

MODSERVE: Scalable and Resource-Efficient Large Multimodal Model Serving

Haoran Qiu
Microsoft Azure Research
haoran.qiu@microsoft.com

Anish Biswas
Microsoft Research India
t-anibiswas@microsoft.com

Zihan Zhao
University of Virginia
rxy6cc@virginia.edu

Jayashree Mohan
Microsoft Research India
jamohan@microsoft.com

Alind Khare
Microsoft M365 Research
alindkhare@microsoft.com

Esha Choukse
Microsoft Azure Research
esha.choukse@microsoft.com

Íñigo Goiri
Microsoft Azure Research
inigog@microsoft.com

Zeyu Zhang
University of Virginia
qxc4fh@virginia.edu

Haiying Shen
University of Virginia
hs6ms@virginia.edu

Chetan Bansal
Microsoft M365 Research
chetanb@microsoft.com

Ramachandran Ramjee
Microsoft Research India
ramjee@microsoft.com

Rodrigo Fonseca
Microsoft Azure Research
fonseca.rodrigo@microsoft.com

Abstract

Large multimodal models (LMMs) demonstrate impressive capabilities in understanding images, videos, and audio beyond text. However, efficiently serving LMMs in production environments poses significant challenges due to their complex architectures and heterogeneous characteristics across their multi-stage inference pipelines.

We present the first comprehensive systems analysis of two prominent LMM architectures, decoder-only and cross-attention, across six representative open-source models, revealing key systems design implications. We also present an in-depth analysis of production LMM inference traces, uncovering unique workload characteristics, including variable, heavy-tailed request distributions and bursty traffic patterns.

Based on these insights, we propose MODSERVE, a modular LMM serving system that decouples stages for independent optimization and adaptive scaling. MODSERVE dynamically reconfigures stages and handles bursty traffic with modality-aware scheduling and autoscaling to meet tail latency SLOs while minimizing costs. MODSERVE achieves 3.3–5.5 \times higher throughput (leading to 25–41.3% cost saving) while meeting SLOs on a 128-GPU cluster with production traces.

1 Introduction

The rapid advancement in generative AI has led to the development of large multimodal models (LMMs) capable of processing inputs across various modalities such as text, image, video, and audio. These models have demonstrated remarkable capabilities in tasks like image captioning [5, 17, 35], visual question answering [46, 47], and multimodal dialogue systems [8, 25, 52]. This has led to a rapid adoption of LMMs in production services, including online applications where latency service-level objectives (SLOs) are critical.

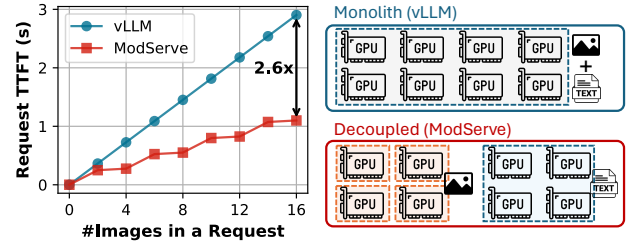


Figure 1. Impact of image workload on LMM inference TTFT for state-of-the-art implementation of Llama3.2-11B on vLLM vs. MODSERVE with an 8-A100 GPU server. The “Monolith” setup deploys the full model using 8 GPUs while the “Decoupled” setup deploys the LLM backend on 4 GPUs and four image encoders on the other 4 GPUs.

Unlike traditional large language models (LLMs) that process purely textual inputs using a single component, a decoder-based transformer architecture [55], LMMs handle fundamentally different types of inputs, each requiring distinct processing approaches. This heterogeneity introduces unique serving complexities that demand novel analysis and serving strategies. For *Image-Text-to-Text* models [21], the inference pipeline consists of multiple specialized stages: **image preprocessing** to transform raw images into tensor representations, **image encoding** to convert these tensors into image tokens, and a **language model backend** that combines text prompts with image tokens to generate text outputs. Currently, these stages are typically served as a monolithic system [4, 23, 56], where all components are integrated within a single serving instance and scaled together as a unified entity.

Unfortunately, existing monolithic inference serving systems fail to account for multimodality, making them unable

to scale effectively while meeting time-to-first-token (TTFT) SLOs, which now include image processing and encoding times. Figure 1 shows how a monolithic deployment struggles to scale as the number of images per request increases (a common scenario in multi-image or video workloads) resulting in sharp TTFT degradation. As a result, image-heavy requests can result in head-of-line (HoL) blocking, reducing system responsiveness and causing overprovisioning.

Our Work. In this paper, we present the first comprehensive systems analysis of two leading LMM architectures: cross-attention (CroAttn) and decoder-only (DecOnly), on both open-source LMMs and novel production LMM inference traces in Azure datacenters. We analyze their multi-stage inference pipelines, performance-resource tradeoffs, and production workload patterns, including variable request rates, diverse multimodal inputs, and bursty traffic. We focus on Image-Text-to-Text but our insights extend to other multimodal scenarios, such as Video-Text-to-Text, where videos are processed as image frame sequences [26], and Audio-Text-to-Text tasks [19], which share similar model architectures with the models we study.

Our analysis identifies three key insights for optimizing LMM inference. First, different LMM inference stages exhibit diverse performance characteristics and varying sensitivity to resource and model configurations (e.g., batching and model sharding), necessitating *decoupled execution*. Second, image encoding is a major bottleneck for TTFT, requiring efficient *encoder parallelization* to reduce both latency and HoL blocking. Finally, production multimodal traffic exhibits distinct bursty patterns driven by increased images per request, highlighting the need for *modality-aware routing* strategies to manage bursts and mitigate tail latency spikes.

Based on these insights, we propose MODSERVE, a novel modular architecture for scalable and resource-efficient LMM serving which directly addresses the challenges identified in our analysis. MODSERVE separates image- and text-specific inference stages into distinct instances for decoupled execution. In MODSERVE, *Image Instances* handle image pre-processing and encoding, while *Text Instances* manage LLM prefill and decoding (Figure 1). Text-only requests are served by *Text Instances*, whereas image-text requests go through *Image Instances* where images are converted to tokens before being forwarded to *Text Instances* for text generation.

MODSERVE’s modular architecture unlocks stage-specific optimizations. MODSERVE manages *Image* and *Text Instances* independently with stage-aware autoscaling, model sharding, and batching. By autoscaling the stages separately, it minimizes resource overprovisioning. For example, during image-driven bursts observed in production traffic, *Image Instances* can scale out independently, making MODSERVE more resource-efficient than monolithic inference systems. To navigate the image encoding bottleneck, MODSERVE parallelizes encoding of a single request across multiple *Image*

Instances (Figure 1), leveraging our finding that the images within a request do not attend to each other during encoding, and hence the requests can be parallelized at the image level.

Further, to manage image-driven bursts, MODSERVE implements modality-aware routing for *Image* and *Text Instances*. For example, images from image-text requests are routed to *Image Instances* with the fewest pending image tokens to encode, reducing HoL blocking and tail latency spikes.

We implement MODSERVE on top of a high-performance inference system, vLLM [23], and demonstrate the effectiveness of MODSERVE through extensive evaluations on a 16-server (128 GPUs) cluster running production Azure LMM inference traces. Compared to state-of-the-art baselines, MODSERVE achieves 3.3–5.5× *higher throughput* under static allocation and *reduces LMM serving cost by 25–41.3%* while meeting the P99 TTFT SLOs.

Summary. This paper makes the following contributions:

- A comprehensive system characterization on LMM serving, examining performance profiles and resource utilization patterns across diverse workloads in both open-source LMM deployments and production environment.
- A large open-source dataset containing sampled production Azure LMM inference traces.
- Design and implementation of MODSERVE, a modular architecture for scalable and efficient LMM serving.
- A thorough evaluation of MODSERVE in a 128-GPU cluster using large-scale production traces.

2 Large Multimodal Models Background

LMMs extend text-centric LLMs by integrating multimodal understanding capabilities for tasks like visual question answering [47] and computer-using agents [37, 40]. Figure 2 shows the typical pipeline of LMM inference in visual understanding tasks [21], which consists of three key stages: (1) *image preprocessing*, where raw images are transformed into uniform-sized tiles; (2) *image encoding*, where an encoder extracts visual features and produces a sequence of image tokens; and (3) *text generation*, where an LLM backend processes the image and text tokens to generate output text tokens. There are two dominant LMM architectures that differ in how the LLM backend handles image tokens and text tokens: (1) *decoder-only* (DecOnly), used in models like DeepSeek’s Janus [6], LLaVA-OneVision [26], InternVL [8], and NVLM-D [11]; and (2) *cross-attention-based* (CroAttn), found in Llama-3.2 Vision [10], NVLM-X [11], and Flamingo [2]. In this work, we analyze six open-source LMMs (listed in Table 1) across these architectures, varying image encoder sizes (400M–6B) and LLM scales (7B–72B).

Image Preprocessing. LMMs typically follow three pre-processing steps on CPU: (1) resize, rescale, pad, and normalize the raw image, (2) segment it into tiles [8, 10, 11] or patches [26], and (3) apply tile/patch-level transformations and sampling. The number of tiles varies, with higher

Table 1. Model configurations for six representative open-source LMMs with an example input image of 896×896 pixels.

