FROM ALIGNMENT DIAGNOSIS TO JOINT SUPERVISION IN MULTIMODAL DRAFT TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal draft training benefits from multimodal (MM) and text-only (TXT) supervision, yet the underlying alignment changes remain poorly understood. We study draft–target disagreement through an acceptance-related risk framework that decomposes distributional error into text-conditioned and visual-response components. Lower overall risk generally accompanies longer accepted sequences, while similar text-conditioned alignment can coexist with markedly different visual-response risk and acceptance. Controlled continuation training shows that high-risk groups outperform random and low-risk groups, and visual-response risk guides MM training more effectively than aggregate risk. However, useful within-modality rankings do not ensure effective cross-modal ranking: raw-score pooling overwhelmingly favors TXT examples in our candidate pools. We therefore compare risks relative to their modality-specific means rather than by absolute magnitude. Mean normalization preserves within-modality rankings and enables mixed continuation-set construction without predefined modality quotas. With an EAGLE draft for LLaVA-1.5-7B, mean-normalized joint construction improves macro-averaged accepted length across five benchmarks from 3.03 with uncalibrated raw-score pooling to 3.60, under the same 34K-example continuation budget and common warm-up.

1 INTRODUCTION

Speculative decoding (SD) accelerates autoregressive generation using a lightweight draft to propose tokens for parallel verification by a larger target model (Leviathan et al., 2023; Chen et al., 2023). Its effectiveness depends on draft–target agreement, rather than the draft's standalone performance (Zhou et al., 2024; Liu et al., 2024b). Efficient decoding is also important for multimodal models (Alayrac et al., 2022; Li et al., 2023), whose autoregressive generation incurs substantial inference latency. Here, alignment concerns both text-conditioned prediction and how the draft and target respond to visual input.

For multimodal target models, draft training can draw on both multimodal (MM) and text-only (TXT) supervision. Language-only drafts can support multimodal targets in some settings (Gagrani et al., 2024), while MSD combines text-only and multimodal supervision through progressive training (Lin et al., 2025). Other approaches, including ViSpec and SpecFLASH, explicitly incorporate visual information into draft-model design (Kang et al., 2025; Wang et al., 2026). Despite these advances, it remains unclear how MM and TXT supervision affect draft–target alignment and whether these alignment differences can guide draft training. We therefore study both the structure of multimodal draft–target disagreement and its relevance to training-data construction.

We characterize draft–target disagreement using total variation distance, which equals the expected single-token rejection probability under standard stochastic speculative sampling when evaluated on the same decoding distributions. We then compare draft and target predictions with and without visual tokens, holding the textual prompt and generation prefix fixed. This separates the distributional error into text-conditioned and visual-response components. Their magnitudes characterize the remaining mismatch under the text condition and the difference between the models' responses to visual input, respectively.

Across the evaluated supervision regimes, lower overall risk generally accompanies longer accepted sequences. TXT-only and Mixed training reach similar text-conditioned risk levels but differ

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.

markedly in visual-response risk and speculative acceptance. Relative to MM-only training, Mixed also reduces both component risks. These observations show how the decomposition characterizes alignment differences that text-conditioned risk alone does not capture, as illustrated schematically in Figure 1(a).

Our central question is whether the structure of current draft–target disagreement can inform subsequent training. We use component risks as diagnostic probes of this relationship: larger mismatch indicates poorer current alignment, but does not by itself imply greater training utility. Controlled continuation experiments show that high-risk groups outperform random and low-risk groups at fixed sample counts and modality composition (Figure 1(b)), with the advantage also observed when either modality is varied separately. For MM data, visual-response ranking further outperforms aggregate-risk ranking. Together, these results illustrate how distinguishing forms of disagreement can guide training within each modality, beyond measuring overall mismatch alone.

The two component risks provide useful training guidance within their respective modalities, but useful within-modality rankings do not automatically translate into effective cross-modal ranking. A larger raw score need not imply greater training priority across modalities: the scores measure different forms of mismatch, and raw-score pooling overwhelmingly favors TXT examples in our candidate pools. We therefore retain within-modality rankings while prioritizing samples whose component risk is large relative to their own modality's average. Mean normalization implements this relative-risk criterion, providing a shared reference for constructing mixed continuation sets without predefined modality quotas. Figure 1(c) illustrates this process: scoring candidates, normalizing scores, and jointly selecting examples for continuation training.

To test whether within-modality risk signals can be combined without calibration, we compare mean-normalized joint construction with Raw-Joint, which pools the unnormalized component scores. With an EAGLE draft (Li et al., 2024a) for LLaVA-1.5-7B (Liu et al., 2024a), mean normalization improves macro-averaged accepted length across five benchmarks from 3.03 to 3.60 under the same 34K continuation budget and common warm-up. It also outperforms fixed 1:1 high-risk grouping (3.45) and composition-matched random grouping (3.35), showing that its advantage is not limited to the uncalibrated pooling baseline.

Our contributions are threefold:

- **Component-wise alignment diagnostics.** We use a decomposition of multimodal draft–target error into text-conditioned and visual-response components to reveal alignment differences across supervision regimes hidden by aggregate risk.
- **Training relevance of component risks.** Controlled continuation experiments show that high component-risk groups outperform random and low-risk groups within both modalities. For MM data, visual-response ranking also outperforms aggregate-risk ranking, linking disagreement structure to training outcomes.
- **Combining supervision through relative risk.** Useful within-modality rankings do not ensure effective cross-modal pooling. A shared ratio-to-average reference enables joint use without predefined modality quotas. Its mean-normalized implementation improves accepted length over raw-score pooling, fixed-ratio high-risk grouping, and composition-matched random baselines in the evaluated settings.

2 Related Work

Speculative decoding. Speculative decoding accelerates autoregressive generation by using a lightweight draft model to propose candidates for parallel verification (Xia et al., 2023; Leviathan et al., 2023). Subsequent work improves candidate generation and draft modeling through multi-head prediction, target-feature-based drafting, dynamic trees, and parallel drafting (Cai et al., 2024; Li et al., 2024a;b; 2025; Chen et al., 2026). Beyond proposal quality, draft–target distributional agreement is also critical to acceptance. DistillSpec improves such agreement through knowledge distillation and studies training-data generation and divergence objectives (Zhou et al., 2024). LK losses further replace proxy divergence objectives with objectives that directly target token acceptance (Samarin et al., 2026). Different from these works, we focus on the internal structure of multimodal draft–target disagreement and its implications for further training.

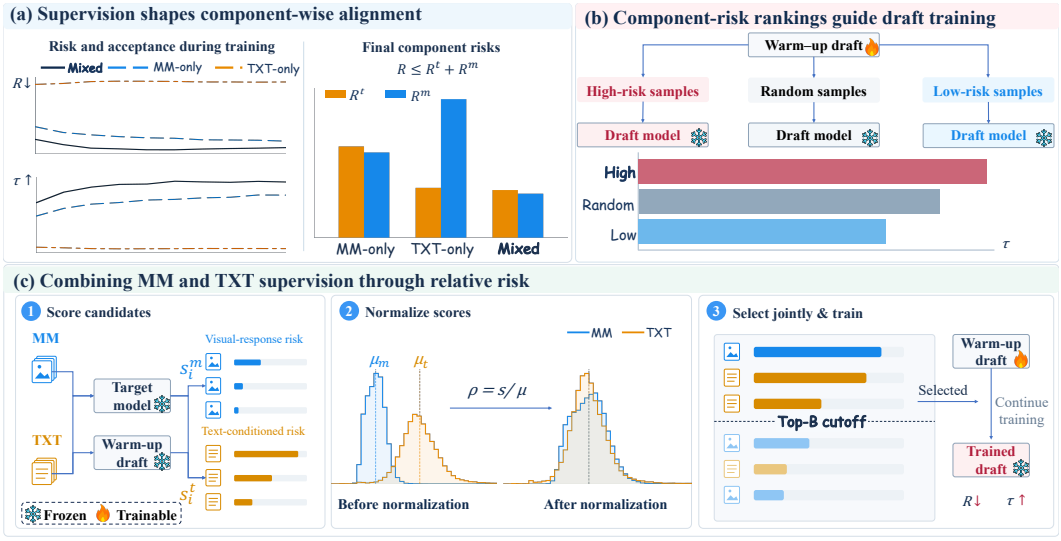


Figure 1: **Overview of the risk framework.** Panels are schematic; quantitative results are reported in the corresponding experimental sections. (a) Supervision regimes differ in component risks and accepted length. (b) High-risk groups yield better continuation-training outcomes than random and low-risk groups. (c) Mean normalization provides a shared relative-risk reference for combining MM and TXT examples into a continuation set without predefined modality quotas.

