SQB-SPEC: ACCELERATING LARGE MULTIMODAL MODELS THROUGH TASK-SPECIFIC SPECULATIVE DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

The deployment of Vision-Language Models (VLMs) is severely limited by the high latency of autoregressive decoding. While speculative decoding offers a lossless acceleration solution, existing monolithic draft models encounter a "representation ceiling" when handling the inherent task heterogeneity of VLMs. In this paper, we propose **SQB-Spec**, a task-specific speculative decoding framework for efficient and robust VLM acceleration. Built upon the EAGLE-3 paradigm, SQB-Spec introduces modality-specific projection pathways for enhanced cross-modal alignment and incorporates a Gated Attention mechanism to mitigate "attention sinks" in long-sequence multimodal drafting. To address task heterogeneity, we further propose the **Specialized Query-projection and Bias (SQB)** mechanism, which modularizes task-specific representational power into lightweight Query-projection (W_Q) branches and linear biases. A phase-level router adaptively activates these pathways, guided by a dual-component objective: a **Historical Load Balancing Loss** for global expert utilization and a **Stride-based Temporal Smoothness Loss** for aligning token-level training with stationary inference. Extensive evaluations on LLaVA-OneVision and Qwen3-VL series demonstrate that SQB-Spec achieves a peak speedup of 2.60x while strictly maintaining the original output distribution, significantly outperforming state-of-the-art baselines across diverse benchmarks including MMMU, DocVQA, and MVBench.

1 INTRODUCTION

Vision-Language Models (VLMs), exemplified by cutting-edge closed-source frontiers such as GPT-5 and Gemini 3 Pro, alongside leading open-source counterparts like LLaVA-OneVision An et al. (2025a); Li et al. (2024a) and Qwen3-VL **?**, have achieved remarkable success across diverse multimodal tasks. These applications span from high-precision document OCR to complex temporal video reasoning, each characterized by highly divergent semantic priors and processing requirements. Despite these advancements, the ubiquitous autoregressive paradigm remains a critical bottleneck, imposing substantial inference latency. In low-concurrency scenarios, sequential decoding exhibits low arithmetic intensity, failing to fully leverage the parallel throughput of modern hardware.

To mitigate this, Speculative Decoding has emerged as an effective approach for inference acceleration. This paradigm employs a draft model with significantly fewer parameters compared to the target model to propose a sequence of candidate tokens, which are subsequently verified by the target model in a single parallelized step. By transitioning to a "propose-verify" workflow, speculative decoding increases arithmetic intensity and leverages the parallel computing capabilities of modern hardware. This is particularly beneficial for edge-side deployment with low query concurrency, where speculative decoding capitalizes on underutilized hardware throughput to reduce latency while strictly maintaining the original output distribution. Recently, several works have emerged in the context of VLM speculative decoding, primarily revolving around two optimization strategies: architectural refinements to improve vision-language alignment within the draft model and input-level techniques such as token pruning to mitigate the computational burden of processing extensive visual sequences.

Despite these advancements, existing VLM-based speculative decoding methodologies typically treat multimodal tasks as a homogenous distribution, overlooking their intrinsic heterogeneity. Since

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

distinct domains, such as high-precision OCR and temporal video reasoning, demand fundamentally different model capabilities Mathew et al. (2021); Wu et al. (2024), a monolithic draft model often encounters a performance ceiling across varied task distributions. Furthermore, the latent semantic requirements within a single response can evolve dynamically as the decoding process unfolds, a phenomenon that static, monolithic architectures fail to capture. Drawing inspiration from the success of Mixture-of-Experts (MoE) Fedus et al. (2022); DeepSeek-AI et al. (2025), we propose a hybrid architectural paradigm: by integrating general-purpose neural components with specialized pathways via an unsupervised router, our approach achieves superior predictive capacity while incurring nearly identical computational overhead compared to a standard backbone. Such a design allows the draft model to adaptively specialize in both diverse task characteristics and shifting contextual nuances with high efficiency.

In this work, we systematically investigate architectural configurations for coupling general-purpose and specialized neural components, leading to the development of the **Specialized Query-projection and Bias (SQB)** mechanism. Guided by our theoretical and empirical analyses, we propose a lightweight router-driven strategy that performs a single classification at the onset of each drafting phase. This routing decision determines the specific neural pathway for all subsequent decoding steps within the phase, ensuring architectural stability while effectively capturing both task-specific features and shifting contextual nuances. To implement this specialization with minimal overhead, we optimize multiple sets of Query projections (W_Q) and linear layer biases ($\mathbf{b}$) corresponding to the router's categories, while maintaining a shared, frozen backbone. This design offers high adaptability and representational power without significant parameter expansion, ensuring that the additional computational latency during the inference forward pass remains negligible. For the primary architecture of our draft model, we build upon EAGLE-3 Li et al. (2025), the state-of-the-art algorithm in LLM speculative decoding. To accommodate the specificities of VLMs, we implement modality-specific neural pathways to process visual and textual tokens independently, thereby capturing the distinct statistical and semantic properties of each modality. Besides, we integrate a gated attention mechanism Qiu et al. (2025) into the draft model's architecture to mitigate this issue, bolstering both inference performance and training stability.

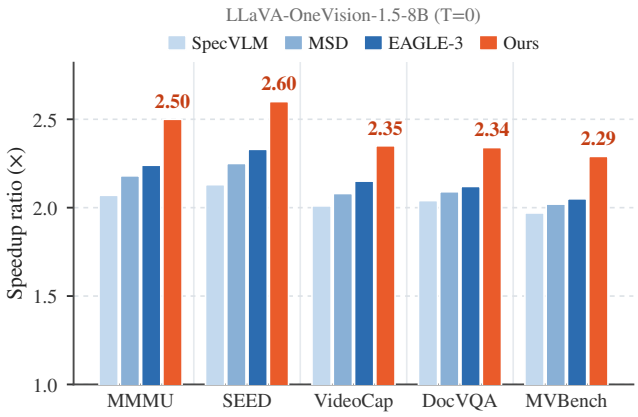


Figure 1: Speedup ratios across diverse multimodal benchmarks. Our method consistently outperforms state-of-the-art baselines, achieving up to 2.60× acceleration while maintaining superior performance across all evaluated tasks.

Extensive evaluations across diverse multimodal benchmarks demonstrate that our approach consistently outperforms existing state-of-the-art methods, achieving speedup ratios of up to **2.60**× while significantly enhancing the average accepted length across both image and video domains (as shown in Figure 1).

Our key contributions can be summarized as follows.

