MASSV: Multimodal Adaptation and Self-Data Distillation for Speculative Decoding of Vision-Language Models

Anonymous ACL submission

Abstract

Speculative decoding significantly accelerates language model inference by enabling a lightweight draft model to propose multiple tokens that a larger target model verifies simultaneously. However, applying this technique to vision-language models (VLMs) presents two fundamental challenges: small language models that could serve as efficient drafters lack the architectural components to process visual inputs, and their token predictions fail to match those of VLM target models that consider visual context. We introduce **M**ultimodal **A**daptation and **S**elf-Data Distillation for **S**peculative Decoding of **V**ision-Language Models (MASSV), which transforms existing small language models into effective multimodal drafters through a two-phase approach. MASSV first connects the target VLM's vision encoder to the draft model via a lightweight trainable projector, then applies self-distilled visual instruction tuning using responses generated by the target VLM to align token predictions. Comprehensive experiments across the Qwen2.5-VL and Gemma3 model families demonstrate that MASSV increases accepted length by up to 30% and delivers end-to-end inference speedups of **up to 1.46x** compared to conventional text-only drafting baselines on visually-grounded tasks.

1 Introduction

Large language models (LLMs) have transformed artificial intelligence by delivering breakthrough capabilities in reasoning (Jaech et al., 2024; DeepSeek-AI et al., 2025), code generation (Hui et al., 2024; Li et al., 2023), and natural language understanding (OpenAI et al., 2023; Gemini Team et al., 2023; Anthropic et al., 2024; Grattafiori et al., 2024). However, these achievements come with substantial computational costs, particularly during inference. The fundamental constraint arises from autoregressive generation, where each token must be predicted sequentially based on all previous tokens, creating an inherent bottleneck that limits parallelization. Speculative decoding (SD) addresses this bottleneck by leveraging smaller draft models to generate multiple candidate tokens autoregressively, which are then verified in parallel by the larger target model (Chen et al., 2023; Leviathan et al., 2023). This technique reduces sequential operations while preserving the original output distribution, effectively amortizing the computational cost and enabling substantial inference speedups without quality degradation.

While SD has been well-studied for text-only models, extending it to vision-language models (VLMs) introduces unique challenges. VLMs process multimodal inputs by mapping image features and text tokens into a joint embedding space, enabling sophisticated visual reasoning capabilities (Radford et al., 2021; Liu et al., 2023). This multimodal conditioning presents two fundamental challenges for SD: (1) architectural incompatibility, as small language models lack the components to process visual inputs, and (2) distribution mismatch, as unimodal draft models cannot effectively capture the visually-grounded nature of the target VLM's outputs. Previous approaches have addressed these challenges either by excluding image tokens entirely or by training small multimodal models from scratch (Gagrani et al., 2024). The former approach fails to leverage visual information, while the latter requires substantial computational resources and may still suffer from distribution misalignment. Lee et al. (2024) explored ensemble-based methods that combine multiple drafting strategies through batch inference. However, these ensemble approaches do not fundamentally address the distribution mismatch between draft and target models, instead relying on averaging predictions from multiple unaligned drafters. Neither of these approaches fully exploit the potential of existing model families or directly optimize

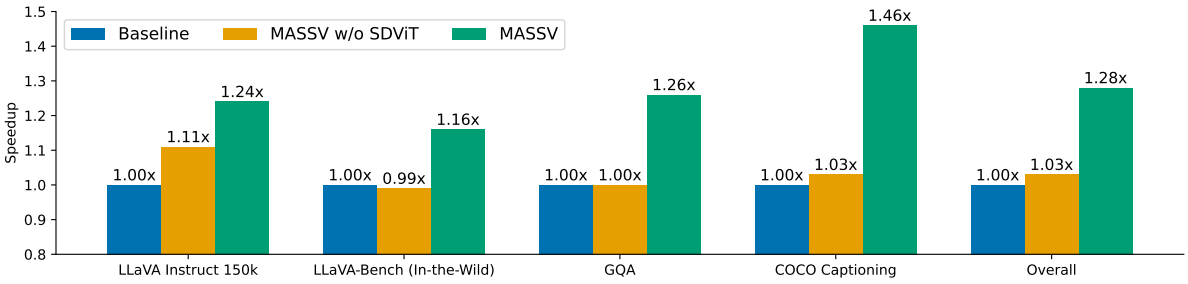


Figure 1: End-to-end wallclock speedups when drafting for Qwen2.5-VL 7B Instruct at temperature $T = 0$ with speculation length $\gamma = 5$. The baseline uses Qwen2.5-1.5B as a text-only drafter (image tokens removed). MASSV consistently yields the highest speedups across all categories, achieving up to 1.46× on COCO captioning and 1.28× overall. The gains are most pronounced for visually grounded tasks, demonstrating the importance of multimodal adaptation and self-distilled visual instruction for accelerating VLM inference.

for the distribution alignment needed for SD.

We introduce **M**ultimodal **A**daptation and **S**elf-Data Distillation for **S**peculative Decoding of **V**ision- Language Models (MASSV), a principled method for adapting smaller language models from the same family as the target VLM to serve as efficient multimodal draft models. Our approach consists of two key components. First, we formulate the multimodal drafting problem as mapping from a target VLM's vision-language embedding space to a draft LM's embedding space, constructing a drafter by connecting the target VLM's vision encoder and multimodal projector to a smaller language model from the same family. Second, we propose a training methodology centered on self-data distillation (Thangarasa et al., 2025; Yang et al., 2024) to align the draft model's distribution with the target model's, specifically optimizing for higher token acceptance rates during SD. As shown in Figure 1, MASSV achieves significant end-to-end speedups, particularly on visually grounded tasks, demonstrating the importance of multimodal adaptation and self-data distillation for improving acceptance rate of draft tokens. Our contributions are as follows:

- We propose MASSV, a comprehensive framework that combines (1) a architectural adaptation connecting target VLM components with smaller language models from the same family, and (2) a self-data distillation technique specifically designed to align multimodal distributions for improved token acceptance.
- We provide extensive empirical evaluations demonstrating significant improvements in acceptance rates across multiple model families, with speedups reaching up to 1.28x overall on multimodal tasks.
- We present detailed ablations revealing that self-data distillation is crucial for multimodal drafting, improving distribution alignment between draft and target models particularly for visually-grounded tasks.

