SPROUT: SYNTAX-PROGRESSIVE DENOISING FOR DETAILED CAPTIONING IN DIFFUSION VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) are increasingly expected to produce long, detailed image descriptions, yet additional detail often comes with increased hallucination, as attributes, relations, and secondary objects are generated from linguistic priors rather than visual evidence. We attribute part of this failure to flat caption supervision, which treats sentence skeletons and fine-grained modifiers as an undifferentiated sequence. We introduce **SPROUT** (**S**yntax-**PR**ogressive **O**rdered **U**nmasking **T**raining), a syntax-progressive framework for detailed captioning in masked diffusion VLMs that turns the denoising axis into a within-example curriculum. We construct a dataset in which each human-written caption is decomposed into five nested levels, yielding a syntactic tier for every word. During training, a tier-conditioned corruption schedule masks fine-detail tokens earlier than skeleton tokens, shifting detail supervision toward contexts where the ground-truth sentence skeleton is visible. SPROUT further incorporates schedule-consistent weighting, same-tier block masking, and termination supervision, while requiring no architectural changes or additional inference-time computation. Across hallucination benchmarks (CHAIR, AMBER, THRONE, and OpenCHAIR) and detailed-captioning benchmarks (Image Paragraph, IIW-400, LN-COCO, and DOCCI-test), SPROUT improves fine-grained faithfulness over both the pretrained backbone and a same-data uniform-SFT baseline while preserving coverage and descriptive quality. It also retains general multimodal capability on MME and MMBench. Tier-stratified analysis further reveals emergent skeleton-first generation under the unchanged confidence decoder.

1 INTRODUCTION

Detailed image captioning has become an important capability of vision-language models (VLMs) (Liu et al., 2023; Li et al., 2024; Bai et al., 2025). Applications in retrieval, accessibility, and multimodal agents increasingly rely on long-form descriptions rather than short image tags. A high-quality detailed caption should cover salient objects, attributes, and relations while remaining faithful to the image. These goals are difficult to satisfy jointly. Longer captions cover more visual content but also increasingly hallucinate objects, attributes, and relations (Rohrbach et al., 2018; Li et al., 2023; Wang et al., 2023a; Kaul et al., 2024). Conversely, methods that reduce hallucination often suppress detail and produce shorter, less informative descriptions. The goal is therefore to improve faithfulness without sacrificing descriptive coverage or caption quality.

Hallucination is not distributed uniformly within a caption. Prior work suggests that models often capture the core scene correctly, while errors concentrate in fine-grained modifiers, secondary objects, relations, and other detailed content (Wang et al., 2023a; Ben-Kish et al., 2024). For example, a model may correctly recognize that a dog is sitting on a couch while inventing the color of an occluded collar. We argue that this pattern is partly caused by the structure of standard caption supervision. A long caption is typically treated as a flat token sequence, giving the same learning signal to the sentence skeleton and to fine-grained modifiers. As a result, detail tokens may be supervised when much of the core scene description is unavailable, encouraging prediction from linguistic regularities rather than from an established visual structure.

A natural alternative is to learn core structure before detail, following the developmental progression from short subject–verb–object utterances to progressively richer modification (Brown, 2009).

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.

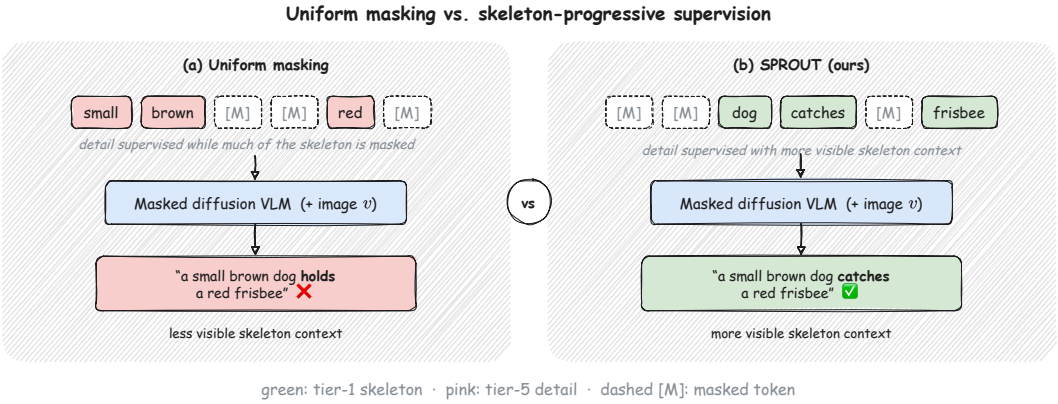

Figure 1: **Uniform masking vs. SPROUT.** Uniform masking may supervise detail words with limited visible ground-truth skeleton context. SPROUT reallocates the same per-token supervision budget toward states with more visible skeleton context, encouraging fine-grained details to be learned after the core scene structure is established.

Conventional curriculum learning follows this idea by presenting easier examples before harder ones (Bengio et al., 2009). However, data-ordering curricula have shown limited gains at scale (Hu et al., 2024; Charpentier et al., 2025). More fundamentally, changing the order of training examples affects when a caption is observed, but not how different parts of that caption are learned. It therefore provides no direct mechanism for enforcing a skeleton-first and detail-later structure within each example. Masked diffusion VLMs (Nie et al., 2026; You et al., 2026) provide a natural within-example curriculum axis. Rather than generating text strictly from left to right, they iteratively denoise a fixed response canvas and are trained on partially observed versions of the same target. Their denoising process often exhibits coarse-to-fine behavior, with global structure emerging before local detail (He et al., 2023; Wang et al., 2023b). This suggests a different form of curriculum. Instead of deciding only which examples are seen first, we can control the contexts in which different parts of the same caption are supervised.

We instantiate this idea as **SPROUT** (**S**yntax-**PR**ogressive **O**rdered **U**nmasking **T**raining), a syntax-progressive framework for detailed captioning in masked diffusion VLMs that turns the denoising axis into a within-example curriculum. SPROUT has two main components. First, we construct a dataset in which each human-written detailed caption is decomposed into five nested levels using a deletion-only and automatically validated procedure. The earliest level in which a word appears defines its syntactic tier, ranging from the tier-1 sentence skeleton to tier-5 fine-grained detail. The intermediate captions are used only to derive per-word tier labels, while the full human-written caption remains the training target. Second, we introduce a tier-conditioned masked-diffusion objective that maps syntactic tiers onto denoising time. Fine-detail tokens receive larger masking probabilities earlier in forward time, while skeleton tokens remain visible for longer. Consequently, detail tokens are supervised more often in corrupted states where the ground-truth tier-1 skeleton is visible. The schedule preserves each token's overall supervision mass while increasing skeleton visibility during detail supervision. We further use same-tier block masking to avoid fragmented contexts within contiguous syntactic units. We also add lightweight termination supervision for variable-length captions. All modifications are restricted to training. At inference time, SPROUT uses the original model architecture and confidence-based decoding algorithm. It requires no tier labels, auxiliary classifiers, additional model passes, or extra decoding steps. SPROUT therefore changes the organization of the training signal while leaving the inference procedure unchanged.

Across hallucination, detailed-captioning, and general VLM benchmarks, SPROUT improves fine-grained faithfulness over both the pretrained backbone and a same-data uniform-SFT baseline while preserving coverage and caption quality. Tier-stratified analysis further reveals an emergent skeleton-first generation pattern under the backbone's unchanged decoder. These results suggest that denoising time can serve as a within-example syntactic curriculum for faithful detailed captioning.

Our contributions are threefold. First, we introduce a detailed-caption dataset with nested syntactic levels and per-word tier labels, together with a tier-stratified hallucination evaluation protocol. Second, we propose SPROUT, a syntax-progressive masked-diffusion training method that aligns

syntactic tiers with denoising time while leaving inference unchanged. Third, we provide theoretical and empirical evidence that reallocating detail supervision toward skeleton-visible contexts improves fine-grained faithfulness without sacrificing descriptive coverage or caption quality.

2 DATASET

SPROUT requires a label absent from existing corpora: for every word in a detailed caption, how essential it is to the syntactic core. We derive this label by decomposing each caption into five nested levels via *deletion only*, and assigning each word the earliest level in which it survives. Because every simpler caption is a subsequence of the richer one and all constraints are programmatically verified, the resulting labels align exactly to the original token grid.

We seed from DOCCI (Onoe et al., 2024): 14,847 human-written descriptions (avg. 122 words) with essentially zero residual hallucination. This choice avoids confounding hallucination reduction with cleanup of noisy model-generated supervision; indeed, mixing GPT-generated captions into training sharply increases CHAIR_s. For each caption L_5, an LLM produces $L_4 \to L_3 \to L_2 \to L_1$ by chained deletion-only simplification, ending in a single $\leq$ 12-word subject–verb–object sentence. Each step is checked against subsequence and length constraints, and captions failing any check are discarded. Full rules, tolerances, retries, and failure analysis appear in Appendix A.5.