- We adapt and significantly enhance the EAGLE-3 framework for the VLM domain by incorporating modality-specific neural pathways and a Gated Attention mechanism.
- We propose the Specialized Query-projection and Bias mechanism, a novel architecture that employs phase-level routing to navigate task hetero-geneity and shifting contexts.
- Extensive evaluations across diverse multimodal bench-marks demonstrate that our approach achieves an outdtanding end-to-end acceleration improvement compared to existing VLM speculative decoding baselines.

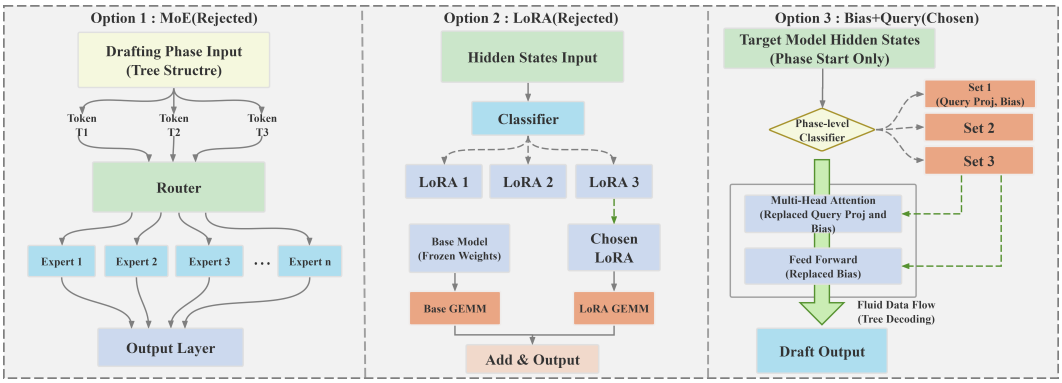


Figure 2: **Structural comparison of different drafting paradigms.** (a) **Monolithic Baseline**: Uses a single, fixed set of projections for all multimodal tasks, leading to a representation bottleneck. (b) **Conventional Adaptation (e.g., MoE/LoRA)**: Introduces task-specific experts or adapters. (c) **Ours (SQB-Spec)**: A lightweight yet adaptive design.

2 PRELIMINARIES

In this section, we establish the theoretical and algorithmic foundations for our work by reviewing the core mechanisms of EAGLE-2 and EAGLE-3. These frameworks represent the current state-of-the-art in speculative decoding for large language models, focusing respectively on inference-time tree optimization and training-time architectural enhancements.

2.1 EAGLE-3: MULTI-LEVEL FEATURE FUSION AND TRAINING-TIME TEST

EAGLE-3 enhances draft model expressivity through multi-level feature fusion and a "train-as-you-test" paradigm. It concatenates target model hidden states from multiple levels into an integrated feature $g = \text{FC}([l; m; h])$. While the initial drafting step uses g, subsequent steps recursively feed the draft model's own output a as a surrogate for g, ensuring semantic continuity. Crucially, EAGLE-3 removes the feature prediction loss ($\mathcal{L}_{fea}$) to focus entirely on distribution approximation, thereby avoiding regression-based error accumulation. During training, simulated "test steps" feed generated features back into the self-attention mechanism, using specialized masks to maintain tree-like dependencies. This strategy enables superior scaling and inference acceleration by bridging the train-inference gap.

3 METHOD

3.1 ANALYSIS OF SPECIALIZED PATHWAYS

To enhance the fitting capacity of the draft model across heterogeneous task distributions without compromising inference efficiency, we explore and analyze three potential architectural paradigms for integrating specialized neural components.

Direction 1: Mixture-of-Experts Architecture. A natural approach to handling multi-domain data is the MoE framework Fedus et al. (2022). However, we identify two critical bottlenecks that render it unsuitable for speculative drafting. First, in the drafting phase, candidate tokens are typically expanded in a tree-based structure rather than a linear chain to maximize acceptance probability Miao et al. (2024). This necessitates the draft model to forward multiple candidate tokens simultaneously. In an MoE setup, this leads to a high probability of activating multiple experts within a single decoding step, significantly increasing the active parameter count and exacerbating the memory bandwidth pressure (HBM-to-SRAM), which ultimately negates the speedup benefits of speculative decoding. Second, empirical studies suggest that MoE architectures yield marginal gains when the backbone parameter count is extremely small Clark et al. (2022), which is a common characteristic of lightweight draft models.

Direction 2: Specialized Low-Rank Adaptation. An alternative is to employ multiple domain-specific LoRA adapters Hu et al. (2021) orchestrated by a routing classifier. However, our hardware profiling reveals a significant latency penalty associated with this approach. As illustrated in Figure 3, the execution time of GEMM kernels with LoRA adapters is nearly 1.3-1.4 times that of standard GEMMs in the low-batch, small-scale regime. This overhead primarily stems from the frequent kernel launch latencies and the suboptimal utilization of Streaming Multiprocessors (SMs) during small-scale matrix multiplications, which fails to fully saturate the hardware's parallel processing capacity. While weight merging could mitigate this, it would require maintaining multiple full-parameter copies in VRAM or performing frequent weight recomputations, both of which impose prohibitive memory and synchronization overhead for practical deployment.

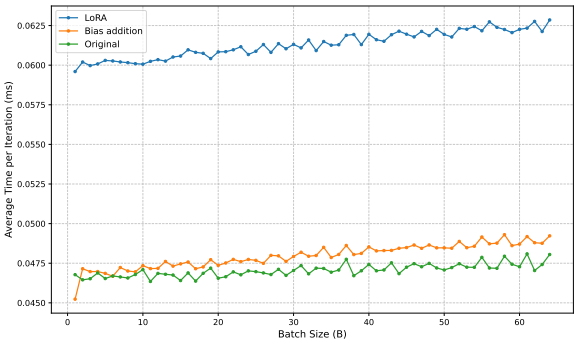


Figure 3: Runtime Comparison: LoRA Variants vs Bias (hidden size=4096, rank=16, dtype=BF16).