2 Preliminaries

We establish the necessary background for our approach. First, we review SD, an inference acceleration technique that uses a smaller draft model to propose tokens that are verified by a larger target model. Second, we describe VLMs, which combine visual encoders with language models to process multimodal inputs. Finally, we discuss how SD has been adapted for VLMs, including the text-only drafting baseline we compare against.

Speculative decoding is a technique for accelerating LLM generation without altering the distribution of the generation output (Leviathan et al., 2023; Chen et al., 2023). In each iteration of the algorithm, a draft model M_q generates multiple draft tokens that are verified in parallel by the target model M_p. The algorithm continues iterating until an end-of-sequence (EOS) token is generated or the max sequence length is reached. Formally, let $X_{1:t} = X_1, X_2, ..., X_t$ be the input sequence for the current iteration. M_q first autoregressively samples γ draft tokens $X_{t+1:t+\gamma}$, where token X_{t+i} is sampled with probability $q(X_{t+i}|X_{1:t+i-1})$. Next, M_p computes the probabilities $p(X_{t+i}|X_{1:t+i-1})$ for $i = 1, 2, ..., \gamma + 1$ in parallel with one forward call. These probabilities are used to evaluate the draft tokens sequentially, with the probability of accepting token X_{t+i} being $\min\left(1, \frac{p(X_{t+i}|X_{1:t+i-1})}{q(X_{t+i}|X_{1:t+i-1})}\right)$. If the token is accepted, it is added to the generation output and the next token is evaluated. Otherwise, if the token is rejected,

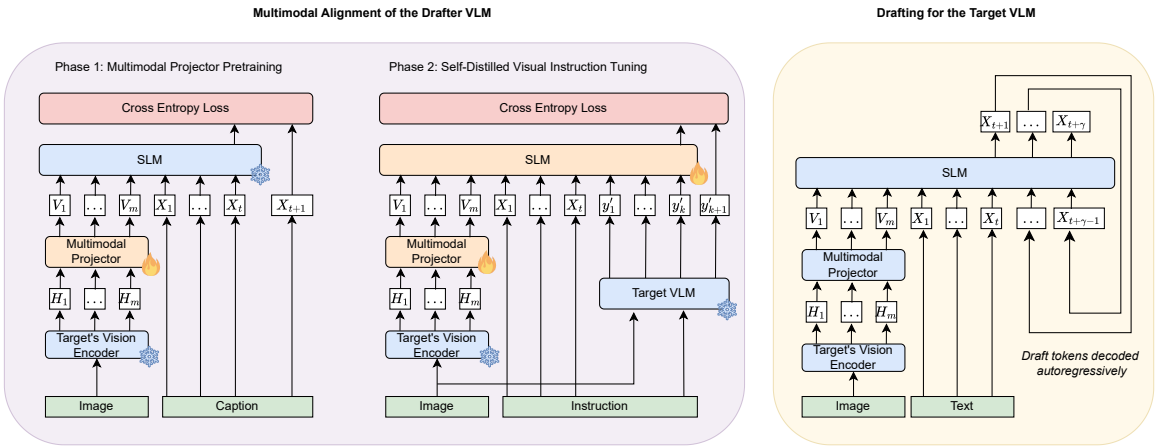


Figure 2: Detailed architecture of MASSV illustrating: (1) the two-phase training methodology consisting of multimodal projector pretraining followed by self-distilled visual instruction tuning, and (2) the deployment configuration for draft token generation during speculative decoding inference. Components marked with the snowflake remain frozen during training to preserve their parameters, while components with the flame are trainable. This architecture enables efficient knowledge transfer from the target vision-language model to the smaller draft model while maintaining alignment in their token distributions.

a new token is sampled from the residual distribution $\text{norm}(\max(p(\cdot|X_{1:t+i-1}) - q(\cdot|X_{1:t+i-1}), 0))$ and the iteration ends. Sampling from the residual distribution ensures the output distribution of the speculative decoding algorithm is the same as the target's output distribution. In the degenerate case where sampling is disabled (temperature = 0), the algorithm simplifies to greedy decoding. The draft model generates tokens by selecting $X_{t+i} = \arg\max_x q(x|X_{1:t+i-1})$. During verification, token X_{t+i} is accepted if and only if $X_{t+i} = \arg\max_x p(x|X_{1:t+i-1})$. If rejected, the token is set to $\arg\max_x p(x|X_{1:t+i-1})$.

Vision-language models (VLMs) process multimodal inputs, consisting of visual and text tokens, by mapping the tokens into a joint embedding space. A VLM consists of three components: a vision encoder ϕ_I, multimodal projector g_θ, and a language model M_p. Given an input consisting of tokens $X_{1:t}$ and visual information I, a VLM first extracts m features $H_{1:m} = \phi_I(I)$ from the image using the vision encoder. These image features are then projected into the joint embeddings space $V_i = g_\theta(H_i)$ for $i \in \{1, ..., m\}$. Finally, the VLM samples the next token X_{t+1} from $p(\cdot|X_{1:t}, V_{1:m})$, where $p(\cdot|\cdot)$ denotes the conditional probability distribution of M_p. Note that directly using SD to accelerate a VLM on multimodal inputs requires the drafter to also be a VLM. However, Gagrani et al. (2024) show that a small language model (SLM) can be used as an effective drafter by conditioning it only on the text tokens in the input. Concretely, given an SLM drafter M_q, the draft token X_{t+i} is sampled from $q(\cdot|X_{1:t+i-1})$ for $i = 1, ..., \gamma$. We refer to this as *text-only drafting* and use it as the baseline in our experiments.

