucial role during pretraining. The Qwen-VL series are pretrained at an image resolution of 448^2, while the LLaVA variants were fine-tuned at extended image resolutions in a post-training phase with far fewer data (665K) compared to Qwen's 1.4B pretraining and 50M fine-tuning samples. Nevertheless, the task-wise adaptive LLaVA variant achieves better overall results than the Qwen-VL series.

The superior performance of the task-wise adaptive LLaVA variant across multiple vision-language tasks demonstrates that, compared to *fixed-resolution* approaches, *adaptive resolution*

Resolution	ViT PE	Projector	LoRA Adapter	VQAv2	GQA	TextVQA
336×336	-	-	-	78.53 (-2.07%)	61.98 (-2.96%)	58.25 (-3.32%)
448×448	✗	✗	✗	77.82 (-2.96%)	61.29 (-4.04%)	56.61 (-6.04%)
448×448	✓	✗	✗	75.32 (-6.07%)	59.98 (-6.09%)	53.44 (-11.30%)
448×448	✗	✓	✗	72.94 (-9.04%)	55.31 (-13.40%)	51.41 (-14.67%)
448×448	✗	✓	✓	79.47 (-0.90%)	63.41 (-0.72%)	58.06 (-3.63%)
336×336	✓	✓	✓	79.33 (-1.07%)	63.33 (-0.85%)	58.19 (-3.42%)
448×448	✓	✓	✓	**80.19**	**63.87**	**60.25**

Table 5: Ablation Analysis of PEFT Components, ✗ and ✓ indicate whether the module is post-trained.

selection is more suitable for real-world applications. So far, we have verified the effectiveness of our task-wise resolution selection strategy through the generalization of the empirical formula and the overall experimental results, answering **RQ1**.

4.4 Ablation Analysis of PEFT Components for Performance

To evaluate the contribution of each component in our PEFT method, we conduct an ablation study (Table 5), examining the impact of tuning three key parameters: position embeddings in the visual encoder, LoRA adapters in the LLM backbone, and projector parameters. We also assess whether performance gains stem from the additional training epoch introduced by post-training by conducting full training at the original resolution (336^2).

Results show that tuning each component is crucial. Tuning only position embeddings or projector parameters leads to significant drops, even compared to training-free positional embedding interpolation. While jointly tuning projector parameters and LoRA adapters improves performance, it remains suboptimal without tuning position embeddings. Additionally, post-training at 336^2 provides only marginal gains over full training or projector + LoRA tuning at 448^2. Notably, on TextVQA, post-training at 336^2 offers no improvement over the original checkpoint, suggesting that gains at 448^2 primarily stem from enhanced perceptual capabilities, not extra training. Overall, our results highlight the importance of each component in PEFT and validate its effectiveness in addressing **RQ2**.

5 Case Study

Table 6 presents two illustrative case studies demonstrating the impact of our heuristics on VLLM performance. Visual inputs (Figs. 7a, 7b, and 7c) are in Appendix E.

As shown in Table 6 (top), we present the VLLM with two images of differing complexities for the same question: "Who is standing?". At the 336^2 resolution, the model correctly identifies the "woman" in the simpler image. However, for the more intricate image with higher complexity, it fails, incorrectly predicting "umpire" instead of "batter". This suggests that more visually complex images may necessitate higher input resolutions for accurate VLLM perception.

Table 6: Case studies: VLLM performance with varying image complexity and question difficulty.

Case 1: Varying Image Complexity (Question: "Who is standing?")			
Image	$C(T)$	Pred. (336^2)	Correct Answer
Fig. 7a	11.35	woman (✓)	woman
Fig. 7b	20.62	umpire (✗)	batter
Case 2: Varying Question Difficulty (Image: Fig. 7c)			
Question	$V(T)$	Pred. (336^2)	Pred. (448^2)
Q1: "Sheet material?"	0.42%	plastic (✓)	plastic (✓)
Q2: "Stoves near tap?"	16.51%	NO (✗)	YES (✓)

The second case (Table 6, bottom) uses a single image (Fig. 7c) but poses two questions of differing difficulty, leading to different uncertainty variances ($V(T)$). For the easier question ("What is the sheet made of?"), the VLLM provides the correct answer ("plastic") at both 336^2 and 448^2 resolutions. However, for the more complex question requiring finer detail ("Are there stoves near the freezer to the right of the tap?"), the model fails at 336^2 but succeeds at the higher 448^2 resolution. This improved performance at higher resolution for the more uncertain (difficult) question aligns with the core intuition behind our $V(T)$ heuristic, as discussed in Section 3.2.3.