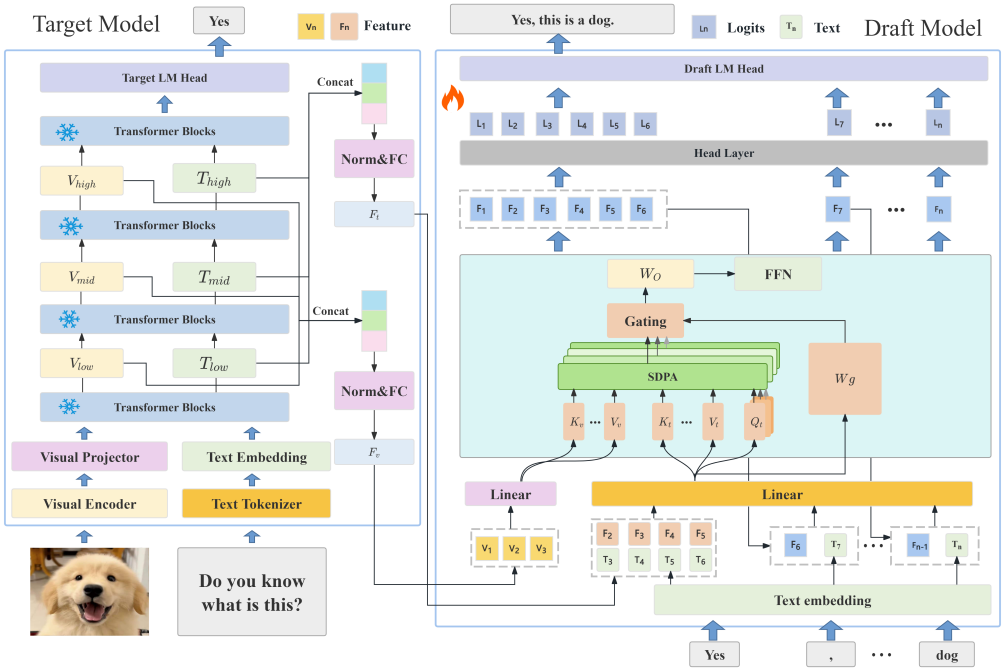
Direction 3: Specialized Query-projection and Bias. We propose **Specialized Query-projection and Bias (SQB)**, training router-specific Query projections (W_Q) and biases across layers (except the LM head) atop a frozen backbone. SQB ensures high hardware efficiency: bias addition is a **point-wise operation** with negligible computational cost. Furthermore, modern backends typically fuse these biases into primary GEMM kernels, eliminating extra kernel launches and maintaining original SM utilization. Despite deploying multiple branches, SQB's parameter-efficiency ensures the total storage remains below the original draft model's size. Regarding expressivity, specialized biases capture domain knowledge via linear shifts Ben Zaken et al. (2022), while W_Q recalibrates attention for modalities like OCR and video by leveraging Query-Key sensitivity to task semantics He et al. (2022). This synergy achieves representational capacity comparable to LoRA with superior hardware efficiency. Based on these analyses, we adopt the SQB strategy as our primary architectural foundation for capturing domain-specific features.

3.2 Model Architecture

Built upon the state-of-the-art speculative decoding framework, EAGLE-3, originally proposed for LLMs, our model introduces three core architectural enhancements tailored for the inherent heterogeneity of vision-language tasks: Router-driven Specialized Query-projection and Bias, Modality-specific Neural Pathways, and Gated Attention Mechanism. The overall framework is illustrated in Figure 4.

Router-driven Specialized Query-projection and Bias. The Router-driven SQB mechanism addresses two distinct dimensions of architectural adaptation: inter-task heterogeneity and intra-task contextual evolution. First, the inherent divergence between multimodal tasks, such as the distinct semantic priors required for document OCR versus temporal video analysis, necessitates specialized processing pathways to maintain high-fidelity predictions. Second, even within a single task, the latent requirements of a response may shift as the decoding sequence progresses, where the evolving context demands a re-calibration of the model's focus. To capture these nuances, we utilize a *phase-level routing* strategy where a lightweight classifier determines the optimal specialized branch at the onset of each drafting phase. This decision triggers the activation of specific Query projection matrices (W_Q) and linear layer biases ($\mathbf{b}$) tailored to the current state, ensuring a precise approximation of the target model distribution. As established in previous sections, this modular specialization provides essential representational flexibility while maintaining a computational profile nearly identical to that of a monolithic backbone.

Modality-specific Neural Pathways. Visual tokens, typically extracted by a pre-trained vision encoder, exhibit distinct statistical distributions and semantic densities compared to textual tokens.


Target Model
Yes
Feature
Target LM Head
Transformer Blocks
Concat
Norm&FC
Visual Projector
Text Embedding
Visual Encoder
Text Tokenizer
Do you know what is this?
Yes, this is a dog.
Logits
Text
Draft Model
Draft LM Head
Head Layer
FFN
Gating
SDPA
Linear
Text embedding
Yes
dog

sive curriculum. In the first stage, we establish foundational linguistic reasoning and multimodal instruction-following capabilities by utilizing text and image-text pairs. The second stage then extends this foundation to complex spatio-temporal reasoning by incorporating video data, which ensures the backbone remains robust when processing high-density visual signals and long-range temporal dependencies. Following the principles of EAGLE-3, both stages employ a Train-at-Test-Time where the draft model utilizes its own previously generated auxiliary features as inputs for subsequent steps. The optimization objective is to minimize the cumulative cross-entropy loss between the target model distribution P_i and the draft model prediction $\hat{P}_i$ over T drafting steps:

$$\mathcal{L}_{\text{reg}} = -\sum_{i=1}^{T} w_{ttt}^{i-1} \sum_{j \in \mathcal{V}} P_{i,j} \log \hat{P}_{i,j}, \tag{1}$$

where $\mathcal{V}$ denotes the vocabulary, T is the drafting length, and $w_{ttt} \in (0, 1)$ represents a discount factor that prioritizes the fidelity of earlier predictions. By the end of these stages, the draft model backbone achieves a high-fidelity approximation of the target model's general behavior across both image and video domains.

Stage 3: Router and Specialization Tuning. In the final stage, the draft model backbone is frozen to preserve the learned multimodal distributions. We optimize the router alongside the specialized query projections and biases using the union of all previously sampled datasets. To promote exploration and enhance generalization across diverse tasks, expert selection is performed via temperature sampling governed by a hyperparameter w_{temp}. Furthermore, to ensure stable phase-level routing and prevent expert collapse, we employ a dual-component auxiliary loss $\mathcal{L}_{\text{router}}$. The first component, **Historical Load Balancing Loss** ($\mathcal{L}_{\text{bal}}$), maintains a uniform utilization of the N expert branches by penalizing experts that are over-represented in the history buffer $\mathcal{C}$, which caches selection frequencies over the past M batches. Let $\bar{\mathcal{C}}_i$ be the historical average frequency for expert i. The loss is defined as:

$$\mathcal{L}_{\text{bal}} = \frac{N}{BO} \sum_{i,b,j}^{N,B,O} \bar{\mathcal{C}}_i S_{b,i,j}, \tag{2}$$

where N denotes the total number of specialized expert branches, B is the batch size, O is the sequence length, and $S_{b,i,j}$ represents the softmax routing probability for expert i at position j in sample b. By treating $\bar{\mathcal{C}}_i$ as a dynamic penalty weight, this objective discourages the router from converging on a subset of heavily used experts, effectively maximizing the collective representational capacity of all specialized branches.