Composed alignments map all levels onto L_5 and assign each word a tier

$$k(j) = \min\{\ell : \text{word } j \text{ appears in } L_\ell\}.$$

For example:

```
  A(1) small(2) brown(2) dog(1) with(3) floppy(3) ears(3)
sitting(1) upright(4) on(1) a(1) gray(4) fabric(4) couch(1)
                ...  warm(5) sunlight(5) ...
```

Tier 1 denotes the syntactic skeleton and tier 5 the finest detail. Punctuation inherits the preceding word's tier, while prompt and padding tokens are assigned tier 0. The intermediate captions $L_1, \ldots, L_4$ never enter the training loss (§3); they only define per-word labels on L_5.

The train split contains 9,207 captions, with mean level lengths 6/18/63/104/123 for $L_1, \ldots, L_5$ and tier distribution 4.9/9.5/36.6/33.9/15.1%. All integrity checks pass with zero violations. Manual inspection and dataset quality analyses are reported in Appendix A.5. We release the dataset, construction and validation code, and a tier-stratified protocol for decomposing CHAIR/AMBER hallucination into object- and attribute-level errors.

3 METHOD

SPROUT aligns syntactic complexity with the denoising axis of a masked diffusion VLM. It uses the tier labels introduced in Section 2 to modify the training corruption process while retaining the original architecture and decoding algorithm. Its core mechanism is a tier-conditioned masking schedule, complemented by same-tier block masking and lightweight termination supervision. Figure 2 summarizes the training procedure.

3.1 MASKED DIFFUSION PRELIMINARIES

Let $x_0 = (x_0^1, \ldots, x_0^L)$ denote the response tokens, and let v denote the visual condition together with the input prompt. Masked diffusion language models (Sahoo et al., 2024; Nie et al., 2026) define an absorbing forward process. At noise level $t \in (0, 1]$, each response token is replaced by $[\text{M}]$ independently with probability t:

$$q_{t|0}\big(x_t^i = [\text{M}] \mid x_0^i\big) = t. \tag{1}$$

A bidirectional transformer p_θ predicts masked tokens with the continuous-time NELBO estimator

$$\mathcal{L}_{\text{uni}} = -\,\mathbb{E}_{t\sim\mathcal{U}(0,1]}\mathbb{E}_{x_t\sim q_{t|0}(\cdot|x_0)}\left[\frac{1}{t}\sum_{i:x_t^i=[\text{M}]}\log p_\theta(x_0^i \mid x_t, v)\right]. \tag{2}$$

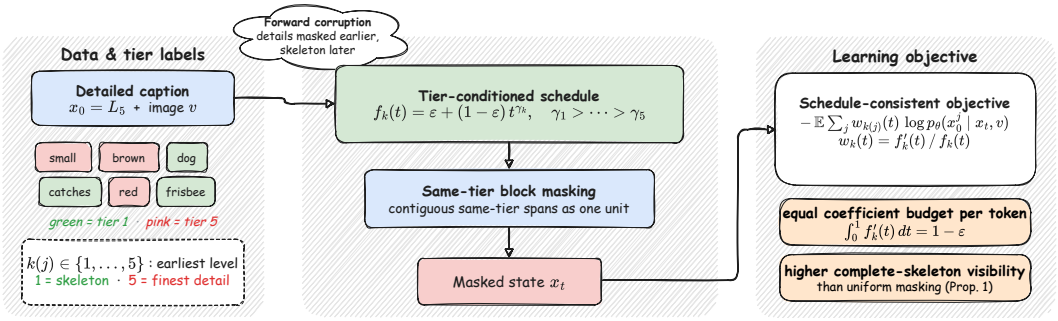

Figure 2: **Overview of SPROUT.** A detailed caption is annotated with deletion-derived syntactic tiers, ranging from the tier-1 sentence skeleton to tier-5 fine details (left). During training, a tier-conditioned forward process masks fine details earlier than skeleton tokens and jointly corrupts contiguous spans within the same tier (middle). The corresponding hazard-weighted objective preserves the total coefficient mass assigned to each token while shifting detail supervision toward states in which the complete tier-1 skeleton is visible (right). Tier labels and the modified corruption process are used only during training, and inference remains unchanged.

Generation starts from an all-[M] canvas and iteratively commits high-confidence predictions.

Uniform corruption treats every position identically, regardless of its linguistic role. Consequently, a detail token is often supervised while a substantial part of the sentence skeleton is also masked. SPROUT changes the distribution of such training contexts without reducing the total loss weight assigned to detail tokens.

3.2 Syntax-progressive denoising

Syntactic tiers. Each word in the full caption $x_0 = L_5$ is assigned a tier $k(j) \in \{1,\dots,5\}$ by the nested simplification procedure in Section 2. Tier 1 contains the subject–verb–object skeleton, whereas higher tiers add progressively finer modifiers and secondary content. The intermediate captions are used only to derive these labels; all training targets remain the original full caption L_5. Untiered response positions use $k = 0$.

Tier-conditioned forward process. We replace the shared masking probability in equation 1 with a tier-specific schedule:

$$q_{t|0}\left(x_t^j = [\mathrm{M}] \mid x_0^j\right) = f_{k(j)}(t), \qquad f_k(t) = \varepsilon + (1-\varepsilon)t^{\gamma_k}. \tag{3}$$

The exponents satisfy $\gamma_1 > \gamma_2 > \cdots > \gamma_5$, with $\gamma_0 = 1$. For $t \in (0,1)$, a smaller exponent gives a larger masking probability; therefore fine-detail tokens are absorbed earlier in forward time and revealed later in reverse time. The endpoint $f_k(1) = 1$ preserves the all-mask initialization used by standard inference. When $\gamma_k \equiv 1$ and $\varepsilon \to 0$, the process reduces to standard masked diffusion.

Schedule-consistent training objective. Changing the forward schedule also changes the variational weight. For the process in equation 3, we optimize

$$\mathcal{L}_{\mathrm{SPROUT}} = -\,\mathbb{E}_{t\sim\mathcal{U}(0,1]}\mathbb{E}_{x_t}\left[\sum_{j:x_t^j=[\mathrm{M}]} w_{k(j)}(t)\log p_\theta(x_0^j \mid x_t, v)\right], \tag{4}$$
$$w_k(t) = \frac{f_k'(t)}{f_k(t)} = \frac{(1-\varepsilon)\gamma_k t^{\gamma_k-1}}{\varepsilon + (1-\varepsilon)t^{\gamma_k}}.$$

The weight $w_k = f_k'/f_k$ is the masking hazard of the tier-specific schedule. Replacing it with $1/f_k$ would introduce a tier- and time-dependent reweighting and would no longer optimize the corresponding absorbing-state variational objective. Appendix A.1.2 derives equation 4. For $\varepsilon > 0$, training omits a boundary reconstruction term that vanishes as $\varepsilon \to 0$; we use $\varepsilon = 10^{-3}$ throughout.

Supervision allocation. Each independently sampled masking variable is a *corruption unit*: a single token under independent masking, or a same-tier span under the block-masked process introduced below. For a tier-k target, the density of its weighted loss contributions at time t is $f'_k(t)$, because it is masked with probability $f_k(t)$ and weighted by $f'_k(t)/f_k(t)$. The schedule therefore redistributes a fixed coefficient budget across noise levels rather than down-weighting fine details.

Proposition 1. *Under the forward process in equation 3 and the objective in equation 4, every tier-k token receives total coefficient mass $\int_0^1 f'_k(t)\,dt = 1 - \varepsilon$. Moreover, consider a tier-k target with $k \geq 2$, and suppose that the tier-1 skeleton of its caption consists of M independently corrupted units. Averaged over the target's weighted supervision events, the probability that all M skeleton units are visible is*

$$G^{\text{all}}_{k\leftarrow 1}(M) = (1-\varepsilon)^M \prod_{r=1}^{M} \frac{r\gamma_1}{\gamma_k + r\gamma_1}. \tag{5}$$

The proof is given in Appendix A.2. Under uniform corruption, equation 5 reduces to $(1-\varepsilon)^M/(M+1)$. Since $\gamma_k < \gamma_1$ for every detail tier, each factor in equation 5 exceeds its uniform counterpart $r/(r+1)$. SPROUT therefore increases the probability that the complete ground-truth tier-1 skeleton is visible during detail supervision for every $M \geq 1$, while preserving the coefficient budget of each detail token. The $M = 1$ case recovers the pairwise rate $(1-\varepsilon)\gamma_1/(\gamma_1+\gamma_k)$.

This result concerns the availability of the ground-truth skeleton in training contexts. It does not imply that confidence decoding follows the same order at test time; we evaluate tier-stratified commit order separately in Section 4.

Schedule parameters. We use $\gamma = (4.5, 3, 1.5, 0.9, 0.6)$. For fixed M, equation 5 depends only on exponent ratios, which control the relative separation between tiers. A common rescaling leaves the visibility profile unchanged. We choose the global scale so that $\min_k \gamma_k = 0.6 > \frac{1}{2}$; under the bounded-loss analysis in Appendix A.3, this gives the hazard-weight component a finite second moment and permits sampling $t \sim \mathcal{U}(0,1]$ without truncating the noise interval. Appendix Table 6 reports complete-skeleton visibility for representative values of M.

Same-tier block masking. Independent token masking can produce fragmented contexts, such as a half-visible noun phrase or modifier. We therefore partition the response into maximal contiguous runs of equal-tier tokens and draw one masking variable per run. All positions in a run are visible or masked together; tier-0 positions remain singleton units. We refer to these runs as *same-tier spans*.