6 Conclusion

In this paper, we take a step towards adapting VLLMs to real-world applications by providing an in-depth investigation of resolution preferences in different vision-language tasks. Based on the findings, we introduce an empirical formula that combines image complexity and uncertainty variance to make task-specific resolution selection without the need for retraining. Additionally, we propose a PEFT approach, enabling extension of the image input resolution for existing VLLM checkpoints. We expect that our research will offer valuable insights for the VLLM research community.

Limitations & Future Work

Our current work has several limitations. Due to computational constraints in an academic environment, we were unable to conduct experiments with larger LLM backbones or retrain models from scratch. This restricts the scope of comparison, particularly against methods requiring extensive pertaining. Moreover, our proposed approach focuses on task-level resolution selection. Future work will explore more granular resolution strategies, such as dynamic sample-level resolution adaptation, which could further improve performance for heterogeneous tasks.

Ethical Statement

This study leverages publicly available datasets (e.g., VQAv2, GQA, TextVQA, OKVQA, MMBench) and pre-trained models (e.g., LLaVA) for evaluation and experimentation. These datasets and models are widely recognized benchmarks in the vision-language research community, distributed under licenses permitting academic and non-commercial use. All artifacts were used in accordance with their intended purposes, without modifications or new data collection. The dataset creators' documentation ensures compliance with ethical guidelines, including the absence of personally identifiable or offensive content.

No ethics review board approval was required, as this research does not involve human subject data or sensitive information. However, we acknowledge that the underlying datasets may contain biases or inaccuracies, which could affect model fairness and generalization. Future research should explore bias mitigation strategies to ensure fair and responsible deployment of vision-language models. The derivative findings, such as task-specific resolution adaptation strategies, remain compatible with the original licenses and intended use.

References

Fuyu-8b: A multimodal architecture for ai agents.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Yitzhak Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *ArXiv*, abs/2308.01390.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Ali Borji. 2021. Enhancing sensor resolution improves cnn accuracy given the same number of parameters or flops. *arXiv preprint arXiv:2103.05251*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2024. Instructblip: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Fernando Roberti De Siqueira, William Robson Schwartz, and Helio Pedrini. 2013. Multi-scale gray level co-occurrence matrices for texture description. *Neurocomputing*, 120:336–345.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. 2023. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36:2252–2274.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Xinpeng Ding, Jianhua Han, Hang Xu, Wei Zhang, and X. Li. 2023. Hilm-d: Towards high-resolution understanding in multimodal large language models for autonomous driving. *ArXiv*, abs/2309.05186.

Danny Driess, F. Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Ho Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Peter R. Florence. 2023. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*.

Qihang Fan, Quanzeng You, Xiaotian Han, Yongfei Liu, Yunzhe Tao, Huaibo Huang, Ran He, and Hongxia Yang. 2024. Vitar: Vision transformer with any resolution. *arXiv preprint arXiv:2403.18361*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. 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.

Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. 2025. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision*, pages 390–406. Springer.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogagent: A visual language model for gui agents. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14281–14290.

Anwen Hu, Haiyang Xu, Jiabo Ye, Mingshi Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *ArXiv*, abs/2403.12895.

Drew A Hudson and Christopher D Manning. 2019. 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.

IDEFICS. 2023. Introducing idefics: An open reproduction of state-of-the-art visual language model. https://huggingface.co/blog/idefics.

Tariq M Khan, Syed S Naqvi, and Erik Meijering. 2022. Leveraging image complexity in macro-level neural network design for medical image segmentation. *Scientific Reports*, 12(1):22286.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *ArXiv*, abs/2305.03726.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023c. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junyan Li, Delin Chen, Tianle Cai, Peihao Chen, Yining Hong, Zhenfang Chen, Yikang Shen, and Chuang Gan. 2024a. Flexattention for efficient high-resolution vision-language models. *ArXiv*, abs/2407.20228.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2024b. Mini-gemini: Mining the potential of multi-modality vision language models. *ArXiv*, abs/2403.18814.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2023d. Monkey: Image resolution and text label are important things for large multi-modal models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26753–26763.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023c. Mm-bench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Tengchao Lv, Yupan Huang, Jingye Chen, Lei Cui, Shuming Ma, Ya-Chi Chang, Shaohan Huang, Wenhui Wang, Li Dong, Weiyao Luo, Shaoxiang Wu, Guoxin Wang, Cha Zhang, and Furu Wei. 2023. Kosmos-2.5: A multimodal literate model. *ArXiv*, abs/2309.11419.

