Length-Adaptive Decoding for Masked Diffusion Machine Translation

Yan Zhan[1] Mengkai Hou[1] Wanting Zhang[2] Zhijun Gao[1]
[1]Peking University [2]BYD Company Limited

Machine translation tests masked diffusion language models (dLLMs) because every source token must be rendered faithfully, while fixed canvas decoding must choose target length before denoising. Existing masked diffusion decoding work mainly studies token unmasking order, leaving this length decision under-explored despite its direct effect on coverage and redundancy. We introduce **Entropy-Valley** (EV), a training-free length selector that scores candidate target canvases by mean predictive entropy from all-mask forward passes and selects the canvas the backbone is most prepared to fill. Relative to a baseline using training corpus length statistics, EV recovers 64.9%, 65.3%, and 33.0% of the COMET-22 gain from reference target lengths on En→Zh, Zh→En, and En→De. Our diagnostics show that denoising-friendly lengths need not match reference lengths. Evaluation by three translation experts supports the En↔Zh adequacy gains, with stronger evidence on Zh→En. Compared with a LLaMA-3-8B autoregressive (AR) model trained on the same fine-tuning data, the EV system ties on En→Zh and leads on Zh→En; an oracle-length diagnostic further shows that, in this masked diffusion MT setting, deciding which tokens to reveal first matters less than how the target length is supplied.

Contact: gaozhijun@pku.edu.cn
Code: github.com/Entropy-Valley/Entropy-Valley
Dataset & Model: huggingface.co/collections/YanZhanPKU/entropy-valley

1 Introduction

Pretrained masked diffusion language models (dLLMs) such as LLaDA-8B (Nie et al., 2025) and Dream-7B (Ye et al., 2025) now rival autoregressive (AR) LLMs on open ended generation and reasoning. Machine translation stresses a different part of the model. A translation system must preserve source content, reorder words across languages, choose suitable morphology, and stop at the right length. If the output is too short, content disappears; if it is too long, the model has room to repeat, hallucinate, or dilute the sentence.

The stopping problem is structural in fixed canvas masked diffusion decoding. An AR decoder generates tokens until it emits EOS. A masked diffusion decoder instead receives a target canvas before denoising begins. During supervised training, EOS is placed after that canvas, so at test time the model tends to fill whatever length it is given. Work on masked iterative generation and masked diffusion decoding often asks which positions to reveal first (Chang et al., 2022; Hong et al., 2026; Aman, 2026). Target length is often supplied either by the reference translation length during evaluation or by a single corpus-level source-to-target ratio. The former is an oracle upper bound rather than an inference procedure; the latter imposes one global length rule on sentences with different compression or expansion needs.

Borrowing length machinery from non autoregressive MT is not a clean fit for this setting. Classic systems predict fertilities or target lengths, sometimes with a small length beam (Gu et al., 2018; Ghazvininejad et al., 2019), or change the decoder into an insertion and deletion architecture (Gu et al., 2019). Those options are reasonable when the model is trained for that purpose, but they do not answer a narrower question: whether a pretrained decoder only masked diffusion backbone can select its canvas at test time without another model or another training objective. Static corpus ratios have the opposite problem. They are easy to deploy, but one ratio cannot account for source sentences with placeholders, numbers, short commands, and idiomatic compression.

Figure 1 shows why this is not a bookkeeping detail. For the source sentence “Tap Reset Now.”, the ratio baseline chooses five target slots and drops the action verb; EV chooses six slots and recovers the verb. For “Under #PRS_ORG#, tap Sign out.”, the shorter canvas loses the placeholder, while the EV canvas preserves it. These are ordinary translation failures: the reader loses an action, an entity marker, or both. They arise before any question about the order in which masks are revealed.

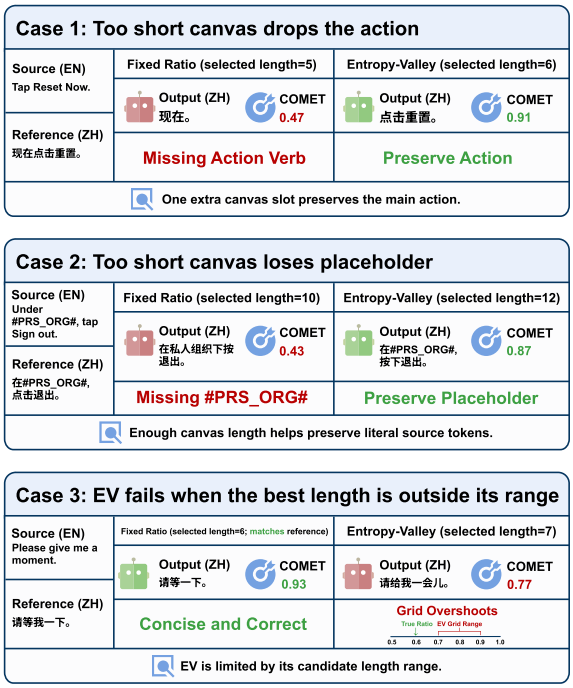

Figure 1 Three En→Zh cases decoded with 32 denoising steps. A short canvas can drop an action verb or a required placeholder; EV chooses a longer canvas in these two cases and recovers the missing content. The third case shows the opposite boundary: the best compression lies below the EV candidate range, so the ratio baseline is better.

The same pattern appears beyond these examples. On WMT22 En→Zh, EV raises placeholder retention from 66.9% to 89.1% and number retention from 77.6% to 81.3% over the ratio baseline (Appendix D). The gain is not just longer output. The right length depends on the source sentence: some inputs need extra room to preserve a placeholder, while high compression cases need a shorter canvas than our current candidate range provides. This points to length selection, not length inflation.

The model already exposes a useful signal before any token is generated. On an all-mask canvas, each candidate length induces a predictive distribution over every target slot. A canvas that is too short must compress too much source content into too few target slots; a canvas that is too long creates extra positions for which the model has no confident token choice. Mean predictive entropy provides a direct way to compare candidate lengths before spending the full decoding budget. Because this signal comes from the translation model itself, it avoids both reference lengths and a separately trained length predictor.

Our approach is to ask the frozen backbone which canvas it is most prepared to fill. Given a source sentence, Entropy-Valley (EV) forms a small set of candidate canvas lengths, runs one all-mask forward pass for each candidate, and scores the canvas by mean predictive entropy. EV then decodes on the canvas with the lowest mean entropy. The method adds no trainable parameters and does not use reference lengths, development set tuning, or an external length predictor. Its target is not exact reference length prediction; it chooses a length that the masked diffusion model can denoise well.

We make four contributions in this LLaDA-8B LoRA-SFT masked diffusion MT setting.

1. We identify target length selection as a measurable bottleneck in fixed canvas masked diffusion MT and include a controlled comparison with decoding-order choices under matched budgets.
2. We introduce Entropy-Valley, which requires no additional training and uses entropy on all-mask canvases to choose a canvas the backbone can denoise well.
3. We show across En→Zh, Zh→En, and En→De that EV recovers 64.9%, 65.3%, and 33.0% of the COMET-22 improvement obtained by supplying reference target lengths, while not simply matching those lengths.
4. We test the length selection finding with statistical tests, bidirectional expert human evaluation, coverage and error type analysis, a matched data LLaMA-3-8B AR baseline, reveal-order controls, and cross backbone checks on Dream-Base and DiffuLLaMA.

2 Related Work

Masked diffusion decoding. Text diffusion models range from continuous-embedding formulations (Li et al., 2022; Gong et al., 2023) to discrete absorbing-state and masked objectives (Austin et al., 2021; Lou et al., 2024; Sahoo et al., 2024). Recent 8B masked diffusion LMs such as LLaDA and Dream report strong open-ended generation and reasoning results (Nie et al., 2025; Ye et al., 2025). In the fixed-canvas MT setting studied here, the target length is supplied before denoising; when this length is unsuitable, coverage and redundancy change before the unmasking order is chosen.

Length selection in machine translation. Parallel MT systems have long treated length as an explicit modeling choice. Early NAT predicts target length before parallel decoding (Gu et al., 2018); CMLM adds a trained length head and small-beam length

search (Ghazvininejad et al., 2019); GLAT changes the training objective (Qian et al., 2021); Levenshtein Transformer avoids a fixed canvas through insertion and deletion operations (Gu et al., 2019). Wang et al. (2021) show that the best candidate from an oracle CMLM length beam can outperform decoding at the exact reference length. These methods are effective when the architecture and objective are built around them. Our setting keeps the pretrained decoder-only masked diffusion backbone fixed after LoRA-SFT, so EV selects length at test time from the model's all-mask entropy rather than adding a length predictor, length beam, or edit-based decoder.

Variable length diffusion language models. Recent work handles unknown generation length in several ways. DAEDAL adjusts an initial length before denoising and can insert masks during denoising (Li et al., 2026); ρ-EOS expands or contracts the sequence during denoising using EOS density (Yang et al., 2026). CAL searches candidate lengths before decoding using calibrated confidence from the first denoising step (Liu et al., 2026), while LR-DLLM corrects length bias in confidence scores across candidate lengths (Cheng et al., 2026). SmartCrop estimates the output length from the prompt and crops the canvas before generation (Rossi et al., 2026). FlexMDM changes training and generation to support mask insertion (Kim et al., 2026); dLLM-Var trains the model to produce EOS and outputs of varying length (Yang et al., 2025). EV uses mean entropy from all mask predictions to choose a canvas for each source sentence before denoising. It does not require a learned length head, selector training, or changes to the denoising schedule.

Autoregressive LLMs for MT. Recent AR LLM-MT systems improve translation through MT-specific continued training, instruction tuning, or preference optimization (Xu et al., 2024a,b; Alves et al., 2024). We use a matched-data LLaMA-3-8B LoRA-SFT baseline to calibrate the masked-diffusion setting against this line of work, but EV targets a different bottleneck: choosing the fixed canvas before denoising.

Reveal order methods. A separate line of masked generation work studies which tokens to reveal, not which canvas to reveal them on. MaskGIT uses confidence-based parallel unmasking (Chang et al., 2022); OeMDM/LoMDM learn generation orders with the backbone (Hong et al., 2026); LogicDiff designs task-specific orders for reasoning (Aman, 2026). These methods are not direct MT length-selection baselines, so we use them to locate EV relative to order scheduling. Our controlled diagnostic asks the complementary question for MT: under matched budgets, how much variation comes from the canvas length source versus the tested unmasking orders (§4.5). EV can be used with standard minimum-entropy decoding, which reveals lower-entropy positions first, because it targets this separate length-selection bottleneck.