The second component is the **Stride-based Temporal Smoothness Loss** ($\mathcal{L}_{\text{smooth}}$), which is designed to enforce local stationarity in expert selection. Instead of using a logarithmic cross-entropy, this loss employs a direct dot product between the current routing scores and a localized pseudo-target to promote temporal consistency. We first aggregate the multi-hot indicator vectors $\mathbb{I} \in \{0, 1\}^N$ of neighboring tokens within a temporal stride K to reflect the neighborhood's expert preference. The smoothness loss is formulated as:

$$\mathcal{L}_{\text{smooth}} = -\frac{1}{BO} \sum_{b,j=1}^{B,O} \left(\sum_{0<|k|\leq K} w_{\text{decay}}^{|k|} \cdot \mathbb{I}_{j+k,\text{idx}} \right) \cdot V_{b,j}, \tag{3}$$

where $w_{\text{decay}} \in (0, 1)$ represents the temporal decay factor and K denotes the radius of the local neighborhood.Here, $V_{b,j}$ denotes the routing probability of the expert selected for the j-th token. The term $\mathbb{I}_{j+k,\text{idx}}$ refers to the indicator value of that same selected expert at the neighboring position $j + k$. By minimizing this objective through a pointwise dot product, the router is encouraged to assign higher probabilities to experts that are consistent with their temporal neighbors, effectively suppressing high-frequency oscillations in expert selection.

The final optimization objective for the router and specialized SQB components is the weighted sum of these two auxiliary losses:

$$\mathcal{L}_{\text{router}} = \mathcal{L}_{\text{bal}} + \alpha \cdot \mathcal{L}_{\text{smooth}}, \tag{4}$$

where α is a hyperparameter balancing the load distribution and temporal consistency constraints. This joint optimization allows the SQB modules to adaptively capture domain-specific features while maintaining the phase-level stationarity required for efficient speculative execution.

4 EXPERIMENTS

4.1 EFFECTIVENESS

Models and Tasks. To evaluate the performance of our method, we select LLaVA-OneVision-1.5-8B-Instruct An et al. (2025b), Qwen-VL family Bai et al. (2025a;b) as the target models due to their robust multimodal capabilities and widespread adoption in the community. To ensure a comprehensive assessment of acceleration consistency, all experiments are conducted under both greedy decoding ($T = 0$) and stochastic decoding ($T = 1$) configurations. Experiments are performed on NVIDIA H100 GPU. We evaluate our approach across a wide range of benchmarks to cover diverse task distributions: **MMMU**Yue et al. (2024), **SEED-Bench**Li et al. (2024b), **DocVQA**Mathew et al. (2021), **VideoDetailedCaption**Wu et al. (2024), **MMMU-Pro** Yue et al. (2025), MME Fu et al. (2025) and **MVBench**Li et al. (2024c).

Metrics. We report two primary metrics to quantify the efficiency of our framework. The first is the Speedup Ratio (SR), defined as the ratio of the wall-clock latency of vanilla autoregressive decoding to that of our speculative decoding approach. The second is the Average Accepted Length (τ), which represents the average number of tokens accepted by the target model per drafting-verification cycle. The metric τ serves as a key indicator of the verification efficiency and the draft model's ability to fit the output distribution of the target model.

Implements. For all three training stages, we maintain a global batch size of 64 and employ the AdamW optimizer with a learning rate of 5e-5, coupled with a cosine decay scheduler. To align the datasets with the target model's output distribution, we utilize the respective target models to regenerate responses for the training samples. **Stage 1:** We train for 1 epoch using the general-purpose text dataset ShareGPT-68K anon8231489123 (2023) and the general-purpose image instruction dataset LLaVA-Instruct-150K Liu et al. (2023). **Stage 2:** Building upon the data used in Stage 1, we incorporate the video dataset Video-R1-COT-165K Feng et al. (2025) and a randomly sampled subset of 68k instances from LLaVA-Video-178K Zhang et al. (2025b). This stage is trained for 4 epochs to enhance visual understanding and reasoning capabilities. **Stage 3:** We utilize the identical dataset as Stage 2 and train for 2 epochs. The hyperparameters are configured as follows: $w_{ttt} = 0.8$, $w_{decay} = 0.6$, $K = 4$, and $\alpha = 0.1$.

Comparisons. We use vanilla autoregressive decoding as the baseline, which serves as the benchmark for speedup ratios (1.00x). We compare our method against several recent speculative decoding methods, including EAGLE-3Li et al. (2025), MASSVGanesan et al. (2025), SpecVLMHuang et al. (2025), and MSDLin et al. (2025). EAGLE-3 is the current state-of-the-art speculative decoding method in the domain of LLMs, which is directly migrated to the VLM domain. MASSV, SpecVLM, and MSD are three specialized VLM speculative decoding frameworks.

Main results. As shown in Table 1, SQB-Spec consistently outperforms competing methods across various models and tasks, achieving speedups of 1.93x-2.60x over vanilla autoregressive decoding. Both the speedup ratio and the average acceptance length significantly surpass those of alternative approaches. This demonstrates that our specialized architecture can precisely capture domain-specific features, enabling a more effective approximation of the target model distribution compared to monolithic architectures, all with negligible overhead.

Furthermore, we observe that speedups for video tasks (e.g., VideoDetailedCaption and MVBench) are generally lower than those for image tasks. This is attributed to the higher number of visual tokens, which increases the proportion of prefill latency. Since speculative decoding cannot accelerate the prefill phase, the overall end-to-end speedup is constrained by this bottleneck. The speedup under Temperature=1 is also notably lower than that under Temperature=0, as the introduction of randomness at Temperature=1 hinders the draft model's predictive capability. The extent of this reduction varies depending on the output diversity of the respective models.

4.2 ABLATION STUDY

1 In this section, we conduct a series of ablation experiments to systematically evaluate the contribution of our proposed components. All experiments are performed using LLaVA-OneVision and Qwen-VL family. We evaluate the performance by reporting the speedup ratio (SR) and the average accepted length (τ) to demonstrate the effectiveness of our method.

Table 1: **Main Results: Efficiency comparison across multiple target models and benchmarks.** We evaluate our method on seven core tasks: MMMU, SEED-Bench, VideoDetailedCaption, DocVQA, MVBench, MME, and MMMU-Pro. We report speedup ratios (SR) and average accepted lengths (τ) compared to the vanilla autoregressive baseline. All methods ensure identical output distributions.