Penousal Machado, Juan Romero, Marcos Nadal, Antonino Santos, João Correia, and Adrián Carballal. 2015. Computerized measures of visual complexity. *Acta psychologica*, 160:43–57.

Louis Mahon and Thomas Lukasiewicz. 2023. Minimum description length clustering to measure meaningful image complexity. *arXiv preprint arXiv:2306.14937*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. 2024a. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. 2024b. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. 2021. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128.

Christoph Redies, Seyed Ali Amirshahi, Michael Koch, and Joachim Denzler. 2012. Phog-derived aesthetic measures applied to color photographs of artworks, natural scenes and objects. In *European conference on computer vision*, pages 522–531. Springer.

Carl F Sabottke and Bradley M Spieler. 2020. The effect of image resolution on deep learning in radiography. *Radiology: Artificial Intelligence*, 2(1):e190015.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *ArXiv*, abs/2405.09818.

Rui Tian, Zuxuan Wu, Qi Dai, Han Hu, Yu Qiao, and Yu-Gang Jiang. 2023. Resformer: Scaling vits with multi-resolution training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22721–22731.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jian-Yuan Sun, Chunrui Han, and Xiangyu Zhang. 2023. Vary: Scaling up the vision vocabulary for large vision-language models. *ArXiv*, abs/2312.06109.

Xiao wen Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Zhe Chen, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Kai Chen, Conghui He, Xingcheng Zhang, Jifeng Dai, Yuxin Qiao, Dahua Lin, and Jiaqi Wang. 2024. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *ArXiv*, abs/2404.06512.

Penghao Wu and Saining Xie. 2023. V*: Guided visual search as a core mechanism in multimodal llms. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13084–13094.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. 2024. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *ArXiv*, abs/2306.13549.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. 2023. Towards perceiving small visual details in zero-shot visual question answering with multimodal llms.

Pan Zhang, Xiao wen Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, Songyang Zhang, Wenwei Zhang, Yining Li, Yang Gao, Peng Sun, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Hang Yan, Conghui He, Xingcheng Zhang, Kai Chen, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *ArXiv*, abs/2407.03320.

Xiangyu Zhao, Xiangtai Li, Haodong Duan, Haian Huang, Yining Li, Kai Chen, and Hua Yang. 2024. Mg-llava: Towards multi-granularity visual instruction tuning. *ArXiv*, abs/2406.17770.

Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Žagar, Walter Zimmer, Hu Cao, and Alois C. Knoll. 2023. Vision language models in autonomous driving: A survey and outlook. *IEEE Transactions on Intelligent Vehicles*.

A Further Details on Related Work

A.1 VLLM Architectures and Resolution Sensitivity

VLLM Architectures. Vision Large Language Models, as one of the most capable and popular solutions to multimodal tasks, extend the reasoning and generating ability of Large Language Models (LLMs) beyond language modalities to encompass inputs such as images, video, and audio (McKinzie et al., 2024b; Tong et al., 2024; Xue et al., 2024). VLLMs can be categorized according to their architecture (Liu et al., 2023b; Driess et al., 2023; fuy; Team, 2024). The encoder-decoder VLLM paradigm, which is the focus of this study, introduces additional multimodal encoders (typically a vision encoder like ViT) and a modality connector to project multimodal features into the spaces interpretable by language models. The implementations of the modality connector vary; common approaches include a projector that directly maps visual features to the language model's embedding space (Liu et al., 2024, 2023a,b), or a resampler that compresses visual features, possibly using cross-gated attention layers, before integrating them into the LLM decoder (Alayrac et al., 2022; Awadalla et al., 2023; Li et al., 2023a). Our work primarily considers LLaVA-style VLLMs, which adopt an encoder-decoder architecture with a projector connector.