Proposition 2. *Consider the span-absorbing process that independently masks each tier-k same-tier span with probability $f_k(t)$. Under the factorized SUBS-style reverse parameterization, equation 4 is the continuous-time NELBO of this process up to the same omitted boundary reconstruction term. In addition, every token retains marginal masking probability $f_{k(j)}(t)$, distinct spans remain independent, and partial masking within a span has probability zero.*

Appendix A.1 gives the proof. Proposition 2 shows that block masking changes the joint distribution of denoising contexts while preserving the per-token schedule and the allocation result of Proposition 1, with M equal to the number of tier-1 same-tier spans.

Termination supervision. Masked diffusion decodes on a fixed response canvas, whereas ground-truth captions have variable length. To explicitly supervise termination, we fill positions after the ground-truth response with $\langle\text{eot}\rangle$ tokens up to the training canvas length. These positions use tier 0 and the standard schedule $f_0(t)$, so they enter equation 4 without syntax-dependent ordering. This component is orthogonal to the tier schedule and is ablated separately.

3.3 Inference

SPROUT changes training only. At test time, we use the backbone's original confidence-based reverse procedure, with no tier labels, schedule computation, auxiliary classifier, extra model passes, or decoding steps. The training objective promotes skeleton-rich contexts for detail prediction but does not force confidence decoding to follow the forward-process marginals. We therefore treat skeleton-first decoding as an empirical property and measure tier-stratified commit order in Section 4. Appendix A.4 examines the training–inference mismatch.

Table 1: We evaluate detailed captions on CHAIR, AMBER, THRONE, and OpenCHAIR. SPROUT reduces hallucination of LLaDA-V across all four benchmarks, ranking best or second-best on 7 of 9 metrics.

	CHAIR (500 COCO)			AMBER		THRONE			OPEN-CHAIR	
Model	$C_i \downarrow$	$C_s \downarrow$	#Obj $\uparrow$	CH $\downarrow$	Cov. $\uparrow$	P $\uparrow$	$F_{0.5} \uparrow$	$F_1 \uparrow$	CH $\downarrow$	Rank $\downarrow$
VISION DIFFUSION LMS										
LaViDa	7.9	**12.6**	2.7	7.5	34.3	71.6	64.5	56.3	27.1	8.8
MMaDA	10.0	26.8	4.8	7.4	47.5	64.1	63.2	61.8	22.8	9.4
LLaDA-O	7.4	31.0	8.7	7.1	62.3	71.9	70.4	68.3	17.4	5.2
LLaDA2.0-Uni	7.8	36.8	8.6	**5.6**	69.5	72.3	72.6	73.1	16.7	4.1
AUTOREGRESSIVE VLMS										
LLaVA-1.5	16.1	50.6	7.1	7.5	49.5	66.4	68.3	71.3	24.5	9.0
InternVL3.5	9.9	36.6	6.0	7.1	62.4	66.2	68.2	71.4	18.7	7.4
Qwen3-VL	9.1	42.6	7.8	7.9	**70.5**	75.2	74.7	73.9	18.6	5.2
DeepSeek-VL	7.3	23.8	5.0	7.5	60.2	69.8	70.0	70.3	19.5	6.4
Fuyu-8B	17.6	60.2	7.3	9.7	48.9	67.4	68.6	70.3	22.1	9.6
LLADA-V + OURS										
LLaDA-V (untrained)	7.4	31.4	9.0	9.3	66.7	67.9	68.9	70.4	29.1	6.8
+ uniform SFT	7.8	18.8	9.1	9.5	59.8	**84.8**	79.2	72.0	13.7	4.2
+ SPROUT (ours)	**6.0** -1.4	28.0 -3.4	**9.3** $+0.3$	7.0 -2.3	67.2 $+0.5$	82.3 $+14.4$	**79.6** $+10.7$	**75.9** $+5.5$	**10.5** -18.6	**1.9** -4.9

4 EXPERIMENTS

Our experiments are organized around four questions. **(RQ1)** Does aligning syntactic complexity with denoising time improve fine-grained faithfulness *without* sacrificing coverage and descriptive detail (§4.2)? **(RQ2)** Does the gain require the denoising axis specifically, or can a data-ordering curriculum over the same captions achieve it (§4.3)? **(RQ3)** Which design choices carry the effect, including constituent block masking and the five-level tier granularity (§4.4)? The role of the schedule exponents γ and the floor ε is analyzed in Appendix A.3. **(RQ4)** Does the skeleton-first generation order predicted in §3.3 actually emerge under unchanged confidence decoding (§4.5)?

4.1 EXPERIMENTAL SETUP

Baselines. (i) *Vision diffusion LMs*: LaViDa (Li et al., 2026), MMaDA (Yang et al., 2026), and LLaDA-O (You et al., 2026); and (ii) *autoregressive VLMs*: LLaVA-1.5 (Liu et al., 2023), InternVL3 (Zhu et al., 2025), Qwen3-VL (Bai et al., 2025), DeepSeek-VL (Lu et al., 2024), and Fuyu-8B (Bavishi et al., 2023). Controlled comparisons use the LLaDA-V backbone: untrained, uniform SFT, and easy→hard / hard→easy curricula on the same data, ordered by an independent DeepSeek complexity score.

Benchmarks. *Hallucination:* $\text{CHAIR}_{i/s}$ (Rohrbach et al., 2018), THRONE (Kaul et al., 2024), AMBER (Wang et al., 2023a), and OpenCHAIR (Ben-Kish et al., 2024). *Detailed captioning:* Image Paragraph (Krause et al., 2017), IIW-400 (Garg et al., 2024), LN-COCO (Pont-Tuset et al., 2020), and held-out DOCCI (Onoe et al., 2024). *General capability:* MME, MMBench, MMStar, MMMU, and MMVP. All models use identical fixed subsets where applicable: 500 COCO images for CHAIR/THRONE, 1,004 AMBER, 1,000 OpenCHAIR and Image Paragraph/LN-COCO, and 520 DOCCI. THRONE, OpenCHAIR, and DCScore use a GPT-5 judge. We pair hallucination metrics with completeness (#Obj, Cover, SPICE) to account for under-description.

4.2 MAIN RESULTS: FAITHFULNESS WITHOUT SACRIFICING COVERAGE (RQ1)

Hallucination. In Table 1, SPROUT improves over the untrained backbone on all nine metrics and attains the best mean rank of all twelve models (1.9, vs. 4.1 for the strongest baseline). The gains do not come from describing less: CHAIR_i drops from 7.4 to 6.0 while #Obj *rises* ($9.0 \rightarrow 9.3$) and AMBER coverage is preserved. Uniform SFT illustrates the opposite trade-off: it reaches the high-

Table 2: Standard captioning metrics on Image Paragraph, IIW-400, LN-COCO, and DOCCI-test. SPROUT reduces hallucination without compromising caption quality.

	Image Paragraph				IIW-400		LN-COCO		DOCCI-test				
Model	B ↑	M ↑	R ↑	S ↑	M ↑	S ↑	M ↑	S ↑	B ↑	M ↑	R ↑	S ↑	Rank ↓
Vision diffusion LMs													
LaViDa	0.0	4.7	14.1	12.7	5.1	7.4	11.3	11.6	0.9	6.5	13.7	8.3	11.8
MMaDA	4.0	12.9	23.9	14.4	6.6	8.1	13.9	14.7	2.0	8.8	19.6	9.2	9.5
LLaDA-O	4.1	12.9	22.5	16.4	11.6	10.4	14.1	11.9	5.3	14.3	22.5	10.8	7.5
LLaDA2.0-Uni	4.3	17.0	23.5	15.8	15.1	12.2	16.1	12.9	5.7	17.6	24.3	12.5	4.5
Autoregressive VLMs													
LLaVA-1.5	5.7	14.6	25.6	16.6	7.8	10.4	14.9	**16.1**	3.0	10.4	21.5	11.8	6.5
InternVL3.5	4.9	18.2	24.5	15.7	12.5	11.5	16.7	13.8	5.7	15.4	23.3	12.3	4.8
Qwen3-VL	3.1	17.9	22.0	13.8	**17.1**	11.5	15.0	10.6	3.9	17.0	22.3	9.2	7.5
DeepSeek-VL	2.9	10.0	20.3	16.6	10.7	**11.7**	**17.1**	15.2	4.7	13.6	23.2	12.9	6.2
Fuyu-8B	4.0	15.9	23.6	15.5	7.3	8.9	15.7	15.5	2.5	9.7	19.4	10.2	8.4
LLaDA-V + Ours													
LLaDA-V (untrained)	4.7	18.7	24.6	16.1	15.1	11.0	16.9	10.9	3.6	16.9	23.7	11.6	5.3
+ uniform SFT	5.8	19.2	26.6	17.1	11.5	10.5	16.8	11.9	4.8	14.8	24.5	15.0	4.1
+ SPROUT (ours)	**6.1** +1.4	**19.6** +0.9	**26.8** +2.2	**18.7** +2.6	12.2 −2.9	11.6 +0.6	**17.1** +0.2	12.7 +1.8	**8.4** +4.8	**17.7** +0.8	**28.8** +5.1	**15.8** +4.2	**2.0** −3.3

est THRONE precision (84.8) by sacrificing recall (F_1 72.0) and coverage (66.7 → 59.8), whereas SPROUT improves precision and both F-scores jointly.