Multimodal draft modeling and training. Existing methods explore visual conditioning, training supervision, and evaluation for multimodal speculative decoding. Language-only drafts can support multimodal targets in some settings (Gagrani et al., 2024), while ViSpec and SpecFLASH explicitly incorporate visual information into draft-model design (Kang et al., 2025; Wang et al., 2026). MSD combines TXT and MM supervision through progressive training (Lin et al., 2025), whereas TIGER couples text-conditioned visual routing with verifier-derived accepted-prefix rewards for draft optimization (Vo et al., 2026). MMSpec provides unified evaluation, distinguishing throughput from latency gains and highlighting the importance of vision awareness under batching (Shen et al., 2026). These advances address how drafts use visual information, learn from supervision, and perform under different inference conditions. A complementary question is how different supervision sources reshape the text-conditioned and visual-response components of draft–target disagreement. We examine whether these component-wise differences provide useful signals for constructing continuation sets within and across modalities.

3 Decomposing Multimodal Draft–Target Risk

Speculative decoding depends on the agreement between the predictive distributions of the draft and target (Leviathan et al., 2023; Zhou et al., 2024). We first relate their overall discrepancy to single-token acceptance under standard stochastic speculative sampling, then decompose it into text-conditioned and visual-response components.

3.1 From Multimodal Risk to Speculative Acceptance

Let I denote the image input, T the textual prompt, and $y_{<j}$ the generation prefix before position j. Under the multimodal condition, the target and draft distributions over the vocabulary $\mathcal{V}$ are

$$p_j = p(\cdot \mid I, T, y_{<j}), \qquad q_j = q(\cdot \mid I, T, y_{<j}). \tag{1}$$

We use unadorned symbols for the multimodal condition and define the total multimodal risk as

$$R_j = \frac{1}{2}\|p_j - q_j\|_1. \tag{2}$$

Under standard stochastic speculative sampling (Chen et al., 2023), a candidate $z \sim q_j$ is accepted with probability $\min\{1, p_j(z)/q_j(z)\}$. Its expected single-token acceptance probability is

$$\alpha_j = \mathbb{E}_{z\sim q_j}\left[\min\left\{1, \frac{p_j(z)}{q_j(z)}\right\}\right] = \sum_{z\in\mathcal{V}} \min\{p_j(z), q_j(z)\} = 1 - R_j. \tag{3}$$

Thus, R_j equals the expected single-token rejection probability when risk computation and decoding use the same transformed distributions. We use it to measure local draft–target alignment.

However, total multimodal risk alone does not reveal how the disagreement is structured relative to a paired text-only condition. We therefore separate the remaining text-conditioned mismatch from the difference between target and draft responses to visual input.

3.2 Text-Conditioned and Visual-Response Components

We construct a text condition by removing visual tokens while retaining the textual prompt T and the same generation prefix $y_{<j}$:

$$p_j^t = p(\cdot \mid T, y_{<j}), \qquad q_j^t = q(\cdot \mid T, y_{<j}). \tag{4}$$

We use this paired text-only condition as a diagnostic reference rather than as an independent task setting. Related image-conditioned versus text-only comparisons have been used to assess visual dependence and mitigate language-prior bias (Favero et al., 2024; Liu et al., 2025).

Based on the target–draft distribution gaps under the multimodal and text conditions, we define three error terms:

$$\mathbf{e}_j = p_j - q_j, \qquad \mathbf{e}_j^t = p_j^t - q_j^t, \qquad \mathbf{e}_j^m = (p_j - p_j^t) - (q_j - q_j^t). \tag{5}$$

Here, $\mathbf{e}_j^t$ describes the remaining text-conditioned mismatch after visual-token removal, whereas $\mathbf{e}_j^m$ captures the difference between the target and draft responses to visual input. The superscripts t and m label the text-conditioned and visual-response components, respectively, rather than the training-data modalities. The unadorned $\mathbf{e}_j$ denotes the total multimodal error. By definition,

$$\mathbf{e}_j = \mathbf{e}_j^t + \mathbf{e}_j^m. \tag{6}$$

We define the corresponding text-conditioned and visual-response risks as

$$R_j^t = \frac{1}{2}\|\mathbf{e}_j^t\|_1, \qquad R_j^m = \frac{1}{2}\|\mathbf{e}_j^m\|_1. \tag{7}$$

The text-conditioned risk R_j^t is the total variation distance between the target and draft distributions under the text condition; thus, both R_j and R_j^t lie in $[0, 1]$. In contrast, R_j^m is the half-ℓ_1 discrepancy between two signed visual responses, rather than a total variation distance. Since $\mathbf{e}_j^m = \mathbf{e}_j - \mathbf{e}_j^t$, the triangle inequality gives $0 \le R_j^m \le R_j + R_j^t \le 2$.

Applying the triangle inequality to Equation (6) gives

$$\underbrace{\frac{1}{2}\|p_j - q_j\|_1}_{R_j} \le \underbrace{\frac{1}{2}\|p_j^t - q_j^t\|_1}_{R_j^t} + \underbrace{\frac{1}{2}\left\|(p_j - p_j^t) - (q_j - q_j^t)\right\|_1}_{R_j^m}. \tag{8}$$

The two component risks therefore provide an upper bound on the total multimodal mismatch. This bound need not be tight, as the error components may partially cancel.

4 How Supervision Shapes Draft–Target Alignment

We first examine whether supplementing multimodal training with text-only supervision improves overall alignment, and whether the improvement is confined to the text-conditioned component. This distinguishes the source of supervision from the type of draft–target mismatch it affects.

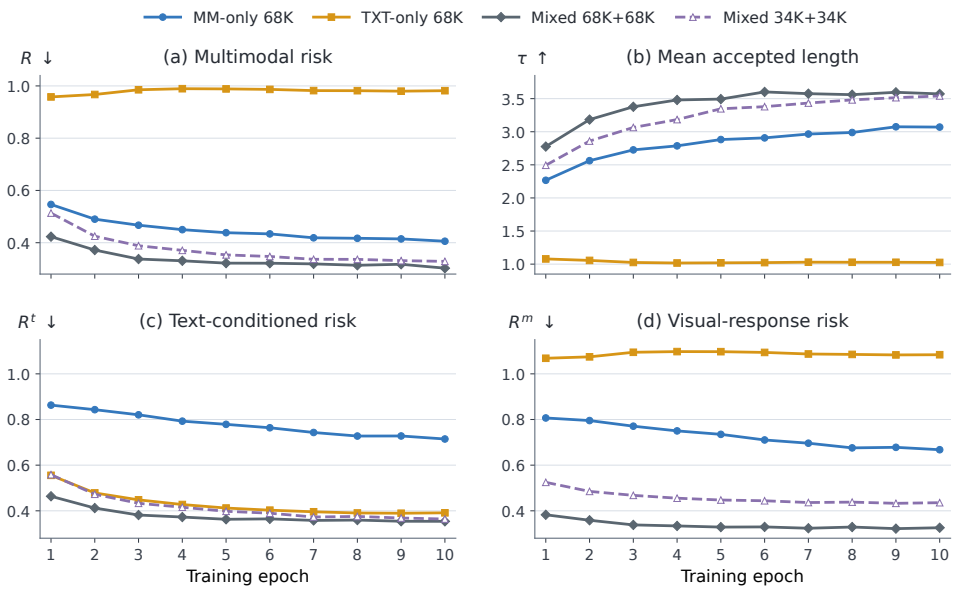


Figure 2: **Risk and accepted length under different supervision regimes.** Solid lines show the three primary regimes; the dashed line shows the Mixed 34K+34K equal-sample-count control. Metrics are macro-averaged across five benchmarks. Panels (c,d) share the same vertical scale.

4.1 Experimental Setup

We use LLaVA-1.5-7B (Liu et al., 2024a) as the target model and train EAGLE drafts (Li et al., 2024a). Multimodal and text-only data come from LLaVA-Instruct-150K (Liu et al., 2023) and ShareGPT, respectively. All runs share the same draft initialization and optimizer settings.

We compare three supervision regimes: **MM-only**, using 68K multimodal instruction samples; **TXT-only**, using 68K text-only instruction samples; and **Mixed**, combining the two sources in two configurations: 68K MM + 68K TXT and 34K MM + 34K TXT.

Mixed 68K+68K combines the same datasets used in MM-only and TXT-only. Mixed 34K+34K randomly samples 34K examples from each dataset and combines them, matching the total sample count of the single-modality regimes. All configurations are trained for ten epochs, with each epoch covering the corresponding training set.