Further Discussion on Resolution Sensitivity in Visual Models. The sensitivity of visual models to input image resolution is a well-established phenomenon. Convolutional Neural Networks (CNNs) inherently leverage inductive biases like local receptive fields and hierarchical feature extraction, tying their performance to spatial information density, where higher resolutions often improve accuracy (Raghu et al., 2021; Borji, 2021; Sabottke and Spieler, 2020). Techniques like dilated convolutions were developed to manage varying receptive field sizes (Chen et al., 2017). Vision Transformers (ViTs), processing images as sequences of patches, also exhibit distinct resolution sensitivities influenced by patch size and pre-training configurations, often struggling with resolutions unseen during training (Fan et al., 2024; Dehghani et al., 2023). Adapting positional embeddings is a common strategy to mitigate this for ViTs (Bai et al., 2023; Li et al., 2023b; Tian et al., 2023). While VLLMs inherit this sensitivity, the interaction with language understanding in multimodal tasks introduces new complexities. Our work aims to quantify and address this specific challenge by proposing a heuristic-driven optimization framework for VLLMs.

A.2 Dynamic Resolution and High-Resolution Techniques in VLLMs

Native Dynamic Resolution VLLMs. A significant line of research focuses on VLLMs with native capabilities to handle dynamic input resolutions, often through architectural innovations or specialized pre-training. For instance, **Qwen2VL** (Wang et al., 2024) employs 2D RoPE for flexible positional encoding. **MiniCPM-V** (Yao et al., 2024) focuses on efficient high-resolution processing, some-times using multi-scale vision encoders. **LLaVA-UHD** (Guo et al., 2025) introduces strategies for ultra-high-definition images and varied aspect ratios, often involving intelligent image slicing. **InternLM-XComposer2-4KHD** (wen Dong et al., 2024) also demonstrates strong capabilities in handling very high resolutions through sophisticated tiling strategies. While these models offer great flexibility, they typically require substantial pretraining and may not explicitly optimize for a single best resolution per task. Our approach, in contrast, focuses on lightweight, post-hoc adaptation of existing VLLMs to a task-specific optimal resolution.

Other High-Resolution Processing Techniques. Beyond models with end-to-end dynamic resolution, other techniques enable VLLMs to process high-resolution information. Some works focus on **using or adapting vision encoders** to directly support higher resolutions within a VLLM framework, such as CogAgent (Hong et al., 2023) with its dense feature integration, or models like MiniGemini (Li et al., 2024b), Kosmos-2.5 (Lv et al., 2023), and Vary (Wei et al., 2023). **Patchification and tiling strategies** are common, where high-resolution images are divided into smaller patches processed by standard encoders, with subsequent feature aggregation; examples include Monkey (Li et al., 2023d), mPLUG-DocOwl (Hu et al., 2024), and LLaVA-NEXT (Liu et al., 2024). **Region-aware processing** aims to focus on salient regions, with methods like V* (Wu and Xie, 2023) selecting relevant regions for fine-grained understanding, MG-LLaVA (Zhao et al., 2024) using multi-grained GNNs, and PS-VLLM (Zhang et al., 2023) progressively selecting visual tokens. To optimize computational costs associated with high resolutions, FlexAttention (Li et al., 2024a) employs dual tokenization for selective processing of high-resolution tokens.

Our work complements these techniques by first providing a mechanism to determine a task-optimal discrete resolution, to which a model (potentially employing some of these techniques) can then be adapted.