Descriptive quality. Table 2 confirms that faithfulness is not bought with quality: SPROUT improves on 11 of 12 metrics over the backbone (best on 9; mean rank 2.0). The contrast is sharpest out of domain — on IIW-400 and LN-COCO, uniform SFT falls *below* the untrained backbone, while SPROUT's only regression anywhere is IIW-400 METEOR (−2.9). How supervision is allocated within each caption thus matters more than the captions themselves.

Table 3: General capability retention. MME p/c, MMBench, MMStar, MMMU, MMVP. Best in bold, second best underlined.

Training strategy	MME_p ↑	MME_c ↑	MMB↑	MMStar↑	MMMU↑	MMVP↑
LLaDA-V (untrained)	**1510.8**	513.9	**83.25**	**60.2**	**49.0**	**29.3**
+ SPROUT	1477.8	**536.8**	81.87	58.9	48.9	**29.3**

General capability. Table 3 verifies that the specialization does not erode general multimodal ability: after SPROUT training, MME-cognition improves (513.9 → 536.8), MMVP is unchanged, and the remaining benchmarks stay within 1.4 points of the backbone. Qualitative examples in Figure 5 illustrate the typical failure mode that is corrected: the backbone commits early to plausible but wrong details (a yellow canopy, a forward-facing subject), while SPROUT first states the correct scene structure and then adds details consistent with it.

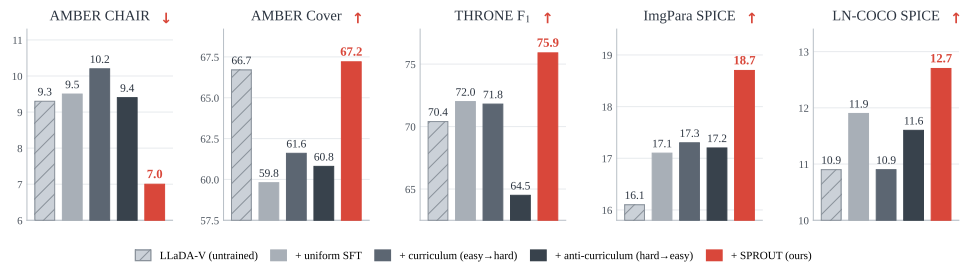

Figure 3: Data-axis curricula vs. SPROUT. Easy→hard and hard→easy data ordering match uniform SFT, while SPROUT's denoising-axis curriculum jointly improves faithfulness and completeness.

4.3 Denoising Axis vs. Data Ordering (RQ2)

If SPROUT's gains merely reflected "easy content first," data-ordering curricula should reproduce them. Figure 3 shows otherwise: ordering captions by independent-LLM complexity scores either way fails to do so. AMBER coverage drops to 61.6/60.8 (vs. 67.2), CHAIR stays at 9.5–10.2 (vs. 7.0), and on THRONE F_1 easy→hard matches uniform SFT (71.8 vs. 72.0) while hard→easy underperforms the backbone (64.5 vs. 70.4), far from SPROUT's 75.9. What matters is not when captions are seen, but which within-caption contexts are supervised along the denoising axis.

4.4 Design-Choice Ablations (RQ3)

Table 4: Ablation of masking schedule and structure. † denotes degenerate under-description.

		AMBER		CHAIR (500 COCO)		THRONE
Schedule	Masking	Cover↑	CHAIR↓	$CHAIR_i$ ↓	#Obj↑	F_1 ↑
Uniform	per-token	59.8	9.5	7.8	9.1	72.0
Uniform	block	66.2	8.2	7.3	10.1	72.9
Tier-cond.	per-token	22.5†	1.7†	1.14†	0.7†	37.1†
Tier-cond.	block (SPROUT)	**67.2**	**7.0**	**6.0**	9.3	**75.9**

Schedule vs. block masking. Table 4 disentangles schedule and structure. Under uniform scheduling, block masking raises AMBER coverage from 59.8 to 66.2. With blocks fixed, tier conditioning improves AMBER CHAIR (8.2→7.0), $CHAIR_i$ (7.29→6.0), and THRONE F_1 (72.9→75.9) at similar coverage (66.2 vs. 67.2). Thus, block masking preserves coverage while tier conditioning adds faithfulness; the per-token tier variant collapses, indicating strong coupling.

Table 5: Ablations of tier construction and granularity with block masking enabled. **(a)** Tier semantics and span structure. **(b)** Tier-granularity ablation. † indicates degenerate under-description.

(a) Tier construction and structure

	CHAIR (500 COCO)		AMBER		THRONE	ImgPara
Model	$CHAIR_i$ ↓	#Obj↑	CHAIR↓	Cover↑	F_1 ↑	SPICE↑
SPROUT	**6.0**	9.3	7.0	**67.2**	**75.9**	**18.7**
Matched random spans	7.4	**10.7**	7.1	66.9	71.1	18.3
Shuffled tiers	3.0†	1.3	2.6†	36.8	54.8	17.2

(b) Tier granularity

	CHAIR (500 COCO)		AMBER		THRONE	ImgPara
Model	$CHAIR_i$ ↓	#Obj↑	CHAIR↓	Cover↑	F_1 ↑	SPICE↑
SPROUT (5-tier)	6.0	9.3	7.0	**67.2**	**75.9**	18.7
2-tier schedule	7.1	**10.4**	7.5	65.2	73.8	**18.8**
3-tier schedule	**5.7**	7.8	**6.3**	63.2	71.4	18.4

Tier construction and structure. Table 5(a) tests whether arbitrary structured corruption can reproduce SPROUT's gains. Length-matched random spans preserve coverage but worsen $CHAIR_i$ and THRONE F_1, while shuffled tiers achieve low CHAIR only by collapsing coverage and object count. Additional controls are reported in Appendix A.6.

Tier granularity. Table 5(b) varies the number of tiers with block masking fixed. Two tiers retain coverage but lower THRONE F_1; three tiers reduce CHAIR at the cost of coverage and F_1. Five tiers provide the best overall faithfulness–coverage balance.

4.5 Emergent Skeleton-First Decoding (RQ4)

SPROUT modifies training only, so any change in generation order must be emergent. Figure 4 reports the mean normalized commit rank of each tier under the unchanged confidence decoder, where lower ranks indicate earlier commitment. After SPROUT training, commit order becomes monotone: skeleton tokens are committed first (T1 rank 0.436), while fine details come last. The untrained backbone instead shows a flatter, non-monotone profile. Reversing γ causes details to unmask before the skeleton and reverses the trend at fine tiers, indicating that the ordering is induced by the training schedule rather than confidence ranking itself. Together with Appendix A.4, these results suggest that SPROUT reshapes the training distribution while leaving inference unchanged.

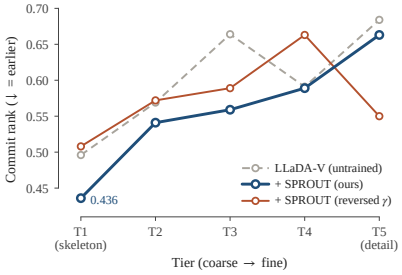

Figure 4: Tiered commit order.

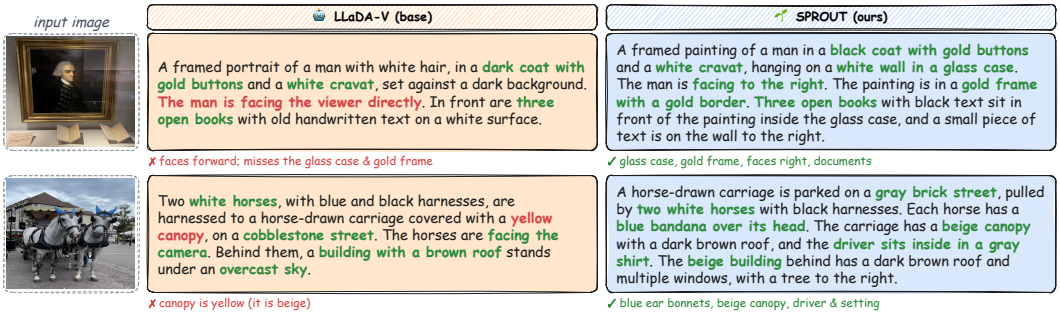

Figure 5: Qualitative comparison between the untrained LLaDA-V backbone and SPROUT.

4.6 Qualitative Analysis

Figure 5 compares captions from the LLaDA-V backbone and SPROUT. The backbone captures the objects but introduces unsupported details, such as an incorrect viewing direction and canopy color. In contrast, SPROUT preserves the scene structure and produces finer details grounded in the image, including the glass case, gold frame, beige canopy, and carriage driver. These examples illustrate that SPROUT improves fine-grained faithfulness without reducing descriptive detail.