LMM Model Name	Abbreviation	Architecture	Tile Size	Image Encoder (#Params)	Total Image Token Size (#Tiles \times #TokensPerTile)	LLM Backend (#Params)	Tensor Parallelism	Average Accuracy (HF-VLM [13])
Llama 3.2 Vision 11B [32]	Llama3.2-11B	Cross-attention	560 \times 560	ViT-H/14 (630M)	$4 \times 1601 \times 1 = 6404$	Llama 3.1 (8B)	TP-4	57.8%
Llama 3.2 Vision 90B [33]	Llama3.2-90B	Cross-attention	560 \times 560	ViT-H/14 (630M)	$4 \times 1601 \times 1 = 6404$	Llama 3.1 (70B)	TP-8	63.4%
LLaVA-OneVision 7B [29]	LLaVA-OV-7B	Decoder-only	384 \times 384	SigLIP (400M)	$10 \times 729 \times 1 = 7290$	Qwen2 (7B)	TP-4	60.1%
LLaVA-OneVision 72B [28]	LLaVA-OV-72B	Decoder-only	384 \times 384	SigLIP (400M)	$10 \times 729 \times 1 = 7290$	Qwen2 (72B)	TP-8	68%
InternVL-2.5 26B [9]	InternVL-26B	Decoder-only	448 \times 448	InternViT (6B)	$5 \times 256 = 1280$	InternLM (20B)	TP-8	71.6%
NVLM-D 72B [12]	NVLM-D-72B	Decoder-only	448 \times 448	InternViT (6B)	$5 \times 256 = 1280$	Qwen2-Instruct (72B)	TP-8	67.6%

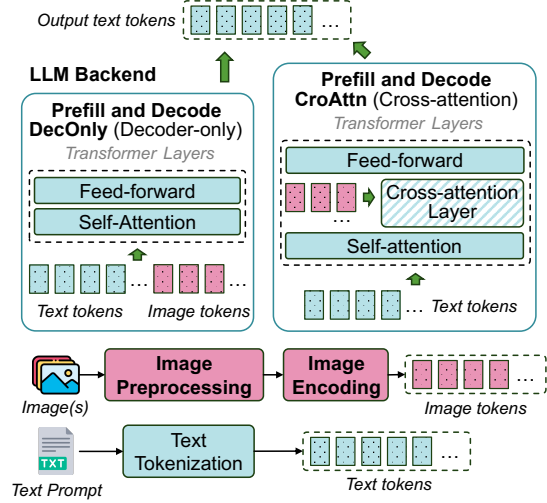
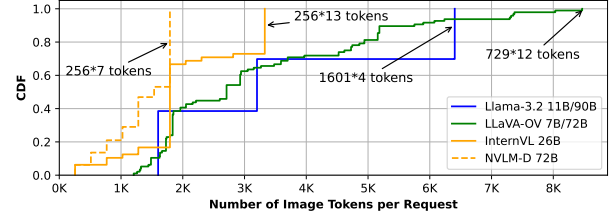

Figure 2. Model architecture for decoder-only and cross-attention-based LMMs in Image-Text-to-Text tasks [21].

image dimensions resulting in more tiles, which ultimately increases the number of image tokens. For example, an image with 896×896 pixels generates 4, 5, or 10 tiles of different sizes after preprocessing for six open-source LMMs (Table 1).

Image Encoding. The image encoder takes processed image tiles as input and produces image tokens that are then passed to the language model backend. Today’s image encoders predominantly use the vision transformer architecture [3] to extract visual features from images. Table 1 shows that different LMMs use different encoders [3, 8, 26, 58], leading to variations in the number of image tokens when running image encoders on the same ShareGPT-4o dataset [7] (Figure 3). This is due to differences in the number of tiles and image tokens generated per tile by each encoder.

Text Generation. Image and text prompt tokens are combined and passed through LLM prefill and decode to generate output tokens, typically using one of two architectures:

Decoder-Only (DecOnly) LMMs. An unmodified LLM backend is reused in DecOnly LMMs (e.g., LLaVA-OV reuses Qwen2 LLM [26]), processing text and image tokens uniformly (“DecOnly” box in Figure 2). While valued for their simplicity


Figure 3. Distribution of image token count (per request) for open-source LMMs on ShareGPT-4o dataset [7]. Different LMMs (e.g., LLaVA-OV 7B and 72B) can share the same image encoder so the number of image tokens is the same.

and unified modality handling, DecOnly models often require long sequences for high-resolution images, resulting in computational inefficiencies during inference.

Cross-Attention (CroAttn) LMMs. Unlike DecOnly LMMs, which leave the LLM backend unchanged, CroAttn-based models (e.g., Llama-3.2 Vision) integrate *cross-attention layers* to process image tokens, treating visual inputs like a “foreign language” in the LLM backend. While more complex to train, they improve inference efficiency by avoiding full image token unrolling in the LLM decoder, making them ideal for high-resolution inputs. Self-attention operates on text tokens, while the cross-attention layer attends to both text and image tokens (“CroAttn” box in Figure 2).

SLO Metrics for LMM Inference. Production LMM serving systems need to satisfy SLOs defined on tail latency (e.g., P99) for worst-case performance. These SLO metrics include *Time to First Token (TTFT)* and *Time Between Tokens (TBT)*. TTFT measures the latency from query (text/images) to the first response token, critical for interactive applications. In contrast to text-centric LLM serving, LMM TTFT includes the following stages of LMMs inference pipeline: (1) image preprocessing, (2) encoding, and (3) language model prefill time. TBT captures the delay between consecutive token generations during decoding, affecting output fluency. As multimodal preprocessing and encoding primarily influence TTFT, in this work, we focus on TTFT while leveraging state-of-the-art techniques [1, 41] to optimize TBT.

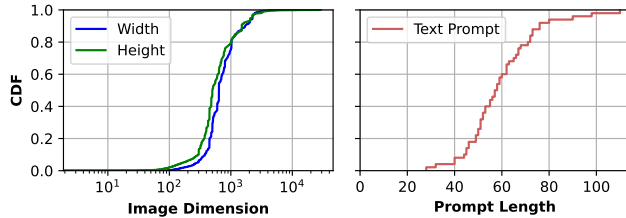


Figure 4. Image dimension distribution and text prompt length distribution of ShareGPT-4o Image dataset [7].

An ideal LMM-serving system should meet TTFT/TBT SLOs while maximizing request *throughput* (i.e., goodput) and compute *utilization* (GPU cost).

LMM Deployments Today. State-of-the-art serving frameworks [4, 23, 56] deploy LMMs as *monolithic* systems to meet latency SLOs. In this setup, all inference components (i.e., image preprocessor, image encoder, and LLM backend) are co-located on the same hardware server as a single unit. These tightly coupled components share uniform batching and model parallelism strategies across the pipeline. Table 1 details the default model parallelism for our open-source LMMs. While this monolithic design is straightforward to implement and common in open-source LMM serving, it limits flexibility and suffers from sharp TTFT degradation under image-heavy workloads (Figure 1).

3 Motivation and LMM Characterization

To further understand the limitations of monolithic deployments and explore unique characteristics that distinguish LMM serving from text-centric LLM serving, we characterize open-source LMMs in the *Image-Text-to-Text* category [21]. We evaluate the performance and resource characteristics of heterogeneous inference stages under varying image inputs and model configurations (Section 3.1). Moreover, to understand multimodal traffic patterns at scale, we analyze sample production traces from one production LMM inference cluster in Azure (Section 3.2).

Characterization Setup. The following is our setup:

Models. We use six representative open-source LMMs across two different architectures (DecOnly and CroAttn) as listed in Table 1. We deploy the models on vLLM [23] in BF16.

Dataset. We use the open-source ShareGPT-4o dataset [7], which includes 50K images of varying resolutions and text prompts from multimodal GPT-4o as shown in Figure 4.

Hardware. Our setup features a DGX-A100 server with 8 NVIDIA A100 GPUs (80GB each) connected via NVLINK [34]. It has 96 AMD Epyc™ 7V12 CPU cores and 1900 GiB DRAM.

3.1 Characterization on Open-Source LMMs

We characterize open-source LMMs to understand how different inference stages impact performance and resource

efficiency. Additionally, we compare DecOnly and CroAttn models to highlight the need for model-specific optimization.

Per-stage Latency Breakdown. Figure 5 plots the split-up of TTFT across the three stages that comprise it; image pre-processing, image encoding, and LLM prefill. There are three key takeaways. First, image preprocessing, which occurs on the CPU, contributes minimally to the overall TTFT, while image encoding time contributes to a major portion in TTFT (especially for CroAttn models). For instance, 79% and 65% of TTFT in Llama3.2-11B and Llama3.2-90B are from image encoding. For DecOnly models such as InternVL-26B and NVLM-D-72B, image encoding latency accounts for 25% and 54% of TTFT. Second, the image encoding time depends on the encoder model size. For instance, scaling from SigLIP-400M (in LLaVA-OV-7B) to InternViT-6B (in InternVL-26B), the median image encoding time increases by 10×. Finally, prefill computation is more efficient in CroAttn models because image tokens are attended to only in the CroAttn layers, as previously discussed in Section 2.