We evaluate each epoch on ChartQA (Masry et al., 2022), MM-Vet (Yu et al., 2024), ScienceQA (Lu et al., 2022), TextVQA (Singh et al., 2019), and NoCaps (Agrawal et al., 2019), reporting macro-averaged R, R^t, R^m, and accepted length τ. Risks are evaluated on fixed response trajectories shared across regimes and checkpoints, and accepted length under a fixed EAGLE decoding configuration. With architecture and decoding settings fixed, higher accepted length requires fewer target-verification cycles for a given output length and is expected to improve decoding speed. Full inference settings appear in Appendix C.1.

4.2 Mixed Supervision Improves Alignment and Drafting

Across the evaluated supervision regimes, lower aggregate risk R generally accompanies greater mean accepted length τ (Figure 2(a,b)). Mixed 68K+68K maintains lower risk and greater accepted length than MM-only across the displayed checkpoints. At epoch 10, it reaches $R = 0.30$ and $\tau = 3.57$, compared with $R = 0.41$ and $\tau = 3.07$ for MM-only. TXT-only instead exhibits high aggregate risk and short accepted sequences, indicating that text-only supervision alone does not reproduce the benefit of mixed training in this setting.

The advantage of Mixed is not solely attributable to its larger sample count. The Mixed 34K+34K control also outperforms MM-only throughout the displayed trajectory, reaching $R = 0.33$ and

$\tau = 3.54$ at epoch 10. Its accepted length is close to that of Mixed 68K+68K, supporting the benefit of combining supervision sources even at the same total sample count as MM-only.

The component risks clarify the benefits of mixed supervision. TXT-only approaches Mixed 68K+68K in text-conditioned risk but retains high visual-response risk and short accepted sequences. Mixed 68K+68K lowers both components relative to MM-only: at epoch 10, R^t and R^m are 0.35 and 0.33, versus 0.71 and 0.67 (Figure 2(c,d)). Mixed 34K+34K likewise lowers both risks, to 0.36 and 0.44. Thus, in this setting, TXT-only training leaves substantial visual-response mismatch, whereas mixed supervision improves both components relative to MM-only. We next test whether component risks also guide training within each modality.

5 Training Relevance of Component Risks

Section 4 shows that component risks distinguish alignment patterns across supervision regimes. We now test whether these signals guide within-modality training: whether higher-risk groups provide more useful continuation supervision and whether visual-response risk outperforms aggregate risk for MM subset construction. All comparisons control sample counts, modality composition, and continuation-training settings.

The common warm-up is the draft obtained after one epoch of Mixed 68K+68K training in Section 4. It is used both to score candidates and to initialize all continuation branches. Each branch uses 17K MM and 17K TXT examples, allowing us to compare risk-based groupings at a fixed total sample count and modality composition.

5.1 Risk-Guided Sample Grouping

For each candidate sample x_i, we average the risks defined in Section 3 over its valid response positions:

$$s_i^t = \frac{1}{L_i}\sum_{j=1}^{L_i} R_{i,j}^t, \qquad s_i^m = \frac{1}{L_i}\sum_{j=1}^{L_i} R_{i,j}^m. \tag{9}$$

where L_i is the number of valid response tokens. For within-modality ranking, we use the text-conditioned score s_i^t for TXT candidates and the visual-response score s_i^m for MM candidates. We rank within each modality, avoiding direct comparisons between the defined risk scales.

We construct three continuation branches: High uses high-score samples from each modality, Low uses the corresponding low-score samples, and Random draws samples randomly and independently from each modality. All branches use identical continuation-training settings and the same warm-up checkpoint. Evaluation follows Section 4.1; we report accepted lengths on individual benchmarks and macro-averaged metrics across all five benchmarks.

Table 1: **Combined within-modality risk grouping.** All continuation runs start from the same warm-up model and use 17K multimodal and 17K text samples.

	Accepted Length					Average			
	ChartQA	MM-Vet	ScienceQA	TextVQA	NoCaps	$\tau \uparrow$	$R \downarrow$	$R^t \downarrow$	$R^m \downarrow$
Warm-up	2.63	2.87	2.68	2.88	2.82	2.78	0.42	0.46	0.38
High	**3.34**	**3.50**	**3.52**	**3.45**	**3.42**	**3.45**	**0.33**	**0.37**	**0.34**
Random	3.10	3.23	3.19	3.29	3.24	3.21	0.36	0.40	0.35
Low	2.81	2.97	2.98	3.02	2.92	2.94	0.40	0.45	0.37

5.2 Training Utility of Component-Risk Rankings

Higher initial component risk corresponds to better continuation-training outcomes. In Table 1, High reaches a macro-averaged accepted length of 3.45, compared with 3.21 for Random and 2.94 for Low. The High–Random–Low ordering holds across all five benchmarks, with lower post-training risks for the higher-risk training groups. Thus, under this protocol, component risks identify useful training groups rather than merely describing the draft's current disagreement.

Varying one modality at a time. Combined grouping changes both modality subsets. To examine whether each component ranking contributes to the observed benefit, we vary the MM or TXT subset while fixing the other to the same RANDOM subset. All remaining training and evaluation settings are unchanged. MM candidates are ranked by s_i^m, and TXT candidates by s_i^t.

Table 2: **Controlled comparisons of component-risk grouping.** MM (TXT) rows vary only the multimodal (text) subset, with the other subset fixed to RANDOM.

	Accepted Length					Average			
	ChartQA	MM-Vet	ScienceQA	TextVQA	NoCaps	$\tau \uparrow$	$R \downarrow$	$R^t \downarrow$	$R^m \downarrow$
MM-HIGH	**3.26**	**3.38**	**3.49**	**3.41**	**3.30**	**3.37**	**0.34**	**0.38**	**0.34**
MM-RANDOM	3.10	3.23	3.19	3.29	3.24	3.21	0.36	0.40	0.35
MM-LOW	2.96	3.06	3.10	3.02	2.97	3.02	0.39	0.43	0.36
TXT-HIGH	**3.27**	**3.43**	**3.34**	**3.30**	**3.39**	**3.35**	**0.35**	**0.38**	**0.34**
TXT-RANDOM	3.10	3.23	3.19	3.29	3.24	3.21	0.36	0.40	0.35
TXT-LOW	3.04	3.20	3.07	3.24	3.23	3.16	0.37	0.41	0.36

Table 2 shows that the High–Random–Low ordering persists in both comparisons. For MM grouping, the macro-averaged accepted lengths are 3.37, 3.21, and 3.02, respectively; the corresponding TXT results are 3.35, 3.21, and 3.16. Thus, each component provides a useful within-modality ranking signal when the other modality subset is held fixed.

5.3 VISUAL-RESPONSE RISK VERSUS AGGREGATE RISK

The preceding experiments show that component-risk rankings can guide continuation-set construction. A remaining question is whether the decomposition provides useful information beyond aggregate draft–target disagreement. For MM samples, we therefore compare visual-response risk s_i^m with the aggregate-risk score

$$s_i = \frac{1}{L_i} \sum_{j=1}^{L_i} R_{i,j}. \tag{10}$$

Both strategies use 17K MM examples and the same 17K high-risk TXT subset, warm-up checkpoint, and continuation-training settings. Only the score used to rank the MM candidates changes. Appendix A.4 visualizes the distinct candidate groups selected by the two signals.

Table 3: **Visual-response versus aggregate risk for MM continuation-set construction.** Both strategies use 17K MM examples and the same 17K high-risk TXT subset. The final four columns report macro-averaged metrics. Bold indicates the better result.

	Accepted Length					Average			
MM ranking signal	ChartQA	MM-Vet	ScienceQA	TextVQA	NoCaps	$\tau \uparrow$	$R \downarrow$	$R^t \downarrow$	$R^m \downarrow$
Aggregate risk s_i	3.27	3.42	3.37	3.38	3.14	3.32	0.34	0.38	0.35
Visual-response risk s_i^m	**3.34**	**3.50**	**3.52**	**3.45**	**3.42**	**3.45**	**0.33**	**0.37**	**0.34**

Table 3 shows that ranking MM by visual-response risk yields higher accepted lengths on all five benchmarks. Macro-averaged accepted length increases from 3.32 to 3.45, while risk decreases from 0.34 to 0.33. Under the same sample budget and TXT supervision, the visual-response component provides a more effective ranking signal than aggregate mismatch in this setting.

This distinction clarifies the role of the decomposition. Aggregate risk measures the draft's overall discrepancy from the target, whereas visual-response risk isolates the difference in their responses to visual input. The comparison shows that the latter distinction is useful for constructing the MM continuation set: the magnitude of total disagreement alone does not provide an equally effective ranking in this experiment. The corresponding comparison with Qwen2.5-VL-3B shows the same qualitative trend (Appendix B.2).

Together, these experiments support using component risks to guide continuation-set construction within each modality. They do not yet establish how scores from the two modalities should be compared. We address this question next by introducing a shared relative-risk reference.