5 Related Work

5.1 Faithful Detailed Image Captioning

Detailed captioning requires both coverage and faithfulness. Recent work improves supervision and evaluation via human-written captions (Onoe et al., 2024), GPT-4V/Gemini-assisted captions and benchmarks (Chen et al., 2024; Dong et al., 2024), dense caption corpora (Singla et al., 2024), and unit-level or preference-based evaluation (Ye et al., 2025; Cheng et al., 2025). Hallucination persists across objects, attributes, relations, and localized spans (Wang et al., 2023a; Kaul et al., 2024; Saito et al., 2026; Wang et al., 2026). Existing mitigation often uses inference-time intervention, reward data, or reduced detail (Leng et al., 2024; Suo et al., 2025; Huang et al., 2024). SPROUT instead restructures within-caption supervision with unchanged inference. Our ablations show GPT-generated augmentation can increase hallucination, motivating human-written seeds.

5.2 Curriculum Learning and Masked Diffusion VLMs

Curriculum learning orders examples from easy to hard (Elman, 1993; Bengio et al., 2009; Kumar et al., 2010; Platanios et al., 2019; Graves et al., 2017). Linguistic curricula mirror child acquisition from simple to richer structures (Brown, 2009), but BabyLM finds limited gains from complexity-based ordering (Hu et al., 2024; Charpentier et al., 2025). Discrete and masked diffusion models generate via iterative denoising (Austin et al., 2021; Li et al., 2022; He et al., 2023; Lou et al., 2023). Masked diffusion uses an absorbing-state NELBO (Sahoo et al., 2024) and scales to LLaDA (Nie et al., 2026), LLaDA-V (You et al., 2026), and LaViDa (Li et al., 2026). Denoising exhibits coarse-to-fine behavior (He et al., 2023; Wang et al., 2023b), while corruption schedules and loss weighting affect training (Ho et al., 2020; Nichol & Dhariwal, 2021; Kingma et al., 2021; Sahoo et al., 2024). SPROUT moves curricula from data order to denoising time, aligning syntactic tiers with token-level corruption while retaining masked diffusion as a special case.

6 Conclusion

We presented SPROUT, a syntax-progressive framework for detailed image captioning with masked diffusion VLMs. By aligning syntactic tiers with denoising time, SPROUT shifts detail supervision toward skeleton-visible contexts while leaving inference unchanged. Across hallucination, detailed-captioning, and general benchmarks, SPROUT improves faithfulness without sacrificing coverage or caption quality. Tier-stratified analysis reveals an emergent skeleton-first generation order, supporting denoising time as a natural axis for within-example curriculum learning.

AI USE STATEMENT

In this work, we used generative AI tools to assist with mathematical derivations, coding, and improving the clarity and readability of the manuscript. All AI-assisted mathematical derivations were independently checked by the authors, all AI-assisted code was verified and tested for correctness, and all AI-assisted text was reviewed and revised by the authors. We did not use generative AI tools to generate research hypotheses, design the methodology or experiments, or analyze and interpret experimental results. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sagnak Tasırlar. Fuyu-8b: A multimodal architecture for ai agents. *URL: https://www. adept. ai/blog/fuyu-8b*, 3, 2023.

Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, and Hadar Averbuch-Elor. Mitigating open-vocabulary caption hallucinations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22680–22698, 2024.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

Roger Brown. *A first language: The early stages*. Harvard University Press-T, 2009.

Lucas Georges Gabriel Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Y Hu, Jing Liu, Jaap Jumelet, Tal Linzen, Aaron Mueller, Candance Ross, et al. Findings of the third babylm challenge: Accelerating language modeling research with cognitively plausible data. In *Proceedings of the First BabyLM Workshop*, pp. 399–420, 2025.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pp. 370–387. Springer, 2024.

Kanzhi Cheng, Wenpo Song, Jiaxin Fan, Zheng Ma, Qiushi Sun, Fangzhi Xu, Chenyang Yan, Nuo Chen, Jianbing Zhang, and Jiajun Chen. Caparena: Benchmarking and analyzing detailed image captioning in the llm era. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14077–14094, 2025.

Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. Benchmarking and improving detail image caption. *arXiv preprint arXiv:2405.19092*, 2024.

Jeffrey L Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99, 1993.

Roopal Garg, Andrea Burns, Burcu Karagol-Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Michael Baldridge, and Radu Soricut. Imageinwords: Unlocking hyper-detailed image descriptions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 93–127, 2024.

Alex Graves, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *international conference on machine learning*, pp. 1311–1320. Pmlr, 2017.

Zhengfu He, Tianxiang Sun, Qiong Tang, Kuanning Wang, Xuan-Jing Huang, and Xipeng Qiu. Diffusionbert: Improving generative masked language models with diffusion models. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 4521–4534, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Michael Y Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. Findings of the second babylm challenge: Sample-efficient pretraining on developmentally plausible corpora. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pp. 1–21, 2024.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024.

Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, CJ Taylor, and Stefano Soatto. Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27228–27238, 2024.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 317–325, 2017.

M Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. *Advances in neural information processing systems*, 23, 2010.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Shufan Li, Konstantinos Kallidromitis, Hritik Bansal, Akash Gokul, Yusuke Kato, Kazuki Kozuka, Jason Kuen, Zhe Lin, Kai-Wei Chang, and Aditya Grover. Lavida: A large diffusion language model for multimodal understanding. *Advances in Neural Information Processing Systems*, 38: 105101–105134, 2026.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35: 4328–4343, 2022.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2023.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *Advances in Neural Information Processing Systems*, 38:50608–50646, 2026.

Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, et al. Docci: Descriptions of connected and contrasting images. In *European Conference on Computer Vision*, pp. 291–309. Springer, 2024.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. Competence-based curriculum learning for neural machine translation. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 1162–1172, 2019.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *European conference on computer vision*, pp. 647–664. Springer, 2020.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, 2018.

Subham S Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184, 2024.

Kuniaki Saito, Risa Shinoda, Shohei Tanaka, Tosho Hirasawa, Fumio Okura, and Yoshitaka Ushiku. Haldec-bench: Benchmarking hallucination detector in image captioning. *arXiv preprint arXiv:2603.15253*, 2026.

Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. From pixels to prose: A large dataset of dense image captions. *arXiv preprint arXiv:2406.10328*, 2024.

Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. Octopus: Alleviating hallucination via dynamic contrastive decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29904–29914, 2025.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023a.

Xinran Wang, Yuxuan Zhang, Xiao Zhang, Haolong Yan, Muxi Diao, Songyu Xu, Zhonghao Yan, Hongbing Li, Kongming Liang, and Zhanyu Ma. Detailverifybench: A benchmark for dense hallucination localization in long image captions. *arXiv preprint arXiv:2604.05623*, 2026.

Yingheng Wang, Yair Schiff, Aaron Gokaslan, Weishen Pan, Fei Wang, Christopher De Sa, and Volodymyr Kuleshov. Infodiffusion: Representation learning using information maximizing diffusion models. In *International conference on machine learning*, pp. 36336–36354. PMLR, 2023b.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *Advances in Neural Information Processing Systems*, 38:138867–138907, 2026.

Qinghao Ye, Xianhan Zeng, Fu Li, Chunyuan Li, and Haoqi Fan. Painting with words: Elevating detailed image captioning with benchmark and alignment learning. *arXiv preprint arXiv:2503.07906*, 2025.

Zebin You, Shen Nie, Xiaolu Zhang, JUN ZHOU, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. Llada-v: Large language diffusion models with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10093–10105, 2026.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Appendix

A.1 Proofs

A.1.1 Setup and notation

Fix a response $x_0 = (x_0^1, \ldots, x_0^L)$ with per-token tiers $k(j)$, and abbreviate $f_j(t) := f_{k(j)}(t)$ for the tier-conditioned schedule of equation 3. Each schedule is continuous and strictly increasing on $[0, 1]$, absolutely continuous on $[0, 1]$, and continuously differentiable on $(0, 1]$, with $f_j(0) = \varepsilon$, $f_j(1) = 1$; write $\alpha_j(t) := 1 - f_j(t)$ for the survival probability.

All corruption processes in this paper are instances of one construction, parameterized by a partition of $\{1, \ldots, L\}$ into *spans*. Given a partition, each span S of tier k carries a single absorption event: with probability $f_k(t)$ every position of S equals $[\mathrm{M}]$ at time t, otherwise the span is intact; absorption is permanent and independent across spans. Equivalently, a span of length r is a super-token over the product vocabulary $\mathcal{V}^r$ subjected to the standard absorbing channel with schedule f_k. Two partitions are used:

- the *trivial partition* into singleton spans recovers the token-factorized process, i.e., the absorbing masking channel

$$q_{t|0}\left(x_t^j \mid x_0^j\right) = \alpha_j(t)\, \delta_{x_0^j}\left(x_t^j\right) + \big(1 - \alpha_j(t)\big)\, \delta_{[\mathrm{M}]}\left(x_t^j\right), \tag{6}$$

  independent across tokens; once masked, a token remains masked, and for $s < t$ the transition kernel has survival ratio $\alpha_j(t)/\alpha_j(s)$ on unmasked tokens;
- the partition into maximal runs of consecutive same-tier tokens (*same-tier spans*), tier-0 positions being singleton spans, gives the *span-absorbing* process of the block-masked sampler.