Insight 1: A major portion of the TTFT is spent on image encoding, particularly for CroAttn models, making image encoding optimization critical to meet TTFT SLOs.

Compute Characteristics of LMM Stages. Image preprocessing on CPU and image encoding on GPU are compute-intensive processes. Figure 6a plots the impact of varying the number of CPU cores on preprocessing latency. Preprocessing is CPU-intensive and benefits from trivially parallelizing across cores. Both stages exhibit linear latency scaling with batch size, saturating compute without significant throughput gains from increased batching as shown in Figures 6b and 6c, respectively. Figure 6d further plots the GPU utilization metrics for a request batch size of one during image preprocessing and image encoding. We observe a consistent SM core activity near 100% during image encoding, with average DRAM utilization below 30%. Image encoding is, therefore, typically compute-bound, resembling the language model’s prefill phase [22]. Moreover, when a request has multiple images in the input prompt, there is no compute dependency between the images during image encoding; hence, image tiles can be parallelized across multiple encoders.

Insight 2: Image preprocessing and encoding are both compute-intensive similar to LLM prefill stage. The independence of image computations in a multimodal request enables parallelization of image preprocessing and encoding across multiple instances.

Compute characteristics of prefill and decode phases of the LLM backend have been well studied; the prefill phase is typically compute-bound, while the decode phase is memory-bound [1, 22, 41]. However, Figure 5 shows that LLM prefill is more efficient in CroAttn models than inDecOnly models,

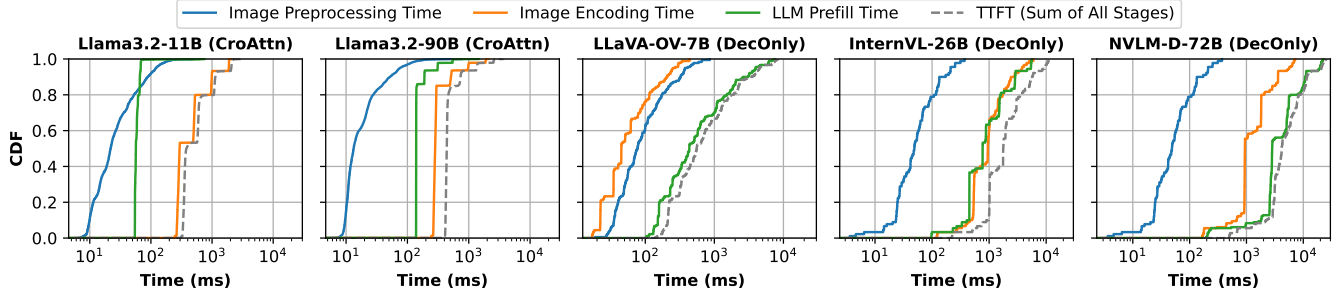
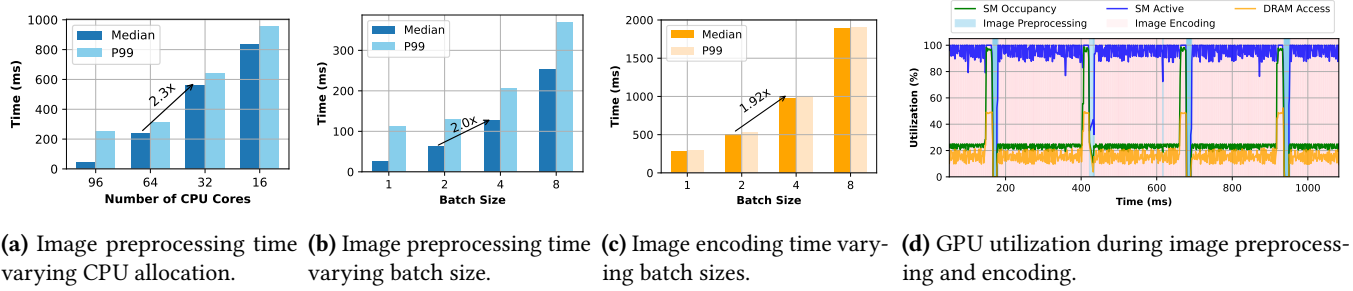


Figure 5. Per-stage request latency breakdown analysis across representative open-source LMMs deployed using default tensor parallelism (TP) as described in Table 1. TTFT (dashed line) is the sum of the latency from each inference stage.



(a) Image preprocessing time varying CPU allocation. (b) Image preprocessing time varying batch size. (c) Image encoding time varying batch sizes. (d) GPU utilization during image preprocessing and encoding.

Figure 6. Compute characteristics of image preprocessing and encoding. Both stages are compute-bound.

resulting in reduced compute boundness and an interesting tradeoff we describe below.

Latency-Accuracy Profiles across LMMs. Figure 7 shows the accuracy versus prefill/TTFT efficiency for different models. When comparing models with similar language model backend sizes across both architectures (e.g., Llama3.2-11B vs. LLaVA-OV-7B and Llama3.2-90B vs. LLaVA-OV-72B vs. NVLM-D-72B), we observe that CroAttn models typically have up to an order of magnitude lower LLM prefill time, leading to lower TTFT. However, the CroAttn models usually achieve 5 points lower accuracy compared to their DecOnly counterparts on the Open VLM leaderboard [13]. For example, Llama3.2-90B scores 63.4, while the similarly sized LLaVA-OV-72B scores 68, but with significantly higher prefill latency and TTFT than Llama3.2-90B.

Insight 3: DecOnly models exhibit 10× worse prefill latency than similar-sized CroAttn models, leading to less TTFT SLO headroom for the image encoding and thus necessitating higher scalability for image workloads.

Impact of Batching. In today’s monolithic deployments, a single batch size is applied across all stages of the LMM on the GPU, which does not strike a balance between latency and throughput given heterogeneous compute characteristics observed across different stages. Figure 8 shows how the batch size affects the median latency of each LMM stage across architectures. As the batch size increases, the latency

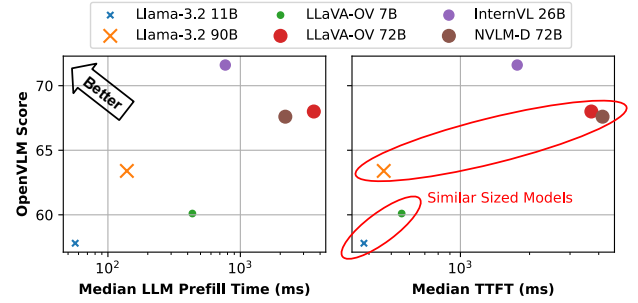


Figure 7. LMM accuracy vs. prefill/TTFT efficiency.

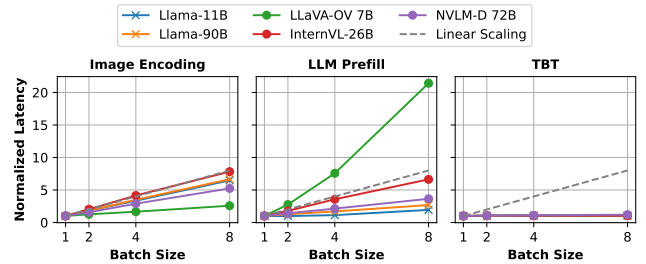


Figure 8. Median latency vs. batch size per LMM stage on GPUs. Latency is normalized to that at batch size one.

grows at varying rates, reflecting each stage’s differing sensitivity to batch size and compute intensity.

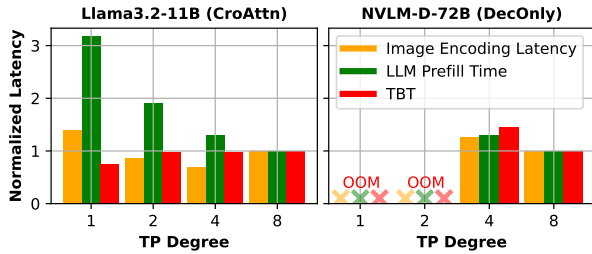


Figure 9. Impact of the tensor parallelism (TP) degree on the median latency of each stage for CroAttn-based and DecOnly LMMs. Latency is normalized to that of TP-8.

Compute-intensive stages like image encoding and LLM prefill (in DecOnly models with longer image token inputs) show limited throughput gains and rising latency beyond small batch sizes. In contrast, the memory-bound decode stage benefits from linear throughput scaling. Due to their low text token count, CroAttn models uniquely gain from prefill batching, diverging from traditional LLM trends where prefill saturates compute even at a batch size of one. Notably, DecOnly model NVLM-D with fewer image tokens also exhibits certain benefits in batching.

Insight 4: *The effectiveness of batching varies for each LMM component and is model-specific. LMM request batching should thus be tailored to each stage.*

Impact of Parallelism. Monolith deployments also limit the flexibility of model sharding within a GPU server which is typically done through tensor parallelism (TP) [53]. Figure 9 shows how increasing TP degrees affects latency across LMM stages. In Llama3.2-11B, the lowest LLM prefill latency occurs at TP-8, image encoding at TP-4, and TBT at TP-1. At TP-8, encoding latency rises due to the tradeoff between compute intensity and inter-GPU communication, making it inefficient to split a small 630M encoder across eight GPUs.