3 Methodology

We introduce a method to adapt an SLM into an effective draft model for LLaVA-style VLMs, which employ a modular architecture of separate vision encoder and language model components connected via a projection layer that maps image features into the language model's embedding space. Our approach integrates the target VLM's frozen vision encoder into the SLM through a randomly initialized MLP-based projector, preserving architectural compatibility while enabling visual processing. We then align the adapted model with the target VLM through a two-phase training protocol: (1) the projector is pretrained on paired image-text data to establish visual grounding; and (2) the model undergoes self-distilled visual instruction tuning to optimize token-level distribution alignment. The overall architecture is shown in Figure 2.

3.1 Architectural Adaptation

Let $M_p^{\text{VLM}} = (\phi_I^p, g_\theta^p, M_p)$ denote the target VLM, where ϕ_I^p is the vision encoder, g_θ^p is the multimodal projector, and M_p is the language model. Let M_q be an SLM from the same model family as M_p. While our method can be applied to any small language model, this work specifically focuses on text-only SLMs from the same model family as

the larger VLM. This choice ensures that the draft model's tokenizer and vocabulary are compatible with those of the target during SD. Although recent work has demonstrated approaches to handle heterogeneous vocabularies (Timor et al., 2025), these techniques trade latency for vocabulary compatibility. Furthermore, existing methods have not demonstrated their effectiveness in handling multiple modalities, as required for VLMs. Due to these limitations and considerations beyond the scope of this work, we leave exploring vocabulary heterogeneity in multimodal SD for future research.

We construct the VLM drafter M_q^{VLM} as follows, $M_q^{\text{VLM}} = (\phi_I^p, g_\psi^q, M_q)$, where ϕ_I^p is the shared vision encoder from the target VLM, g_ψ^q is a randomly initialized multimodal projector, and M_q is the draft SLM. The projector g_ψ^q has the same architecture as g_θ^p, but its output dimension d_{out}^q is set to match the embedding dimension of M_q, $g_\psi^q : \mathbb{R}^{d_{\text{vis}}} \to \mathbb{R}^{d_{\text{emb}}^q}$ where d_{vis} is the vision encoder's output dimension and d_{emb}^q is the embedding dimension of M_q. We choose to share the vision encoder between the target and the drafter, since this ensures that the drafter and target process the same visual features $H_{1:m} = \phi_I^p(I)$ for a given image input I. This architectural choice also reduces compute cost by avoiding redundant vision encoding operations.

3.2 Multimodal Alignment

Multimodal Projector Pretraining Following Liu et al. (2024b), we first pretrain the multimodal projector g_ψ^q by training the VLM drafter with the vision encoder and SLM backbone frozen. Given a pretraining dataset $D_{\text{pre}} = \{(I_j, C_j)\}_{j=1}^N$ of image-caption pairs, we optimize,

$$\mathcal{L}_{\text{pre}}(\psi) = -\sum_{j=1}^{N}\sum_{i=1}^{|C_j|} \log q_\psi(c_j^i | c_j^{1:i-1}, V_j), \quad (1)$$

where $V_j = g_\psi^q(\phi_I^p(I_j))$ are the projected visual features, c_j^i is the i-th token of caption C_j, and q_ψ denotes the distribution of the draft VLM with projector parameters ψ. Only ψ is updated during this phase while ϕ_I^p and M_q remain frozen.

Self-Distilled Visual Instruction Tuning (SDViT) In this phase, we introduce SDViT, an approach that employs SDD to align the drafter's distribution with the target's multimodal distribution. Let $D = \{(I_i, X_i, y_i)\}_{i=1}^n$ be a visual instruction dataset, where I_i is the image input, X_i is the text instruction, and y_i is the reference response. The original SDD formulation by Thangarasa et al. (2025); Yang et al. (2024) generates target outputs using task-specific contexts and templates. In contrast, for SD, our objective is to align the drafter's token-level predictions with the target's. Therefore, we directly use the target VLM to generate responses, $y_i' = \text{sample}_{\text{top-p}}(p(\cdot|I_i, X_i))$, where p denotes the target VLM's distribution conditioned on both image I_i and text instruction X_i. This creates a self-distilled dataset $D' = \{(I_i, X_i, y_i')\}_{i=1}^n$. We then fine-tune the drafter with its vision encoder frozen to minimize,

$$\mathcal{L}_{\text{SDViT}}(\theta) = -\sum_{i=1}^{n}\sum_{k=1}^{|y_i'|} \log q_\theta(y_i'^k | y_i'^{1:k-1}, X_i, V_i), \quad (2)$$

where $V_i = g_\psi^q(\phi_I^p(I_i))$ are the projected visual features, $y_i'^k$ is the k-th token of the target's response, and q_θ denotes the drafter's distribution with parameters $\theta = \{\psi, \theta_q\}$ (projector and SLM parameters). In contrast to generic visual instruction tuning with fixed dataset labels, our self-distillation strategy trains the drafter on the target's actual outputs, directly optimizing for the acceptance mechanism in SD. SDViT addresses this through diverse sampling, where the target VLM generates responses across different temperature values with top-p sampling, creating a varied dataset that better represents the full response distribution. Specifically, draft tokens are accepted with probability $\min\left(1, \frac{p(X_t|X_{1:t}, I)}{q(X_t|X_{1:t}, I)}\right)$. By training on the target's outputs rather than generic labels, we maximize the overlap between the drafter's distribution q and the target's distribution p, leading to higher token acceptance rates during inference. Our results in Section 4.2 show that this alignment translates to improved token acceptance rates during SD.