6 Combining MM and TXT Supervision through Relative Risk

Section 5 shows that component-risk rankings can guide continuation-set construction within each modality. Combining these rankings requires a shared reference for the two component scores. We adopt a relative-risk criterion: samples receive higher priority when their component discrepancy is larger relative to the average of their own modality. Mean normalization provides a simple implementation of this criterion.

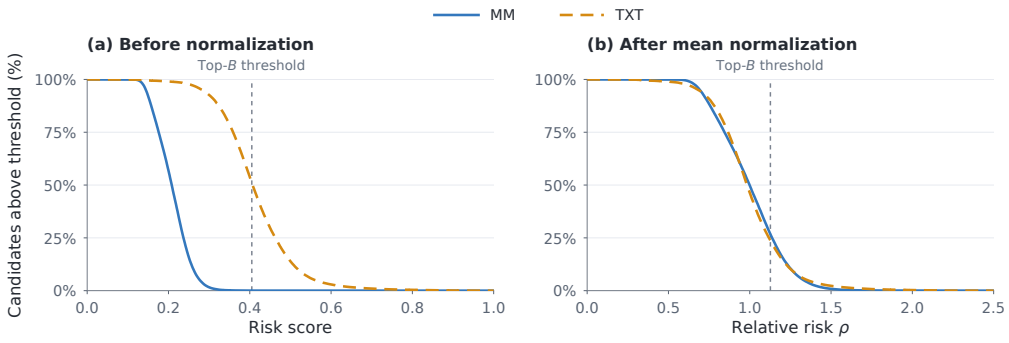


Figure 3: **Cross-modal risk distributions before and after mean normalization.** Each curve shows the fraction of candidates within a modality exceeding a given score threshold. The left panel uses raw scores and the right panel uses mean-normalized scores. Dashed vertical lines mark the respective joint Top-34K thresholds.

6.1 A Shared Relative-Risk Reference

MM samples are scored by s_i^m and TXT samples by s_i^t. These scores measure different forms of draft–target disagreement, so their raw magnitudes do not establish a common scale of training utility. A relatively high score within one modality can remain numerically smaller than scores from the other. Direct pooling can therefore favor a source because of its score scale rather than prioritize relatively large disagreement within each source. In our candidate pools, the mean scores are $\mu_m = 0.2082$ and $\mu_t = 0.4139$. Raw-score pooling selects only 18 MM examples among the top 34K candidates (Figure 3). This motivates a relative-risk reference that accounts for source-specific scales while preserving within-modality rankings.

We instead express each score relative to the average component risk of its own pool:

$$\rho_i = \begin{cases} s_i^m/\mu_m, & x_i \text{ is multimodal,} \\ s_i^t/\mu_t, & x_i \text{ is text-only.} \end{cases} \tag{11}$$

The pool means are computed using the common warm-up model. For positive pool means, $\rho_i = 1$ indicates that a sample's component risk equals its modality-specific average. This transformation preserves zero component discrepancy and within-pool rankings. It is also invariant to multiplying all scores in either pool by a positive constant. Thus, it provides a shared ratio-to-average interpretation without requiring the normalized distributions to coincide. We use this interpretation as a criterion for combining the pools, rather than assuming that equal normalized scores imply equal training utility across modalities.

To evaluate this criterion at a fixed continuation-set size, we jointly rank candidates by ρ_i and construct the continuation set from the top B examples. We call this procedure mean-normalized risk selection with joint pooling (MNRS-Joint). Its modality composition is determined by the normalized scores and the budget, rather than predefined modality quotas. At $B = 34$K, the resulting set contains 18,173 MM and 15,827 TXT examples. This composition follows from the candidate scores and budget, not from tuning against test-set performance.

Table 4: **Combining MM and TXT supervision under a shared relative-risk criterion.** Each continuation-training method uses 34K samples; the common warm-up model is shown for reference. The final four columns report macro-averaged metrics. Bold indicates the best displayed value, including ties.

	Accepted Length					Average			
Method	ChartQA	MM-Vet	ScienceQA	TextVQA	NoCaps	$\tau \uparrow$	$R \downarrow$	$R^t \downarrow$	$R^m \downarrow$
Warm-up	2.63	2.87	2.68	2.88	2.82	2.78	0.42	0.46	0.38
Raw-Joint	3.06	3.02	3.26	3.01	2.78	3.03	0.39	0.41	0.37
Fixed 1:1 High	3.34	3.50	3.52	3.45	3.42	3.45	0.33	**0.37**	**0.34**
Random-Matched	3.30	3.42	3.37	3.33	3.33	3.35	0.34	0.38	**0.34**
TXT→MM	2.98	3.53	3.20	3.40	3.59	3.34	0.35	0.41	0.35
MNRS-JOINT	**3.51**	**3.66**	**3.61**	**3.54**	**3.67**	**3.60**	**0.32**	**0.37**	**0.34**

6.2 EVALUATING RELATIVE-RISK-GUIDED TRAINING

We evaluate whether the two component rankings remain useful when combined through the relative-risk criterion. All methods use a 34K continuation set and follow the warm-up, scoring, and evaluation protocol in Section 5. Raw-Joint ranks the pooled unnormalized scores directly. Fixed 1:1 High uses 17K high-risk examples from each modality. Random-Matched uses the modality counts of MNRS-JOINT but draws examples randomly. Motivated by the text-then-multimodal training schedule used in ViSpec (Kang et al., 2025), we also include a TXT→MM baseline under our common draft architecture and warm-up, matching the modality composition and total sample exposure of MNRS-JOINT.

Table 4 shows that MNRS-JOINT reaches a macro-averaged accepted length of 3.60, compared with 3.45 for Fixed 1:1 High and 3.35 for Random-Matched. The first comparison evaluates a score-dependent modality composition against fixed equal quotas; the second holds the modality composition fixed and compares risk-ranked with randomly drawn examples. Raw-Joint reaches 3.03, indicating a limitation of directly pooling unnormalized scores in the current candidate pools. Relative to these three baselines, MNRS-JOINT also reduces total risk R, reaching 0.32 compared with 0.33, 0.34, and 0.39, respectively. The supplementary sequential-training baseline, TXT→MM, reaches an accepted length of 3.34, compared with 3.60 for MNRS-JOINT, favoring mixed over sequential continuation under this protocol.

Additional comparisons support mean normalization as a simple implementation of the relative-risk criterion. It yields smaller cross-modal distribution distances and higher observed accepted length than RMS and z-score normalization in the evaluated setting (Appendix A.3). The distributional comparison is descriptive rather than an explanation of the training gains.

The near-balanced composition under Mixed warm-up is not intrinsic to mean normalization. With MM-only and TXT-only warm-up drafts, the selected MM shares change to 66.58% and 21.40%, respectively, while the advantage over composition-matched random training persists (Appendix A.2). Thus, the observed benefit is not confined to a near-equal modality mixture. Additional budget comparisons appear in Appendix A.1.

7 CONCLUSION

We analyze multimodal draft–target disagreement through text-conditioned and visual-response components. Across the evaluated supervision regimes, similar text-conditioned alignment can coexist with different visual-response risks and accepted lengths, while mixed supervision lowers both component risks relative to MM-only. Controlled continuation experiments connect these diagnostics to training outcomes: high-risk groups outperform random and low-risk groups, and visual-response ranking outperforms aggregate-risk ranking for MM samples. To combine the modality-specific rankings, mean normalization implements a shared relative-risk criterion without predefined modality quotas. Results on LLaVA-1.5-7B and Qwen2.5-VL-3B support the training relevance of disagreement structure and its use in constructing mixed continuation sets.

AI USE STATEMENT

We used generative AI tools to assist with translation and manuscript preparation, including improving grammar, clarity, conciseness, and organization. We also used these tools to support literature retrieval and discovery by suggesting search terms and potentially relevant publications. AI-generated suggestions served as auxiliary input rather than authoritative sources. The authors retain responsibility for the final text, the accuracy and relevance of citations, and all scientific claims and conclusions.

REFERENCES

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8948–8957, 2019.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736. Curran Associates, Inc., 2022. doi: 10.52202/068431-1723. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf.

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, 2025. URL https://arxiv.org/abs/2502.13923.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D. Lee, Deming Chen, and Tri Dao. Medusa: Simple LLM inference acceleration framework with multiple decoding heads. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5209–5235. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/cai24b.html.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. Accelerating large language model decoding with speculative sampling, 2023. URL https://arxiv.org/abs/2302.01318.

Jian Chen, Yesheng Liang, and Zhijian Liu. Dflash: Block diffusion for flash speculative decoding, 2026. URL https://arxiv.org/abs/2602.06036.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14303–14312, June 2024.