In contrast, NVLM-D-72B, with a larger 6B image encoder, sees a 1.3 \times latency reduction when increasing TP from 4 to 8. However, this comes with diminishing returns relative to resource cost. To balance throughput and latency, operators can deploy two TP-4 encoders for higher throughput or one TP-8 encoder for lower latency, both using eight GPUs.

Insight 5: *Treating the image encoder and LLM backend as a monolith limits parallelism flexibility and degrades performance. Decoupling them enables independent scaling and optimized efficiency through pipelining.*

3.2 Production LMM Trace Analysis

Building on the insights from open-source LMM characterization, we further analyze multimodal traffic patterns at scale, leveraging production traces from one of Azure’s LMM inference clusters. The traces capture a sample of multi-tenant

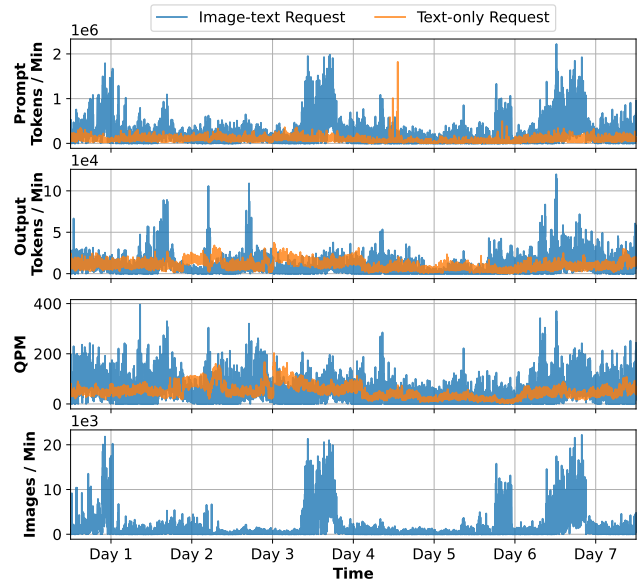


Figure 10. Aggregated prompt/out token rate, request arrival rates in queries per minute (QPM), and image rate for a production LMM inference cluster in one week.

traffic, including both text-only and image-text requests. Our study focuses on (1) temporal and burstiness patterns and (2) heterogeneity of multimodal requests. We plan to open-source these production traces.

Temporal Patterns and Burstiness. Figure 10 shows the traffic of text-only and image-text requests separately to understand their dynamic behavior and overall impact on the system. The traces are collected over a span of one week. To understand the traffic patterns, we report the timeline of prompt (input) token rate, output token rate, request arrival rate, and input image rate. Our analysis reveals two key characteristics in production LMM inference:

- **Diverse Arrival Patterns.** Image-text requests show up to 5 \times higher prompt token rates than text-only requests. In addition, their peak and trough occurrences are largely independent, showing minimal correlation.
- **Image-Driven Bursts.** Image-text requests experience significant burstiness, not only due to higher request arrival rates but also increased images per request (e.g., video workloads). As a result, existing LLM traffic prediction methods [50] (which work well for workloads with diurnal patterns) have a high average error rate of 79%.

Request Heterogeneity. Figure 11a shows that prompt lengths vary significantly across modalities. Both image-text and text-only requests follow a heavy-tailed power-law distribution ($\alpha=4.4$ and 2.9, respectively) where a higher α means a heavier tail with more extreme events occurring more frequently. In addition, image-text requests have longer median prompts due to image tokens but shorter tails than

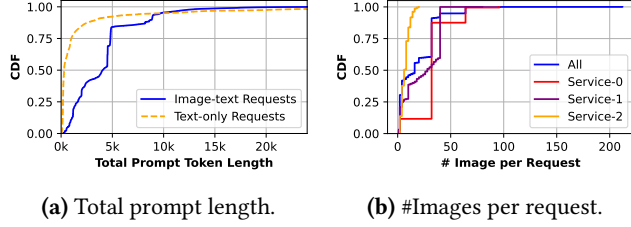


Figure 11. LMM input characterization in production.

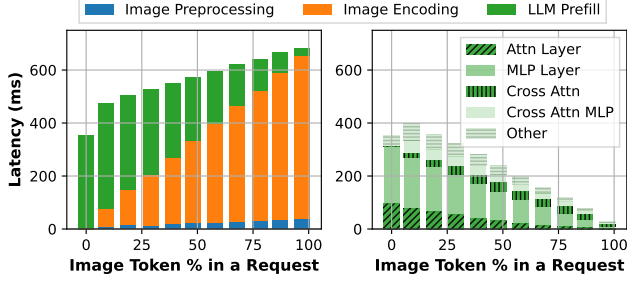


Figure 12. Llama3.2-11B (CroAttn) TTFT breakdown (left) and LLM prefill time breakdown (right) under various image-to-text token ratios in a request.

text-only requests. Figure 11b shows that the number of images per request also varies significantly with a heavy tail. In addition, among the top three services issuing text-image requests, we observe high inter-service variability. Some services process $16\times$ more images per request than others.

Comparing the image dimension distribution in our production traces with that of ShareGPT-4o image dataset [7], we observe similar distributions, with median image width and height around 500 pixels and P95 exceeding 1000 pixels.

Insight 6: Production LMM image traffic exhibits bursty behavior independent of traffic patterns of text requests. Serving systems must dynamically scale resources to handle modality-specific bursts efficiently.

Impact of Mixed Modality. Given LMM requests’ input heterogeneity, Figure 12 shows how varying image token percentages within a single request affects TTFT and LLM prefill time in a CroAttn model Llama3.2-11B, with detailed latency breakdowns. DecOnly models have no prefill time variation with varying token ratios as image and text tokens are treated in the same manner. We fix the total context length of each request at 16K tokens while varying the percentage of image tokens by adjusting the number of images (0–10 images in each case with 1601 tokens per image).

TTFT increases with the percentage of image tokens in a request due to the increased image encoding computation, resulting in a $1.5\times$ TTFT degradation when transitioning from text-only to image-only inputs. However, this latency gain is significantly lower than DecOnly models because CroAttn

models attend to image tokens only within the CroAttn layers, resulting in reduced LLM prefill time (shown in green) and partially offsetting the overhead from image encoding. The right figure further illustrates this by breaking down the layer-wise LLM prefill time, highlighting a reduction in self-attention compute (i.e., “Attn Layer” and “MLP Layer”) as the proportion of image tokens increases. Although the cross-attention computation peaks at the 50% image tokens (due to the dependency on both image and text tokens), it contributes much less than self-attention computation because there are only 4 CroAttn layers (out of 40 layers).

Insight 7: DecOnly models maintain consistent prefill times regardless of token modality, making total token count the key factor for request routing. In contrast, CroAttn models experience reduced prefill latency as the image token percentage increases, requiring a modality-aware routing strategy that balances both text and image token load.

4 MODSERVE Design and Implementation

Based on our insights from the characterization study of open-source LMM benchmarks and production LMM workloads, we propose MODSERVE, a novel decoupled architecture for scalable and resource-efficient LMM serving.

The key idea in MODSERVE is to separate image- and text-specific inference stages into distinct instances, given the need to optimize each stage separately (Insight 1 and 3) and enable seamless interaction between stages. Unlike monolithic infrastructures, MODSERVE enables independent optimization of each stage, improving resource efficiency while meeting performance SLOs. This decoupling also enables modality-aware request serving, addressing tail latency, heterogeneous bursts, and resource contention.

Overview. Figure 13 shows MODSERVE’s design. A pool of *Image Instances* handles image preprocessing and encoding of image-text requests. The resulting image tokens are passed to a pool of *Text Instances*, which performs LLM prefill and decode operations. Text-only requests bypass the image components and are queued directly at the *Text Instances*. Two pools are managed by the *Image and Text Pool Managers*.

MODSERVE adopts a hierarchical architecture inspired by DynamoLLM [50]. Onboarding any new LMMs (e.g., Llama3.2-11B) starts with an offline profiling phase to build model-stage profiles that capture how model configurations and load impact performance (Section 4.1). MODSERVE uses these profiles to guide model configuration and instance scaling. After model deployment, MODSERVE then reconfigures resources periodically to adapt to workload patterns, scaling for increased image-text requests (or vice versa) (Section 4.2). For each request, MODSERVE selects the optimal LLM Text or Image Instances for execution (Section 4.3).

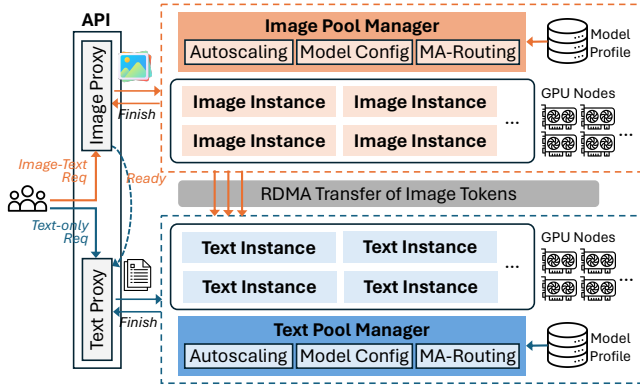


Figure 13. Overview of the MODSERVE architecture.