Target Model	Method	MMMU		SEED		VideoDetCap		DocVQA		MVBench		MME		MMMU-Pro		Average	
		SR	τ	SR	τ	SR	τ	SR	τ	SR	τ	SR	τ	SR	τ	SR	τ
Temperature=0																	
LLaVA-OneVision-1.5-8B-Instruct	MASSV	1.39x	3.65	1.41x	3.77	1.36x	3.53	1.36x	3.41	1.33x	3.29	1.40x	3.71	1.33x	3.52	1.37	3.55
	SpecVLM	2.07x	3.86	2.13x	4.15	2.01x	3.73	2.04x	3.79	1.97x	3.68	2.10x	4.02	1.99x	3.72	2.04	3.85
	Vispec	2.14x	3.94	2.20x	4.12	2.05x	3.80	2.09x	3.86	2.02x	3.70	2.17x	4.03	2.05x	3.76	2.10	3.89
	MSD	2.18x	3.46	2.25x	3.55	2.08x	3.34	2.09x	3.32	2.02x	3.28	2.21x	3.58	2.09x	3.31	2.13	3.41
	EAGLE-3	2.24x	3.68	2.33x	3.81	2.15x	3.45	2.12x	3.46	2.05x	3.32	2.30x	3.79	2.15x	3.50	2.19	3.57
	Ours w/o SQB	2.33x	3.92	2.43x	4.04	2.23x	3.68	2.20x	3.73	2.13x	3.52	2.39x	3.98	2.24x	3.74	2.28	3.80
	EAGLE-3 w SQB	2.46x	4.48	2.55x	4.61	2.32x	4.43	2.31x	4.46	2.25x	4.29	2.52x	4.57	2.38x	4.30	2.40	4.45
	Ours	**2.50x**	**4.59**	**2.60x**	**4.75**	**2.35x**	**4.52**	**2.34x**	**4.55**	**2.29x**	**4.41**	**2.57x**	**4.68**	**2.42x**	**4.44**	**2.44**	**4.56**
Qwen2.5-VL-7B	MASSV	1.13x	2.89	1.16x	2.97	1.12x	2.81	1.10x	2.73	1.07x	2.63	1.13x	2.93	1.09x	2.85	1.11	2.83
	SpecVLM	1.70x	3.10	1.75x	3.28	1.67x	3.01	1.69x	3.06	1.61x	2.93	1.70x	3.23	1.61x	2.93	1.68	3.08
	Vispec	1.75x	3.16	1.79x	3.25	1.70x	3.05	1.71x	3.13	1.65x	3.00	1.79x	3.20	1.67x	2.99	1.72	3.11
	EAGLE-3	1.90x	3.12	1.94x	3.16	1.81x	2.96	1.90x	3.07	1.80x	2.90	1.95x	3.20	1.82x	3.00	1.87	3.06
	EAGLE-3 w SQB	2.03x	3.55	2.07x	3.60	1.93x	3.38	2.02x	3.49	1.91x	3.55	2.08x	3.65	1.95x	3.40	2.00	3.52
	Ours	**2.09x**	**3.76**	**2.13x**	**3.82**	**1.99x**	**3.62**	**2.08x**	**3.71**	**1.97x**	**3.52**	**2.14x**	**3.87**	**2.01x**	**3.60**	**2.06**	**3.70**
Qwen3-VL-4B	MASSV	1.30x	3.26	1.35x	3.44	1.30x	3.17	1.31x	3.02	1.28x	2.94	1.31x	3.30	1.26x	3.20	1.30	3.19
	SpecVLM	1.95x	3.44	2.02x	3.69	1.89x	3.32	1.92x	3.40	1.85x	3.35	2.00x	3.58	1.88x	3.33	1.93	3.44
	Vispec	2.02x	3.55	2.08x	3.70	1.93x	3.46	1.98x	3.47	1.92x	3.37	2.06x	3.67	1.95x	3.39	1.99	3.52
	EAGLE-3	2.05x	3.30	2.09x	3.34	1.95x	3.13	2.04x	3.24	1.94x	3.07	2.10x	3.37	1.98x	3.17	2.02	3.23
	EAGLE-3 w SQB	2.18x	3.82	2.22x	3.87	2.08x	3.64	2.17x	3.75	2.05x	3.54	2.23x	3.92	2.10x	3.66	2.15	3.74
	Ours	**2.39x**	**4.06**	**2.43x**	**4.12**	**2.28x**	**3.95**	**2.36x**	**4.02**	**2.24x**	**3.88**	**2.44x**	**4.16**	**2.31x**	**3.90**	**2.35**	**4.01**
Qwen3-VL-8B	MASSV	1.21x	3.15	1.25x	3.23	1.20x	2.99	1.18x	2.90	1.18x	2.82	1.22x	3.20	1.17x	3.01	1.20	3.04
	SpecVLM	1.82x	3.28	1.89x	3.62	1.77x	3.21	1.78x	3.22	1.73x	3.14	1.86x	3.42	1.73x	3.24	1.80	3.30
	Vispec	1.88x	3.34	1.93x	3.51	1.78x	3.30	1.82x	3.32	1.79x	3.19	1.90x	3.47	1.81x	3.26	1.84	3.34
	EAGLE-3	1.98x	3.21	2.01x	3.25	1.88x	3.04	1.98x	3.15	1.88x	2.98	2.02x	3.28	1.90x	3.08	1.95	3.14
	SQB-Spec w/o	2.05x	3.46	2.08x	3.50	1.94x	3.29	2.05x	3.41	1.94x	3.22	2.09x	3.53	1.97x	3.30	2.02	3.39
	EAGLE-3 w SQB	2.12x	3.73	2.16x	3.78	2.02x	3.55	2.11x	3.66	1.99x	3.45	2.17x	3.83	2.04x	3.57	2.09	3.65
	Ours	**2.19x**	**3.97**	**2.23x**	**4.03**	**2.04x**	**3.89**	**2.19x**	**3.94**	**2.06x**	**3.84**	**2.25x**	**4.07**	**2.11x**	**3.82**	**2.15**	**3.94**
Qwen3-VL-32B	MASSV	1.12x	3.01	1.12x	3.10	1.09x	2.93	1.10x	2.77	1.09x	2.66	1.12x	3.07	1.05x	2.89	1.10	2.92
	SpecVLM	1.66x	3.13	1.71x	3.36	1.61x	3.11	1.65x	3.13	1.57x	3.00	1.67x	3.32	1.59x	3.08	1.64	3.16
	Vispec	1.69x	3.24	1.75x	3.36	1.65x	3.12	1.67x	3.19	1.63x	3.03	1.74x	3.31	1.64x	3.10	1.68	3.19