3 Method: Entropy-Valley

3.1 Overview

Given a source sentence $\mathbf{x}$, our goal is to choose a target canvas length before masked diffusion decoding begins. The output of the length selector is $L^\star$; the translation is then produced by the same minimum-entropy decoding (MED) schedule used by the baseline, only on the selected canvas. As shown in Figure 2, Entropy-Valley (EV) follows four operations. First, EV builds a small set of candidate canvas lengths from a fixed source-to-target ratio set. Second, it probes each candidate length once with an all-mask forward pass and scores the canvas by mean predictive entropy. Third, EV selects the lowest-entropy canvas, or entropy valley, and then runs standard MED decoding on that canvas. EV changes only the target length supplied to the decoder, not the backbone, training loss, or unmasking schedule.

3.2 Problem formulation

A masked diffusion language model receives a source prompt and a target canvas $\mathbf{y} \in (\mathcal{V} \cup \{[\textsc{Mask}]\})^L$. It denoises the canvas over T steps under an order schedule π. In our main experiments, π is MED, which reveals lower-entropy positions first. Training appends EOS to each target sequence. At inference, the allocated canvas includes a final slot designated for EOS, and the decoded string is truncated at the first EOS. The canvas length L must be supplied before decoding.

We study test-time length selection for a frozen LoRA-SFT masked diffusion MT system. A length selector takes $\mathbf{x}$ and returns $\widehat{L}(\mathbf{x}) \in \mathbb{N}_+$ from a finite candidate set

$$\mathcal{C}(\mathbf{x}) = \{\max\{1, \lfloor r|\mathbf{x}| \rfloor\} + 1 : r \in \mathcal{R}\}.$$

Here $|\mathbf{x}|$ is the source-side tokenizer length under

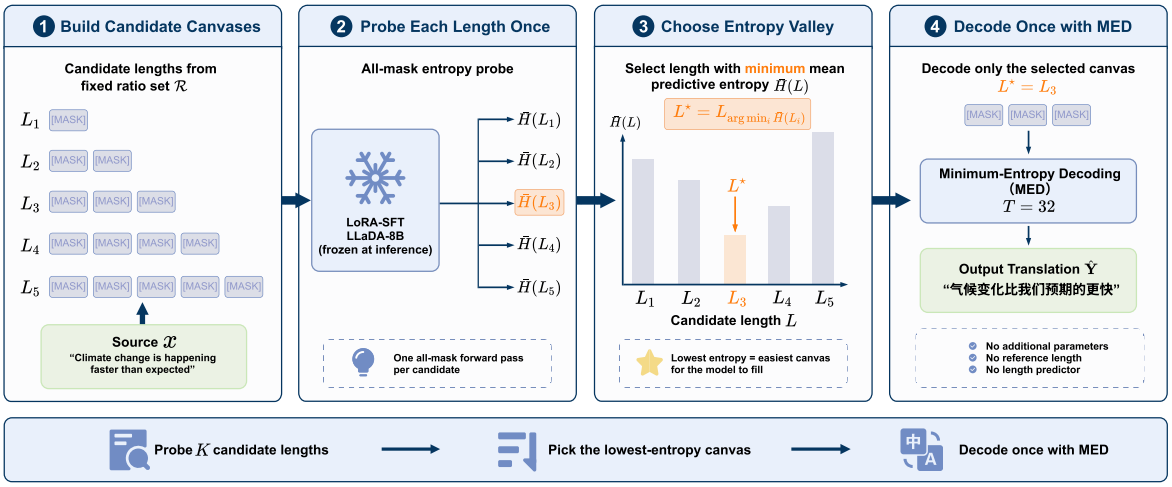

Figure 2 Entropy-Valley method overview. Given a source sentence, EV uses five candidate ratios from the fixed set $\mathcal{R}$, yielding up to five distinct target canvases after duplicate lengths are removed. It then probes each candidate length once with an all-mask forward pass through the frozen LoRA-SFT LLaDA-8B model and records the mean predictive entropy $\bar{H}(L)$ over all slots except the designated EOS slot. EV selects the entropy valley, $L^\star = \arg\min_L \bar{H}(L)$, and decodes only this selected canvas with minimum-entropy decoding (MED) for T=32 steps. The procedure adds no parameters, does not use the reference length of the current sentence, and requires no length predictor.

the same tokenizer used by the backbone, and the final $+1$ reserves the designated EOS slot. At inference, the selector does not use the reference length of the current sentence or a separately trained length predictor. It scores the predefined candidates using the frozen model's all-mask uncertainty:

$$L^\star = \arg\min_{L \in \mathcal{C}(\mathbf{x})} \bar{H}(L),$$

$$\bar{H}(L) = \frac{1}{L-1} \sum_{i=1}^{L-1} H\big(p_\theta(y_i \mid \mathbf{x}, [\textsc{Mask}]^L)\big) .$$

The target is not to recover L^*. EV chooses the candidate length that the current backbone appears most prepared to denoise.

3.3 Candidate canvas construction

A fixed ratio is cheap, but it assigns the same compression rule to short commands, entity-heavy sentences, and idiomatic translations. Probing every possible length would require no additional training, but it would spend computation on canvases that are implausible for the language pair. EV instead searches a small set of plausible sentence-level lengths.

The median ratios below are WMT19 training corpus statistics and serve as reference scales. The ratio grid is set separately for each direction. Diagnostic range comparisons on WMT22 subsets informed the reported compact grids. Once chosen, each grid was used for every sentence and decoding method in that direction:

Direction	Median ratio	$\mathcal{R}$
En→Zh	0.80	$\{0.70, 0.75, 0.80, 0.85, 0.90\}$
Zh→En	1.24	$\{1.00, 1.10, 1.20, 1.30, 1.40\}$
En→De	1.48	$\{1.50, 1.60, 1.70, 1.80, 1.90\}$

For a source sentence $\mathbf{x}$, EV maps the ratios to canvas lengths using the formula above and removes duplicates. The five ratios can thus yield fewer than five distinct lengths for short source sentences. These grids provide a compact set of plausible canvases, while selection for each sentence comes from the frozen model's all-mask entropy.

The candidate set gives EV only one role at test time: choose among a few plausible canvases for the current source. It does not use the reference length of the current sentence or any manual adjustment during decoding. The range still matters; Appendix Table 6 reports the candidate-width control.

3.4 Entropy-Valley scoring

The decoder must choose L before generation. A useful score should be available then and should come from the same model that fills the canvas. On an all-mask canvas, a too-short length forces source content into too few slots, while a too-long length leaves positions with diffuse alternatives. Mean predictive entropy exposes this mismatch without running the full denoising loop.

For each $L \in \mathcal{C}(\mathbf{x})$, EV runs one forward pass on $[\,\text{prompt}(\mathbf{x})\,]\|[\text{MASK}]^L$ and computes $\bar{H}(L)$. The final slot is excluded because it is designated for EOS and typically has entropy close to zero. Averaging over the remaining $L-1$ slots compares candidates of different lengths without allowing this slot to dominate the score. EV selects the entropy valley, $L^\star = \arg\min_{L \in \mathcal{C}(\mathbf{x})} \bar{H}(L)$.

The score makes EV a denoising-friendly canvas selector rather than a reference-length regressor. The chosen length need not match the reference exactly if it gives the backbone a canvas it can fill reliably (details in §4.5).

3.5 Decoding algorithm

Algorithm 1 gives the full inference procedure. The only difference from the fixed-ratio baseline is the loop over candidate lengths. Once $L^\star$ is selected, decoding uses the standard MED schedule for the same T steps used by the baseline.

Algorithm 1 Entropy-Valley length selection and decoding

Require: Source sentence $\mathbf{x}$; frozen model p_θ; fixed ratio set $\mathcal{R}$; decoding steps T; MED schedule π.
Ensure: Translation decoded from the selected canvas.

```
Form candidate lengths C(x) ← {max{1, ⌊r|x|⌋} + 1 : r ∈ R}; deduplicate.
for L ∈ C(x) do
    Build the all-mask input [ prompt(x) ] || [MASK]^L.
    Record q_1^L, ..., q_L^L with one forward pass.
    Set H̄(L) to the mean entropy of q_1^L, ..., q_{L-1}^L.
end for
Select L* ← arg min_{L∈C(x)} H̄(L).
Decode [ prompt(x) ] || [MASK]^{L*} with MED schedule π for T steps.
Return the decoded target string, truncated at the first EOS as in the shared decoding protocol.
```

3.6 Cost and implementation details

Entropy-Valley uses five ratios and probes at most five distinct candidate lengths before decoding. At the default T=32, this adds at most five probe passes to the 32 denoising passes used by the baseline; fewer are needed when multiple ratios map to the same integer length. Candidates can be evaluated sequentially with no extra peak memory, as in our setup, or batched if memory permits. EV adds no trainable parameters, alignment model, or separately trained length predictor. Appendix Table 12 controls for the extra forward pass budget, and Appendix B records the canvas conventions, external length property, and length versus order protocol.

4 Experiments

The experiments test six claims about Entropy-Valley: the length problem is measurable, EV improves over a fixed corpus ratio, the gain is not just reference-length matching, the candidate design matters, automatic gains align with human judgments on En↔Zh, and the method has clear boundary cases across language pairs and backbones. We first fix the protocol, then report the main LLaDA results, matched-data AR calibration, cross-backbone scope, analysis and controls, human evaluation, and a brief discussion of what EV selects. Additional paired significance tests are in Appendix C.

4.1 Setup

We LoRA-SFT LLaDA-8B-Base (Nie et al., 2025) on 200k WMT19 training pairs per direction (En↔Zh and En→De), train three independent runs for each direction, and evaluate on all 2,037 WMT22 samples per direction with COMET-22 (Rei et al., 2022) and sacreBLEU (Post, 2018). Decoding uses T=32 MED with EOS truncation; full protocol, significance procedure, hyperparameters, and compute are in Appendix A.

4.2 Main Results

Figure 3A gives the COMET overview, and Table 1 reports the corresponding COMET and sacreBLEU values. With the backbone, training data, and MED schedule fixed, the comparison isolates the target-canvas selector. Replacing the training-corpus Ratio with Entropy-Valley raises COMET-22 by $+0.0172$ on En→Zh and $+0.0165$ on Zh→En, closing $64.9\pm7.4\%$ and $65.3\pm0.8\%$ of the COMET gap between the length oracle and the fixed-ratio baseline. On En→De, EV gains $+0.0070$ COMET and closes $33.0\pm8.4\%$ of the same gap. The corresponding sacreBLEU gains are $+1.85$, $+1.63$, and $+0.82$, respectively.