4.1 Offline LMM Profile Generation

When onboarding a new LMM, MODSERVE generates resource-performance profiles by characterizing the image encoder and LLM backend independently. This profiling runs controlled inference workloads with varying model parallelisms (e.g., TP-2 and TP-8), batch sizes, and load (i.e., image tokens per second for image encoders and prompt tokens per second for LLM backends) to capture per-stage performance characteristics. To efficiently model performance across different load conditions, MODSERVE profiles a set of representative load levels (up to the maximum throughput) and extrapolates the behavior for intermediate loads. The resulting profiles take load, parallelism, and batch size as inputs to predict key performance metrics, including encoding latency for *Image Instances* and prefill time and TBT for *Text Instances*. The *Pool Managers* use these profiles to guide operational decisions (Section 4.2): (1) pool autoscaling to meet latency SLOs without overprovisioning, (2) model sharding that selects optimal TP degrees, and (3) max batch sizing for each stage.

Since multiple LMMs may share the same image encoder or LLM backend, MODSERVE minimizes overhead by reusing model profiles across deployments. These profiles are cached in cluster-local storage and synchronized via a global repository, enabling efficient sharing across clusters.

4.2 Decoupled Resource Management

MODSERVE’s decoupled approach to resource management stems from our insights on stage-specific performance disparities in batching (Insight 4), independent scaling benefits (Insight 5), and modality-specific traffic patterns (Insight 6). Specifically, MODSERVE periodically reconfigures resources (i.e., every five minutes to match the autoscaling overhead) to align with workload demands.

The *Image Pool Manager* maintains a pool of *Image Instances*, which preprocess images on CPU and encode image workloads on GPU for image-text requests. Meanwhile, the *Text Pool Manager* manages a pool of *Text Instances* responsible for the prefill and decode stages of both image-text

and text-only requests. Based on model profiles (Section 4.1), each manager independently optimizes pool autoscaling, model sharding, and max batch sizing to minimize costs while meeting performance SLOs.

Before their online operation, the initial number of *Image Instances* (N_i) is determined using the median image QPS multiplied by the median image encoding latency. The number of *Text Instances* (N_t) is set as N_i divided by the median number of images per request, based on historical LMM inference traces. If no history is available, MODSERVE initially overprovisions resources to ensure reliability.

Token-Aware Pool Autoscaling. The *Pool Managers* dynamically scale the number of *Image* and *Text Instances* based on real-time workload demands. For example, a surge in image-heavy requests leads to more image preprocessors and encoders, while an increase in text requests or requests’ prompt lengths triggers the *Text Pool Manager* to scale LLM replicas to handle variations in prefill.

The number of replicas per stage is computed as $\lceil \frac{ML}{MC} \rceil$ where ML is the modality-specific load (e.g., prompt tokens/sec for *Text Instances*, image tokens/sec for *Image Instances*) and MC is the maximum capacity each stage can handle without violating SLOs, based on the offline LMM profiling data. Unlike traditional web service autoscaling, which reacts to request rates, MODSERVE optimizes scaling based on token throughput (tokens/s), capturing variations in both request rates and request sizes (Insight 6 and Figure 11a).

For *Image Instances*, image token counts are precomputed based on static mapping from image dimensions (Figure 3). Autoscaling of *Text Instances* is based on text token load in CroAttn models but total tokens in DecOnly models due to homogeneous self-attention across modalities. Advanced autoscaling hysteresis prevention techniques [60] can be employed to avoid excessive scaling actions caused by transient workload fluctuations but are not covered in this paper.

Model Sharding. The *Pool Managers* also determine instance sharding for optimal tensor parallelism (TP) for image encoders and LLM backends. Our characterization (Section 3.1) shows image encoders achieve peak throughput with lower TP than LLMs. Therefore, the model sharding degree for each instance is configured separately for maximum throughput while ensuring SLO attainment on TTFT and TBT. By decoupling the components, MODSERVE ensures independent sharding, optimizing parallelism without unnecessary synchronization overhead.

When scaling beyond a single GPU server, MODSERVE prioritizes autoscaling over pipeline parallelism (PP) [48] to maximize throughput, seamlessly transitioning to batch-level optimizations as needed.

Identifying Max Batch Size. For each stage, the maximum batch size is configured to maximize throughput while meeting latency SLOs. Batch sizing decisions are guided by the offline model-stage profiles, which predict their impact on

encoding and decoding latencies. *Image Instances* may forgo batching as small max batch sizes often achieve optimal GPU utilization (*Insight 3*). In contrast, *Text Instances* batch requests when beneficial, optimizing token throughput during prefill/decode based on TTFT and TBT SLOs, particularly for CroAttn LMMs (*Insight 4*).

4.3 Modality-Aware LMM Request Serving

For each incoming LMM request, MODSERVE dynamically routes and schedules workloads to balance load across *Image* and *Text Instances*. The *Pool Managers* optimize this process to minimize queueing delays and improve TTFT latency.

Request Routing Across Instances. To mitigate tail TTFT latency surges caused by modality-specific bursts and queueing delays, MODSERVE employs a modality-aware routing strategy that balances image and text workloads independently. Traditional request-level LLM load balancing (e.g., round-robin, memory-based [51]) overlooks the computational intensity of image encoding (*Insight 2*), making them vulnerable to load imbalances during image bursts (*Insight 6*), leading to high tail latencies.

Instead, MODSERVE routes requests by input modality. Image-text requests are assigned to *Image Instances* with the least image-token load. Large requests (i.e., those with more images) are consequently distributed across multiple *Image Instances* for parallel processing and encoding (*Insight 2*), preventing degraded batching performance that would occur if all images were routed to a single instance. This effectively enables a form of request chunking [1], where images in a large request can be processed in an interleaved manner with other requests, reducing HoL blocking and improving scheduling flexibility.

To route traffic between Text Instances, text-only requests and image-text requests with completed image tokens are directed to the *Text Instance* with the least total pending tokens (text+image) for DecOnly models and the least total pending text tokens for CroAttn models because of the attention mechanism difference between the two model architectures (*Insight 7*). Modality-aware routing enables parallel image encoding and dynamically adapts to image or text traffic bursts, reducing queueing delays and improving TTFT, particularly at the tail.

Instance Request Scheduling. At the instance level, MODSERVE minimizes resource contention between image-text and text-only requests with priority scheduling based on modality and prompt size. While decoupling isolates image and text processing, contention can still arise in *Text Instances*, where both request types share prefill processing. This issue is particularly pronounced in DecOnly LMMs, which exhibits lower efficiency during the prefill stage (Figure 8). Performance degradation occurs from increased batching latency for all requests, while non-batched processing

introduces HoL blocking and high queueing delays at tails due to request heterogeneity (Figure 11).

To address these challenges, MODSERVE replaces traditional FIFO scheduling—which may exacerbate HoL blocking [42, 45]—with an SLO-driven scheduling strategy that can prioritize shorter requests (e.g., text-only queries or small image-text requests with tight SLOs) to maintain low latency.

Pool Managers continuously monitor SLO attainment and trigger pool autoscaling when the rate falls below a predefined threshold (default 0.99 with a sensitivity study in Section 5.3), ensuring adaptive resource allocation under dynamic workloads (especially in cases of unpredictable traffic). MODSERVE can work with state-of-the-art batch scheduling techniques [1, 41] to optimize TBT during the decode stage, which we leave to future work as we do not observe TBT degradation in LMM characterization (*Insight 4*).

Image Token Transfer. Once image processing is complete, MODSERVE transfers image tokens from *Image Instances* to *Text Instances* via a pull-based RDMA mechanism. Unlike a push-based approach that immediately sends image tokens upon availability, this strategy reduces synchronization overhead and optimizes load balancing.

Since image-text requests may involve multiple images processed across different *Image Instances*, the transfer follows a many-to-one pattern, aggregating all image tokens before text processing begins. This defers transfer until all tokens are ready, allowing for better scheduling decisions by routing requests to the least-loaded *Text Instance*. Meanwhile, queuing at the *Text Instance* overlaps with token transfer, effectively hiding transfer latency.

MODSERVE colocates *Image* and *Text Instances* in the same server when each *Text Instance* is not taking up all GPUs on a server to avoid image transfer overhead and unallocated idle GPUs. For example, it may place one TP-4 Text Instance and two TP-2 Image Instances within the same 8-GPU server. Unlike monolithic deployments, colocated instances remain independently configurable and can serve corresponding stages of different requests independently.

4.4 Implementation

We implement MODSERVE using 5,000 lines of Python code. We base the *Text Instance* on vLLM [23] (v0.7.2), a state-of-the-art generative model inference platform, and build the *Image Instance* on HuggingFace Transformers [20]. The modular architecture of MODSERVE enables easy integration with other serving engines (e.g., TensorRT [38] and DeepSpeed [4]). We use numactl to restrict CPU and memory usage of image preprocessing to a single NUMA node, which reduces memory access latency and performance variation. To ensure efficient GPU-to-GPU memory transfer of image tokens, we use PyTorch’s distributed communication with the NCCL backend and GPU Direct RDMA. The P99 transfer latency of image tokens per request is 5 ms.