4 Empirical Results

4.1 Experimental Setup

Draft and Target Models. Our evaluation leverages two distinct model families: the Qwen2.5-VL Instruct (Bai et al., 2025) and instruction-tuned Gemma3 (Gemma Team et al., 2025). Specifically, for Qwen2.5-VL, we set the 7B model as our primary target, applying MASSV to Qwen2.5-1.5B Instruct. Similarly, for Gemma3, we target the 12B IT variant and adapt Gemma3-1B IT using MASSV. We selected these specific SLMs because they are from the same model families as the larger target models and were readily available as checkpoints

Target Model	Method	LLaVa 150k	LLaVA-Bench	GQA	COCO	Overall
Temperature = 0						
Qwen2.5-VL 7B Instruct	Baseline	2.37 (1.00x)	2.61 (1.00x)	2.59 (1.00x)	2.21 (1.00x)	2.46 (1.00x)
	MASSV	3.21 (1.24x)	3.12 (1.16x)	3.28 (1.26x)	3.26 (1.46x)	**3.20**$_{\uparrow 0.74}$ (**1.28x**)
Qwen2.5-VL 32B Instruct	Baseline	2.46 (1.00x)	2.70 (1.00x)	2.79 (1.00x)	2.48 (1.00x)	2.61 (1.00x)
	MASSV	3.12 (1.26x)	2.90 (1.07x)	3.19 (1.13x)	3.09 (1.23x)	**3.04**$_{\uparrow 0.43}$ (**1.17x**)
Gemma3-12B IT	Baseline	2.71 (1.00x)	2.72 (1.00x)	2.75 (1.00x)	2.84 (1.00x)	2.76 (1.00x)
	MASSV	3.30 (1.19x)	3.00 (1.11x)	3.07 (1.18x)	3.41 (1.24x)	**3.19**$_{\uparrow 0.43}$ (**1.18x**)
Gemma3-27B IT	Baseline	2.49 (1.00x)	2.70 (1.00x)	2.61 (1.00x)	2.73 (1.00x)	2.65 (1.00x)
	MASSV	3.00 (1.20x)	2.84 (1.05x)	2.86 (1.09x)	3.24 (1.20x)	**2.99**$_{\uparrow 0.34}$ (**1.14x**)
Temperature = 1						
Qwen2.5-VL 7B Instruct	Baseline	2.47 (1.00x)	2.75 (1.00x)	2.63 (1.00x)	2.41 (1.00x)	2.58 (1.00x)
	MASSV	3.35 (1.26x)	2.98 (1.09x)	3.19 (1.19x)	3.31 (1.35x)	**3.18**$_{\uparrow 0.60}$ (**1.22x**)
Qwen2.5-VL 32B Instruct	Baseline	2.48 (1.00x)	2.69 (1.00x)	2.75 (1.00x)	2.56 (1.00x)	2.63 (1.00x)
	MASSV	3.01 (1.25x)	2.87 (1.09x)	3.00 (1.09x)	3.04 (1.19x)	**2.97**$_{\uparrow 0.34}$ (**1.15x**)
Gemma3-12B IT	Baseline	2.67 (1.00x)	2.79 (1.00x)	2.78 (1.00x)	2.94 (1.00x)	2.82 (1.00x)
	MASSV	3.08 (1.13x)	2.82 (1.05x)	3.01 (1.10x)	3.37 (1.16x)	**3.06**$_{\uparrow 0.24}$ (**1.11x**)
Gemma3-27B IT	Baseline	2.57 (1.00x)	2.67 (1.00x)	2.63 (1.00x)	2.73 (1.00x)	2.67 (1.00x)
	MASSV	2.81 (1.09x)	2.62 (1.02x)	2.82 (1.07x)	3.13 (1.15x)	**2.84**$_{\uparrow 0.17}$ (**1.08x**)

Table 1: Mean accepted lengths (τ) and speedups across model families, tasks, and temperatures ($T \in \{0, 1\}$) with speculation length $\gamma = 5$. Values show tokens accepted per target VLM forward pass, with speedup ratios in parentheses (normalized to baseline). MASSV consistently outperforms the text-only baseline (Gagrani et al., 2024), achieving substantial gains on visually-grounded tasks like COCO captioning (+47.5% at $T = 0$: $2.21 \rightarrow 3.26$) and improving overall acceptance (+30.1% for Qwen2.5-VL 7B: $2.46 \rightarrow 3.20$). MASSV delivers practical efficiency with $1.28\times$ end-to-end speedup for Qwen2.5-VL 7B at $T = 0$ and generalizes effectively to larger models without requiring direct alignment.

on HuggingFace. We utilize *text-only drafting* with the off-the-shelf SLM as our baseline (1.00x).

Drafter Training for Multimodal Adaptation. The draft model training process consists of two distinct phases and requires only moderate compute infrastructure, achievable with standard research hardware (e.g., four-GPU server with current-generation accelerators). Initially, we pretrain each drafter for one epoch on the LLaVA-Pretrain-LCS-558K [1] dataset, using a global batch size of 256 and a learning rate of 1 x 10^{-4}. Subsequently, we fine-tune the models on data distilled from the LLaVA-mix-665K [2] dataset for another epoch with a batch size of 128 and learning rate 2 x 10^{-5}. See Appendix A for more details.

Evaluation Tasks. We conduct evaluations using four multimodal benchmarks: LLaVA Instruct 150k (Liu et al., 2023), LLaVA-Bench (In-the-Wild) [3], GQA (Hudson and Manning, 2019), and image captioning prompts from COCO Test 2017 (Lin et al., 2015). Performance is measured by mean accepted length (τ), which quantifies the average number of tokens accepted per forward pass of the target model, directly correlating to speedup independent of hardware. Evaluation settings and prompts for GQA reasoning and COCO Captioning tasks are provided in Appendix B.

