# HiViS: Hiding Visual Tokens from the Drafter for Speculative Decoding in Vision-Language Models

Supplementary Material

## 7. Evaluation Under ViSpec Settings

we additionally evaluate HiViS and our reproduced EAGLE-2 under the exact settings used by ViSpec. Following their protocol, we prepend the same system prompt: *A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human's questions.* and include all task-specific instructions provided in the dataset. During speculative decoding, we adopt ViSpec's draft token tree configuration, using a total of 30 draft tokens, a tree depth of 4, and 8 nodes selected during each expansion step. Under this evaluation setup, without any additional training, HiViS still consistently surpasses ViSpec in both speedup ratio and average acceptance length across all tested tasks. Full results are reported in Table 1.

Table 1. Speedup ratio ($SR$) and average acceptance length ($\tau$) across benchmarks under ViSpec's settings.

| Model | Methods | ChartQA | | VQAv2 | | ScienceQA | | TextVQA | | MME | | MMVet | | SEED-Bench | | GQA | | Avg | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $SR$ | $\tau$ | $SR$ | $\tau$ | $SR$ | $\tau$ | $SR$ | $\tau$ | $SR$ | $\tau$ | $SR$ | $\tau$ | $SR$ | $\tau$ | $SR$ | $\tau$ | $SR$ | $\tau$ |
| T = 0 | | | | | | | | | | | | | | | | | | | |
| LLaVA-Next-7B | EAGLE-2 | 1.37× | 2.75 | 1.94× | 3.66 | 1.52× | 2.80 | 1.52× | 2.96 | 1.53× | 2.91 | 1.50× | 2.80 | 1.67× | 3.64 | 1.55× | 3.56 | 1.58× | 3.14 |
| | ViSpec | 1.75× | 3.98 | 2.02× | 3.82 | 2.36× | 3.87 | 2.10× | 3.99 | 1.84× | 3.80 | 1.81× | 3.81 | 1.78× | 4.08 | 1.63× | 3.86 | 1.91× | 3.90 |
| | HiViS | **1.80×** | **4.03** | **2.19×** | **4.34** | **2.52×** | **4.05** | **2.19×** | **4.07** | **1.89×** | **3.81** | **1.90×** | **3.96** | **1.85×** | **4.17** | **1.72×** | **4.15** | **2.01×** | **4.07** |
| LLaVA-Next-13B | EAGLE-2 | 1.50× | 2.78 | 1.80× | 3.67 | 1.69× | 2.80 | 1.63× | 2.08 | 1.66× | 2.92 | 1.63× | 2.82 | 1.75× | 3.47 | 1.72× | 3.61 | 1.67× | 3.02 |
| | ViSpec | 1.84× | 3.91 | 2.47× | 3.88 | 2.39× | 3.71 | 2.16× | 3.78 | 2.00× | 3.83 | 1.93× | 3.66 | 1.94× | 4.08 | 1.83× | 3.93 | 2.07× | 3.85 |
| | HiViS | **1.88×** | **4.01** | **2.61×** | **4.27** | **2.56×** | **4.00** | **2.32×** | **4.03** | **2.07×** | **3.93** | **2.02×** | **3.91** | **1.97×** | **4.13** | **1.87×** | **4.22** | **2.16×** | **4.06** |
| Qwen2.5-VL-7B | EAGLE-2 | 1.46× | 2.44 | 2.00× | 3.47 | 1.57× | 2.57 | 2.34× | 2.44 | 1.58× | 2.54 | 1.53× | 2.37 | 1.71× | 3.06 | 1.99× | 3.43 | 1.77× | 2.79 |
| | ViSpec | 1.85× | 3.56 | 1.88× | 3.22 | 1.87× | 3.17 | 2.94× | 3.22 | 1.80× | 3.09 | 1.78× | 3.08 | 1.74× | 3.25 | 1.93× | 3.30 | 1.97× | 3.24 |
| | HiViS | **1.97×** | **3.85** | **2.13×** | **3.78** | **2.04×** | **3.47** | **3.04×** | **3.34** | **1.89×** | **3.25** | **1.89×** | **3.30** | **1.85×** | **3.51** | **2.15×** | **3.74** | **2.12×** | **3.53** |
| T = 1 | | | | | | | | | | | | | | | | | | | |
| LLaVA-Next-7B | EAGLE-2 | 1.22× | 2.34 | 1.48× | 2.92 | 1.30× | 2.33 | 1.21× | 2.50 | 1.31× | 2.38 | 1.30× | 2.46 | 1.44× | 2.80 | 1.42× | 2.79 | 1.34× | 2.57 |
| | ViSpec | 1.45× | 3.18 | 1.82× | 3.07 | 1.93× | 2.96 | 1.50× | 2.84 | 1.57× | 2.99 | 1.51× | 3.01 | 1.53× | 3.15 | 1.43× | 3.13 | 1.59× | 3.04 |
| | HiViS | **1.60×** | **3.30** | **1.88×** | **3.26** | **2.14×** | **3.06** | **1.62×** | **2.84** | **1.63×** | **3.04** | **1.60×** | **3.06** | **1.55×** | **3.16** | **1.50×** | **3.26** | **1.69×** | **3.12** |
| LLaVA-Next-13B | EAGLE-2 | 1.42× | 2.50 | 1.60× | 3.01 | 1.56× | 2.29 | 1.42× | 2.62 | 1.51× | 2.50 | 1.50× | 2.47 | 1.58× | 2.85 | 1.47× | 2.96 | 1.51× | 2.65 |
| | ViSpec | 1.69× | 3.28 | 2.14× | 3.19 | 1.85× | 2.94 | 1.73× | 2.97 | 1.73× | 3.09 | 1.77× | 3.07 | 1.69× | 3.23 | 1.64× | 3.17 | 1.78× | 3.12 |
| | HiViS | **1.76×** | **3.39** | **2.28×** | **3.40** | **2.16×** | **3.13** | **1.83×** | **3.07** | **1.78×** | **3.16** | **1.81×** | **3.20** | **1.89×** | **3.29** | **1.75×** | **3.37** | **1.91×** | **3.25** |
| Qwen2.5-VL-7B | EAGLE-2 | 1.43× | 2.42 | 1.63× | 2.64 | 1.39× | 2.28 | 1.79× | 1.99 | 1.35× | 2.14 | 1.33× | 2.04 | 1.40× | 2.41 | 1.61× | 2.68 | 1.49× | 2.33 |
| | ViSpec | 1.78× | 3.41 | 1.57× | 2.65 | 1.57× | 2.67 | 2.15× | 2.44 | 1.51× | 2.47 | 1.54× | 2.50 | 1.39× | 2.54 | 1.57× | 2.63 | 1.64× | 2.66 |
| | HiViS | **1.89×** | **3.61** | **1.74×** | **2.92** | **1.73×** | **2.88** | **2.19×** | **2.47** | **1.59×** | **2.59** | **1.63×** | **2.65** | **1.43×** | **2.74** | **1.72×** | **2.93** | **1.74×** | **2.85** |