	EAGLE-3	2.26x	3.12	2.29x	3.17	2.15x	2.97	2.22x	3.08	2.14x	2.94	2.30x	3.20	2.17x	3.01	2.22	3.07
	Ours	**2.52x**	**3.84**	**2.55x**	**3.89**	**2.38x**	**3.74**	**2.45x**	**3.82**	**2.36x**	**3.73**	**2.58x**	**3.93**	**2.44x**	**3.72**	**2.47**	**3.81**
Temperature=1																	
LLaVA-OneVision-1.5-8B-Instruct	MASSV	1.33x	3.49	1.36x	3.61	1.32x	3.40	1.30x	3.24	1.28x	3.15	1.34x	3.55	1.28x	3.36	1.32	3.40
	SpecVLM	1.91x	3.59	1.96x	3.82	1.78x	3.48	1.80x	3.51	1.83x	3.39	1.93x	3.72	1.84x	3.45	1.86	3.57
	Vispec	2.04x	3.66	2.08x	3.84	1.94x	3.51	1.96x	3.56	1.92x	3.42	2.05x	3.74	1.94x	3.48	1.99	3.60
	MSD	2.07x	3.33	2.15x	3.38	1.99x	3.21	2.02x	3.23	1.96x	3.13	2.09x	3.44	1.99x	3.19	2.04	3.27
	EAGLE-3	2.12x	3.47	2.21x	3.58	2.03x	3.24	2.07x	3.31	2.01x	3.14	2.15x	3.62	2.04x	3.31	2.09	3.38
	Ours	**2.36x**	**4.08**	**2.45x**	**4.22**	**2.22x**	**4.02**	**2.27x**	**4.04**	**2.23x**	**3.95**	**2.42x**	**4.15**	**2.28x**	**3.92**	**2.32**	**4.05**
Qwen2.5-VL-7B	MASSV	1.04x	2.88	1.10x	2.97	1.07x	2.77	1.05x	2.67	1.03x	2.58	1.09x	2.96	1.02x	2.77	1.06	2.80
	SpecVLM	1.52x	2.92	1.58x	3.18	1.44x	2.88	1.45x	2.90	1.45x	2.78	1.54x	3.00	1.45x	2.81	1.49	2.92
	Vispec	1.63x	3.00	1.66x	3.19	1.56x	2.92	1.59x	2.90	1.54x	2.85	1.65x	3.03	1.54x	2.85	1.60	2.96
	EAGLE-3	1.85x	3.18	1.89x	3.20	1.76x	3.02	1.85x	3.12	1.76x	2.96	1.90x	3.25	1.78x	3.05	1.83	3.11
	EAGLE-3 w SQB	1.98x	3.52	2.02x	3.57	1.88x	3.35	1.97x	3.46	1.87x	3.42	2.03x	3.62	1.90x	3.38	1.95	3.47
	Ours	**2.04x**	**3.72**	**2.08x**	**3.78**	**1.94x**	**3.58**	**2.03x**	**3.67**	**1.93x**	**3.39**	**2.09x**	**3.83**	**1.96x**	**3.56**	**2.01**	**3.65**
Qwen3-VL-4B	MASSV	1.23x	3.16	1.28x	3.30	1.25x	3.04	1.22x	2.93	1.18x	2.83	1.24x	3.16	1.18x	3.05	1.23	3.07
	SpecVLM	1.78x	3.25	1.86x	3.43	1.69x	3.18	1.71x	3.15	1.71x	3.02	1.80x	3.32	1.74x	3.15	1.76	3.21
	Vispec	1.90x	3.27	1.94x	3.47	1.84x	3.20	1.83x	3.23	1.81x	3.04	1.94x	3.41	1.81x	3.18	1.87	3.26
	EAGLE-3	2.02x	3.30	2.04x	3.33	1.92x	3.15	2.00x	3.24	1.91x	3.09	2.03x	3.36	1.95x	3.17	1.98	3.23
	EAGLE-3 w SQB	2.13x	3.62	2.15x	3.66	2.03x	3.48	2.11x	3.57	2.02x	3.42	2.14x	3.69	2.06x	3.49	2.09	3.56
	Ours	**2.30x**	**4.02**	**2.33x**	**4.06**	**2.19x**	**3.90**	**2.27x**	**3.96**	**2.18x**	**3.85**	**2.31x**	**4.09**	**2.22x**	**3.88**	**2.26**	**3.97**
Qwen3-VL-8B	MASSV	1.14x	2.98	1.18x	3.14	1.13x	2.92	1.13x	2.82	1.13x	2.75	1.16x	3.04	1.11x	2.93	1.14	2.94
	SpecVLM	1.68x	3.09	1.71x	3.32	1.53x	3.01	1.58x	3.05	1.60x	2.91	1.66x	3.23	1.59x	3.00	1.62	3.09
	Vispec	1.77x	3.15	1.83x	3.34	1.67x	3.01	1.71x	3.05	1.66x	2.92	1.79x	3.17	1.69x	2.99	1.73	3.09
	EAGLE-3	1.94x	3.23	1.97x	3.25	1.83x	3.07	1.92x	3.17	1.83x	3.02	1.95x	3.28	1.86x	3.08	1.90	3.16
	Ours	**2.09x**	**3.78**	**2.12x**	**3.79**	**1.96x**	**3.68**	**2.07x**	**3.72**	**1.97x**	**3.63**	**2.13x**	**3.83**	**2.01x**	**3.65**	**2.05**	**3.73**
Qwen3-VL-32B	MASSV	1.08x	2.83	1.07x	2.93	1.04x	2.79	1.04x	2.63	1.00x	2.57	1.07x	2.92	1.04x	2.77	1.05	2.78
	SpecVLM	1.55x	2.93	1.56x	3.18	1.43x	2.82	1.42x	2.84	1.48x	2.81	1.53x	3.08	1.49x	2.84	1.49	2.93
	Vispec	1.61x	2.98	1.67x	3.15	1.53x	2.93	1.58x	2.97	1.52x	2.78	1.62x	3.09	1.54x	2.82	1.58	2.96
	EAGLE-3	2.06x	3.18	2.05x	3.20	1.96x	3.02	1.98x	3.12	1.92x	2.97	2.07x	3.23	1.99x	3.04	2.00	3.11
	Ours	**2.24x**	**3.68**	**2.23x**	**3.68**	**2.11x**	**3.55**	**2.14x**	**3.62**	**2.08x**	**3.55**	**2.25x**	**3.72**	**2.16x**	**3.55**	**2.17**	**3.62**