The paired tests give the clearest support on En↔Zh, where EV remains above Ratio under both paired bootstrap and Wilcoxon tests (Appendix Table 11). For En→De, the mean over three runs is positive, but the sentence-level evidence is weaker. This direction is a boundary case for length selection within the fixed LLaDA system.

The extra EV probes do not explain the main gains. Appendix Table 12 gives Ratio the same or more forward passes and recovers at most 0.001 COMET, far below the En↔Zh gains in Table 1.

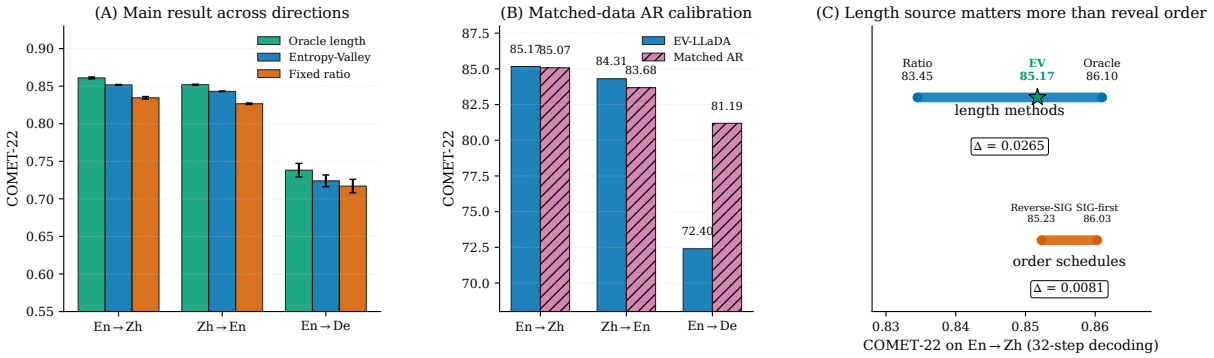

Figure 3 Length-adaptive decoding for masked-diffusion MT. (A) Mean COMET-22 over three runs for a reference-length upper bound, EV, and a corpus-ratio baseline across three directions. EV closes 64.9% on En→Zh, 65.3% on Zh→En, and 33.0% on En→De. (B) Matched-data AR calibration across three directions. (C) Controlled En→Zh diagnostic: with reference target lengths fixed, the tested source-guided unmasking schedules span 0.0081 COMET. With the schedule fixed to MED, three length choices span 0.0265 COMET. The order-schedule sweep uses reference lengths, so it is a diagnostic rather than a deployable alternative to EV.

		COMET-22		sacreBLEU	
Direction	**Method**	Mean ± Std	Gap Closure	Mean ± Std	Gap Closure
En→Zh	Oracle length	0.8610 ± 0.0013	100%	40.81 ± 0.23	100%
	Fixed ratio 0.8	0.8345 ± 0.0017	0% (baseline)	36.72 ± 0.22	0%
	Entropy-Valley	**0.8517 ± 0.0006**	**64.9 ± 7.4%**	**38.57 ± 0.13**	45.3%
Zh→En	Oracle length	0.8519 ± 0.0007	100%	27.93 ± 0.23	100%
	Fixed ratio 1.2	0.8266 ± 0.0010	0% (baseline)	23.65 ± 0.21	0%
	Entropy-Valley	**0.8431 ± 0.0004**	**65.3 ± 0.8%**	**25.28 ± 0.33**	38.1%
En→De	Oracle length	0.7382 ± 0.0090	100%	22.55 ± 0.77	100%
	Fixed ratio 1.8	0.7170 ± 0.0090	0% (baseline)	20.73 ± 0.68	0%
	Entropy-Valley	**0.7240 ± 0.0078**	**33.0 ± 8.4%**	**21.55 ± 0.85**	45.1%

Table 1 Main LLaDA-8B+LoRA results on WMT22 with 32-step MED decoding. Values are mean ± std over three runs. Length Oracle uses the reference target length at decode time and is an upper bound. Gap closure is (method − Ratio)/(Oracle − Ratio). Paired tests are in Appendix Table 11.

4.3 Reference and matched-data baselines

Figure 3B and Table 2 use a matched autoregressive system as a calibration point, not as a target-canvas baseline. With the same fine-tuning data, LLaDA with Entropy-Valley is essentially tied with LLaMA-3-8B + LoRA-SFT on En→Zh and is higher on Zh→En. This calibrates the masked-diffusion setting without changing the claim: EV addresses canvas selection within a fixed backbone.

En→De gives the boundary. The matched autoregressive system is about 7.4 COMET points above the LLaDA Length Oracle row and about 8.8 points above LLaDA with Entropy-Valley. The gap remains even when the reference target length is supplied, so the En→De shortfall is not mainly a target-canvas selection error. EV is scoped to fixed-backbone canvas selection.

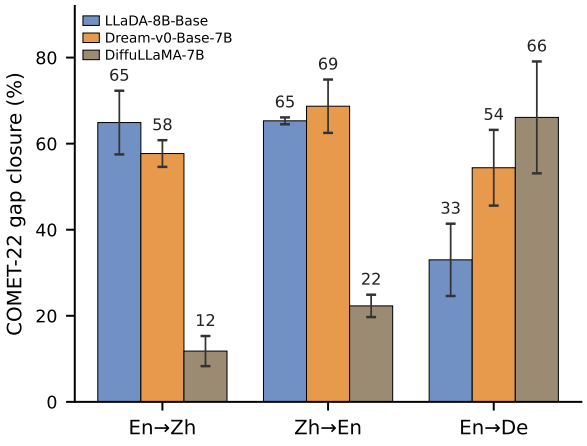

Figure 4 Cross-backbone scope check using COMET-22 gap closure (%) for three masked-diffusion backbones and three directions. Gap closure is measured between Ratio and Length Oracle for each backbone–direction pair. Bars show means over three runs; error bars show standard deviations across runs. Full COMET-22 and sacreBLEU values are in Appendix G (Tables 17 and 18).

System	Size	Setting	En→Zh	Zh→En	En→De
LLaDA-8B masked diffusion					
Fixed ratio	8B	LoRA-SFT	83.45	82.66	71.70
Entropy-Valley	8B	LoRA-SFT	**85.17**	**84.31**	72.40
Oracle length[†]	8B	LoRA-SFT	86.10	85.19	73.82
Matched-data autoregressive baseline					
LLaMA-3-8B + LoRA-SFT	8B	LoRA-SFT	85.07±0.03	83.68±0.07	**81.19±0.11**

Table 2 Matched-data COMET-22 calibration on WMT22. LLaDA rows show the Table 1 means over three runs scaled ×100; the matched AR row reports mean±std over three runs after training on the same fine-tuning data and evaluating on the same test sets. [†]Oracle length supplies the reference target length at decode time and is an upper bound, not a deployable method.

4.4 Scope across masked-diffusion backbones

We repeat the Length Oracle, Ratio, and Entropy-Valley protocol on Dream-Base and DiffuLLaMA to test if the length-selection pattern is specific to LLaDA. These runs use the same 200k SFT data scale, MED decoding protocol, and three directions as the main evaluation. Figure 4 summarizes the cross-backbone gap-closure pattern.

Across all tested directions and backbones, Entropy-Valley remains above Ratio, but the amount of oracle-gap closure changes by model family. This supports the entropy-based length criterion beyond LLaDA while keeping the claim scoped to the tested systems.

The variation is not a causal attribution. Dream-Base and DiffuLLaMA change tokenizer, initialization, pretraining mixture, and architecture at the same time. Their larger En→De closures do not identify which factor helps, and DiffuLLaMA's smaller Chinese-side closures are consistent with limited Chinese tokenizer coverage.

4.5 Controls for the EV gain

Entropy-Valley changes only the target-canvas choice. The controls below ask whether the main gains instead come from reveal order, reference-length matching, candidate-window tuning, a better global ratio, or metric-only artifacts.

Length choice versus reveal order. The order diagnostic fixes the target length to the reference length and changes only the reveal schedule. On En→Zh, the EV–Ratio difference is larger than the full span across the tested reveal orders. Strict left-to-right and random order also score below MED. In this controlled setting, target canvas choice produces the larger variation among the evaluated decoding decisions, while poor reveal orders can still hurt (Appendix Table 15).

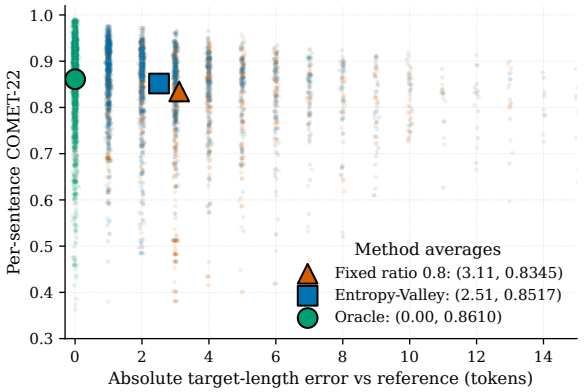

Figure 5 Quality need not track reference length. Target length error and COMET-22 on En→Zh WMT22 (N=2037). Points show sentence scores; MAE values use the same outputs, and centroid heights use the main results in Table 1.

Not reference matching. EV does not aim to predict the reference length. Figure 5 compares target length error with COMET-22 on En→Zh. EV remains far from Length Oracle in length error, yet recovers about 65% of the Length Oracle–Ratio COMET gap. This pattern supports the method framing: entropy selects a canvas the backbone can denoise well, rather than a human reference length.

Candidate range. The candidate range defines what EV can choose. Appendix Table 6 shows that wider windows are not uniformly better, supporting a compact fixed window rather than unrestricted length search.

Fixed ratio sweep. EV is not just a better global ratio. Appendix Table 7 sweeps fixed ratios around each corpus median on 500-sentence diagnostic subsets. EV remains above the best fixed ratio in all three directions. A single ratio applies the same compression rule to every source sentence; EV changes the canvas per sentence, which is why retuning one ratio for the whole corpus does not reproduce the gain.

Diagnostic	N	Ratio	EV
Placeholder retention	58	66.9%	**89.1%**
Number retention	290	77.6%	**81.3%**
$r \geq 0.8$ COMET	1005	0.8056	**0.8402**
$r < 0.6$ COMET	236	0.8449	**0.8503**

Table 3 Coverage and length buckets on En→Zh. Retention rows measure whether literal source items such as placeholders and numbers are preserved. COMET rows group sentences by r, the reference target length divided by the source length under the LLaDA tokenizer. Appendix D gives the definitions, confidence intervals, and Zh→En companion rows.