The reverse model uses the mean-prediction SUBS parameterization (Sahoo et al., 2024): given x_t, the network outputs a distribution over the clean value of every masked position, unmasked positions are copied, and a masked position never transitions to another masked state. For a general partition, the reverse kernel unmasks a span jointly with span prediction $p_\theta(x_0^S \mid x_t, v) = \prod_{j \in S} p_\theta(x_0^j \mid x_t, v)$; this within-span conditional independence is exactly what the factorized denoising head provides, and for singleton spans it is the standard token-wise SUBS kernel.

Standard MDLM as a special case. When $\gamma_k \equiv 1$, $f_k(t) = \varepsilon + (1 - \varepsilon)t$ with hazard $w(t) = (1 - \varepsilon)/(\varepsilon + (1 - \varepsilon)t)$; as $\varepsilon \to 0$ this recovers $f(t) = t$, $w(t) = 1/t$, i.e., Eqs. equation 1–equation 2. With block masking enabled, the uniform block-masked objective is itself the NELBO of the corresponding span-absorbing process by Proposition 2(i), up to the $O(\varepsilon)$ boundary term made explicit in equation 11 (exact as $\varepsilon \to 0$).

A.1.2 Proof of Proposition 2(i); the hazard-weighted NELBO equation 4 as the singleton case

We follow the discrete-to-continuous argument of Sahoo et al. (2024), carried out once for an arbitrary fixed partition into spans; the same-tier-span partition yields Proposition 2(i), and the trivial

partition yields the token-factorized NELBO used in the main text. Partition $[0,1]$ into T intervals with $t_i = i/T$, $s_i = t_{i-1}$. The discrete NELBO is

$$\mathcal{L}_T = \underbrace{\mathbb{E}_q\big[-\log p_\theta(x_0 \mid x_{t_1}, v)\big]}_{\text{reconstruction}} + \sum_{i=2}^{T} \underbrace{\mathbb{E}_q\, D_{\mathrm{KL}}\big[q(x_{s_i} \mid x_{t_i}, x_0) \,\big\|\, p_\theta(x_{s_i} \mid x_{t_i})\big]}_{\mathcal{L}_i} + \text{const}, \tag{7}$$

the constant containing the deterministic prior term at $t = 1$.

Per-span decomposition. Because the forward process factorizes over spans and the span-level SUBS parameterization predicts masked spans conditionally independently given x_t, each KL decomposes, $\mathcal{L}_i = \sum_S \mathcal{L}_i^S$. If a span S of tier k is intact at t_i, the posterior is a point mass and $\mathcal{L}_i^S = 0$. If S is masked at t_i, the posterior over its state at s_i is supported on the two points $\{[\mathrm{M}]^{|S|}, x_0^S\}$ with

$$q\Big(x_{s_i}^S = x_0^S \,\Big|\, x_{t_i}^S = [\mathrm{M}]^{|S|}, x_0\Big) = \frac{\alpha_k(s_i) - \alpha_k(t_i)}{1 - \alpha_k(t_i)} = \frac{f_k(t_i) - f_k(s_i)}{f_k(t_i)}. \tag{8}$$

Define the schedule increment and one-step unmasking probability

$$\Delta f_{k,i} := f_k(t_i) - f_k(s_i), \qquad r_{k,i} := \frac{\Delta f_{k,i}}{f_k(t_i)}.$$

Under the SUBS reverse parameterization, conditional on $x_{t_i}^S = [\mathrm{M}]^{|S|}$, the reverse kernel assigns probability $1 - r_{k,i}$ to remaining masked and probability $r_{k,i} p_\theta(x_0^S \mid x_{t_i}, v)$ to the clean span value. Therefore the per-span KL is exactly

$$\mathcal{L}_i^S = \mathbb{E}_q\left[\mathbb{1}\Big[x_{t_i}^S = [\mathrm{M}]^{|S|}\Big]\, \frac{\Delta f_{k,i}}{f_k(t_i)} \big(-\log p_\theta(x_0^S \mid x_{t_i}, v)\big)\right]. \tag{9}$$

By the factorized head,

$$-\log p_\theta(x_0^S \mid x_t, v) = \sum_{j \in S} -\log p_\theta(x_0^j \mid x_t, v).$$

Continuous-time limit and estimator. For a span S of tier k, define

$$h_S(t) := \mathbb{E}_q\left[\mathbb{1}\Big[x_t^S = [\mathrm{M}]^{|S|}\Big]\, \big(-\log p_\theta(x_0^S \mid x_t, v)\big)\right].$$

The KL sum is therefore the Riemann–Stieltjes sum

$$\sum_{i=2}^{T} \frac{h_S(t_i)}{f_k(t_i)} \big(f_k(t_i) - f_k(s_i)\big).$$

The omitted first interval contributes at most $O(f_k(t_1) - f_k(0)) = O(T^{-\gamma_k})$ under bounded log-loss, and hence vanishes as $T \to \infty$. Thus the sum converges to

$$\int_{(0,1]} \frac{h_S(t)}{f_k(t)}\, df_k(t).$$

This formulation does not require f_k' to remain bounded at $t = 0$. In particular, for the power schedule $f_k(t) = \varepsilon + (1 - \varepsilon) t^{\gamma_k}$ with any $\gamma_k > 0$, f_k is absolutely continuous on $[0,1]$ and

$$df_k(t) = f_k'(t)\, dt$$

almost everywhere. Therefore

$$\mathcal{L}_\infty = \int_0^1 \sum_S \frac{f_k'(t)}{f_k(t)}\, \mathbb{E}_q\left[\mathbb{1}\Big[x_t^S = [\mathrm{M}]^{|S|}\Big]\, \big(-\log p_\theta(x_0^S \mid x_t, v)\big)\right] dt. \tag{10}$$

For completeness, the endpoint $t = 0$ causes no difficulty even when $\gamma_k < 1$ and $f_k'(t)$ diverges as $t \downarrow 0$. Under bounded log-loss, $h_S(t)/f_k(t)$ is bounded because $\Pr(x_t^S = [\mathrm{M}]^{|S|}) = f_k(t)$. Hence the contribution of $(0, \delta]$ is bounded by a constant times

$$f_k(\delta) - f_k(0) = (1 - \varepsilon)\delta^{\gamma_k} \longrightarrow 0 \qquad (\delta \downarrow 0).$$

Thus the Stieltjes integral over $(0,1]$ is well defined and equals the Lebesgue integral in equation 10, including for the sublinear schedules used in our experiments.

Boundary reconstruction term. For $\varepsilon > 0$ the reconstruction term of equation 7 does not vanish in the limit: as $T \to \infty$, $t_1 \to 0$ and each span is masked at t_1 with probability $f_k(t_1) \to \varepsilon$, so the reconstruction term converges to

$$R_\varepsilon := \mathbb{E}\Big[\sum_{S:\, x_{0^+}^S = [\mathrm{M}]^{|S|}} -\log p_\theta\big(x_0^S \mid x_{0^+}, v\big)\Big], \tag{11}$$

where the boundary state x_{0^+} masks each span independently with probability ε. The exact continuous-time NELBO is therefore $\mathcal{L}_{\mathrm{NELBO}} = \mathcal{L}_\infty + R_\varepsilon + \mathrm{const}$. Each span contributes to R_ε with probability ε, so $R_\varepsilon = O(\varepsilon)$ whenever the per-token log-loss is bounded (as for any softmax head with bounded logits); at $\varepsilon = 0$ it vanishes and the identity $\mathcal{L}_{\mathrm{NELBO}} = \mathcal{L}_\infty + \mathrm{const}$ is exact. Training optimizes $\mathcal{L}_\infty$ only, omitting R_ε ($\varepsilon = 10^{-3}$ in our experiments); every "exact NELBO" statement in this paper is accordingly to be read modulo this explicitly bounded boundary term, as stated in Proposition 2(i) and §3.2.

Since all tokens of a masked span are masked simultaneously, the hazard attached to the span distributes over its tokens, and sampling $t \sim U(0,1]$ and x_t from the forward process yields the unbiased single-sample estimator of $\mathcal{L}_\infty$

$$\widehat{\mathcal{L}} = \sum_{j:\, x_t^j = [\mathrm{M}]} \frac{f_j'(t)}{f_j(t)} \big(-\log p_\theta(x_0^j \mid x_t, v)\big), \tag{12}$$

which is equation 4 with x_t drawn from the span-absorbing process of the chosen partition. For the same-tier-span partition this proves Proposition 2(i). For the trivial partition ($|S| \equiv 1$) it establishes the token-factorized NELBO stated in the main text; the power-schedule form of the weights and the $\gamma_k = 1$, $\varepsilon \to 0$ reduction to equation 2 follow by substitution. □

Remark: why $1/f_k$ is not the variational weight. The correct weight is the hazard f_k'/f_k; the inverse masking probability $1/f_k$ differs by $1/f_k'(t) = t^{1-\gamma_k}/((1-\varepsilon)\gamma_k)$, dependent on both tier and time, coinciding only for the unsmoothed linear schedule. Training with $1/f_k$ optimizes a different weighted denoising objective, not the absorbing-state NELBO of equation 3.