4.2 Results

Our results demonstrate MASSV's significant improvements over the text-only baseline across all evaluated settings (see Table 1). At temperature $T = 0$, MASSV achieves a noticeable increase in mean accepted length (MAL), most notably improving by 30.1% (from 2.46 to 3.20) for the Qwen2.5-VL 7B Instruct model. Similarly, at $T = 1$, MASSV attains a MAL improvement of 23.3% (from 2.58 to 3.18). These improvements are consistent across different downstream tasks, with the largest relative gains observed in visually intensive tasks such as COCO captioning. For instance, MASSV increases MAL by 47.5% (2.21 to 3.26) on COCO captioning tasks at $T = 0$, highlighting the importance of multimodal drafting for visually-grounded generations. Moreover, MASSV consistently outperforms the baseline on the Gemma3 family despite their significant architectural differences (e.g., dynamic visual token count in Qwen2.5-VL versus interleaved sliding window attention in Gemma3). Specifically, MASSV improves MAL by 15.6% (2.76 to 3.19) on Gemma3-12B IT at $T = 0$, demonstrating its

[1] https://huggingface.co/datasets/liuhaotian/LLaVA-Pretrain

[2] https://huggingface.co/datasets/liuhaotian/LLaVA-Instruct-150K/blob/main/llava_v1_5_mix665k.json

[3] https://huggingface.co/datasets/liuhaotian/llava-bench-in-the-wild

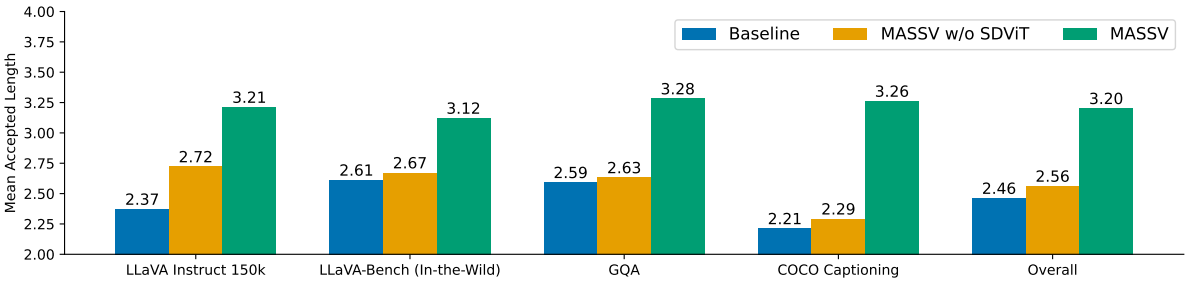


Figure 3: Mean accepted lengths when drafting for Qwen2.5-VL 7B Instruct at temperature $T = 0$ with speculation length $\gamma = 5$. The baseline uses Qwen2.5-1.5B as a text-only drafter (image tokens removed). MASSV achieves a substantial improvement in token acceptance across all tasks, increasing overall mean accepted length from 2.46 to 3.20 (+30.1%).

effectiveness across diverse VLMs.

Generalization to Larger Model Variants. We also evaluated MASSV on larger variants within each model family, specifically Qwen2.5-VL 32B and Gemma3-27B. Although we did not directly apply SDViT to these larger targets, we hypothesized that MASSV, when applied to smaller distilled versions (7B and 12B), could still benefit their larger counterparts due to their shared architecture and distillation lineage. Our empirical results confirm this hypothesis, demonstrating that MASSV provides meaningful gains even when scaling up within the same model family. This finding is particularly impactful as it allows substantial computational and time savings by enabling MASSV adaptation on smaller, more efficient targets, which can subsequently generalize to larger models.

End-to-end Inference Speedups. The mean accepted length improvements translate directly to substantial wall-clock speedups during inference. MASSV achieves an overall end-to-end speedup of 1.28× for Qwen2.5-VL 7B Instruct at temperature $T = 0$, with even higher speedups on specific tasks such as COCO captioning (1.46×). These speedups remain consistent across model families, with Gemma3-12B IT achieving 1.18× acceleration. Notably, MASSV demonstrates effective scalability to larger models, achieving 1.17× speedup for Qwen2.5-VL 32B and 1.14× for Gemma3-27B, despite not requiring direct alignment on these larger targets. These results show that MASSV's improved token acceptance rates translate to meaningful practical efficiency gains.

5 Ablation Studies

We investigate the impact of self-distilled visual instruction tuning on distribution alignment, and we examine whether multimodal capability provides meaningful benefits over text-only drafting.

5.1 Effect of Self-Distilled Visual Instruction Tuning

We assess the role of self-distilled distillation in our method by comparing drafters trained with SDViT versus standard fine-tuning on a vanilla dataset. Specifically, we adapt Qwen2.5-1.5B Instruct and Gemma3-1B IT into drafters for Qwen2.5-VL 7B Instruct and Gemma3-12B IT, respectively. Figure 3 demonstrates the efficacy of MASSV with SDViT (green bar) for Qwen2.5-VL 7B Instruct across diverse multimodal benchmarks. MASSV exhibits substantial performance gains, most prominently in COCO Captioning where the mean accepted length increases from 2.21 to 3.26 tokens (+47.5%). Table 2 summarizes our comprehensive ablation study on SDViT across both target models: Qwen2.5-VL 7B Instruct and Gemma3-12B IT. The quantitative evaluation results demonstrate the critical importance of self-distilled visual instruction tuning for effective multimodal SD. For the Gemma3 architecture, without SDViT (denoted as $\text{MASSV}_{\text{w/o SDViT}}$), the Gemma3-1B IT draft model exhibits a significant performance regression, with mean accepted length deteriorating to 2.33 compared to the baseline's 2.74 (a 13% decrease in acceptance rate). This indicates that naive architectural adaptation without distribution alignment can be notably detrimental to performance. In contrast, when enhanced with SDViT, the model achieves a mean accepted length of 3.14, representing a substantial 14.6% improvement over the baseline and a 1.18x speedup. These results highlight the critical role of distribution alignment in multimodal SD.