Effectiveness of the SQB Module. To verify the effectiveness of our SQB module, we conduct a comparative experiment by removing this component (denoted as **Ours w/o SQB**). The results, as shown in Table 1 at $T = 0$, demonstrate that the full model consistently outperforms the variant across all benchmarks. This improvement indicates that our phrase-level segmentation enables the draft model to dynamically select the optimal Q and bias for different tasks, thereby significantly enhancing both SR and τ.

Effectiveness of Architectural Refinement. To evaluate the efficacy of our architectural refinements, we further train a variant by integrating the SQB module into the EAGLE-3 backbone (**EAGLE-3 w SQB**), following the same configurations as our main experiments. As illustrated in Table 1, results collectively demonstrate that our refinements significantly enhance the draft model's predictive accuracy, leading to superior performance in both SR and τ.

Evaluation of Router Training Strategies. To investigate the effectiveness of our training objectives, we conduct ablation experiments by individually removing the $\mathcal{L}_{\text{smooth}}$) and the temperature sampling strategy during the training of the SQB module on our proposed backbone. The resulting acceleration

Table 2: **Ablation Study on Router Training Strategies.** We evaluate the impact of the stride-based temporal smoothness loss ($\mathcal{L}_{\text{smooth}}$) and temperature sampling (w_{temp}) on LLaVA, Qwen-VL families at $T = 0$. SR denotes speedup ratio and τ represents average accepted length. The "Ours" row corresponds to the full proposed framework.

Target Model	Method	MMMU		SEED		MVBench		MME		Average	
		SR	τ	SR	τ	SR	τ	SR	τ	SR	τ
LLaVA-OneVision-1.5-8B	w/o $\mathcal{L}_{\text{smooth}}$	2.34x	3.92	2.45x	4.08	2.14x	3.81	2.20x	3.79	2.28x	3.90
	w/o w_{temp}	2.44x	4.31	2.54x	4.45	2.23x	4.19	2.31x	4.12	2.38x	4.27
	Ours	**2.50x**	**4.59**	**2.60x**	**4.75**	**2.29x**	**4.41**	**2.40x**	**4.50**	**2.45x**	**4.56**
Qwen2.5-VL-7B	w/o $\mathcal{L}_{\text{smooth}}$	2.00x	3.33	2.04x	3.40	1.89x	3.24	1.94x	3.28	1.97x	3.31
	w/o w_{temp}	2.08x	3.66	2.12x	3.74	1.96x	3.57	2.00x	3.62	2.04x	3.65
	Ours	**2.13x**	**3.88**	**2.17x**	**3.94**	**2.01x**	**3.76**	**2.05x**	**3.82**	**2.09x**	**3.85**
Qwen3-VL-4B	w/o $\mathcal{L}_{\text{smooth}}$	2.16x	3.72	2.20x	3.80	2.04x	3.61	2.08x	3.66	2.12x	3.70
	w/o w_{temp}	2.24x	3.98	2.28x	4.06	2.11x	3.88	2.16x	3.92	2.20x	3.96
	Ours	**2.30x**	**4.22**	**2.34x**	**4.28**	**2.16x**	**4.08**	**2.21x**	**4.13**	**2.25x**	**4.18**
Qwen3-VL-8B	w/o $\mathcal{L}_{\text{smooth}}$	2.06x	3.41	2.10x	3.48	1.94x	3.32	1.98x	3.36	2.02x	3.39
	w/o w_{temp}	2.14x	3.75	2.18x	3.82	2.01x	3.65	2.06x	3.69	2.10x	3.73
	Ours	**2.19x**	**3.97**	**2.23x**	**4.03**	**2.06x**	**3.84**	**2.11x**	**3.88**	**2.15x**	**3.93**
Qwen3-VL-32B	w/o $\mathcal{L}_{\text{smooth}}$	1.90x	3.15	1.94x	3.22	1.79x	3.08	1.84x	3.12	1.87x	3.14
	w/o w_{temp}	1.97x	3.44	2.01x	3.50	1.85x	3.35	1.89x	3.39	1.93x	3.42
	Ours	**2.02x**	**3.65**	**2.06x**	**3.70**	**1.89x**	**3.52**	**1.94x**	**3.57**	**1.98x**	**3.61**

performance is summarized in Table 2. **Role of Temporal Smoothness.** As shown in the table, removing $\mathcal{L}_{\text{smooth}}$ leads to a clear performance drop in both **SR** and τ. This is because the lack of temporal constraints causes high-frequency oscillations in expert assignments, creating a misalignment between fragmented token-level training and our stationary phase-level execution. **Impact of Temperature Sampling.** Similarly, replacing stochastic temperature sampling with deterministic *argmax* selection impairs the router's robustness and overall efficiency. These results confirm that temperature sampling serves as a critical modality-aware dropout, enabling the router to generalize effectively across diverse and ambiguous multimodal task domains.

5 Related Work

Speculative Decoding in LLMs. Speculative decoding accelerates inference by decoupling drafting and verification stages Leviathan et al. (2023); Chen et al. (2023). To eliminate the overhead of auxiliary models, self-speculative strategies such as LayerSkip Elhoushi et al. (2024) and Self-Speculative Decoding Zhang et al. (2024) leverage internal layers of the target model, while Lookahead Fu et al. (2023) and REST He et al. (2024) utilize Jacobi iterations or retrieval-based caches for drafting. Alternatively, Medusa Cai et al. (2024) integrates multiple MLP heads for parallel token prediction. A prominent direction involves enhancing consistency through feature-level prediction. The EAGLE series Li et al. (2024d) emphasizes feature-level consistency by predicting continuous embeddings and utilizing dynamic tree topologies. Advancing this approach, EAGLE-3 Li et al. (2025) incorporates multi-layer feature fusion and self-simulation training to mitigate error accumulation, establishing a new scaling law for inference acceleration that outperforms previous feature-level benchmarks such as HASS Zhang et al. (2025a).

6 Conclusion

In this paper, we demonstrate that the proposed split-path strategy provides significant architectural value by decoupling modality-specific processing and enhancing speculative efficiency. However, certain limitations remain. Specifically, the additional architectural complexity may introduce non-negligible synchronization overhead, and the fixed branching mechanism may lack the flexibility required to adapt to varying input complexities, potentially leading to the loss of fine-grained cross-modal nuances. Future work will explore dynamic branch allocation and adaptive multi-scale feature fusion to further refine the balance between inference speed and reasoning precision in diverse multimodal scenarios.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. In this study, no human subjects or animal experimentation was involved. All datasets used were sourced in compliance with relevant usage guidelines, ensuring no violation of privacy. We have taken care to avoid any biases or discriminatory outcomes in our research process. No personally identifiable information was used, and no experiments were conducted that could raise privacy or security concerns. We are committed to maintaining transparency and integrity throughout the research process.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure that the results presented in this paper are reproducible. All code and datasets have been made publicly available in an anonymous repository to facilitate replication and verification. The experimental setup, including training steps, model configurations, and hardware details, is described in detail in the paper. We have also provided a full description of our method, to assist others in reproducing our experiments. Additionally, all training datasets and benchmarks are publicly available, ensuring consistent and reproducible evaluation results. We believe these measures will enable other researchers to reproduce our work and further advance the field.