Two consequences. First, toggling block masking changes only the corruption sampler — loss expression and weights are literally identical — yet each configuration optimizes the NELBO of its own forward process, exactly up to the shared $O(\varepsilon)$ boundary term equation 11. Second, the two NELBOs bound the likelihoods of *different* generative processes (token-wise vs. span-wise reverse unmasking); the block-masked objective is not a bound for the token-factorized process, and we never claim it is. Confidence decoding follows the ancestral reverse kernel of neither process; this shared approximation is treated in Appendix A.4.

A.1.3 Proof of Proposition 2(ii)–(iii)

Let S be a same-tier span of length r and tier k; block masking draws $u_S \sim U(0,1)$ and masks all of S iff $u_S < f_k(t)$.

(ii) Marginals and independence. For $j \in S$, $\Pr(x_t^j = [\mathrm{M}]) = \Pr(u_S < f_k(t)) = f_k(t)$, identical to the token-factorized marginal. Hence $\mathbb{E}[\mathbb{1}[x_t^j = [\mathrm{M}]] w_{k(j)}(t)] = f_{k(j)}(t) w_{k(j)}(t) = f_{k(j)}'(t)$, preserving the coefficient budget in Proposition 1(a). Distinct spans use independent u_S, so the complete-skeleton calculation in Proposition 1(b) also applies unchanged, with M equal to the number of tier-1 spans.

(iii) Fragmentation. Within S all tokens share one indicator, so $\Pr\big(0 < \#\{j \in S : x_t^j = [\mathrm{M}]\} < r\big) = 0$; under independent masking the all/none events have probabilities $f_k(t)^r$ and $(1 - f_k(t))^r$, leaving fragmentation probability $1 - f_k(t)^r - (1 - f_k(t))^r$. □

What is preserved and what changes. Parts (ii)–(iii) concern *which tokens are selected and how they are weighted*; they do not assert that the objective value or its gradients are unchanged — the joint law of x_t, hence the distribution of denoising contexts entering $-\log p_\theta(x_0^j \mid x_t, v)$, differs by construction. Within a span indicators are perfectly correlated; across spans independent. Block

masking thus spends the same per-token coefficient budget on span-coherent states, and by part (i) the resulting loss remains an exact variational objective up to the $O(\varepsilon)$ boundary term equation 11; whether the changed context distribution improves the learned model is settled by the ablation in §4.

A.2 Supervision allocation and complete-skeleton visibility

Over the joint randomness of $t \sim U(0,1]$ and forward masking, the density with which a tier-k token contributes a weighted prediction at level t is

$$\mathbb{E}\Big[\mathbb{1}[x_t^j = [\mathrm{M}]]w_k(t)\Big] = f_k(t)\frac{f_k'(t)}{f_k(t)} = f_k'(t). \tag{13}$$

Thus the unnormalized effective training measure is $d\mu_k(t) = f_k'(t)dt$, and its normalized density is

$$\bar{\mu}_k(t) = \frac{f_k'(t)}{1-\varepsilon}. \tag{14}$$

As $\varepsilon \to 0$, this converges to the $\mathrm{Beta}(\gamma_k, 1)$ density $\gamma_k t^{\gamma_k - 1}$. Proposition 2(ii) ensures that the same measure applies under same-tier block masking.

Proof of Proposition 1. For part (a),

$$\int_0^1 f_k'(t)dt = f_k(1) - f_k(0) = 1 - \varepsilon.$$

For part (b), fix a supervised tier-k token j, $k \geq 2$, and let $S_1, \ldots, S_M$ be the independent masking units that constitute the complete tier-1 skeleton. Under token-wise masking these units are singleton tokens; under block masking they are the tier-1 same-tier spans. The target unit containing j is distinct from all tier-1 units. Therefore, conditional on t and on the event that j is supervised, each S_r is visible with probability $1 - f_1(t)$ and the M visibility indicators are independent. Hence

$$\Pr(S_1, \ldots, S_M \text{ all visible} \mid t, j \text{ supervised}) = \big(1 - f_1(t)\big)^M. \tag{15}$$

Averaging equation 15 against the normalized weighted-event density equation 14 gives

$$G_{k\leftarrow 1}^{\mathrm{all}}(M) = \int_0^1 \big(1 - f_1(t)\big)^M \bar{\mu}_k(t)dt \tag{16}$$

$$= (1-\varepsilon)^M \gamma_k \int_0^1 t^{\gamma_k - 1}\big(1 - t^{\gamma_1}\big)^M dt. \tag{17}$$

With $u = t^{\gamma_1}$ and $a = \gamma_k/\gamma_1$,

$$G_{k\leftarrow 1}^{\mathrm{all}}(M) = (1-\varepsilon)^M \frac{\gamma_k}{\gamma_1}\int_0^1 u^{a-1}(1-u)^M du \tag{18}$$

$$= (1-\varepsilon)^M \frac{\gamma_k}{\gamma_1} B(a, M+1). \tag{19}$$

Finally,

$$aB(a, M+1) = \frac{\Gamma(a+1)\Gamma(M+1)}{\Gamma(a+M+1)} = \prod_{r=1}^{M}\frac{r}{a+r},$$

which yields the product form in equation 5. □

Uniform schedule and strict improvement. When $\gamma_1 = \gamma_k = 1$,

$$G_{k\leftarrow 1}^{\mathrm{all,uni}}(M) = \frac{(1-\varepsilon)^M}{M+1}. \tag{20}$$

For SPROUT, $\gamma_k < \gamma_1$ for $k \geq 2$, and therefore

$$\frac{r\gamma_1}{\gamma_k + r\gamma_1} > \frac{r}{r+1} \qquad (r = 1, \ldots, M).$$

Multiplying these inequalities proves $G_{k\leftarrow 1}^{\mathrm{all}}(M) > G_{k\leftarrow 1}^{\mathrm{all,uni}}(M)$ for every $M \geq 1$. For $M = 1$, the result reduces to the pairwise expression $(1-\varepsilon)\gamma_1/(\gamma_1 + \gamma_k)$.

Table 6: Complete tier-1 skeleton visibility $G^{\text{all}}_{k\leftarrow 1}(M)$ as $\varepsilon \to 0$ for a skeleton containing M independent tier-1 spans. Uniform corruption gives $1/(M+1)$. SPROUT uses $\gamma = (4.5, 3, 1.5, 0.9, 0.6)$. The $M = 1$ row is the previous pairwise special case; larger M measures visibility of the full skeleton.

M	Uniform	Tier 2	Tier 3	Tier 4	Tier 5
1	0.500	0.600	0.750	0.833	0.882
2	0.333	0.450	0.643	0.758	0.827
4	0.200	0.316	0.534	0.676	0.767
6	0.143	0.251	0.474	0.629	0.730
8	0.111	0.211	0.435	0.597	0.705
10	0.091	0.184	0.406	0.573	0.686

Corpus-level rate. The number of tier-1 spans varies across captions. Because every tier-k token has the same integrated coefficient budget, the rate for a uniformly selected tier-k supervision event is estimated by the token-weighted average

$$\overline{G}^{\text{all}}_{k\leftarrow 1} = \frac{\sum_{x\in\mathcal{D}} N_k(x) G^{\text{all}}_{k\leftarrow 1}(M_1(x))}{\sum_{x\in\mathcal{D}} N_k(x)}, \tag{21}$$

where $M_1(x)$ is the number of tier-1 spans and $N_k(x)$ is the number of tier-k target tokens in caption x. This is the quantity to report when summarizing complete-skeleton visibility over the training corpus.

A.3 Estimator variance and the role of ε

The hazard $w_k = f_k'/f_k$ diverges as $t \to 0$ whenever $\gamma_k < 1$. Unbiasedness is unaffected, but the variance of the single-sample estimator is not. Write the per-token weight contribution as $Z_j = \mathbb{1}[x_t^j = [\mathrm{M}]]\, w_{k(j)}(t)$, $t \sim U(0,1]$, treating the log-loss factor as bounded for this analysis.

Proposition 3 (Second moment of the hazard weight)**.** *For $f_k(t) = \varepsilon + (1-\varepsilon)t^{\gamma_k}$, $\varepsilon \in (0,1)$, $\mathbb{E}[Z_j^2] = \int_0^1 f_k'(t)^2/f_k(t)\, dt$ is finite iff $\gamma_k > \frac{1}{2}$; at $\varepsilon = 0$ the condition tightens to $\gamma_k > 1$.*

Proof. $\mathbb{E}[Z_j^2] = \int_0^1 f_k w_k^2\, dt = \int_0^1 f_k'^2/f_k\, dt$ with $f_k'^2 = (1-\varepsilon)^2\gamma_k^2 t^{2\gamma_k-2}$. Since $\varepsilon \le f_k \le 1$,

$$(1-\varepsilon)^2\gamma_k^2 \int_0^1 t^{2\gamma_k-2} dt \le \int_0^1 \frac{f_k'^2}{f_k}\, dt \le \frac{(1-\varepsilon)^2\gamma_k^2}{\varepsilon} \int_0^1 t^{2\gamma_k-2} dt,$$

and $\int_0^1 t^{2\gamma_k-2} dt < \infty$ iff $\gamma_k > \frac{1}{2}$. At $\varepsilon = 0$, the integrand is $\gamma_k^2 t^{\gamma_k-2}$, integrable iff $\gamma_k > 1$. □

Two design consequences. First, the floor ε is not merely numerical: it relaxes the finite-variance threshold from $\gamma_k > 1$ to $\gamma_k > \frac{1}{2}$, which gives the hazard-weight component a finite second moment for the sublinear exponents $\gamma_4 = 0.9$ and $\gamma_5 = 0.6$. Second, the threshold constrains — as a lower bound, without uniquely determining — the one degree of freedom that the complete-skeleton visibility leaves open: for every fixed M, equation 5 depends only on exponent *ratios*, so the visibility profile determines γ only up to a global scale. We select the scale so that $\min_k \gamma_k = 0.6$, a fixed margin above $\frac{1}{2}$, giving $\gamma = (4.5, 3, 1.5, 0.9, 0.6)$; the scale is thus chosen to satisfy the variance constraint rather than tuned on downstream performance. Consequently, $t \sim U(0,1]$ is sampled without truncation, and the budget invariance and complete-skeleton visibility of Proposition 1 hold for the deployed estimator as stated (modulo the $(1-\varepsilon)$ factor and the $O(\varepsilon)$ boundary term of equation 11). ε remains an explicit bias–variance knob.