Distribution Analysis. To understand the mechanism behind these improvements, we analyze the distribution alignment between drafters and targets. For each multimodal input, we compute

Target Model	Method	τ	Speedup
Qwen2.5-VL 7B Instruct	Baseline	2.46	1.00x
	$\text{MASSV}_{\text{w/o SDViT}}$	2.56	1.04x
	MASSV	3.20	1.28x
Gemma3-12B IT	Baseline	2.74	1.00x
	$\text{MASSV}_{\text{w/o SDViT}}$	2.33	0.87x
	MASSV	3.14	1.18x

Table 2: Ablation results on the effect of SDViT on drafting performance. Qwen2.5-1.5B Instruct and Gemma3-1B IT are the base SLMs used to create drafters for Qwen2.5-VL 7B Instruct and Gemma3-12B IT, respectively. The reported mean accepted lengths (τ) are measured on the overall multimodal speculative decoding benchmark dataset at temperature = 0.

the Total Variation Distance (TVD) between the drafter's and target's output token distributions. The TVD measures the maximum difference between two probability distributions: $\text{TVD}(P, Q) = \frac{1}{2}\sum_{x\in\mathcal{X}} |P(x) - Q(x)|$, where P and Q are the target and drafter token distributions, respectively, and $\mathcal{X}$ is the vocabulary. TVD is particularly relevant in the context of SD, as it bounds the expected probability that tokens proposed by the draft model will be rejected by the target model. By minimizing TVD through our SDViT approach, we directly optimize for higher token acceptance rates, which explains the improved mean accepted length observed in our experiments. For discrete distributions like token probabilities, TVD ranges from 0 (identical distributions) to 1 (completely disjoint distributions). Figure 4 shows the resulting distribution. The drafter trained with SDViT produces significantly more tokens with output distributions closely matching the target. This demonstrates that SDViT enables the drafter to more faithfully reproduce the target model's token-level behavior. These results indicate that: (1) SDD substantially improves distribution alignment between drafter and target, and (2) distribution alignment contributes more to drafting performance than raw multimodal capability.

5.2 Text-Only vs Multimodal Drafting

Given that distribution alignment appears more important than multimodal capability, we investigate whether multimodal processing provides any benefit over text-only drafting. This question is particularly relevant since text-only drafting could offer computational advantages by avoiding visual encoding operations during the draft phase.

We evaluate our VLM drafters in text-only mode by discarding visual tokens from the input, thereby using only the language model backbone of our

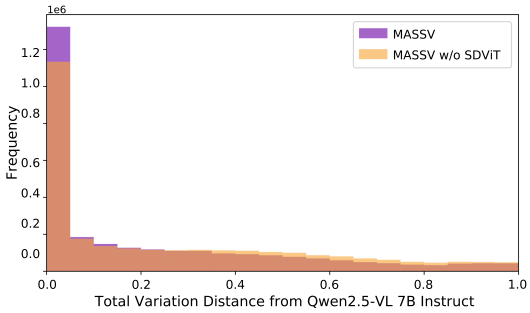


Figure 4: Histogram of total variation distances (TVD), comparing the Qwen2.5-1.5B drafters trained with (purple) and without (orange) self-distilled visional instruction (SDViT) against the Qwen2.5-VL 7B target model on our multimodal SD benchmark. MASSV yields a highly skewed distribution concentrated at low TVD values, indicating tighter alignment with the target's token distribution. In contrast, $\text{MASSV}_{\text{w/o SDViT}}$ produces a broader, heavier-tailed distribution, reflecting reduced alignment. The left-skewed shape of the MASSV distribution quantitatively suggests that SDViT narrows the distributional gap between draft and target.

Target Model	Method	τ
Qwen2.5-VL 7B Instruct	Text-Only	2.84
	Multimodal	3.20
Gemma3-12B IT	Text-Only	2.99
	Multimodal	3.19

Table 3: Ablation results on the performance of text-only drafting. The VLM drafter's language model backbone serves as a text-only drafter by discarding all visual tokens. Mean accepted lengths (τ) are measured on the overall benchmark dataset at temperature = 0.

adapted drafter. This approach mirrors the baseline strategy used in prior work (Gagrani et al., 2024), where standard SLMs trained from scratch serve as drafters for VLM targets without processing any visual information. Table 3 shows that multimodal drafting consistently outperforms text-only drafting across both model families. The improvements are substantial: 12.7% higher mean accepted length for Qwen2.5-VL (3.20 vs. 2.84) and 6.7% higher for Gemma3 (3.19 vs. 2.99). These gains demonstrate that while distribution alignment is the primary factor in drafting performance, incorporating visual information provides additional benefits for predicting the target VLM's outputs. The advantage of multimodal drafting likely stems from its ability to condition token predictions on the actual visual content, particularly for visually-grounded tokens such as object names, spatial relationships, and visual attributes. While text-only drafting must rely solely on linguistic patterns and context, multimodal drafting can leverage direct visual evidence to better

predict the target VLM's outputs. Based on these observations, we focus exclusively on multimodal drafting in our main experiments (Section 4). This choice ensures we capture the full benefits of visual information while maintaining strong distribution alignment through SDViT. As we demonstrate across multiple model families and tasks, this combination of multimodal capability and distribution alignment yields consistent improvements in SD performance.

6 Related Work

Speculative decoding has emerged as a promising technique for accelerating LLM inference without compromising output quality. This approach leverages smaller, faster draft models to autoregressively generate multiple candidate tokens, which are then verified in parallel by the larger target model in a single forward pass (Leviathan et al., 2023; Chen et al., 2023). The theoretical foundations of this technique were established by identifying conditions under which speculative proposals can preserve the original model's output distribution (Leviathan et al., 2023). Recent advancements include tree-structured variants (Li et al., 2024b,a; Wang et al., 2025; Chen et al., 2024), self-drafting (Elhoushi et al., 2024; Zhang et al., 2024; Liu et al., 2024a; Xia et al., 2025), N-gram-based (Stewart et al., 2024; Ou et al., 2024) and retrieval-based (He et al., 2024; Yang et al., 2023) that further enhance inference efficiency. However, these approaches have primarily focused on text-only models, where the draft and target operate within the same modality space.