B More Implementation Details

B.1 Vision-Language Tasks

Science-QA (Lu et al., 2022), a multimodal science question answering benchmark featuring over 21k multiple-choice questions on diverse topics. The visual component includes natural images and diagrams, testing the model's ability to integrate both textual and visual information for coherent reasoning and explanation generation. *Vizwiz* (Gurari et al., 2018), a dataset derived from real-world images paired with spoken questions from visually impaired individuals. This task assesses a model's ability to process low-quality, unstructured images and generate accurate responses to conversational queries. *VQAv2* (Goyal et al., 2017), an expanded version of the original Visual Question Answering (VQA) dataset, designed to reduce language biases. It challenges models to deeply understand visual content in order to answer questions about pairs of semantically similar yet visually distinct images. *TextVQA* (Singh et al., 2019), a dataset focusing on a model's capacity to read and reason about textual elements in images, evaluating its ability to integrate Optical Character Recognition (OCR) with visual reasoning to answer questions. *OKVQA* (Marino et al., 2019), a benchmark that requires models to leverage external knowledge beyond image and question analysis, necessitating access to and reasoning with unstructured knowledge sources for accurate answers. *GQA* (Hudson and Manning, 2019), a dataset designed for real-world visual reasoning and compositional question answering, requiring models to demonstrate strong multi-modal understanding, logical reasoning, and the ability to answer questions that necessitate connecting information across both visual and linguistic domains. *MMBench* (Liu et al., 2023c), a comprehensive multimodal evaluation set with over 2,974 multiple-choice questions across 20 ability dimensions, providing a robust assessment of various vision-language skills, such as reasoning, comprehension, and explanation generation. *MMBench-CN*, a variant of MMBench focusing on tasks involving Chinese text and images, evaluating the model's proficiency in processing and understanding multilingual data.

B.2 Baseline Methods

In addition to the original LLaVA model, we compare our method with several state-of-the-art approaches, including BLIP-2 (Li et al., 2023c), InstructBLIP (Dai et al., 2024) (with LLM backbones at two scales), Shikra (Chen et al., 2023), and IDEFICS (IDEFICS, 2023) (also with LLM backbones at two scales), as well as Qwen-VL and Qwen-VL-Chat (Bai et al., 2023). The results for these baseline methods, along with LLaVA with the Vicuna-13B backbone, are cited from previous work (Liu et al., 2023a). For LLaVA with a Vicuna-7B backbone, we report our reproduced results across different vision-language tasks.

As a training-free baseline to extend the image input resolution, we apply positional embedding interpolation to extend the position embeddings of the vision encoder in LLaVA. This technique, widely used for Vision Transformers in VLLMs (Bai et al., 2023; Li et al., 2023b), allows models to handle higher image input resolutions than their original training resolution. We evaluate the performance of this extension without any additional training of the projector and the LLM backbone.

B.3 Method details

Image Complexity Heuristic Approach Image complexity for vision-language tasks is calculated using an open-source tool[3]. We utilize the author-recommended hyperparameters: the number of clusters is set to 8, and the subsample rate is 0.8. To reduce computational overhead, the input image resolution is set to 112×112, and two cluster levels are used, with their combined scores yielding the final complexity value. The complexity scores are normalized via min-max scaling, where the minimum and maximum values are computed from 100 sampled images from the ImageNet dataset (Deng et al., 2009).

RandAugment Perturbation on Image Input When assessing model variance across different resolutions, we apply random perturbations to each input image using the RandAugment algorithm, implemented via an existing tool[4]. For each image, we perform three random augmentations. To mitigate the effects of randomness and enhance result stability, we repeat the variance measurement process three times, each using a different random seed. The final uncertainty variance is obtained by averaging the results from these three iterations.

B.4 More Parameter-Efficient Fine-Tuning Details

The standard training hyperparameters are largely preserved, as outlined in Table 7, with two notable adjustments for image resolutions of 560^2 and 672^2: (1) The learning rate is reduced from $2e-5$ to $1e-5$ to prevent training loss explosion observed with the original rate. (2) The maximum number of tokens is increased from 2048 to 3072 and 4096, respectively, to accommodate the increased number of image tokens.

[3]https://github.com/Lou1sM/meaningful_image_complexity

[4]https://github.com/TorchSSL/TorchSSL/blob/main/datasets/augmentation/randaugment.py

Table 7: Hyperparameters at two training stages

Hyperparameter	batch size	lr	lr schedule	weight decay	epoch	optimizer	max tokens
Stage 1	256	1e-3	cosinie decay	0	1	AdamW	2048
Stage 2	128	2e-4					

Table 8: Training time cost

Resolution	224×224	336×336	448×448	560×560	672×672
Training Time Cost	11h 50m	16h 17m	24h 7m	32h 29min	124h 44m

Post-training experiments are conducted on eight NVIDIA GeForce RTX 4090 GPUs, with training time costs detailed in Table 8. Due to GPU memory limitations, DeepSpeed ZeRO-3 was employed for training at the resolution of 672^2, while ZeRO-2 was used for other resolutions. This accounts for the significant increase in training time between 672^2 and 560^2.