Mukul Gagrani, Raghavv Goel, Wonseok Jeon, Junyoung Park, Mingu Lee, and Christopher Lott. On speculative decoding for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 8285–8289, June 2024.

Jialiang Kang, Han Shu, Wenshuo Li, Yingjie Zhai, and Xinghao Chen. Vispec: Accelerating vision-language models with vision-aware speculative decoding. 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, pp. 115511–115532. Curran Associates, Inc., 2025. doi: 10.52202/085713-3852. URL

https://proceedings.neurips.cc/paper_files/paper/2025/file/a7bfdee9544cea324cf183ac03c7d5c0-Paper-Conference.pdf.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19274–19286. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/leviathan23a.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. EAGLE: Speculative sampling requires rethinking feature uncertainty. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28935–28948. PMLR, 21–27 Jul 2024a. URL https://proceedings.mlr.press/v235/li24bt.html.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. EAGLE-2: Faster inference of language models with dynamic draft trees. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7421–7432, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.422. URL https://aclanthology.org/2024.emnlp-main.422/.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. Eagle-3: Scaling up inference acceleration of large language models via training-time test. 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, pp. 136737–136756. Curran Associates, Inc., 2025. doi: 10.52202/085713-4562. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/c7b5a35ea98b62512a869c19ea7b03cb-Paper-Conference.pdf.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, June 2024a.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 125–140, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73010-8.

Xiaoxuan Liu, Lanxiang Hu, Peter Bailis, Alvin Cheung, Zhijie Deng, Ion Stoica, and Hao Zhang. Online speculative decoding. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning*

Research, pp. 31131–31146. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/liu24y.html.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 2507–2521. Curran Associates, Inc., 2022. doi: 10.52202/068431-0182. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/11332b6b6cf4485b84afadb1352d3a9a-Paper-Conference.pdf.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Alexander Samarin, Sergei Krutikov, Anton Shevtsov, Sergei Skvortsov, Filipp Fisin, and Alexander Golubev. Lk losses: Direct acceptance rate optimization for speculative decoding, 2026. URL https://arxiv.org/abs/2602.23881.

Hui Shen, Xin Wang, Ping Zhang, Yunta Hsieh, Qi Han, Zhongwei Wan, Ziheng Zhang, Jingxuan Zhang, Jing Xiong, Ziyuan Liu, Yifan Zhang, Hangrui Cao, Chenyang Zhao, and Mi Zhang. Mmspec: Benchmarking speculative decoding for vision-language models, 2026. URL https://arxiv.org/abs/2603.14989.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Quynh Vo, Cong-Duy Nguyen, Ponhvoan Srey, Luu Anh Tuan, and Thong Nguyen. Tiger: Text-conditioned visual gated routing with acceptance alignment for multimodal speculative decoding, 2026. URL https://arxiv.org/abs/2607.11131.

Zihua Wang, Ruibo Li, Haozhe Du, Joey Tianyi Zhou, Yu Zhang, and Xu Yang. Specflash: A latent-guided semi-autoregressive speculative decoding framework for efficient multimodal generation, 2026. URL https://arxiv.org/abs/2505.12728.

Heming Xia, Tao Ge, Peiyi Wang, Si-Qing Chen, Furu Wei, and Zhifang Sui. Speculative decoding: Exploiting speculative execution for accelerating seq2seq generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3909–3925, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.257. URL https://aclanthology.org/2023.findings-emnlp.257/.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-vet: Evaluating large multimodal models for integrated capabilities. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57730–57754. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/yu24o.html.

Yongchao Zhou, Kaifeng Lyu, Ankit Singh Rawat, Aditya Krishna Menon, Afshin Rostamizadeh, Sanjiv Kumar, Jean-François Kagy, and Rishabh Agarwal. Distillspec: Improving speculative decoding via knowledge distillation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=rsY6J3ZaTF.

Table 5: **Risk-guided selection across training-data budgets.** Budget denotes the total number of selected examples. Fixed 1:1 High and Random use equal modality counts; Random-Matched matches the modality counts of MNRS-JOINT. The final four columns report macro-averaged metrics. Bold indicates the best displayed value within each budget, including ties.

		Accepted Length					Average			
Budget	Method	ChartQA	MM-Vet	ScienceQA	TextVQA	NoCaps	$\tau \uparrow$	$R \downarrow$	$R^t \downarrow$	$R^m \downarrow$
17K	Random	2.67	2.85	2.77	2.87	2.92	2.82	0.43	0.46	0.38
	Fixed 1:1 High	2.76	3.07	2.92	3.03	3.16	2.99	0.40	0.44	0.37
	Random-Matched	2.79	2.95	2.68	2.94	2.89	2.85	0.42	0.45	0.38
	MNRS-JOINT	**2.85**	**3.19**	**2.94**	**3.08**	**3.33**	**3.08**	**0.38**	**0.42**	**0.36**
34K	Random	3.10	3.23	3.19	3.29	3.24	3.21	0.36	0.40	0.35
	Fixed 1:1 High	3.34	3.50	3.52	3.45	3.42	3.45	0.33	**0.37**	**0.34**
	Random-Matched	3.30	3.42	3.37	3.33	3.33	3.35	0.34	0.38	**0.34**
	MNRS-JOINT	**3.51**	**3.66**	**3.61**	**3.54**	**3.67**	**3.60**	**0.32**	**0.37**	**0.34**
68K	Random	3.16	3.29	3.23	3.29	3.29	3.25	0.35	0.40	0.35
	Fixed 1:1 High	3.41	3.56	3.58	3.45	3.33	3.47	0.33	0.37	**0.33**
	Random-Matched	3.34	3.46	3.52	3.36	3.12	3.36	0.35	0.38	0.34
	MNRS-JOINT	**3.66**	**3.68**	**3.81**	**3.56**	**3.41**	**3.62**	**0.32**	**0.35**	**0.33**

A SUPPLEMENTARY EXPERIMENTS AND ANALYSES

A.1 RISK-GUIDED SELECTION UNDER DIFFERENT DATA BUDGETS

We evaluate within-modality risk selection and cross-modal joint selection under total budgets of 17K, 34K, and 68K samples. Fixed 1:1 High selects equal numbers of multimodal and text-only examples using visual-response and text-conditioned risks, respectively, while Random samples the same number of examples from each modality. MNRS-JOINT selects examples jointly by mean-normalized component risks, whereas Random-Matched randomly selects the same modality counts as MNRS-JOINT. All branches start from the common warm-up checkpoint and continue training for ten epochs.

The MM:TXT counts selected by MNRS-JOINT are 8,821:8,179 at 17K, 18,173:15,827 at 34K, and 35,384:32,616 at 68K. These compositions follow from joint ranking at each budget, without predefined modality quotas. The 34K results reproduce the corresponding comparisons reported in the main paper.

Within-modality selection. Table 5 shows that Fixed 1:1 High outperforms Random on all five benchmarks at every evaluated budget. Its macro-averaged accepted-length gains are 0.17, 0.24, and 0.21 at 17K, 34K, and 68K, respectively, computed from unrounded values. It also obtains lower values of all three macro-averaged risks at each budget. Since these comparisons hold sample count and modality composition fixed, they support the training utility of component-risk selection across the evaluated budgets.

Cross-modal joint selection. MNRS-JOINT achieves the highest accepted length on every benchmark and the lowest macro-averaged risks at each budget. Relative to Random-Matched, its macro-averaged accepted-length gains are 0.23, 0.25, and 0.26 at 17K, 34K, and 68K, respectively, computed from unrounded values. Because modality counts are identical within each comparison, these gains support risk-guided selection beyond the selected composition alone. Compared with Fixed 1:1 High, MNRS-JOINT provides additional gains of 0.09, 0.15, and 0.16, respectively. The advantage of mean-normalized joint selection therefore extends beyond the 34K budget used in the main experiments.

Effect of increasing the budget. Increasing the training-data budget improves macro-averaged accepted length across all evaluated strategies, with generally lower post-training risks. However, the additional gains become smaller at larger budgets. For MNRS-JOINT, accepted length increases

Table 6: **Joint selection with different warm-up drafts.** Each warm-up is trained for one epoch, followed by ten continuation epochs on 34K selected samples. Random-Matched matches the modality counts of MNRS-JOINT within each setting. The final four columns report macro-averaged metrics. Bold indicates the best displayed value within each setting, including ties.