Multi-candidate decoding. A heavier decode-many-then-score alternative does not explain the result. Appendix Table 8 decodes several neighboring canvas lengths to completion and then selects one by average log-probability. Unlike EV's up to five one-pass probes (§3.6), this alternative pays for multiple completed decodes and is lower-scoring in the tested setting.

Direct diffusion length baselines. On independent runs for each direction using the same translation setup, EV exceeds DAEDAL (Li et al., 2026) in both directions and CAL (Liu et al., 2026) on En→Zh. EV and CAL are comparable on Zh→En. EV uses lower measured inference cost than CAL in this setup (Appendix Table 9).

Coverage and limits. The automatic gains align with source coverage, not only with average COMET. Table 3 focuses on diagnostics tied to the length-selection story. On En→Zh, EV preserves more placeholders and numbers than Ratio. These items are sensitive to canvas length because a short canvas can drop literal content before reveal order matters.

The same buckets show the current boundary. The largest COMET gain appears in the dominant $r \geq 0.8$ bucket, where the default candidate grid is well aligned. The $r < 0.6$ bucket still improves on average, but individual failures can occur when the needed compression lies below the grid. Figure 1 shows such a case, and Appendix E gives the full failure breakdown.

4.6 Human evaluation on En↔Zh

The human study checks whether the automatic En↔Zh gains are also reflected in human adequacy judgments. Three experts bilingual in Chinese and English compared anonymised Entropy-Valley and Ratio outputs. We use 100 stratified WMT22 sentences for En→Zh and another 100 for Zh→En,

Direction	ΔAdequacy EV − Ratio	ΔFluency EV − Ratio	Preference EV / Ratio / Tie
En→Zh	+0.18	+0.18	28 / 19 / 53
Zh→En	**+0.50**	+0.11	**45 / 25 / 30**

Table 4 Human evaluation on 100 stratified WMT22 sentences per direction. Score differences are EV − Ratio, averaged over three bilingual expert annotators; preference is the per-sentence majority vote.

with the source and gold translation shown. The full protocol, rater statistics, and metric correlation results are reported in Appendix I.

Human judgments follow the automatic pattern, with clearer evidence on Zh→En. The main signal is adequacy rather than fluency: EV raises adequacy by +0.50 on Zh→En and wins more majority preferences than Ratio (45 vs. 25). On En→Zh, the adequacy margin is smaller (+0.18), and most sentences are ties under majority preference. Fluency changes are small in both directions. This pattern matches the length-selection story: EV helps preserve source content rather than improving fluency across the board.

4.7 Discussion: what Entropy-Valley selects

The diagnostics identify EV as a canvas selector, not a reference-length predictor. Figure 5 shows that high translation quality does not require matching the reference length exactly. Table 3 gives the complementary mechanism: when a fixed ratio allocates too few slots, the selected canvas can preserve source-side placeholders and numbers that would otherwise be dropped. The useful signal is the backbone's uncertainty about how well it can fill a candidate canvas before decoding begins.

The boundary comes from the candidate set. The fixed ratio set keeps EV free from extra training, but it also limits the canvases EV can choose at inference time. If the needed compression falls outside this set, as in Figure 1, the entropy score cannot select the missing canvas. Figure 4 shows a second boundary: EV can select a better canvas for each tested backbone, but gains remain bounded by the backbone and tokenizer.

5 Conclusion

This paper identifies target-canvas selection as a structural bottleneck in fixed-canvas masked-diffusion MT. The issue is not only how to reveal masked positions, but which canvas the decoder

is asked to fill before denoising begins. A short canvas can drop source content; an overlong canvas can create unsupported slots. The central takeaway is that masked-diffusion MT needs an explicit test-time length decision.

Entropy-Valley turns this decision into a model-internal query. By probing a small set of all-mask canvases, EV chooses the candidate the backbone appears most prepared to denoise, without adding a length head or using reference lengths. Across the tested directions, this simple selector improves over a fixed corpus ratio, aligns with human adequacy judgments on En↔Zh, and preserves literal source content that short canvases can lose. In the controlled En→Zh comparison, target canvas choice produces the larger variation among the evaluated decoding decisions, while strict left-to-right and random reveal orders also reduce quality.

Limitations

Our deepest evaluation uses LLaDA-8B-Base. Dream-Base and DiffuLLaMA extend the same protocol to two additional backbones, and EV remains above Ratio in every tested cell. Absolute scores and oracle-gap closures still vary with the backbone and tokenizer, so broader generalization requires more model families and language pairs.

En→De is the clearest boundary for the current LLaDA system. EV-LLaDA improves over Ratio, but the sentence-level evidence is weaker than on En↔Zh (Appendix C). The matched AR baseline is about 8.8 COMET points above EV-LLaDA and about 7.4 points above even the LLaDA length oracle on the 0–100 scale, so the En→De gap is not mainly a target-canvas selection error. The step-budget sweep gives the same warning from another angle: at $T \geq 64$, the fixed En→De ratio slightly exceeds EV (Appendix Figure 6 and Table 10).

The human evaluation is limited to En↔Zh. These judgments support the automatic adequacy gains most clearly on Zh→En, while En→Zh is positive but weaker under preference testing. We did not run human evaluation for En→De or De→Fr, so conclusions for those directions rely on automatic metrics; the En→De analysis also includes matched AR calibration.

The fixed candidate ratio set is another intended constraint. The reported main grids are direction specific and were informed by comparisons on WMT22 diagnostic subsets. The De→Fr transfer uses one fixed grid without changes, but broader transfer of the procedure for choosing the grid remains open. The fixed set keeps EV training-free and prevents an unrestricted length search, but it also limits outlier sentences whose target length requires stronger compression. With $\mathcal{R}=\{0.70, \ldots, 0.90\}$ on En→Zh, EV cannot choose a canvas below $0.70|\mathbf{x}|$; Appendix E gives such a failure case. A natural next step is a compression-aware candidate generator that expands or shifts $\mathcal{R}$ when the entropy curve suggests that the fixed window is insufficient.

Ethical Considerations

EV changes only inference-time canvas selection and does not alter the backbone or training objective. The resulting system still inherits risks from the backbone, translation fine-tuning data, and decoding procedure. Because backbone pretraining data and tokenization vary across model families and scales, deployment should evaluate fairness and robustness for the intended language pairs and user domains. Automatic quality gains do not remove errors in individual sentences, so use in medical, legal, and other high risk settings should retain human review.

We use public WMT news benchmarks and evaluation packages under their public research terms or licenses. WMT news data may contain names or sensitive events already present in public news text. We release the processed experiment datasets and trained LoRA adapters described in Appendix A. The datasets remain subject to the original WMT research terms, and the adapters retain the applicable base model license. Backbone weights are obtained from their original providers. We do not collect private user data or attempt to identify individuals.

The three evaluators were professional translators bilingual in Chinese and English. They rated both En→Zh and Zh→En batches using public WMT22 sentences and outputs with system identities hidden. They were informed of the study purpose, worked independently, consented to the use of ratings without personal identifiers and aggregate statistics for research reporting, and were compensated at local professional translator hourly rates. The task collected no sensitive personal data and, under the rules of our institution, did not require formal ethics review.

References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

Shaik Aman. 2026. LogicDiff: Logic-guided denoising improves zero-shot reasoning in masked diffusion language models. *Preprint*, arXiv:2603.26771.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, volume 34, pages 17981–17993.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. 2022. MaskGIT: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325.

Zicong Cheng, Ruixuan Jia, Jia Li, Guo-Wei Yang, Meng-Hao Guo, and Shi-Min Hu. 2026. Improving variable-length generation in diffusion language models via length regularization. *Preprint*, arXiv:2602.07546.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 6112–6121.

Shansan Gong, Shivam Agarwal, Yizhe Zhang, Jiacheng Ye, Lin Zheng, Mukai Li, Chenxin An, Peilin Zhao, Wei Bi, Jiawei Han, Hao Peng, and Lingpeng Kong. 2025. Scaling diffusion language models via adaptation from autoregressive models. In *International Conference on Learning Representations*.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2023. DiffuSeq: Sequence to sequence text generation with diffusion models. In *International Conference on Learning Representations*.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *International Conference on Learning Representations*.

Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. Levenshtein transformer. In *Advances in Neural Information Processing Systems*, volume 32.

Chunsan Hong, Sanghyun Lee, and Jong Chul Ye. 2026. Unifying masked diffusion models with various generation orders and beyond. In *International Conference on Machine Learning*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Jaeyeon Kim, Cheuk-Kit Lee, Carles Domingo-Enrich, Yilun Du, Sham Kakade, Timothy Ngotiaco, Sitan Chen, and Michael S. Albergo. 2026. Any-order flexible length masked diffusion. In *International Conference on Learning Representations*.

Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jiaqi Wang, and Dahua Lin. 2026. Beyond fixed: Training-free variable-length denoising for diffusion large language models. In *International Conference on Learning Representations*.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022. Diffusion-LM improves controllable text generation. In *Advances in Neural Information Processing Systems*, volume 35, pages 4328–4343.

Hengchang Liu, Zhao Yang, and Bing Su. 2026. Diffusion LMs can approximate optimal infilling lengths implicitly. *Preprint*, arXiv:2602.00476.

AI @ Meta Llama Team. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Aaron Lou, Chenlin Meng, and Stefano Ermon. 2024. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Proceedings of the 41st International Conference on Machine Learning*.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Large language diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, pages 50608–50646.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Lihua Qian, Hao Zhou, Yu Bao, Mingxuan Wang, Lin Qiu, Weinan Zhang, Yong Yu, and Lei Li. 2021. Glancing transformer for non-autoregressive neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1993–2003.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 578–585.

Vittorio Rossi, Giacomo Cirò, Davide Beltrame, Luca Gandolfi, Paul Röttger, and Dirk Hovy. 2026. Diffusion language models are natively length-aware. *Preprint*, arXiv:2603.06123.

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

Minghan Wang, Jiaxin Guo, Yuxia Wang, Yimeng Chen, Chang Su, Hengchao Shang, Min Zhang, Shimin Tao, and Hao Yang. 2021. How does length prediction influence the performance of non-autoregressive translation? In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 205–213.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. A paradigm shift in machine translation: Boosting translation performance of large language models. In *International Conference on Learning Representations*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024b. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In *Proceedings of the 41st International Conference on Machine Learning*.

Jingyi Yang, Yuxian Jiang, and Jing Shao. 2026. ρ-eos: Training-free bidirectional variable-length control for masked diffusion LLMs. *Preprint*, arXiv:2601.22527.