Multimodal Speculative Decoding. Extending speculative decoding to vision-language models introduces fundamental challenges absent in unimodal settings. Gagrani et al. (2024) conducted initial explorations in this domain by evaluating several draft model variants with the LLaVA-7B architecture (Liu et al., 2024b). Their analysis across image question-answering, captioning, and reasoning tasks revealed modest token acceptance rates, with the multimodal variant achieving only marginal improvements over text-only counterparts. Detailed traces demonstrated that while drafters successfully predicted function words and repeated tokens, they struggled with visually-grounded content, highlighting two fundamental challenges: (1) architectural misalignment between drafters and vision-language targets, and (2) distributional divergence between text-only priors and visually-informed outputs. Lee et al. (2024) introduced a batch-based approach that combines predictions from multiple drafting methods to increase the likelihood of token acceptance. While their ensemble technique improves empirical performance without parameter overhead, it operates primarily as a post-hoc aggregation mechanism rather than addressing the underlying distributional divergence between individual drafters and the target model. Our MASSV framework directly addresses these limitations through principled vision-language alignment techniques.

Draft Model Alignment. Self-distillation uses a model's own outputs as training targets, extending traditional knowledge distillation approaches. While Yang et al. (2024) showed self-distillation can bridge distribution gaps during language model fine-tuning and Thangarasa et al. (2025) demonstrated its effectiveness in mitigating catastrophic forgetting in pruned models, we extend these insights to multimodal drafting. In particular, SD^2 (Lasby et al., 2025) apply SDD to fine-grained sparse draft models, aligning them closely with their original dense counterparts and yielding substantially higher mean accepted lengths than undistilled sparse drafters. Unlike previous work, we explicitly formulate self-distillation as an optimization for token acceptance probability in the speculative decoding framework. By training our draft model on responses generated by the target VLM itself rather than fixed dataset labels, we align the draft model's distribution with that of the target's.

7 Conclusion

In this work, we present MASSV, a method to transform SLMs into highly efficient speculative drafters for VLMs. MASSV addresses challenges like architectural incompatibility and distribution mismatch by grafting the frozen vision encoder of the target VLM onto the draft model via a trainable projector and aligning the drafter's token distribution through fine-tuning on self-generated vision-language data. Across both Qwen2.5-VL and Gemma3 model families, MASSV increases mean accepted length by 16-30% with end-to-end inference speedups of up to 1.46x. Ablation studies show that SDD is critical for distribution alignment, and full multimodal drafting consistently outperforms text-only approaches. Given its generalizability and demonstrated performance gains, MASSV presents a readily deployable solution for significantly accelerating VLM inference across diverse architectures and tasks.

Limitations

While this work establishes a comprehensive framework for constructing a drafter VLM using an SLM from the same family as the target VLM, there is scope for exploring the use of SLMs that come from a different model family. We chose to focus on using SLMs from the same model family as the target, since this ensures that the draft model's tokenizer and vocabulary are compatible with those of the target during speculative decoding. Overcoming this limitation would allow producing efficient multimodal draft models for a wider range of multimodal target models. Another limitation of our method is that it is applicable specifically for the VLM architecture. While we chose to focus on this architecture due to its widespread use, there is scope for exploring the construction of multimodal drafters for multimodal targets that have different architectures.

References

Anthropic and 1 others. 2024. Introducing the next generation of claude. https://www.anthropic.com/news/claude-3-family.

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, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv*.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. 2023. Accelerating large language model decoding with speculative sampling. *arXiv*.

Zhuoming Chen, Avner May, Ruslan Svirschevski, Yu-Hsun Huang, Max Ryabinin, Zhihao Jia, and Beidi Chen. 2024. Sequoia: Scalable and robust speculative decoding. In *NeurIPS*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv*.

Mostafa Elhoushi, Akshat Shrivastava, Diana Liskovich, Basil Hosmer, Bram Wasti, Liangzhen Lai, Anas Mahmoud, Bilge Acun, Saurabh Agarwal, Ahmed Roman, Ahmed Aly, Beidi Chen, and Carole-Jean Wu. 2024. LayerSkip: Enabling early exit inference and self-speculative decoding. In *ACL*.

Mukul Gagrani, Raghavv Goel, Wonseok Jeon, Junyoung Park, Mingu Lee, and Christopher Lott. 2024. On speculative decoding for multimodal large language models. *arXiv*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, and 1 others. 2023. Gemini: A family of highly capable multimodal models. *arXiv*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *arXiv*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, and 1 others. 2024. The llama 3 herd of models.

Zhenyu He, Zexuan Zhong, Tianle Cai, Jason Lee, and Di He. 2024. REST: Retrieval-based speculative decoding. In *ACL*.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, Kai Dang, Yang Fan, Yichang Zhang, An Yang, Rui Men, Fei Huang, Bo Zheng, Yibo Miao, Shanghaoran Quan, and 5 others. 2024. Qwen2.5-coder technical report.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Mike Lasby, Nish Sinnadurai, Valavan Manohararajah, Sean Lie, and Vithursan Thangarasa. 2025. Sd^2: Self-distilled sparse drafters. *arXiv*.

Minjae Lee, Wonjun Kang, Minghao Yan, Christian Classen, Hyung Il Koo, and Kangwook Lee. 2024. In-batch ensemble drafting: Toward fast and robust speculative decoding for multimodal language models. *OpenReview*.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. In *ICML*.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo,

Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, and 48 others. 2023. Starcoder: may the source be with you! *arXiv*.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. 2024a. Eagle-2: Faster inference of language models with dynamic draft trees. *arXiv*.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. 2024b. Eagle: Speculative sampling requires rethinking feature uncertainty. *arXiv*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *arXiv*.

Fangcheng Liu, Yehui Tang, Zhenhua Liu, Yunsheng Ni, Duyu Tang, Kai Han, and Yunhe Wang. 2024a. Kangaroo: Lossless self-speculative decoding for accelerating LLMs via double early exiting. In *NeurIPS*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *CVPR*.

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

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jie Ou, Yueming Chen, and Wenhong Tian. 2024. Lossless acceleration of large language model via adaptive n-gram parallel decoding. *arXiv*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *PMLR*.

Lawrence Stewart, Matthew Trager, Sujan Kumar Gonugondla, and Stefano Soatto. 2024. The n-grammys: Accelerating autoregressive inference with learning-free batched speculation. *arXiv*.

Vithursan Thangarasa, Ganesh Venkatesh, Mike Lasby, Nish Sinnadurai, and Sean Lie. 2025. Self-data distillation for recovering quality in pruned large language models. In *MLSys*.