		Accepted Length					Average			
Warm-up	Method	ChartQA	MM-Vet	ScienceQA	TextVQA	NoCaps	$\tau \uparrow$	$R \downarrow$	$R^t \downarrow$	$R^m \downarrow$
MM-only	Random-Matched	2.83	3.20	2.97	3.11	3.23	3.07	0.38	0.42	0.37
	MNRS-JOINT	**2.98**	**3.28**	**3.13**	**3.28**	**3.46**	**3.23**	**0.36**	**0.41**	**0.36**
TXT-only	Random-Matched	3.28	3.26	3.57	3.19	3.11	3.28	0.35	0.39	0.34
	MNRS-JOINT	**3.49**	**3.46**	**3.94**	**3.35**	**3.23**	**3.50**	**0.33**	**0.36**	**0.33**

from 3.08 at 17K to 3.60 at 34K, followed by a smaller improvement to 3.62 at 68K. These results support the effectiveness of risk-guided selection across the evaluated budgets. Since all branches train for ten epochs, larger datasets also entail greater training exposure; the cross-budget comparisons do not hold training compute fixed.

A.2 RELATIVE-RISK-GUIDED TRAINING WITH DIFFERENT WARM-UP DRAFTS

We examine whether the benefits of mean-normalized joint selection extend beyond the Mixed warm-up used in the main experiments. We obtain alternative warm-up drafts after one epoch of MM-only or TXT-only training. For each warm-up, we recompute candidate risks and modality-specific normalization means using the corresponding checkpoint, then apply joint selection under the same 34K sample budget. Both MNRS-JOINT and Random-Matched continue training from that checkpoint for ten epochs, with Random-Matched using the same modality counts as MNRS-JOINT.

The selected composition varies with the warm-up draft. MM-only warm-up yields 22,638 MM and 11,362 TXT samples (66.58:33.42), whereas TXT-only warm-up yields 7,277 MM and 26,723 TXT samples (21.40:78.60). For reference, the Mixed warm-up used in the main paper yields 18,173 MM and 15,827 TXT samples (53.45:46.55). Thus, the selected modality ratio depends on the risk distributions induced by the warm-up checkpoint.

Selection benefits across warm-up settings. Table 6 shows that MNRS-JOINT outperforms Random-Matched on all five benchmarks under both alternative warm-up settings. With MM-only warm-up, macro-averaged accepted length increases from 3.07 to 3.23; with TXT-only warm-up, it increases from 3.28 to 3.50. The corresponding gains are 0.16 and 0.21, computed from unrounded values. All three macro-averaged risks also decrease under both settings, indicating improved overall, text-conditioned, and visual-response alignment. Because each comparison fixes the starting checkpoint and selected modality counts, these improvements support the value of risk-guided selection beyond initialization or data composition alone.

Checkpoint-dependent modality composition. The near-balanced composition obtained with Mixed warm-up is therefore a checkpoint-specific outcome, not a constraint imposed by mean normalization. Across the three evaluated warm-up regimes, the selected compositions differ substantially, yet MNRS-JOINT outperforms its composition-matched random baseline in each setting. These results support risk-guided construction across different resulting mixtures, without establishing that the selected ratios are optimal.

A.3 ALTERNATIVE CROSS-MODAL CALIBRATION STRATEGIES

We compare mean normalization with RMS and z-score normalization using the same warm-up draft and candidate pools. Calibration statistics are computed exclusively from the complete valid candidate pools, without using benchmark results. We examine both the discrepancy between the calibrated score distributions and the downstream performance of the resulting 34K continuation sets.

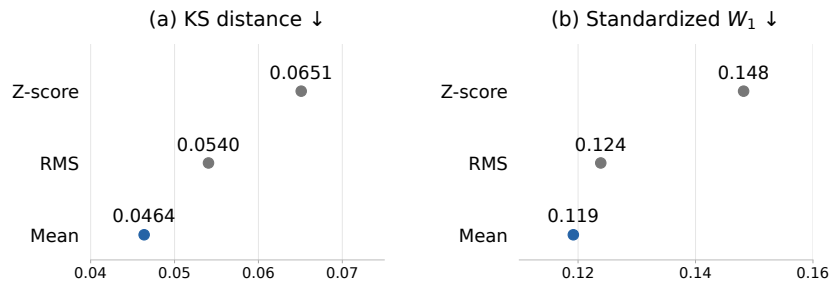


Figure 4: **Cross-modal score distribution discrepancies under alternative calibration methods.** (a) Kolmogorov–Smirnov distance. (b) Standardized Wasserstein-1 distance. Each method is fitted separately to the complete MM and TXT candidate pools, with 68K valid examples per modality. Distances are computed over these complete pools and are independent of the continuation-set budget. Lower values indicate smaller distribution discrepancies; the panels use separate horizontal scales.

Calibration methods. For each modality $k \in \{m, t\}$, let μ_k, σ_k, and $r_k = \sqrt{\mathbb{E}_k[(s^k)^2]}$ denote the mean, population standard deviation, and root mean square of its candidate scores. We compare three calibration rules:

$$\text{Mean: } \frac{s_i^k}{\mu_k}, \qquad \text{RMS: } \frac{s_i^k}{r_k}, \qquad \text{Z-score: } \frac{s_i^k - \mu_k}{\sigma_k}. \tag{12}$$

Here, each expression gives the calibrated score ρ_i^k, and $\mathbb{E}_k$ denotes the empirical average over valid candidates in modality k.

All statistics are computed separately for each modality, with equal weight assigned to each valid sample. Missing scores are not imputed. Negative z-scores are retained without clipping.

For positive denominators, all three transformations preserve within-modality rankings. Mean normalization provides a ratio-to-average interpretation, whereas RMS normalization scales scores by their root mean square. Z-score normalization adjusts both location and dispersion.

Full-pool distribution comparison. We compare the empirical distributions of the calibrated MM and TXT scores over the complete candidate pools, containing 68K valid examples per modality. For each calibration method, let $\widehat{F}_m$ and $\widehat{F}_t$ denote the corresponding empirical cumulative distribution functions. We measure their discrepancy using the Kolmogorov–Smirnov distance,

$$D_{\mathrm{KS}} = \sup_x \left| \widehat{F}_m(x) - \widehat{F}_t(x) \right|, \tag{13}$$

and the Wasserstein-1 distance,

$$W_1 = \int_{\mathbb{R}} \left| \widehat{F}_m(x) - \widehat{F}_t(x) \right| \, \mathrm{d}x. \tag{14}$$

Because the calibration methods produce different output scales, we standardize W_1 by the root mean within-modality variance:

$$\widetilde{W_1} = \frac{W_1}{\sqrt{\big(\mathrm{Var}(\rho^m) + \mathrm{Var}(\rho^t)\big)/2}}. \tag{15}$$

The variances are computed over the complete calibrated pools using population normalization. Both distances are evaluated directly from the sample scores rather than histogram approximations.

Figure 4 shows that mean normalization yields the smallest discrepancies among the three methods on the current candidate pools. Its KS distance is 0.0464, compared with 0.0540 for RMS and 0.0651 for z-score normalization. The corresponding standardized Wasserstein-1 distances are 0.119, 0.124, and 0.148. These results characterize the empirical score distributions; they do not establish equal training utility across modalities or determine downstream performance.

Table 7: **LLaVA performance with alternative cross-modal calibration methods.** All methods use the same candidate pools, warm-up checkpoint, and 34K continuation-set budget. MNRS-JOINT uses mean normalization. The final four columns report macro-averaged metrics. Bold indicates the best displayed value, including ties.

	Accepted Length					Average			
Method	ChartQA	MM-Vet	ScienceQA	TextVQA	NoCaps	$\tau \uparrow$	$R \downarrow$	$R^t \downarrow$	$R^m \downarrow$
Z-score Joint	3.38	3.55	3.46	3.40	3.51	3.46	0.33	**0.37**	**0.34**
RMS Joint	3.41	3.57	3.50	3.43	3.56	3.49	0.33	**0.37**	**0.34**
MNRS-JOINT	**3.51**	**3.66**	**3.61**	**3.54**	**3.67**	**3.60**	**0.32**	**0.37**	**0.34**

Shared continuation-training protocol. For each method, we jointly rank MM and TXT candidates by their calibrated scores and use the top 34K examples to construct the continuation set. The modality composition follows from the ranking, without predefined modality quotas. All methods use the same warm-up checkpoint and ten-epoch continuation-training protocol. Mean normalization yields 18,173 MM and 15,827 TXT examples; RMS yields 18,403 MM and 15,597 TXT examples; and z-score normalization yields 18,799 MM and 15,201 TXT examples. Their MM shares are 53.45%, 54.13%, and 55.29%, respectively.

Downstream performance. Table 7 reports the LLaVA results. Mean normalization achieves a macro-averaged accepted length of 3.60, compared with 3.49 for RMS and 3.46 for z-score normalization. It also achieves the highest accepted length on each of the five benchmarks and a lower macro-averaged total risk in the reported results.