Yicun Yang, Cong Wang, Shaobo Wang, Zichen Wen, Biqing Qi, Hanlin Xu, and Linfeng Zhang. 2025. Diffusion LLM with native variable generation lengths: Let [EOS] lead the way. *Preprint*, arXiv:2510.24605.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025. Dream 7b: Diffusion large language models. *Preprint*, arXiv:2508.15487.

A Reproducibility and Configuration

Unless stated otherwise, the LLaDA experiments use the following configuration. Subset analyses that change N are identified at the relevant table.

Backbone. LLaDA-8B-Base (Nie et al., 2025); Transformer, 8.02B parameters. We use the released weights without modification. LoRA (Hu et al., 2022) targets seven modules per Transformer block: the attention projection matrices $\{q, k, v, o\}$ and the FFN matrices $\{\text{ff\_proj}, \text{up\_proj}, \text{ff\_out}\}$ (no adaptation of embeddings): rank 64, α=128, dropout 0.05. Trainable parameters $\approx$ 157M (1.95% of the backbone). bf16 training and inference throughout.

Training. AdamW optimiser (β_1=0.9, β_2=0.95, weight decay 0.01); cosine learning-rate schedule with 5% warm-up, peak LR 2×10^{-4}; three epochs per direction with a global batch size of 128 sequences (gradient accumulation =4 across 8×H20 GPUs, per-device micro-batch =4). Maximum canvas length 1024 tokens (source prompt + target canvas). Training loss is the standard masked-diffusion cross-entropy with a uniform masking schedule sampled per minibatch.

Hyperparameter	Value
Optimizer	AdamW
AdamW betas	(0.9, 0.95)
Weight decay	0.01
Peak learning rate	2×10^{-4}
Scheduler	Cosine decay
Warm-up	5% of training steps
Epochs	3 per direction
Global batch size	128 sequences
Gradient accumulation	4 steps
Per-device micro-batch	4 sequences
Hardware	8×H20-96GB GPUs
Precision	bf16 training and inference
Maximum canvas length	1024 tokens
LoRA rank / alpha / dropout	64 / 128 / 0.05
Model selection	Fixed final checkpoint per run; no EV-specific dev tuning

Table 5 Training and adaptation hyperparameters for the LLaDA and matched AR LoRA-SFT runs.

Matched AR baseline. The matched autoregressive baseline uses LLaMA-3-8B-Base (Llama Team, 2024) with LoRA rank 64, α=128, dropout 0.05, and the corresponding seven target module families $\{q, k, v, o, \text{gate}, \text{up}, \text{down}\}$_proj. Here matched means that the AR and LLaDA systems use the same SFT pairs, prompt family, LoRA scale, and evaluation data; the module names follow each backbone's architecture. The AR baseline decodes with beam size 4 and native EOS termination.

Data and runs. En↔Zh uses the HuggingFace `wmt19/zh-en` training split, and En→De uses `wmt19/de-en`. We apply only monolingual length filters (English and German: 5–200 whitespace-tokenised words; Chinese: 2–600 characters), then deterministically sample 200k training pairs and 2k development pairs per setting. The WMT19 development pairs are not used to set the reported ratio grids or decoding hyperparameters. Three independent training runs per direction yield nine checkpoints. En→De is included as a typologically more distant check, not as a SoTA claim.

Data license and privacy. All training and test examples come from public WMT benchmark datasets distributed for machine-translation research. The WMT news benchmarks are public datasets and may contain names or sensitive events already present in public news text. We do not collect private user data, do not attempt to identify individuals, and do not redistribute raw corpora. Human-evaluation forms contain only public WMT sentences and anonymised system outputs. We use released model weights, evaluation packages, and WMT benchmark data under their respective public research terms or licenses; we do not redistribute the raw WMT corpora or third-party model weights in the public release.

Prompt templates. Source is formatted as a zero-shot instruction:

Direction	Prompt template
En→Zh	`Translate English to Chinese.\n\nEnglish: {src}\nChinese:`
Zh→En	`Translate Chinese to English.\n\nChinese: {src}\nEnglish:`
En→De	`Translate English to German.\n\nEnglish: {src}\nGerman:`

Default decoding. Default decoder is MED at T=32. EOS is not constrained; the decoded string is truncated at the first EOS.

Compared length methods. The main text compares Length Oracle (reference length, upper bound), Ratio (a fixed training-corpus ratio baseline set before evaluation), and Entropy-Valley. The fixed ratios are 0.8 for En→Zh, 1.2 for Zh→En, and 1.8 for En→De. The Pareto sweep uses $T \in \{2, 4, 8, 16, 32, 64, 128\}$.

Evaluation. We evaluate on the WMT22 News Translation test set, N=2037 per direction. En→Zh and Zh→En use the same WMT22 sentence pairs with source and target roles swapped; En→De uses the WMT22 En–De

test set. COMET-22 uses `unbabel-comet` v2.2 with `Unbabel/wmt22-comet-da`. BLEU denotes sacreBLEU throughout the paper. The En→Zh signature is `nrefs:1|case:mixed|eff:no|tok:zh|smooth:exp|version:2.4.0`; other directions use `tok:13a`.

Significance. Paired bootstrap 10k resamples, Wilcoxon signed-rank (pratt), paired Cohen's d.

Compute. About 420 H20-96GB GPU-hours for the full experimental grid. At T=32 on a single H20 with batch size 16, wall-clock decoding is roughly 1.5 s per sentence for Ratio and 1.7 s per sentence for Entropy-Valley on En→Zh. The EV probe cap adds at most 15.6% to the forward pass count at T=32, 7.8% at T=64, and 3.9% at T=128; duplicate integer lengths can reduce the realized count.

Resource availability. We release the implementation, experiment configurations, decoding and evaluation scripts, and human evaluation templates at https://github.com/Entropy-Valley/Entropy-Valley. The processed experiment datasets and trained LoRA adapters for En→Zh, Zh→En, and En→De are available at https://huggingface.co/collections/YanZhanPKU/entropy-valley. The repository documents the required third-party backbones and applicable resource licenses.

B Fixed canvas protocol and the length and order dissociation

Two protocol assumptions underlie the main text: the target canvas is an external decoding input, and any comparison of length choices needs the reveal order held fixed.

Canvas convention. Source conditioning is injected via prompt concatenation, so the model input is $[\,\text{prompt}(\mathbf{x})\,] \,\|\, [\textsc{Mask}]^L$. Decoding runs for T steps; each step unmasks a fraction of the canvas under schedule π. MED reveals the K_t positions with lowest entropy at each step.

Length is external. LLaDA-style fixed-canvas decoding allocates the target canvas before denoising rather than learning an AR-style stopping rule. The decoder fills the mask slots it is given and has no separate process for choosing how many slots to allocate. Under our shared protocol, L is supplied before denoising and decoded strings are truncated at the first EOS.

Dissociating length from order. Diffusion MT quality depends on both the chosen canvas length L and the order schedule π. For order comparisons we fix L to the reference length L^*; for length comparisons we fix π to MED at T=32. This separates target canvas choice from reveal order in the comparison reported in §4.5.

Notation and length unit. Let $\mathbf{x}$ be the source sentence, $\mathbf{y}$ the target canvas, and L^* the reference length. Length is the number of target canvas slots, i.e., the L in $[\,\text{prompt}(\mathbf{x})\,] \,\|\, [\textsc{Mask}]^L$, and excludes the source prompt. The final slot follows the EOS inclusive training format, and the decoded string is truncated at the first EOS. EV excludes this designated EOS slot when averaging entropy, so the score uses the first $L-1$ slots. Candidate ratios are computed against source tokenizer length under the same tokenizer.

Length and order diagnostic. On En→Zh at T=32 (N=2037), the static-ratio baseline leaves a 0.0265 COMET-22 gap to the length oracle. With L held at the reference length, the tested source-guided order schedules span 0.0081 COMET-22. The EV–Ratio difference is 2.1× the tested order span; the reference-length-inclusive span is 3.3× the order span. Table 15 also reports strict left-to-right and random controls.

C Controls for the EV gain

The controls below examine the candidate range, a global ratio, decode and rerank, decoding budget, probe cost, and sentence-level variation. Subset controls use fixed WMT22 diagnostic subsets with the same prompt, decoding, and evaluation protocol as §4.1.

C.1 Selector checks

These checks use an En→Zh run trained under the same configuration. On a dense sweep over ratios $[0.40, 1.50]$ for 500 sentences, aggregate mean entropy reaches its minimum at ratio 0.70 (3.7695) and rises to 5.3031 at ratio 1.50. In the full test set, EV selects all five nominal ratios.

A realized-ratio reassignment preserves the selected ratio distribution while changing the pairing between each source and its canvas. EV remains $+0.0040$ COMET above the sentencewise mean reassignment (95% CI $[0.0023, 0.0058]$). Reassignment within source-length deciles gives the same conclusion. Length-normalized alternatives behave similarly, while unnormalized summed entropy performs worse. These controls support source-conditioned canvas selection and mean entropy as the selection rule.

C.2 Candidate range sensitivity

The WMT19 training-corpus median provides a reference scale, while diagnostic range comparisons on WMT22 subsets informed the reported ratio grid. Table 6 asks how narrowing or widening that grid changes the result. On these 200-sentence subsets, the En→Zh grid has the lowest length MAE and the highest sacreBLEU, and the En→De grid has the highest subset COMET. The Zh→En subset prefers a wider grid than the reported one, showing sensitivity to higher target length variance.

Range	K	Ratios	MAE	sBLEU	COMET
En→Zh (N=200)					
Reported (narrow)	5	[0.70, 0.90]	**2.71**	**38.73**	0.8562
Medium	7	[0.60, 1.00]	2.84	38.49	**0.8594**
Wide	9	[0.50, 1.10]	3.06	38.17	0.8588
Full	12	[0.50, 1.20]	3.22	37.58	0.8577
Extra	15	[0.40, 1.50]	3.89	35.98	0.8476
En→De (N=200)					
Reported (narrow)	5	[1.50, 1.90]	4.08	22.29	**0.7305** (56.0% gap)
Medium	7	[1.30, 2.10]	**3.90**	**22.36**	0.7300 (53.4%)
Wide	9	[1.20, 2.20]	3.92	22.32	0.7267 (36.1%)
Full	11	[1.00, 2.40]	3.98	21.97	0.7268 (36.6%)
Zh→En (N=200)					
Reported (narrow)	5	[1.10, 1.30]	—	27.64	0.8546
Medium	7	[1.05, 1.35]	—	27.62	0.8554
Wide	9	[1.00, 1.40]	—	**27.77**	0.8566
Full	12	[0.95, 1.50]	—	27.63	**0.8586**

Table 6 Candidate-range control for Entropy-Valley on fixed 200-sentence WMT22 diagnostic subsets. The WMT19 training-corpus median provides a reference scale, while diagnostic range comparisons on WMT22 subsets informed the reported ratio grid. The En→Zh grid minimizes length MAE and gives the highest sacreBLEU, while the medium range gives the highest subset COMET; the En→De grid gives the highest subset COMET. The Zh→En subset benefits from a wider range.