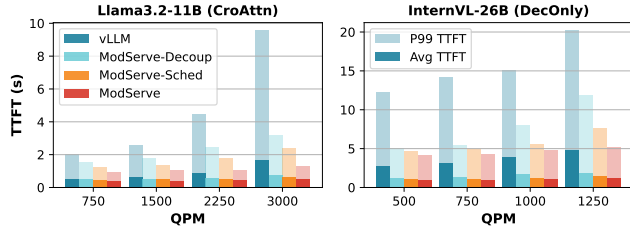


Figure 14. TTFT comparison with fixed 16 servers (128 GPUs) without autoscaling.

We implement the *Image and Text Pool Managers* as lightweight gRPC servers (hosted on dedicated VMs) with low memory and compute requirements, drawing inspiration from DynamoLLM [50]. For failure detection and recovery, the *Pool Managers* use a simple heartbeat-based membership management protocol [14]. However, MODSERVE can be easily extended to adopt more robust leader election (e.g., Raft [39]) and fault-tolerance algorithms [54].

5 Evaluation

5.1 Experimental Setup

Models and Workloads. We use Llama3.2-11B and InternVL-26B as representative models for CroAttn and DecOnly LMMs, respectively. To ensure realistic workload distribution, we reuse the LMM dataset from Section 3.1 and adopt the inter-arrival timestamps of requests and the number of images associated with each request (ranges from 0 to 16) from the production LMM inference trace (Section 3.2).

Hardware. We evaluate MODSERVE on a cluster with 16 DGX-A100 servers [34] (128 GPUs). Each server has the same configuration as the server used in our characterization study (Section 3.1). The GPUs within a server are connected with NVLINK 3.0 while cross-server connection is via InfiniBand.

Baselines and Systems. We compare MODSERVE against the state-of-the-art generative model inference serving system, vLLM [23], which supports LMM inference as a monolithic setup. We also evaluate MODSERVE with a few variants of MODSERVE implemented on top of vLLM: (1) vLLM with decoupled Image/Text Instances (i.e., MODSERVE-Decoup), (2) MODSERVE-Decoup plus modality-aware scheduling (i.e., MODSERVE-Sched), and (3) MODSERVE-Sched plus modality-aware routing (i.e., MODSERVE), for ablation study.

SLO Definition. We define the SLO metrics for LMM inference based on the TTFT/TBT during the isolated run of a single text-only request and text-image request (with one image) on the monolith baseline setup. Then, we scale the SLO metrics with a constant factor (i.e., SLO factor) to evaluate how MODSERVE performs under tight/relaxed SLOs (Section 5.3). The SLOs are defined on P99 tail latency.

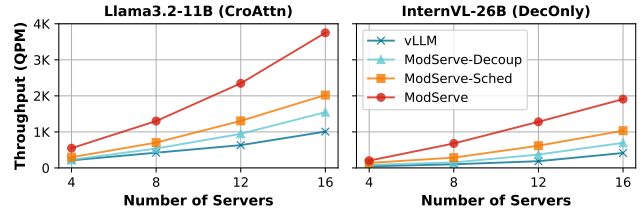


Figure 15. Maximum load meeting SLO.

5.2 End-to-end Performance

Static Resource Allocation. We begin by evaluating MODSERVE under a static resource allocation setup, where a fixed number of servers remain active at all times without autoscaling. This setup isolates the benefits of decoupling, modality-aware request scheduling, and routing from pool autoscaling (which we explore independently). Figure 14 shows the average and tail (P99) TTFT achieved by MODSERVE and the baselines when serving different input loads over fixed resources (16 servers with 128 GPUs in total). In this setup, vLLM (monolith) deploys 32 instances (each with TP-4) while the other approaches (decoupled) deploy 20 Text Instances (TP-4) and 48 Image Instances (TP-1).

Compared to vLLM, statically decoupling (MODSERVE-Decoup) improves the average and P99 TTFT by 27% and 42% (for Llama3.2), 46% and 47% (for InternVL). This is because monolithic deployments process all modalities on shared GPU resources, leading to contention and inefficient utilization under imbalanced modality traffic. In addition, MODSERVE-Decoup with the same number of GPUs can deploy 16 extra Image Instances and enables image encoding parallelization that reduces TTFT significantly compared to the monolithic deployment on vLLM.

MODSERVE shows a more pronounced TTFT improvement over the monolith baseline when serving InternVL. This is because the monolith deployment faces resource contention with DecOnly models due to their high prefill latency (Insight 3), which contends with image encoding. Additionally, InternVL’s image encoder has higher batching performance degradation (Insight 4) and thus benefits more from parallelization. Adding modality-aware request scheduling (MODSERVE-Sched) further reduces the average and P99 TTFT by 12% and 25%, modality-aware routing (MODSERVE) reduces the average and P99 TTFT by 14% and 32%, as it reduces HoL blocking and mitigates tail latency spikes.

Overall, MODSERVE achieves the lowest TTFT across all load levels, demonstrating the effectiveness of modular inference pipelines. We observe similar TBT performance in all approaches due to its compute insensitivity (as indicated by Figure 8). Figure 15 further evaluates the maximum throughput under the TTFT and TBT SLO when varying the static resource allocation from 4 to 16 servers. MODSERVE achieves

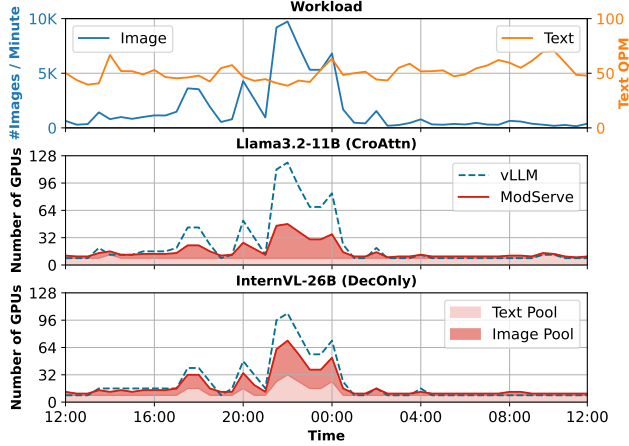


Figure 16. GPU allocation with autoscaling (up to 16 servers) during a one-day interval on the production traces.

a 3.3 \times and 5.5 \times throughput improvement over vLLM (monolith) for Llama3.2 and InternVL, respectively, which confirms that DecOnly models benefit more from decoupling.

Resource Allocation with Autoscaling. We now assess how MODSERVE and vLLM (monolith) baseline handle image-driven bursts seen in the production trace (Figure 10). Fundamentally, to serve traffic bursts, a system needs to scale up the resources to meet the workload demand while scaling down to avoid overprovisioning. Therefore, we enable autoscaling in both MODSERVE and vLLM and evaluate them on a one-day interval of the production trace that contains an image-driven burst. For a fair comparison, both MODSERVE and vLLM (monolith) use similar SLO-driven autoscaling heuristics based on offline model profiling (Section 4.2).

Figure 16 compares the number of GPUs used by MODSERVE and vLLM (monolith) to serve the image-driven burst in the production trace. MODSERVE takes 41.3% and 25% fewer GPUs compared to vLLM to serve Llama-3.2 (CroAttn) and InternVL (DecOnly) models respectively while meeting the tail latency SLOs. MODSERVE’s cost reduction is higher for Llama-3.2 (CroAttn) model as the increase in image tokens caused by image-driven bursts not overwhelming LLM backend in CroAttn models as observed in its latency profile (Figure 12). However, in InternVL (DecOnly), the LLM backend’s latency increases with the increase in image tokens due to homogeneous self-attention. Therefore, to meet SLOs, MODSERVE scales up the number of Text Instances for InternVL more than for Llama-3.2 during image-driven bursts (light pink in Figure 16). Overall, MODSERVE’s stage-aware autoscaling prevents unnecessarily scaling up LLM backends (done by vLLM due to monolith deployment) during image-driven bursts and prevents resource over-provisioning.

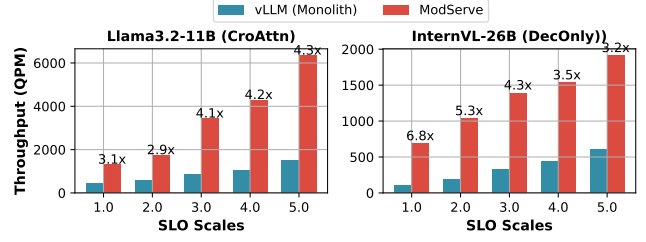
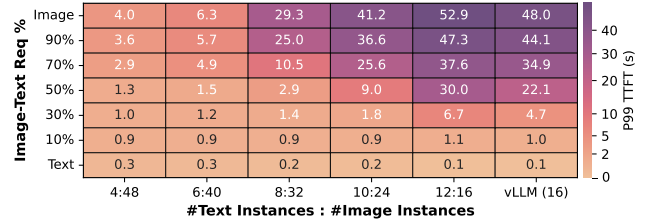
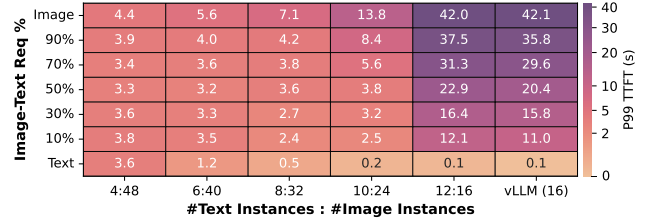


Figure 17. Throughput impact varying the SLO scale.