Because these transformations preserve within-modality rankings, they affect the jointly constructed training sets through the allocation of examples across modalities. The comparison therefore evaluates each calibration together with the dataset composition it induces; it does not isolate calibration under a fixed modality composition.

In this setting, mean normalization yields both smaller full-pool distribution discrepancies and better observed drafting performance than the two alternatives. These observations support its use as a simple implementation of the relative-risk criterion. They do not establish that smaller distribution distances cause better training outcomes or that mean normalization is universally optimal.

A.4 COMPARING MM RISK RANKINGS

To characterize how the two MM ranking signals differ, we compare aggregate risk s_i and visual-response risk s_i^m on the same 68K MM candidate pool used for the LLaVA experiments. Both scores are computed using the common warm-up draft, and each ranking selects the top 17K candidates.

Figure 5 shows that visual-response ranking retains candidates whose aggregate risk falls below the aggregate-selection threshold. Conversely, aggregate ranking includes candidates whose visual-response risk falls below the component-selection threshold. The two signals therefore prioritize distinct groups in the paired risk space. This visualization characterizes the selection differences; their training relevance is assessed by the controlled continuation comparison in Table 3.

B ADDITIONAL RESULTS ON QWEN2.5-VL-3B

The Qwen2.5-VL-3B (Bai et al., 2025) experiments follow the same risk definitions, candidate data sources, and evaluation protocol as the LLaVA setting, with draft dimensions matched to the target architecture. We evaluate the three primary supervision regimes and the sample-level and joint-selection comparisons described below.

B.1 SUPERVISION-LEVEL RISK ANALYSIS

Figure 6(a,b) shows multimodal risk and accepted length across MM-only, TXT-only, and Mixed training. As in the LLaVA experiments, lower multimodal risk is generally associated with longer

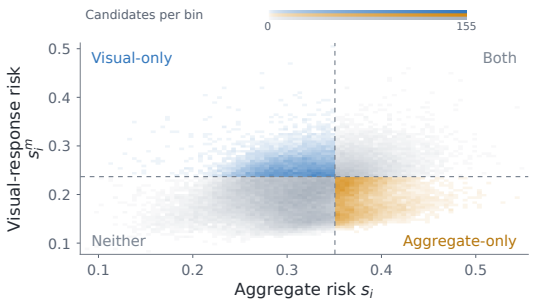


Figure 5: **Aggregate versus visual-response risk for MM candidate selection.** The plot summarizes 68K MM candidates in equal-area bins of width 0.005 along each axis. Color intensity encodes candidate counts using a shared square-root scale. Dashed lines mark the respective Top-17K thresholds. Blue and orange denote regions selected only by visual-response and aggregate risk, respectively; gray denotes selection by both or neither.

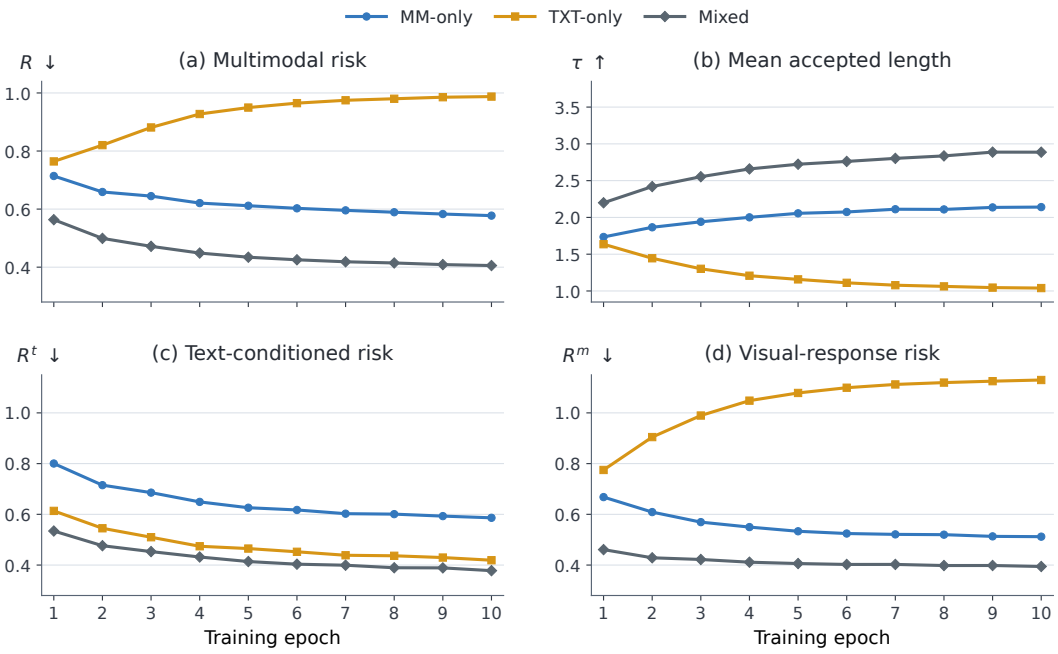


Figure 6: **Risk and accepted length under different supervision regimes with Qwen2.5-VL-3B.** All metrics are macro-averaged across five benchmarks. Panels (c,d) share the same vertical scale.

accepted sequences. Mixed training maintains the lowest overall risk and the highest accepted length, whereas TXT-only training exhibits high multimodal risk and low acceptance throughout the training trajectory.

The component trajectories in Figure 6(c,d) further reproduce the alignment pattern observed with LLaVA. TXT-only training approaches Mixed training in text-conditioned risk R^t, but maintains substantially higher visual-response risk R^m. Thus, similar text-conditioned alignment can again coexist with markedly different visual-response alignment and speculative acceptance.

B.2 Sample-Level Training Utility

We next examine whether the sample-level findings also transfer to the Qwen2.5-VL-3B setting. Following the LLaVA experiments, HIGH combines MM samples with high visual-response risk

Table 8: **Sample-level risk grouping with Qwen2.5-VL-3B.** All continuation branches use the same training budget and modality composition. Warm-up is included for reference. Bold indicates the best result among continuation branches.

	Accepted Length					Average			
	ChartQA	MM-Vet	ScienceQA	TextVQA	NoCaps	$\tau \uparrow$	$R \downarrow$	$R^t \downarrow$	$R^m \downarrow$
Warm-up	2.28	2.20	2.28	1.97	2.26	2.20	0.56	0.53	0.46
HIGH	**2.40**	**2.35**	**2.46**	**2.12**	**2.47**	**2.36**	**0.53**	**0.51**	**0.46**
RANDOM	2.27	2.13	2.22	1.89	2.22	2.15	0.58	0.56	0.48
LOW	2.25	2.01	2.12	1.86	1.98	2.04	0.59	0.57	0.48

Table 9: **Visual-response versus aggregate risk for MM continuation-set construction with Qwen2.5-VL-3B.** The TXT subset, MM budget, warm-up checkpoint, and continuation-training settings are fixed. The final four columns report macro-averaged metrics. Bold indicates the better result.

	Accepted Length					Average			
MM ranking signal	ChartQA	MM-Vet	ScienceQA	TextVQA	NoCaps	$\tau \uparrow$	$R \downarrow$	$R^t \downarrow$	$R^m \downarrow$
Aggregate risk s_i	2.27	2.25	2.30	1.98	2.27	2.21	0.56	0.52	0.47
Visual-response risk s_i^m	**2.40**	**2.35**	**2.46**	**2.12**	**2.47**	**2.36**	**0.53**	**0.51**	**0.46**

and TXT samples with high text-conditioned risk; LOW uses the corresponding low-risk samples, while RANDOM samples independently from both candidate pools. All three branches use the same continuation-training budget and modality composition.

Table 8 shows a consistent separation in downstream performance. HIGH achieves the highest accepted length on all five benchmarks, reaching a macro-average of 2.36, compared with 2.15 for RANDOM and 2.04 for LOW, while also obtaining the lowest post-training risks among the continuation branches. Both RANDOM and LOW fall below the warm-up model's mean accepted length of 2.20, showing that continuation training does not necessarily improve acceptance under this protocol.

The High–Random–Low ordering holds across all five benchmarks, reproducing the sample-level training-utility pattern observed in the main LLaVA experiments.

Visual-response risk versus aggregate risk. We repeat the MM ranking comparison in Table 3 using Qwen2.5-VL-3B, keeping the TXT subset, MM budget, warm-up checkpoint, and continuation-training settings fixed. As shown in Table 9, visual-response risk yields higher accepted length on all five benchmarks, increasing the macro-average from 2.21 to 2.36 while reducing multimodal risk from 0.56 to 0.53. This reproduces the qualitative trend observed with LLaVA-1.5-7B and provides additional evidence for component-wise MM selection.