C.3 Fixed ratio and multiple candidate controls

Two simpler alternatives could in principle explain the gain: retune the single global ratio per direction, or decode several neighbouring lengths and rerank. Neither closes the gap. Sweeping post-hoc fixed ratios on the 500-sentence subsets (Table 7) leaves EV above the best fixed ratio in all three directions, because per-sentence selection cannot be replaced by one corpus-wide ratio. The heavier alternative in Table 8 decodes five neighbouring canvases to completion and selects by average log-probability; it pays roughly $3.5\times$ the cost of Ratio and still scores below EV on the same subset.

Method	sacreBLEU	COMET-22
En→Zh (N=500, 32 steps)		
Oracle	40.40	0.8617
Fixed 0.70	31.73	0.8114
Fixed 0.75	34.59	0.8208
Fixed 0.80	35.84	0.8281
Fixed 0.85	36.84	0.8354
Fixed 0.90	36.93	0.8400
Entropy-Valley	**38.48**	**0.8557**
En→De (N=500, 32 steps)		
Oracle	22.04	0.7433
Fixed 1.4	16.88	0.6930
Fixed 1.5	18.57	0.7089
Fixed 1.6	19.28	0.7134
Fixed 1.7	20.12	0.7187
Fixed 1.8	19.64	0.7219
Fixed 1.9	19.91	0.7227
Fixed 2.0	18.67	0.7186
Fixed 2.1	18.05	0.7086
Fixed 2.2	17.29	0.7038
Entropy-Valley	**21.24**	**0.7327**
Zh→En (N=500, 32 steps)		
Oracle	28.42	0.8568
Fixed 1.0	19.48	0.8099
Fixed 1.1	21.71	0.8196
Fixed 1.2	24.46	0.8298
Fixed 1.3	**25.57**	0.8372
Fixed 1.4	24.62	0.8390
Entropy-Valley	25.45	**0.8453**

Table 7 Fixed-ratio sweep on fixed 500-sentence WMT22 diagnostic subsets. EV is compared with fixed ratios around the corpus median for all three directions. On Zh→En, the best post-hoc fixed ratio is 1.4 at 0.8390 COMET-22; EV reaches 0.8453, exceeding it by +0.0063 COMET and closing 35.4% of the oracle-length gap on this subset.

Method	sacreBLEU	COMET-22	Cost
Oracle length	40.40	0.8617	1.00×
Fixed ratio 0.8	35.84	0.8281	1.00×
EV probe+decode	**38.48**	**0.8557**	1.15×
Decode 5 neighbours	37.85	0.8404	∼ 3.5×

Table 8 Multi-candidate length selection on the same En→Zh WMT22 subset used in Table 7 (N=500, 32 steps). The alternative decodes five candidate canvas lengths obtained as integer offsets around the EV-selected length, $\{L_{\text{EV}}-2, L_{\text{EV}}-1, L_{\text{EV}}, L_{\text{EV}}+1, L_{\text{EV}}+2\}$ (with length collisions deduplicated), to completion and selects by average log-probability; its cost counts all completed decodes.

C.4 Direct diffusion length baselines

These controls use one run per direction under the main configuration. Within each direction, every method uses the same model, inputs, and evaluation code. We compare EV with DAEDAL (Li et al., 2026) and CAL (Liu et al., 2026).

Dir.	Method	COMET	EV−method	Calls	Slots	Latency (s)
En→Zh	Ratio	0.833721	+0.010930	32.00	512.66	0.644
En→Zh	**EV**	**0.844651**	–	36.09	588.32	0.724
En→Zh	DAEDAL	0.828043	+0.016608 ([0.012954, 0.020210]; $p < 2\times10^{-4}$)	34.58	1175.53	1.798
En→Zh	CAL	0.837897	+0.006754 ([0.003726, 0.009880]; $p < 2\times10^{-4}$)	39.25	628.26	0.773
Zh→En	Ratio	0.818169	+0.016416	32.00	618.43	0.635
Zh→En	**EV**	**0.834585**	–	36.43	708.98	0.721
Zh→En	DAEDAL	0.812488	+0.022097 ([0.018849, 0.025375]; $p < 2\times10^{-4}$)	37.49	1533.16	1.972
Zh→En	CAL	0.833352	+0.001233 ([−0.001248, 0.003686]; $p = 0.3342$)	39.42	786.51	0.770

Table 9 Direct diffusion length comparisons on WMT22 (N=2037 per direction). Confidence intervals and paired bootstrap p-values are shown for DAEDAL and CAL. Calls and Slots are per-sentence averages; latency is measured in seconds per sentence in the reported setup. EV is above DAEDAL in both directions and CAL on En→Zh. The EV and CAL interval on Zh→En includes zero. EV uses fewer calls and slots than CAL, with lower latency in this setup.

C.5 Decoding budget robustness

The next control varies the decoding-step budget $T \in \{2, 4, 8, 16, 32, 64, 128\}$ on the full WMT22 test set for one analyzed run (N=2037). On En→Zh, EV is at least as high as Ratio at every step budget. On En→De, EV leads at small T and falls slightly below Ratio at $T \geq 64$; this is the boundary recorded in the Limitations section.

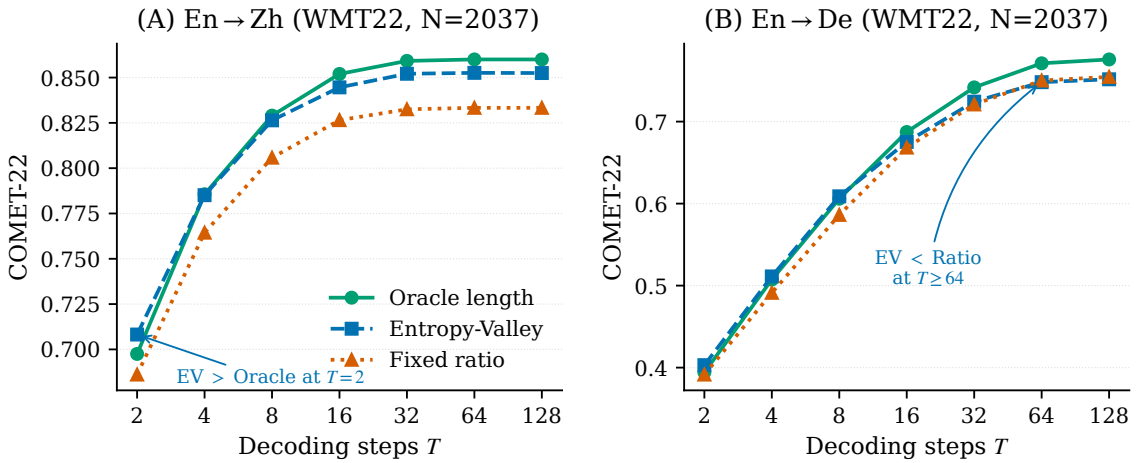

Figure 6 Cost–quality Pareto. COMET-22 vs decoding steps $T \in \{2, 4, 8, 16, 32, 64, 128\}$ on WMT22 (N=2037, one analyzed run). (A) En→Zh: EV is at least as high as Ratio at every T. (B) En→De: EV is higher at small T and slightly lower than Ratio at $T \geq 64$. Oracle is a length oracle, not a decode oracle.

	En→Zh				En→De			
Steps	Oracle	EV	Ratio	Gap%	Oracle	EV	Ratio	Gap%
2	0.6975	0.7082	0.6861	**194%**	0.3942	0.4030	0.3915	**426%**
4	0.7856	0.7851	0.7644	98%	0.5074	0.5112	0.4912	**123%**
8	0.8290	0.8264	0.8058	89%	0.6060	0.6089	0.5862	**115%**
16	0.8520	0.8446	0.8265	71%	0.6873	0.6752	0.6683	36%
32	0.8592	0.8521	0.8325	73%	0.7417	0.7243	0.7209	16%
64	0.8600	0.8526	0.8333	72%	0.7711	0.7481	0.7501	−10%
128	0.8600	0.8525	0.8333	72%	0.7758	0.7516	0.7545	−14%

Table 10 Full cost–quality Pareto on WMT22 (N=2037, one analyzed run). Gap% is computed relative to the Ratio and Oracle length columns at the same step budget; values can exceed 100% because Oracle fixes reference length but does not optimize decoding under small T.

C.6 Paired significance and compute budget

The final controls address statistical noise and probe cost. Table 11 reports paired bootstrap and Wilcoxon tests for the En↔Zh and En→De main comparisons. The En→Zh comparison is reliable in run A, and Zh→En is reliable across all three runs. En→De passes Wilcoxon but not bootstrap, matching the weaker sentence-level evidence noted in §4.2. Table 12 evaluates Ratio at EV's maximum total forward pass budget, T=37, and at a larger T=40 budget. The extra budget recovers at most 0.001 COMET, well below the En↔Zh EV gains, so the gain is the length choice rather than the additional compute.

Dir.	Comp.	Δ	95% CI	Boot. p	Wilc. p
En→Zh (A)	EV–Ratio	+0.0196	[+0.0166, +0.0226]	$< 10^{-4}$	4.1×10^{-32}
En→Zh (A)	Oracle–EV	+0.0071	[+0.0046, +0.0096]	$< 10^{-4}$	1.3×10^{-14}
En→Zh (A)	Oracle–Ratio	+0.0266	[+0.0233, +0.0300]	$< 10^{-4}$	3.4×10^{-58}
En→De (A)	EV–Ratio	+0.0034	[−0.0016, +0.0086]	0.091	8.3×10^{-3}
En→De (A)	Oracle–EV	+0.0173	[+0.0121, +0.0225]	$< 10^{-4}$	1.6×10^{-9}
En→De (A)	Oracle–Ratio	+0.0208	[+0.0155, +0.0261]	$< 10^{-4}$	6.0×10^{-14}
Zh→En (A)	EV–Ratio	+0.0174	[+0.0150, +0.0198]	$< 10^{-4}$	6.9×10^{-45}
Zh→En (B)	EV–Ratio	+0.0154	[+0.0131, +0.0178]	$< 10^{-4}$	1.9×10^{-41}
Zh→En (C)	EV–Ratio	+0.0167	[+0.0143, +0.0192]	$< 10^{-4}$	2.1×10^{-45}

Table 11 Sentence-level paired COMET-22 tests for LLaDA on WMT22 (N=2037 throughout). Positive Δ means the first method is better. Paired bootstrap uses 10k resamples; Wilcoxon uses the signed-rank test with Pratt handling of zeros. The table reports run A for En→Zh and En→De, plus all three Zh→En runs. Oracle rows show the run A gap to the reference-length upper bound.