Remark: truncated sampling at scales violating the threshold. For grids with $\min_k \gamma_k \le \frac{1}{2}$ — e.g. the same ratios at the $\times\frac{2}{3}$ scale $(3, 2, 1, 0.6, 0.4)$ — the plain estimator has infinite variance and one must intervene. Truncating $t \sim U[t_{\min}, 1]$ removes, for tier k, exactly the supervision mass $\int_0^{t_{\min}} f_k' = (1-\varepsilon)t_{\min}^{\gamma_k}$: a tier-dependent relative budget bias of $t_{\min}^{\gamma_k}$, largest for the smallest exponents. It also changes complete-skeleton visibility to

$$G^{\text{all,trunc}}_{k\leftarrow 1}(M) = (1-\varepsilon)^M \frac{\gamma_k}{\gamma_1} \frac{B(a, M+1) - B_{t_{\min}^{\gamma_1}}(a, M+1)}{1 - t_{\min}^{\gamma_k}}, \qquad a = \frac{\gamma_k}{\gamma_1}, \tag{22}$$

where B_z is the lower incomplete beta function. Thus truncation both breaks exact budget invariance and perturbs the visibility profile, especially for the finest tiers. Alternatively, importance sampling $t \sim q$ with $q(t) \propto t^{\gamma_{\min}-1}$ on $(0, 1]$ and each summand of equation 12 rescaled by $1/q(t)$ is unbiased for the full-interval objective with finite second moment for every tier when $\varepsilon > 0$, since the second-moment integrand $f_k'^2/(f_k q) = O(t^{2\gamma_k - 1 - \gamma_{\min}})$ near 0 is integrable because $2\gamma_k > \gamma_{\min}$. We avoid both interventions by the scale choice above.

A.4 Relation between training and inference

Two gaps, one inherited. Let π_t^{tr} denote the training distribution over partially masked states at level t induced by the (block-masked) forward process, and π^{dec} the distribution over states visited by N-step confidence decoding. Standard MDLM inference already departs from ancestral sampling of the learned reverse process — confidence ranking is a deterministic selection heuristic — so a gap exists independently of SPROUT. SPROUT adds a schedule-specific gap: under π_t^{tr} the expected masked fraction of tier k is $f_k(t)$, so along reverse time the tiers unmask at different rates, whereas confidence decoding imposes no per-tier marginals. We therefore treat whether the learned confidence profile respects this ordering as an empirical question, rather than an assumption.

Schedule-consistent reverse marginals. A decoder is *schedule-consistent* if, on the reverse grid $t_n = 1 - n/N$, the fraction of tier-k positions still masked after step n is $f_k(t_n)$ in expectation. Ancestral sampling of the span-absorbing reverse process is schedule-consistent by construction; confidence decoding need not be. The tier-stratified commit-order statistics of §4 test whether this schedule-induced ordering emerges under confidence decoding. We compare SPROUT against the untrained backbone and a reversed-schedule control to separate schedule-induced ordering from the ordering produced by confidence ranking alone.

A.5 Dataset construction and quality control

This appendix expands §2: the per-level deletion rules, the validation protocol with its tolerances, and an analysis of selection bias. Here, "syntactic" denotes the deletion-derived skeleton-to-detail hierarchy, not an explicit parse-tree representation.

Per-level deletion rules. Each step conditions only on the previous level and is instructed to delete, never rewrite: L_4 drops camera/lighting/atmosphere framing (a regex pre-pass removes the view-framing phrase opening 70.5% of DOCCI captions); L_3 drops secondary-object attributes and peripheral objects; L_2 keeps only the main subject(s) with one or two core attributes; L_1 is a single ≤ 12-word subject–verb–object sentence. Chaining is essential: a one-shot "give me all five levels" prompt met the subsequence constraint for only 3% of captions, versus 95.5% for the chained procedure.

Validation protocol. Each step is accepted only if two conditions hold. First, every content word of L_k maps, essentially in order, to a position of L_{k+1} after normalization for case, punctuation, hyphens, and light stemming, tolerating bounded local reorderings ($\leq 30\%$, for example "puppy standing" becomes "standing puppy") and a bounded number of skipped glue words ($\leq 25\%$). Second, word counts decrease strictly across levels within prescribed bands. Violations are retried up to four times with the specific error appended to the prompt. Captions that still fail are dropped, retaining $14{,}176$ of $14{,}847$ (95.5%). The tolerance values were fixed once on a 200-caption development set before large-scale generation and not tuned thereafter. These tolerances are safe for training because the intermediate texts never enter the loss (§3); their only effect is on the per-word labels.

Tier transfer. Word tiers transfer to BPE tokens by word-boundary alignment, so each sub-word inherits the tier of the word it belongs to. All tier labels lie in the valid range $\{1, \ldots, 5\}$, with zero out-of-range labels across the $14{,}176$ retained captions.

Manual inspection. We manually inspected the induced tier assignments and verified that they follow the intended skeleton-to-detail progression, with lower tiers capturing the core scene structure and higher tiers adding progressively finer attributes, modifiers, and secondary content.

Selection bias from dropped captions. The 4.5% of captions dropped by validation could in principle be the syntactically atypical ones, which would make the training distribution more regular than DOCCI. Comparing the retained and dropped sets on their original DOCCI text shows the two are closely matched: mean length 122.6 versus 129.1 words, dependency-tree depth 8.38 versus 8.32, and 11.73 versus 11.96 clauses per caption. The dropped set is thus not systematically simpler than the retained set, so validation introduces negligible selection bias.

A.6 Additional Tier-Construction Controls

We further evaluate alternative tier constructions beyond the controls reported in Table 5(a). Table 7 reports the complete comparison. Frequency- and POS-based tiers exhibit the same under-description failure mode as shuffled tiers: although they obtain low CHAIR scores, AMBER coverage and object count drop sharply, together with substantial degradation in THRONE F_1. In contrast, the deletion-derived tiers retain descriptive coverage while achieving strong fine-grained faithfulness. These results further show that arbitrary tier assignments do not reproduce SPROUT's faithfulness–coverage balance.

Table 7: Additional tier-construction controls. Low CHAIR values marked by † coincide with severe under-description.

	CHAIR (500 COCO)		AMBER		THRONE	ImgPara
Model	$CHAIR_i \downarrow$	#Obj↑	CHAIR↓	Cover↑	$F_1 \uparrow$	SPICE↑
SPROUT (deletion tiers)	6.0	9.3	7.0	**67.2**	**75.9**	**18.7**
Matched random spans	7.39	**10.7**	7.1	66.9	71.1	18.3
Shuffled tiers	2.98†	1.3	2.6†	36.8	54.8	17.2
Frequency-based tiers	2.94†	1.6	2.6†	39.5	64.5	15.9
POS-based tiers	2.38†	1.4	2.5†	37.5	57.9	17.9

A.7 Additional Qualitative Results.

Figures 6 and 7 provide additional comparisons between the untrained LLaDA-V backbone and SPROUT across diverse scenes. The backbone often captures the main scene correctly but introduces plausible unsupported objects, attributes, or background details. In contrast, SPROUT produces descriptions that are more closely grounded in the image while retaining comparable descriptive detail. These examples complement the qualitative results in Figure 5 and show that the reduction in hallucination extends across different scene types.

A.8 Implementation Details

We use LLaDA-V as the base model and keep its SigLIP2 vision encoder frozen during training. The training set contains 13,810 DOCCI examples, including 9,207 tier-annotated captions and 4,603 question-answer pairs. The five tiers L1–L5 use schedule exponents $\gamma = (4.5,\ 3.0,\ 1.5,\ 0.9,\ 0.6)$, respectively. We train for 3 epochs on 8 NVIDIA A100 40GB GPUs using AdamW with an effective batch size of 128 and a peak learning rate of 1×10^{-5}. At inference, we generate 256 tokens each step. CHAIR and THRONE are evaluated on 500 COCO validation images, and GPT-5 is used as the judge for THRONE and OpenCHAIR. Other benchmarks are evaluated with lmms-eval.



Figure 6: Additional qualitative comparison results.



Figure 7: Additional qualitative comparison results.