(a) Llama3.2-11B CroAttn LMM.



(b) InternVL-26B DecOnly LMM.

Figure 18. Impact of image request percentage (Y-axis) and instance allocation (X-axis), i.e., #Text Instances (TP4) : #Image Instances (TP1) on 8 servers (64 GPUs).

5.3 Sensitivity Study

Impact of SLO Scale. Figure 17 shows the maximum throughput MODSERVE can achieve when changing the SLO scale (higher values refer to more relaxed SLOs). As the SLO scale increases, MODSERVE consistently outperforms the vLLM, achieving up to 4.3 \times higher throughput for Llama-3.2 and 6.8 \times for InternVL. This trend highlights that MODSERVE better utilizes resources under the same latency requirements.

Impact of Image-to-Text Instance Ratio. Figure 18 shows the effect of varying the ratio of Image and Text Instances on 64 GPUs (8 servers) along the X-axis, in comparison to vLLM monolith with 16 instances. For instance, “4:48” denotes a configuration with 4 Text Instances (TP-4) and 48 Image Instances (TP-1). As the ratio of Text Instances increases, we observe that MODSERVE consistently achieves superior TTFT performance compared to vLLM (monolith) until the ratio reaches 10:24. However, at 12:16, the decoupled configuration contains the same number of image encoders but 4 fewer LLM backends, resulting in inferior performance. Moreover,

reducing image encoders below the monolith baseline contradicts the core goal of decoupling to scale up/out the image encoders independently for multimodal processing.

Impact of Image:Text Request Ratio. Figure 18 also shows the impact of varying image-text request percentages in the workload (Y-axis). As this percentage increases from 10% to 90% (more image-heavy), TTFT for Llama-3.2 (CroAttn) increases. InternVL (DecOnly) follows a similar trend, except at lower *Text Instance* ratios (e.g., 4:48), where P99 TTFT decreases from 3.8 to 3.3 seconds due to reduced text load. This stems from DecOnly models’ poor prefill efficiency. For the same reason, at low image-text request percentages (e.g., 10%), InternVL sees a lower P99 TTFT as more *Text Instances* help distribute the text-heavy load.

On the other hand, across all image-text request percentages, increasing the number of *Text Instances* raises P99 TTFT in Llama3.2 due to a reduced number of *Image Instances*, leading to longer image encoding times. However, regardless of distribution, MODSERVE outperforms the monolith baseline (by up to 18.4× for Llama3.2 and 9.2× for InternVL) when Image:Text Instance ratio exceeds 2.4, demonstrating its efficiency handling multimodal workloads.

6 Related Work

LMM Characterization. Lee *et al.* [24] provides a comprehensive characterization of multimodal generation models at Meta, while we focus on multimodal inputs. Hou *et al.* [15] focus on traditional multimodal models employing small-scale convolutional neural networks. In contrast, our work presents a detailed analysis of multimodal input workloads on both open-source LMM models and production traces, highlighting their unique execution and workload patterns.

LMM Serving. Recent research has introduced several techniques to optimize LMM serving by addressing key inefficiencies in inference computation and memory usage. Inf-MLLM [36] employs token caching strategies and attention bias to maintain performance with long contexts while reducing KV cache memory consumption. Elastic Cache [31] utilizes an importance-driven cache merging strategy to prune KV caches efficiently during inference. Dynamic-LLaVA [18], VTW [30], and QueCC [27] present various vision token sparsification and compression techniques to dynamically reduce redundancy in vision tokens. These optimizations primarily operate at the model level, trading off computational overhead with model performance. They are orthogonal to our proposed system-level design for SLO-driven LMM serving that does not impact model performance, which can further benefit from such model-level advancements, e.g., faster image encoding through token compression.

Text-Centric LLM Serving. Recent studies have delved into disaggregating LLM prefill and decode phases for text-only LLM serving. Examples include Splitwise [41], DistServe [59], Mooncake [43], and MemServe [16]. Other optimizations for

LLM serving include key-value cache management [23], continuous batching [57], request scheduling [1, 42, 45, 51], and energy optimization [44, 49, 50]. While these optimizations can be applied in MODSERVE to enhance LLM backend prefill and decode efficiency, our work focuses on the unique characteristics of multimodal models.

7 Conclusion

We present the first comprehensive systems analysis of LMMs on both open-source models and production LMM inference traces. Our insights lead to the design of MODSERVE, a scalable and resource-efficient LMM-serving framework that decouples inference stages for dynamic reconfiguration, adaptive scaling, and modality-aware scheduling. Evaluations show that MODSERVE achieves 25–41% cost savings compared to the state-of-the-art while efficiently serving production-scale LMM inference workloads.

References

- [1] Amey Agrawal, Nitin Kedia, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav Gulavani, Alexey Tumanov, and Ramachandran Ramjee. 2024. Taming Throughput-Latency Tradeoff in LLM Inference with Sarathi-Serve. In *Proceedings of the 18th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*.
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. *2024 Conference on Neural Information Processing Systems (NeurIPS 2024)* 35 (2022), 23716–23736.
- [3] Dosovitskiy Alexey. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929* (2020).
- [4] Reza Yazdani Aminabadi, Samyam Rajbhandari, Ammar Ahmad Awan, Cheng Li, Du Li, Elton Zheng, Olatunji Ruwase, Shaden Smith, Minjia Zhang, Jeff Rasley, et al. 2022. DeepSpeed-Inference: Enabling efficient inference of transformer models at unprecedented scale. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*.
- [5] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. 2022. VisualGPT: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, 18030–18040.
- [6] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling.
- [7] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites. *arXiv preprint arXiv:2404.16821* (2024).
- [8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*, 24185–24198.
- [9] Chen, Zhe and Wu, Jiannan and Wang, Wenhai and Su, Weijie and Chen, Guo and Xing, Sen and Zhong, Muyan and Zhang, Qinglong and Zhu, Xizhou and Lu, Lewei and others. 2024. HuggingFace Model:

- OpenGVLab/InternVL2_5-26B. https://huggingface.co/OpenGVLab/InternVL2_5-26B.
- [10] Jianfeng Chi, Ujjwal Karn, Hongyuan Zhan, Eric Smith, Javier Rando, Yiming Zhang, Kate Plawiak, Zacharie Delpierre Coudert, Kartikeya Upasani, and Mahesh Pasupuleti. 2024. Llama Guard 3 Vision: Safeguarding Human-AI Image Understanding Conversations. *arXiv preprint arXiv:2411.10414* (2024).
- [11] Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. NVLM: Open Frontier-Class Multimodal LLMs. *arXiv:2409.11402* [cs.CL] <https://arxiv.org/abs/2409.11402>
- [12] Dai, Wenliang and Lee, Nayeon and Wang, Boxin and Yang, Zhuolin and Liu, Zihan and Barker, Jon and Rintamaki, Tuomas and Shoeybi, Mohammad and Catanzaro, Bryan and Ping, Wei. 2024. HuggingFace Model: nvidia/NVLM-D-72B. <https://huggingface.co/nvidia/NVLM-D-72B>.
- [13] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. 2024. VLMEvalKit: An open-source toolkit for evaluating large multi-modality models. https://huggingface.co/spaces/opencompass/open_vlm_leaderboard. In *Proceedings of the 32nd ACM International Conference on Multimedia*.
- [14] Indranil Gupta, Tushar D Chandra, and Germán S Goldszmidt. 2001. On Scalable and Efficient Distributed Failure Detectors. In *Proceedings of the Twentieth Annual ACM Symposium on Principles of Distributed Computing (PODC)*.
- [15] Xiaofeng Hou, Cheng Xu, Jiacheng Liu, Xuehan Tang, Lingyu Sun, Chao Li, and Kwang-Ting Cheng. 2022. Characterizing and understanding end-to-end multi-modal neural networks on GPUs. *IEEE Computer Architecture Letters* 21, 2 (2022), 125–128.
- [16] Cunchen Hu, Heyang Huang, Junhao Hu, Jiang Xu, Xusheng Chen, Tao Xie, Chenxi Wang, Sa Wang, Yungang Bao, Ninghui Sun, et al. 2024. MemServe: Context caching for disaggregated LLM serving with elastic memory pool. *arXiv preprint arXiv:2406.17565* (2024).
- [17] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2023. PromptCap: Prompt-guided image captioning for VQA with GPT-3. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV 2023)*. 2963–2975.
- [18] Wenxuan Huang, Zijie Zhai, Yunhang Shen, Shaoshen Cao, Fei Zhao, Xiangfeng Xu, Zheyu Ye, and Shaohui Lin. 2024. Dynamic-LLaVA: Efficient Multimodal Large Language Models via Dynamic Vision-language Context Sparsification. *arXiv preprint arXiv:2412.00876* (2024).
- [19] HuggingFace. 2024. Audio-Text-to-Text Models. https://huggingface.co/models?pipeline_tag=audio-text-to-text.
- [20] HuggingFace. 2024. HuggingFace Transformers. <https://huggingface.co/docs/transformers/en/index>.
- [21] HuggingFace. 2024. Image-Text-to-Text Models. https://huggingface.co/models?pipeline_tag=image-text-to-text.
- [22] Aditya K Kamath, Ramya Prabhu, Jayashree Mohan, Simon Peter, Ramachandran Ramjee, and Ashish Panwar. 2024. Pod-attention: Unlocking full prefill-decode overlap for faster LLM inference. *arXiv preprint arXiv:2410.18038* (2024).
- [23] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the Symposium on Operating Systems Principles (SOSP)*.
- [24] Yejin Lee, Anna Sun, Basil Hosmer, Bilge Acun, Can Balioglu, Changhan Wang, Charles David Hernandez, Christian Puhersch, Daniel Haziza, Driss Guessous, et al. 2024. Characterizing and Efficiently Accelerating Multimodal Generation Model Inference. *arXiv preprint arXiv:2410.00215* (2024).
- [25] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv preprint arXiv:2408.03326* (2024).
- [26] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv:2408.03326* [cs.CV] <https://arxiv.org/abs/2408.03326>
- [27] Kevin Y Li, Sachin Goyal, Joao D Semedo, and J Zico Kolter. 2024. Inference Optimal VLMs Need Only One Visual Token but Larger Models. *arXiv preprint arXiv:2411.03312* (2024).
- [28] Li, Bo and Zhang, Yuanhan and Guo, Dong and Zhang, Renrui and Li, Feng and Zhang, Hao and Zhang, Kaichen and Li, Yanwei and Liu, Ziwei and Li, Chunyuan. 2024. HuggingFace Model: Imms-lab/llava-onevision-qwen2-72b-ov-sft. <https://huggingface.co/Imms-lab/llava-onevision-qwen2-72b-ov-sft>.
- [29] Li, Bo and Zhang, Yuanhan and Guo, Dong and Zhang, Renrui and Li, Feng and Zhang, Hao and Zhang, Kaichen and Li, Yanwei and Liu, Ziwei and Li, Chunyuan. 2024. HuggingFace Model: Imms-lab/llava-onevision-qwen2-7b-ov. <https://huggingface.co/Imms-lab/llava-onevision-qwen2-7b-ov>.
- [30] Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. 2024. Boosting Multimodal Large Language Models with Visual Tokens Withdrawal for Rapid Inference. *arXiv preprint arXiv:2405.05803* (2024).
- [31] Zuyan Liu, Benlin Liu, Jiahui Wang, Yuhao Dong, Guangyi Chen, Yongming Rao, Ranjay Krishna, and Jiwen Lu. 2025. Efficient inference of vision instruction-following models with elastic cache. In *European Conference on Computer Vision (ECCV)*.
- [32] Meta AI. 2024. HuggingFace Model: meta-llama/Llama-3.2-11B-Vision-Instruct. <https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct>.
- [33] Meta AI. 2024. HuggingFace Model: meta-llama/Llama-3.2-90B-Vision-Instruct. <https://huggingface.co/meta-llama/Llama-3.2-90B-Vision-Instruct>.
- [34] Microsoft Azure. 2024. Azure VM NDm-A100-v4 sizes series. <https://learn.microsoft.com/en-us/azure/virtual-machines/sizes/gpu-accelerated/ndma100v4-series>.
- [35] Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. ClipCap: CLIP Prefix for Image Captioning. *arXiv preprint arXiv:2111.09734* (2021).
- [36] Zhenyu Ning, Jieru Zhao, Qihao Jin, Wenchao Ding, and Minyi Guo. 2024. Inf-MLLM: Efficient Streaming Inference of Multimodal Large Language Models on a Single GPU. *arXiv preprint arXiv:2409.09086* (2024).
- [37] Runliang Niu, Jindong Li, Shiqi Wang, Yali Fu, Xiyu Hu, Xueyuan Leng, He Kong, Yi Chang, and Qi Wang. 2024. ScreenAgent: A Vision Language Model-Driven Computer Control Agent. *arXiv preprint arXiv:2402.07945* (2024).
- [38] NVIDIA. 2024. NVIDIA TensorRT. <https://github.com/NVIDIA/TensorRT>.
- [39] Diego Ongaro and John Ousterhout. 2014. In Search of An Understandable Consensus Algorithm. In *Proceedings of the 2014 USENIX Conference on USENIX Annual Technical Conference (ATC)*.
- [40] OpenAI. 2025. Computer-Using Agent: Introducing a universal interface for AI to interact with the digital world. <https://openai.com/index/computer-using-agent>.
- [41] Pratyush Patel, Esha Choukse, Chaojie Zhang, Aashaka Shah, Ñiño Goiri, Saeed Maleki, and Ricardo Bianchini. 2024. Splitwise: Efficient generative LLM inference using phase splitting. In *Proceedings of the ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*.
- [42] Archit Patke, Dharmath Reddy, Saurabh Jha, Haoran Qiu, Christian Pinto, Chandra Narayanaswami, Zbigniew Kalbarczyk, and Ravishankar Iyer. 2024. Queue Management for SLO-Oriented Large Language Model Serving. In *Proceedings of the ACM Symposium on Cloud*

Computing (SoCC).

- [43] Ruoyu Qin, Zheming Li, Weiran He, Mingxing Zhang, Yongwei Wu, Weimin Zheng, and Xinran Xu. 2024. Mooncake: A KVCache-centric Disaggregated Architecture for LLM Serving. *arXiv preprint arXiv:2407.00079* (2024).
- [44] Haoran Qiu, Weichao Mao, Archit Patke, Shengkun Cui, Saurabh Jha, Chen Wang, Hubertus Franke, Zbigniew Kalbarczyk, Tamer Başar, and Ravishankar K. Iyer. 2024. Power-aware Deep Learning Model Serving with μ -Serve. In *USENIX Annual Technical Conference (USENIX ATC 2024)*.
- [45] Haoran Qiu, Weichao Mao, Archit Patke, Shengkun Cui, Saurabh Jha, Chen Wang, Hubertus Franke, Zbigniew T. Kalbarczyk, Tamer Başar, and Ravishankar K. Iyer. 2024. Efficient Interactive LLM Serving with Proxy Model-based Sequence Length Prediction. In *The 5th International Workshop on Cloud Intelligence / AIOps at ASPLOS 2024*.
- [46] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A benchmark for visual question answering using world knowledge. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- [47] Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023)*.
- [48] Mohammad Shazeer et al. 2020. Megatron-LM: Training Multi-Billion Parameter Language Models Using Pipeline Parallelism. *arXiv preprint arXiv:1909.08053* (2020). <https://arxiv.org/abs/1909.08053>
- [49] Jovan Stojkovic, Chaojie Zhang, Íñigo Goiri, Esha Choukse, Haoran Qiu, Rodrigo Fonseca, Josep Torrellas, and Ricardo Bianchini. 2025. TAPAS: Thermal-and Power-Aware Scheduling for LLM Inference in Cloud Platforms. *arXiv preprint arXiv:2501.02600* (2025).
- [50] Jovan Stojkovic, Chaojie Zhang, Íñigo Goiri, Josep Torrellas, and Esha Choukse. 2025. DynamoLLM: Designing LLM Inference Clusters for Performance and Energy Efficiency. In *Proceedings of the IEEE International Symposium on High Performance Computer Architecture (HPCA)*.
- [51] Biao Sun, Ziming Huang, Hanyu Zhao, Wencong Xiao, Xinyi Zhang, Yong Li, and Wei Lin. 2024. Llumnix: Dynamic Scheduling for Large Language Model Serving. *arXiv preprint arXiv:2406.03243* (2024).
- [52] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).
- [53] vLLM. 2024. Distributed Inference and Serving. https://docs.vllm.ai/en/latest/serving/distributed_serving.html.
- [54] Cheng Wang, Xusheng Chen, Weiwei Jia, Boxuan Li, Haoran Qiu, Shixiong Zhao, and Heming Cui. 2018. PLOVER: Fast, Multi-core Scalable Virtual Machine Fault-tolerance. In *Proceedings of the 15th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*.
- [55] A Waswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A Gomez, L Kaiser, and I Polosukhin. 2017. Attention is all you need. In *2017 Conference on Neural Information Processing Systems (NIPS 2017)*.
- [56] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. <https://www.aclweb.org/anthology/2020.emnlp-demos.6>
- [57] Gyeong-In Yu, Joo Seong Jeong, Geon-Woo Kim, Soojeong Kim, and Byung-Gon Chun. 2022. Orca: A Distributed Serving System for Transformer-Based Generative Models. In *Proceedings of the 16th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*.
- [58] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid Loss for Language Image Pre-Training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*.
- [59] Yinmin Zhong, Shengyu Liu, Junda Chen, Jianbo Hu, Yibo Zhu, Xuanzhe Liu, Xin Jin, and Hao Zhang. 2024. DistServe: Disaggregating Prefill and Decoding for Goodput-optimized Large Language Model Serving. In *Proceedings of the 18th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*.
- [60] Ding Zou, Wei Lu, Zhibo Zhu, Xingyu Lu, Jun Zhou, Xiaojin Wang, Kangyu Liu, Kefan Wang, Renen Sun, and Haiqing Wang. 2024. OptScaler: A Collaborative Framework for Robust Autoscaling in the Cloud. *Proceedings of the VLDB Endowment* 17, 12 (Aug. 2024), 4090–4103. doi:10.14778/3685800.3685829