Direction	Method	T	Probe cap	Fwd cap	COMET-22	sacreBLEU
En→Zh, one analyzed run, WMT22 N=2037						
En→Zh	Ratio 0.8	32	0	32	0.8325	36.53
En→Zh	**Entropy-Valley**	**32**	**≤ 5**	**≤ 37**	**0.8521**	**38.56**
En→Zh	Ratio 0.8	37	0	37	0.8334	36.66
En→Zh	Ratio 0.8	40	0	40	0.8333	36.71
Zh→En, one analyzed run, WMT22 N=2037						
Zh→En	Ratio 1.2	32	0	32	0.8260	23.53
Zh→En	**Entropy-Valley**	**32**	**≤ 5**	**≤ 37**	**0.8434**	**25.16**
Zh→En	Ratio 1.2	37	0	37	0.8266	23.71
Zh→En	Ratio 1.2	40	0	40	0.8269	23.73

Table 12 Forward-pass budget control. Ratio is evaluated at the default budget (T=32), EV's maximum total forward pass budget (T=37), and a larger budget (T=40). Duplicate integer candidate lengths can make EV's realized budget smaller than the cap.

D Coverage losses from short canvases

The analysis in this section shows that EV preserves source-side content that short Ratio canvases can omit. The evidence differs across directions and length buckets.

Error-type categories. For every WMT22 source sentence we extract three categories whose retention can be detected by literal substring match in the target output: anonymisation **placeholders** (`#PRS_ORG#`, `#DATE#`, etc.), **numbers** (integers, decimals, percentages, dollar amounts), and **capitalised named-entity tokens** (heuristic; En→Zh only, since Chinese sources do not use capitalisation). Each sentence is also bucketed by the LLaDA-token-level reference-to-source ratio $r = |L^*|/|\mathbf{x}|$. Retention rates are averaged per sentence over the three runs used in §4.2, and the EV−Ratio difference is reported with paired bootstrap 95% CIs (10k resamples). The outputs are not re-decoded for this analysis.

Literal retention. On En→Zh, placeholder retention rises from 66.9% to 89.1% (+22.1 pp) and number retention from 77.6% to 81.3% (+3.7 pp), both with bootstrap CIs strictly above zero (Table 13). Capitalised-NE retention moves little because En→Zh often transliterates entities; this category is heuristic rather than diagnostic. On Zh→En, Ratio already retains 96.6% of placeholders and 80.7% of numbers, so the EV gains of +2.9 and +1.9 pp are too small to account for the Zh→En COMET improvement on their own. The Zh→En mechanism must involve non-literal content that our error-type categories do not directly probe.

Length buckets. The largest En→Zh COMET gain (+0.0346) is in the dominant $r \geq 0.8$ bucket where the fixed candidate grid is well aligned. The $r < 0.6$ bucket still improves on average (+0.0054), but individual sentences in this regime can fail when the needed compression lies below the grid; §E gives such a case.

The middle block of Table 13 gives the bucket values used for the fixed-grid boundary discussion in §5.

E Fixed grid wins and boundary failures

The cases that follow turn Figure 1 into one win, one win, and one loss for the fixed candidate grid. Each case lists the source, reference length, the canvases chosen by Ratio and EV, and the resulting COMET. The Chinese outputs are rendered as raster images to avoid font issues.

Coverage. Source: *"Tap Reset Now."*; reference length L^*=6. Ratio's L=5 canvas collapses to a single time adverb (COMET 0.47); Entropy-Valley's L=6 canvas recovers the action verb (COMET **0.91**, Δ=+0.44).

Placeholder preservation. Source: *"Under #PRS_ORG#, tap Sign out."*; reference length L^*=12. Ratio's L=10 canvas drops the #PRS_ORG# placeholder (COMET 0.43); Entropy-Valley's L=12 canvas preserves it verbatim (COMET **0.87**, Δ=+0.43).

Candidate-grid boundary. Source: *"Please give me a moment."*; reference length L^*=6, true compression ratio $r \approx 0.6$. Ratio's L=6 canvas saturates (COMET **0.93**); Entropy-Valley's L=7 canvas overshoots (COMET 0.77). The needed canvas lies below the fixed grid $[0.70, 0.90]$, so the entropy score cannot reach it.

Aggregate losses. The three cases generalise to the full failure population in Table 14. In the analyzed run, EV scores below Ratio on 596 En→Zh sentences; the dominant cause is a within-grid error, where EV picks a neighbouring canvas although a better canvas is represented in the grid. Most losses are selection errors within the fixed grid rather than grid-boundary failures.

F Reveal order controls and negative diagnostics

These diagnostics test whether the evaluated reveal-order schedules or a chat-aligned masked-diffusion backbone close the Ratio–Oracle gap under the

Category	N	Ratio	EV	Δ (pp)	95% CI (pp)
En→Zh — error-type preservation rate					
Placeholder (#PRS_ORG#)	58	66.9%	89.1%	+22.1*	[+12.4, +32.5]
Number / percentage / decimal	290	77.6%	81.3%	+3.7*	[+1.4, +6.1]
Capitalized named entity (heuristic)	879	11.2%	11.9%	+0.7*	[+0.0, +1.4]
En→Zh — length-compression buckets (LLaDA token ratio)					
Bucket	**N**	**Ratio COMET**	**EV COMET**	**Δ COMET**	**95% CI**
$r < 0.6$	236	0.8449	0.8503	$+0.0054^*$	$[+0.0016, +0.0093]$
$0.6 \leq r < 0.8$	796	0.8630	0.8678	$+0.0048^*$	$[+0.0016, +0.0081]$
$r \geq 0.8$	1005	0.8056	0.8402	$+0.0346^*$	$[+0.0295, +0.0401]$
Zh→En — error-type preservation rate					
Placeholder (#PRS_ORG#)	58	96.6%	99.4%	+2.9	[-1.1, +8.6]
Number / percentage / decimal	303	80.7%	82.6%	+1.9	[0.0, +4.0]

Table 13 Coverage and error categories. Top: source-side literal retention rate, averaged over three runs. Middle: En→Zh COMET-22 grouped by $r = |L^*|/|\mathbf{x}|$, the reference target length divided by source length under the LLaDA tokenizer. Bottom: Zh→En literal retention. Δ is in percentage points for retention rows; * marks an interval whose unrounded lower endpoint is above zero.

Likely cause	N	Share	Mean ΔCOMET
Overshoot beyond grid	129	21.6%	−0.0278
Under-compression	126	21.1%	−0.0206
Within-grid wrong pick	339	56.9%	−0.0247
Very short source	2	0.3%	—

Table 14 Failure buckets for the 596 En→Zh cases where EV trails Ratio in one analyzed run (WMT22, N=2037). Buckets are assigned by mutually exclusive rules.

same fixed-canvas protocol. Neither does so in the tested settings.

F.1 Source-guided order schedules under matched length

Holding L at the reference length removes length choice from the comparison, so any remaining variation comes from the reveal order. Under this protocol at T=32 on full WMT22 En→Zh (N=2037), six source-guided variants span 0.0081 COMET, and none reliably improves over MED. Strict left-to-right and random order are additional controls, and both score below MED. The result covers the evaluated schedules rather than all reveal policies.

SIG-first might still help on subsets where source-guided unmasking has the strongest case, such as dates, enumerations, named entities, or numbers. On a 400-sentence challenge subset stratified by these categories, SIG-first averages −0.0004 COMET relative to MED (Table 16), and a sentence-level SIG-Concentration score does not identify a useful subset either (Spearman $\rho = -0.0675$ with SIG-first benefit, $p = 0.13$). On this subset MED is again at least as good as the alternatives, so the negative result for the tested

Schedule	COMET-22	Δ vs MED	Paired bootstrap p
MED (baseline)	0.8597	—	—
SIG-first	0.8603	+0.0006	0.62
Reverse-SIG	0.8523	−0.0075	$< 10^{-4}$
Hybrid MED→SIG, p=0.25	0.8582	−0.0015	0.21
Hybrid MED→SIG, p=0.50	0.8575	−0.0022	0.011
Hybrid MED→SIG, p=0.75	0.8591	−0.0007	0.28
Entropy-weighted SIG	0.8591	−0.0006	0.38
Additional reveal-order controls			
Strict L2R	0.8510	−0.0088	–
Random	0.8539	−0.0059	$< 10^{-4}$
Source-guided span (max−min)	**0.0081**	—	—

Table 15 Reveal-order controls on WMT22 En→Zh (N=2037, T=32) with reference length and the tokens-per-step schedule fixed. The six source-guided variants span 0.0081 COMET. Strict left-to-right and random order are additional controls and both score below MED. A paired bootstrap value is not reported for Strict L2R.

schedules holds at the category level as well.

Category	N	SIG-first	MED	Δ (SIG−MED)
Dates	32	0.8652	0.8665	−0.0013
Enumeration	75	0.8560	0.8533	+0.0027
Named entities	280	0.8627	0.8629	−0.0002
Numbers	123	0.8679	0.8724	−0.0045
Overall	**400**	**0.8640**	**0.8644**	**−0.0004**

Table 16 Source-guided order diagnostic on an En→Zh challenge subset (N=400, one analyzed run, T=32 MED vs. SIG-first under matched oracle length). SIG-first reveals high-source-dependence slots first; MED is the confidence-first schedule used in the main experiments.

F.2 Dream-Instruct protocol mismatch

We also tested DREAM-V0-INSTRUCT-7B. Under the LLaDA-compatible MED fixed-canvas protocol used everywhere else in this paper, it does not produce usable MT outputs, which is why the cross-backbone check in §G uses DREAM-V0-BASE-7B

instead. We tested three configurations:

1. **LoRA-SFT with plain prompts.** LoRA (rank 64) on 200k En→Zh pairs using the LLaDA-compatible plain-text template (9,660 logged steps). Loss plateaued near the random baseline $\log|V| \approx 11.93$ (Qwen2.5 vocabulary $|V|$=152,064): final loss 12.65 and last-10-step mean 12.77 ± 0.88, with the mean over steps ≥ 1000 at 13.22. Under this LoRA setup, training does not override Dream-Instruct's chat-aligned predictive distribution.