Nadav Timor, Jonathan Mamou, Daniel Korat, Moshe Berchansky, Oren Pereg, Gaurav Jain, Roy Schwartz, Moshe Wasserblat, and David Harel. 2025. Accelerating llm inference with lossless speculative decoding algorithms for heterogeneous vocabularies. *arXiv*.

Jikai Wang, Yi Su, Juntao Li, Qingrong Xia, Zi Ye, Xinyu Duan, Zhefeng Wang, and Min Zhang. 2025. Opt-tree: Speculative decoding with adaptive draft tree structure. *arXiv*.

Heming Xia, Yongqi Li, Jun Zhang, Cunxiao Du, and Wenjie Li. 2025. SWIFT: On-the-fly self-speculative decoding for LLM inference acceleration. In *ICLR*.

Nan Yang, Tao Ge, Liang Wang, Binxing Jiao, Daxin Jiang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Inference with reference: Lossless acceleration of large language models. *arXiv*.

Zhaorui Yang, Qian Liu, Tianyu Pang, Han Wang, H. Feng, Minfeng Zhu, and Wei Chen. 2024. Self-distillation bridges distribution gap in language model fine-tuning. In *ACL*.

Jun Zhang, Jue Wang, Huan Li, Lidan Shou, Ke Chen, Gang Chen, and Sharad Mehrotra. 2024. Draft & verify: Lossless large language model acceleration via self-speculative decoding. In *ACL*.

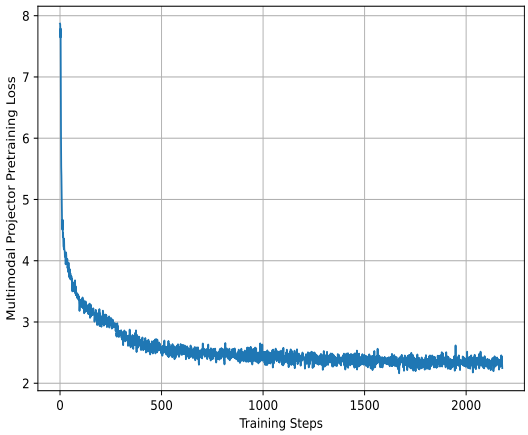
A Additional Training Details

The training curves presented in Figure 5 illustrate the convergence patterns for both phases of the MASSV methodology described in Section 3. In Phase 1 (Multimodal Alignment), the multimodal projector pretraining loss exhibits rapid convergence within the first 500 steps, starting from approximately 8.0 and stabilizing around 2.5 by step 2000. This demonstrates effective knowledge transfer from the target VLM's vision encoder to the draft model via the trainable projector. Phase 2 (Self-Distilled Visual Instruction Tuning) shows a more gradual optimization process with the loss starting at approximately 2.6 and stabilizing around 1.1 with minor fluctuations across 5000 training steps. These training dynamics align with our experimental setup where each drafter was first pretrained for one epoch on the LLaVA-Pretrain-LCS-558K dataset (batch size 256, learning rate 10^{-3}), followed by fine-tuning on data distilled from LLaVA-mix-665K (batch size 128, learning rate 2×10^{-5}) using the target VLM. The convergence patterns show successful training of both the multimodal projector and subsequent distribution alignment through self-distilled visual instruction tuning.

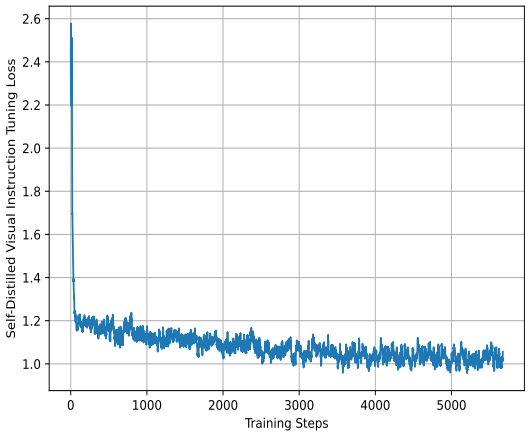
B Additional Evaluation Details

Inference Settings. During inference, all drafters run on a single H100 GPU, with speculation length set to $\gamma = 5$. We evaluate performance at sampling temperatures $T \in \{0, 1\}$.

Prompt Templates. The following prompt templates were used during the evaluations described in Section 4.1. The GQA prompt explicitly requests reasoning explanations alongside answers, evaluating the model's visual reasoning capabilities. The COCO Captioning prompt elicits detailed image descriptions without stylistic constraints. These standardized prompts ensure consistent evaluation across all model variants (baseline, MASSV without SDViT, and full MASSV), enabling fair comparison of mean accepted length and end-to-end speedup metrics. By maintaining these consistent prompt templates, we facilitate meaningful performance comparison not only within our experimental framework but also with previously published results in multimodal speculative decoding research.



(a) Phase 1: Multimodal Alignment



(b) Phase 2: Self-Distilled Visual Instruction Tuning

Figure 5: Training loss curves obtained during the two-phase MASSV training process when adapting Qwen2.5-1.5B Instruct into a VLM drafter for Qwen2.5-VL 7B Instruct. (a) shows the cross-entropy loss during multimodal projector pretraining, which rapidly decreases from ∼8.0 to ∼2.5 within 2000 steps, indicating efficient adaptation of the trainable projector. (b) displays the loss trajectory during fine-tuning with self-generated target VLM responses, with stable convergence around 1.1 across 5000 training steps, demonstrating successful token distribution alignment between the draft and target models.

Prompt for COCO Captioning Evaluation
Examine the provided image carefully and generate a comprehensive description. Please include relevant details about objects, their spatial relationships, activities, attributes, and any other notable visual elements.

Prompt for GQA Evaluation
For the following question, provide a detailed explanation of your reasoning process. Please analyze the visual elements systematically and articulate each step of your thought process leading to the final answer.
$\{\{Question\}\}$