AI USE STATEMENT

Large Language Models (LLMs) were used to aid in the writing and polishing of the manuscript. Specifically, we used an LLM to assist in refining the language, improving readability, and ensuring clarity in various sections of the paper. The model helped with tasks such as sentence rephrasing, grammar checking, and enhancing the overall flow of the text.

It is important to note that the LLM was not involved in the ideation, research methodology, or experimental design. All research concepts, ideas, and analyses were developed and conducted by the authors. The contributions of the LLM were solely focused on improving the linguistic quality of the paper, with no involvement in the scientific content or data analysis.

The authors take full responsibility for the content of the manuscript, including any text generated or polished by the LLM. We have ensured that the LLM-generated text adheres to ethical guidelines and does not contribute to plagiarism or scientific misconduct.

REFERENCES

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. Llava-onevision-1.5: Fully open framework for democratized multimodal training, 2025a. URL https://arxiv.org/abs/2509.23661.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. Llava-onevision-1.5: Fully open framework for democratized multimodal training, 2025b. URL https://arxiv.org/abs/2509.23661.

anon8231489123. ShareGPT: Vicuna unfiltered dataset. https://huggingface.co/datasets/anon8231489123/ShareGPT_Vicuna_unfiltered, 2023. Dataset available on Hugging Face.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang,

Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. BitFit: Simple parameter-efficient fine-tuning for Transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1–9. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.acl-short.1.

Tianle Cai, Yuhong Li, Zhengyang Geng, Haotian Peng, Jason D. Lee, Deming Chen, and Tri Dao. Medusa: Simple LLM-free framework for accelerating LLM generation with multiple decoding heads. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*. PMLR, 2024. URL https://proceedings.mlr.press/v235/cai24b.html.

Charlie Chen, Sebastian Borgeaud, Jean-Baptiste Alayrac, Katie Buchatskaya, Jean-Baptiste Lan, Zhitao Liang, David Agapiou, Eric Huang, Anton Belov, Aurelia Guy, et al. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*, 2023. URL https://arxiv.org/abs/2302.01318.

Aidan Clark, Diego De Las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, et al. Unified scaling laws for routed language models. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 4052–4086. PMLR, 2022.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

Mostafa Elhoushi, Akshat Shrivastava, Diana Liskovich, Basil Hosmer, Bram Wasti, Liangzhen Lai, Anas Mahmoud, Bilge Acun, Saurabh Agarwal, Ahmed Roman, Ahmed A Aly, Beidi Chen, and Carole-Jean Wu. LayerSkip: Enabling early exit inference and self-speculative decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12613–12633, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.681.

William Fedus, Barret Zoph, and Noam Shazeer. Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing video reasoning in MLLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. Mme: A comprehensive evaluation benchmark for multimodal large language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-4899. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/d79a27cf2772fe00be7f341efc0eb517-Paper-Datasets_and_Benchmarks_Track.pdf.

Yichao Fu, Peter Huang, Zeyu Wang, Hua Huang, and Zhihao Jia. Break the sequential dependency of LLM inference: Decoding with lookahead prophecies. *arXiv preprint arXiv:2312.12728*, 2023. URL https://arxiv.org/abs/2312.12728.

Mugilan Ganesan, Shane Segal, Ankur Aggarwal, Nish Sinnadurai, Sean Lie, and Vithursan Thangarasa. Massv: Multimodal adaptation and self-data distillation for speculative decoding of vision-language models, 2025. URL https://arxiv.org/abs/2505.10526.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=0RDcd5Axok.

Zhenick He, Zexuan Zhong, Tianle Cai, Jason Lee, and Jian Tang. REST: Retrieval-based speculative decoding. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1581–1595, Mexico City, Mexico, June 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.naacl-long.88.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. URL https://arxiv.org/abs/2106.09685.

Haiduo Huang, Fuwei Yang, Zhenhua Liu, Xuanwu Yin, Dong Li, Pengju Ren, and Emad Barsoum. SpecVLM: Fast speculative decoding in vision-language models. *arXiv preprint arXiv:2509.11815*, 2025. URL https://arxiv.org/abs/2509.11815.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pp. 19274–19286. PMLR, 2023. URL https://proceedings.mlr.press/v202/leviathan23a.html.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024a. URL https://arxiv.org/abs/2408.03326.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13299–13308. IEEE, 2024b.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22195–22206, 2024c.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. EAGLE-2: Faster inference of language models with dynamic draft trees. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024d.

Yuhui Li, Fangyun Wei, Zhang Chao, and Hongyang Zhang. EAGLE-3: Scaling up inference acceleration of large language models via training-time test. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Luxi Lin, Zhihang Lin, Zhanpeng Zeng, and Rongrong Ji. Speculative decoding reimagined for multimodal large language models. *arXiv preprint arXiv:2505.14260*, 2025. URL https://arxiv.org/abs/2505.14260.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. LLaVA visual instruct pretrain LCS-558k. https://huggingface.co/datasets/liuhaotian/LLaVA-Pretrain, 2023.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2200–2209, 2021.

Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Zeyu Wang, Zhengxin Zhang, Rae Ying Yee Wong, Alan Zhu, Lijie Yang, Xiaoxiang Shi, Chunan Shi, Zhuoming Chen, Daiyaan Arfeen, Reyna Abhyankar, and Zhihao Jia. SpecInfer: Accelerating large language model serving with tree-based speculative inference and verification. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3 (ASPLOS)*, pp. 1225–1243. ACM, 2024. doi: 10.1145/3620666.3651335.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free, 2025. URL https://arxiv.org/abs/2505.06708.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025.

Jun Zhang, Jue Wang, Huan Li, Lidan Shou, Ke Chen, Gang Chen, and Sharad Mehrotra. Draft& verify: Lossless large language model acceleration via self-speculative decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11263–11282. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.607. URL http://dx.doi.org/10.18653/v1/2024.acl-long.607.

Lefan Zhang, Xiaodan Wang, Yanhua Huang, and Ruiwen Xu. Learning harmonized representations for speculative sampling. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025a. URL https://openreview.net/forum?id=T9u56s7mbk.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data, 2025b. URL https://arxiv.org/abs/2410.02713.