In the ablation study (Section 4.4), we separately fine-tune only the projector and only the position embeddings, using the stage 1 setting for consistency with the goals of the different training stages. The corresponding hyperparameters are also detailed in Table 7.

C Impact of Statistical Distributions on Empirical Formula Performance

To evaluate the extent to which the statistical distributions of complexity $C(T)$ and uncertainty variance $V(T)$ influence the performance of the empirical formula, we present the standard deviations of $C(T)$ and $V(T)$ for each vision-language task, along with their respective ratios to the mean values. These statistics are detailed in Table 9.

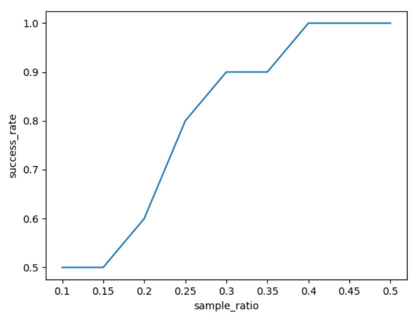


Figure 5: Relationship between sampling ratio and the success rate of the empirical formula.

Table 9: Statistical characteristics of $C(T)$ and $V(T)$ in each task. SD represents Standard Deviation, and Ratio indicates the ratio of the standard deviation to the mean.

Task	$C(T)$ SD	$C(T)$ Ratio	$V(T)$ SD	$V(T)$ Ratio
ScienceQA-IMG	3.3633	0.2384	0.4398	2.5466
Vizwiz	2.4405	0.1541	0.3383	6.0196
VQAv2	2.2005	0.1242	0.7925	4.2562
GQA	1.6582	0.0910	1.2595	4.9103
TextVQA	2.3057	0.1318	0.5258	3.3405
OKVQA	2.1958	0.1224	0.5487	3.7711
MMBench	3.5426	0.2196	1.2040	2.8915
MMBench-CN	3.5482	0.2197	1.0840	2.8310

The results indicate that $C(T)$ exhibits relatively low variance across tasks, whereas $V(T)$ shows substantially higher variability. This observation justifies our decision to adopt task-wise selection instead of sample-wise selection, as the higher variability in $V(T)$ at the sample level complicates consistent prediction.

To further assess the influence of $C(T)$ and $V(T)$ variance on the effectiveness of the empirical formula, we conducted an additional experiment. Specifically, we randomly sampled subsets of varying proportions from the original dataset and computed the average $C(T)$ and $V(T)$ values for these subsets to estimate task-level statistics. We then evaluated the empirical formula, previously tuned using a hyperparameter k on three reference tasks, to predict the optimal resolution across all tasks under these conditions.

The sampling proportions vary from 10% to 50%, with each experiment repeated 10 times using different random seeds. The success rate was defined as the percentage of instances where the empirical formula accurately predicted the optimal resolution for all tasks. The results, presented in Figure 5, reveal the following key findings: (1) At a sampling ratio of 40%, the success rate reaches 100%, demonstrating the empirical formula's robustness in predicting the optimal resolution. (2) At a sampling ratio of 10%, the success rate drops to 50%, indicating that a smaller subset size introduces variability that adversely affects prediction accuracy.

These findings highlight that while reducing the dataset size can lower computational costs, excessively small subsets may lead to suboptimal predictions. Moreover, the current approach relies on random sampling; future exploration of more advanced sampling strategies that select representative samples could potentially achieve high success rates with smaller subsets.

D Reference Tasks

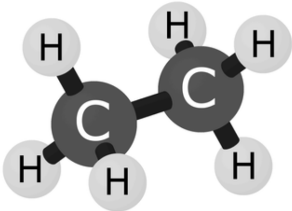
We utilize three reference tasks to determine the hyperparameter in Equation 1. Figure 6 presents three image samples from each reference task.

E Case Study Images

Figure 7 provide the visual inputs referenced in the Case Study (Section 5, Table 6).

F Acknowledgment of AI Assistance in Writing and Revision

We utilized LLMs for revising and enhancing writing of this paper.



(a) Single and simple object: Ethane is (). A. an elementary substance B. a compound

(b) Middle-level complexity: Are all the animals the same?

(c) Multiple objects: What is the brand being advertised?

Figure 6: We select three reference tasks with images in different levels of complexity to optimize the hyperparameter in Equation 1.

(a)

(b)

(c)

Figure 7: Three case study images