B.3 CROSS-MODAL RISK CALIBRATION AND JOINT SELECTION

Finally, we evaluate cross-modal joint selection with Qwen2.5-VL-3B. The scale mismatch between the MM and TXT component scores is particularly pronounced in this setting: directly ranking the raw scores selects only 5 multimodal examples among the top 34K candidates, with the remaining 33,995 examples coming from the TXT pool. This imbalance substantially degrades continuation training, with Raw-Joint reaching a macro-averaged accepted length of only 1.70, below the warm-up model's 2.20.

Mean normalization addresses this cross-modal scale imbalance and yields 23,420 MM and 10,580 TXT examples under the same 34K budget. Fixed 1:1 High corresponds to the HIGH component-risk branch in Table 8. As shown in Table 10, MNRS-JOINT achieves the highest macro-averaged accepted length of 2.51 and the lowest macro-averaged risks. It improves substantially over Raw-Joint (1.70), while also outperforming Fixed 1:1 High (2.36) and Random-Matched (2.22). Because

Table 10: **Cross-modal risk calibration and joint selection with Qwen2.5-VL-3B.** Each continuation-training selection strategy uses 34K samples; the common warm-up model is shown for reference. The final four columns report macro-averaged metrics. Bold indicates the best displayed value, including ties.

	Accepted Length					Average			
Method	ChartQA	MM-Vet	ScienceQA	TextVQA	NoCaps	$\tau \uparrow$	$R \downarrow$	$R^t \downarrow$	$R^m \downarrow$
Warm-up	2.28	2.20	2.28	1.97	2.26	2.20	0.56	0.53	0.46
Raw-Joint	2.26	1.50	2.33	1.27	1.15	1.70	0.73	**0.47**	0.77
Fixed 1:1 High	2.40	2.35	2.46	2.12	2.47	2.36	0.53	0.51	0.46
Random-Matched	2.23	2.19	2.30	1.98	2.39	2.22	0.56	0.53	0.47
MNRS-JOINT	**2.50**	**2.50**	**2.64**	**2.20**	**2.70**	**2.51**	**0.50**	**0.47**	**0.44**

Random-Matched uses exactly the same MM:TXT composition as MNRS-JOINT, their comparison further shows that the improvement is not explained by modality composition alone.

Overall, the Qwen2.5-VL-3B experiments reproduce the three main observations from the LLaVA setting: (1) lower overall draft–target risk accompanies longer accepted sequences; (2) component risks distinguish examples with different downstream training utility and provide additional information beyond aggregate disagreement; and (3) mean-normalized calibration enables effective joint use of heterogeneous MM and TXT risk signals under a shared training budget.

C EXPERIMENTAL DETAILS

Code availability. An anonymized code package is provided in the supplementary material, including implementations for risk computation, continuation-set construction, draft training, and evaluation, along with experiment configurations.

C.1 INFERENCE AND EVALUATION SETTINGS

Decoding configuration. For the main LLaVA setting, we use LLaVA-1.5-7B as the target model and a single-layer EAGLE draft, with the same dynamic-tree decoding implementation across all compared methods. Both target and draft models run in FP16 during inference with a batch size of one on a single GPU. We use greedy decoding (temperature 0) and a draft tree with 32 nodes, including the root, a maximum expansion depth of 6, and a branching factor of 4. The generation and total sequence-length thresholds are 256 and 4,096 tokens, respectively. Generation stops upon a termination token or when a length threshold is reached, with stopping conditions checked after each verification cycle. The final cycle may therefore extend beyond the configured length thresholds.

Evaluation protocol. We evaluate on ChartQA, MM-Vet, ScienceQA, TextVQA, and NoCaps, with all methods using the same image-containing evaluation examples and generation seed 123. For each benchmark, accepted length τ is computed as the total number of generated tokens divided by the total number of target-verification cycles. The generated-token count includes the additional target-verified token emitted in each cycle. We report the unweighted average of the five benchmark-level accepted lengths.

Risk computation. For both LLaVA and Qwen, we compute full-vocabulary probabilities from unscaled target and draft logits, applying softmax and risk calculations in FP32 without top-k or top-p filtering. Draft predictions use teacher-forced target features rather than draft-generated feature rollouts. For the paired text condition, we remove the image input and its marker and recompute target features and draft states without image-conditioned caches, keeping the response tokens fixed. Candidate scores use masked assistant responses from the original datasets, with sequences truncated to 2,048 tokens, whereas benchmark risks use fixed response trajectories shared across drafts. Risks are averaged over response tokens within each example, then equally over examples within each benchmark and across the five benchmarks; all components use identical positions and weights. These distributional diagnostics are computed separately from greedy decoding, and their association with accepted length is evaluated empirically.

C.2 BENCHMARK-SPECIFIC PROMPTS

All benchmarks use the target model's chat template without a system message. The image precedes the task-specific text:

```
USER: <image>
{task-specific text} ASSISTANT:
```

We use the test splits of ChartQA and MM-Vet and the validation splits of ScienceQA, TextVQA, and NoCaps. For ScienceQA, we evaluate only image-bearing examples. The task-specific text is specified below; all methods use the same templates.

ChartQA. We append an explanation instruction to the original question:

```
{question}
Please answer with an explanation.
```

MM-Vet. We use the original question without additional instructions:

```
{question}
```

ScienceQA. We provide the question and labeled answer choices, requesting both an answer and a short explanation:

```
You are given an image and a multiple-choice question.
Choose the best answer from the choices.
Question: {question}
Choices:
(A) {choice_a}
(B) {choice_b}
...
({last_letter}) {last_choice}
Please follow this format:
Answer: (X) <choice>
Reasoning: <short explanation>
```

The dataset's reference solution, lecture, and hint are not included in the input.

TextVQA. We append a fixed OCR-oriented instruction to the question:

```
{question}
Perform an OCR task on the provided image. Extract the
text accurately and provide a detailed explanation of
the process. Ensure the response is comprehensive and
well-structured.
```

Dataset-provided OCR tokens are not included in the prompt.

NoCaps. We use the same captioning instruction for every image:

```
Describe the image in one sentence, paying particular
attention to uncommon objects.
```

Ground-truth answers and reference captions are excluded from all prompts.

C.3 TRAINING SETTINGS

Draft architecture and objective. We use an EAGLE-style draft model built from a lightweight single-layer autoregressive decoder. The draft takes cached target hidden states together with shifted token embeddings as input, while the target model, token embedding, and language-model head remain frozen. Model-specific hidden dimensions and attention configurations follow the corresponding target architecture.

Table 11: **Draft training hyperparameters.** Settings are shared across experiments unless specified.

Hyperparameter	Value
Optimizer	AdamW
Peak learning rate	3×10^{-5}
Adam coefficients	$(0.9,\ 0.95)$
Weight decay	0
Learning-rate schedule	Logarithmic warm-up, linear decay
Warm-up steps	2,000
Configured schedule horizon	800,000 optimizer steps
Gradient clipping	0.5
Maximum sequence length	2,048 tokens
Effective batch size	16 (Section 4); 8 (Sections 5 and 6)
Training precision	BF16
Hidden-state noise	Uniform, coefficient 0.2
Gradient checkpointing	Enabled
Random seed	123

We optimize the trainable draft parameters with a combination of a SmoothL1 hidden-state regression loss and a soft-target cross-entropy loss, weighted by 1.0 and 0.1, respectively. The regression loss aligns draft hidden states with target features, while the cross-entropy term aligns their token distributions through the frozen language-model head. Both losses are averaged over supervised response positions retained after truncation, with the same loss weights across all experiments.

Optimization. Table 11 lists the training hyperparameters. The effective batch size is fixed within each experiment, using gradient accumulation when needed. The scheduler horizon specifies the learning-rate decay schedule rather than the actual number of training steps; training duration follows the epoch counts below.

Training schedules and initialization. Data compositions and selection strategies are described in Sections 4, 5, and 6. The supervision-regime experiments in Section 4 train each draft for ten epochs from the same initialization. The experiments in Sections 5 and 6 are separate continuation branches, rather than successive training stages. Unless otherwise specified in the warm-up ablation, each branch starts from the same one-epoch Mixed warm-up checkpoint and trains for ten additional epochs on its selected dataset. When initializing each branch, we load the draft weights. The learning-rate schedule continues from the warm-up endpoint, while the AdamW (Loshchilov & Hutter, 2019) optimizer states are reset. Subsequent resumes restore the complete training state.

The TXT→MM baseline starts from the same warm-up checkpoint and uses the same selected examples as MNRS-JOINT, training on the text subset for ten epochs and then on the multimodal subset for ten epochs. Each selected sample is presented ten times in either schedule, matching total sample exposure between sequential and joint training. Optimizer and scheduler states are retained across the TXT-to-MM transition.