## 8. Mixed vs. All-Multimodal Training for HiViS

We conduct an additional experiment by replacing the text-only dataset in HiViS's training with an equally size of samples' multimodal dataset. As shown in Figure 1, the mixed-dataset HiViS (multimodal + text) outperforms the all-multimodal variant on most tasks, and maintains only small gaps on the remaining ones. This outcome reflects a property of HiViS: since the drafter in HiViS operates purely in the fused language space and never observes raw visual tokens, its performance is primarily governed by how well it models the target VLM's next-token distribution rather than how well it processes visual sequences. Text-only dataset, which contain longer sequences and a much richer vocabulary, provides stronger supervision for learning long-range language dependencies, which in turn yields a drafter that is more robust across tasks.

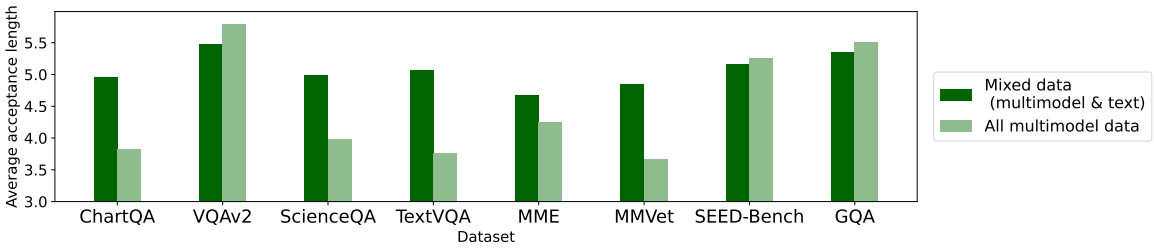


Figure 1. Comparison of average acceptance length between HiViS trained on mixed data (multimodal + text) and HiViS trained on multimodal-only data across several benchmarks, evaluated with a draft tree depth of 6 and temperature of 0.