2. **LoRA-SFT with Qwen chat template.** Training loss reaches a minimum of 11.9162 at step 2,120, indistinguishable from the uniform-distribution baseline $\log|V| \approx 11.93$; the decoder still emits `<|im_end|>` at the canvas tail and most outputs are truncated to the empty string.

3. **Zero-shot MED decoding.** Without any SFT, EOS confidence at the canvas tail is uniformly high across candidate lengths, so the entropy signal is uninformative under this protocol. On WMT22 En→Zh (N=500), the empty-output rate is 97.4% (EV), 96.6% (Ratio), and 94.8% (Oracle); non-empty outputs are short repetition patterns (e.g. "...." , "*992255*") with mean length 0.17–0.97 characters versus a $\sim$ 25-character reference.

The shared failure mode is a chat-aligned EOS at the canvas tail that no length choice can repair. Dream-Base and DiffuLLaMA do not have this failure mode under the same MED protocol, which is why the cross-backbone scope check is run there.

G Scope across masked diffusion backbones

The cross-backbone check asks whether the gain reported on LLaDA is specific to that backbone. We re-run the Length Oracle, Ratio, and Entropy-Valley protocol of §4.2 on two other pretrained masked-diffusion models, DREAM-V0-BASE-7B (Ye et al., 2025), a Qwen2.5-7B initialization continued-pretrained on 580B tokens with a masked-diffusion objective, and DIFFULLAMA-7B (Gong et al., 2025), a Llama-2-7B initialization continued-pretrained with the MDLM absorbing-diffusion objective. The protocol is held fixed: 200k SFT pairs per direction, LoRA (r=64, α=128, dropout 0.05; per-backbone 7-module target set), T=32 MED decoding, and three runs per direction. We use Dream-Base rather than Dream-Instruct because the latter is not usable under this protocol (§F.2).

Interpretation across backbones. EV remains above Ratio in every Dream-Base and DiffuLLaMA cell of Table 17, with the same pattern in the sacreBLEU companion view (Table 18). The size of the gain, and the share of the Ratio–Oracle gap it closes, change with the backbone: Dream-Base closes between 54% and 69% across directions; DiffuLLaMA closes more on En→De (66.1%) than on the Chinese-side directions, consistent with its 32k Llama-2 tokenizer covering Chinese less well than the 152k Qwen2 tokenizer used by Dream-Base. These runs change tokenizer, initialization, pretraining mixture, and architecture together, so they check scope but do not attribute the variation to a single factor.

H Additional translation direction

We evaluate De→Fr to add a direction with no English on either side. Two nested WMT19 training pools use the same 500 update budget and candidate set for the direction. Ratio is 1.0, and the fixed EV candidate set is $\{0.8, 0.9, 1.0, 1.1, 1.2\}$. The trained systems and candidate set are evaluated on a WMT19 held out set and transferred unchanged to FLORES devtest.

I Bilingual human evaluation protocol

The protocol below supports the En↔Zh human study in Table 4 and the per-annotator breakdown in Figure 7 and Table 20. The same three professional translators bilingual in Chinese and English (A, B, C) rated two direction-specific packs of 100 WMT22 sentences each. The public repository includes the written guidelines, row-level CSV form templates, and system-to-slot mapping examples.

Form schema. Each direction uses one spreadsheet with one row per sentence and eleven columns: `ID`, `Source`, `Reference`, anonymised outputs in `System A` and `System B`, four 1–5 rating cells for adequacy and fluency, `Preference` (A / B / Tie), and free-text `Notes`. An independent Bernoulli(0.5) flip maps each system to slot A or B for every row; annotators never see the mapping. Each direction contains 25 EV-best cases by COMET, 25 Ratio-best cases, 25 near ties with

Backbone	Direction	Oracle COMET	Ratio COMET	EV COMET	Δ EV−Ratio	COMET Gap closure
LLaDA-8B-Base (paper main results, §4.2; reproduced for cross-backbone comparison)						
LLaDA-8B-Base	En→Zh	0.8610±0.0013	0.8345±0.0017	**0.8517±0.0006**	+0.0172	**64.9±7.4%**
LLaDA-8B-Base	Zh→En	0.8519±0.0007	0.8266±0.0010	**0.8431±0.0004**	+0.0165	**65.3±0.8%**
LLaDA-8B-Base	En→De	0.7382±0.0090	0.7170±0.0090	**0.7240±0.0078**	+0.0070	**33.0±8.4%**
Dream-v0-Base-7B (Qwen2.5-7B continued-pretrained; Qwen2 tokenizer, 152k vocab)						
Dream-Base	En→Zh	0.8414±0.0020	0.8002±0.0016	**0.8239±0.0004**	+0.0238	**57.7±3.1%**
Dream-Base	Zh→En	0.8490±0.0008	0.8303±0.0009	**0.8431±0.0009**	+0.0128	**68.7±6.2%**
Dream-Base	En→De	0.7271±0.0017	0.6904±0.0028	**0.7106±0.0026**	+0.0201	**54.4±8.8%**
DiffuLLaMA-7B (Llama-2-7B continued-pretrained; Llama-2 tokenizer, 32k vocab)						
DiffuLLaMA	En→Zh	0.7390±0.0041	0.6010±0.0021	**0.6172±0.0035**	+0.0162	11.8±3.5%
DiffuLLaMA	Zh→En	0.8261±0.0016	0.6612±0.0040	**0.6980±0.0026**	+0.0368	22.3±2.6%
DiffuLLaMA	En→De	0.6909±0.0369	0.6439±0.0348	**0.6749±0.0306**	+0.0310	**66.1±13.0%**

Table 17 Cross-backbone COMET-22 results for Entropy-Valley (WMT22, N=2037, three runs per aggregate cell, T=32 MED). Values are mean ± std across runs. COMET-22 gap closure is (EV − Ratio)/(Oracle − Ratio) per run, then mean ± std. EV is above Ratio in all six Dream-Base and DiffuLLaMA aggregate COMET cells; paired tests for every run give $p < 10^{-3}$ in all 18 Dream-Base and DiffuLLaMA comparisons, with positive 95% bootstrap confidence intervals. Absolute scores and closures vary by backbone.

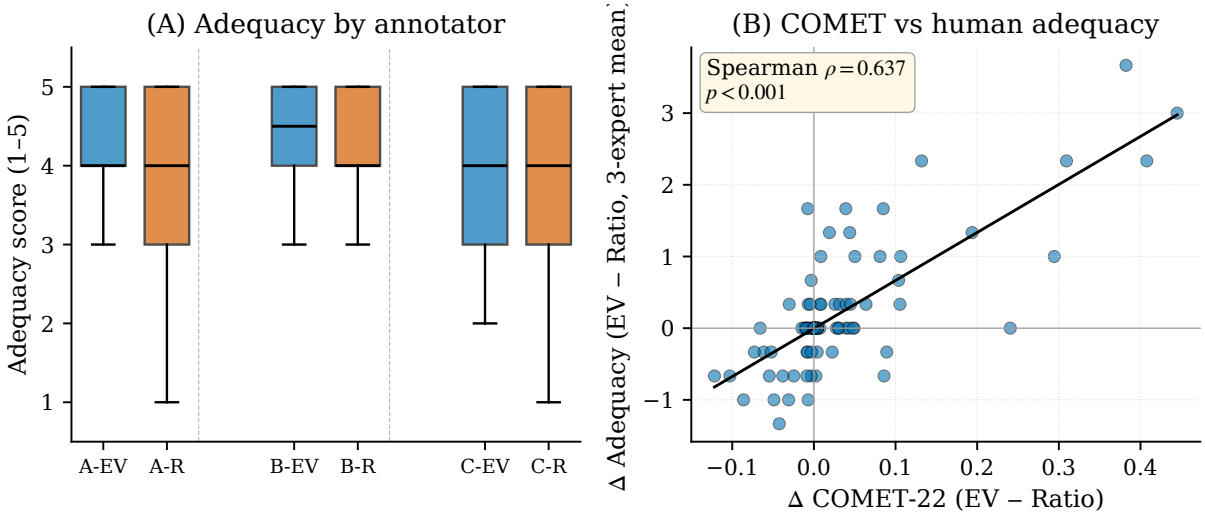

Figure 7 Human evaluation on En→Zh. (A) Per-annotator adequacy boxplots. (B) Sentence-level ΔCOMET-22 vs 3-expert mean ΔAdequacy (EV − Ratio); Spearman $\rho = 0.637$ ($p < 0.001$). The corresponding Zh→En correlation is reported in Table 20.

Backbone	Direction	Oracle	Ratio	EV
LLaDA-8B-Base	En→Zh	40.81±0.23	36.72±0.22	**38.57±0.13**
LLaDA-8B-Base	Zh→En	27.93±0.23	23.65±0.21	**25.28±0.33**
LLaDA-8B-Base	En→De	22.55±0.77	20.73±0.68	**21.55±0.85**
Dream-Base	En→Zh	36.30±0.22	29.37±0.20	**32.83±0.28**
Dream-Base	Zh→En	27.24±0.28	23.56±0.04	**25.78±0.20**
Dream-Base	En→De	21.64±0.16	17.31±0.25	**20.06±0.17**
DiffuLLaMA	En→Zh	24.91±0.16	7.49±0.12	**7.92±0.10**
DiffuLLaMA	Zh→En	22.63±0.30	9.36±0.32	**10.80±0.14**
DiffuLLaMA	En→De	19.72±2.64	14.95±1.89	**17.94±1.99**

Table 18 Companion sacreBLEU view of Table 17 (WMT22, N=2037, three runs per cell). Values are mean ± std across runs.

Evaluation set	Pool	Ratio	EV	ΔCOMET (95% CI)
WMT19 held out	50k	0.7428	0.7514	+0.0086 [0.0034, 0.0139]
WMT19 held out	200k	0.7386	0.7485	+0.0100 [0.0050, 0.0152]
FLORES devtest	50k	0.7398	0.7481	+0.0082 [0.0020, 0.0143]
FLORES devtest	200k	0.7315	0.7403	+0.0088 [0.0024, 0.0152]

Table 19 De→Fr results for two nested training pools. WMT19 held out contains N=1512 sentences, and FLORES uses the full N=1012 De→Fr devtest. Both pools use a fixed 500 update budget and one candidate set for the direction. EV exceeds Ratio in all four settings, with positive paired confidence intervals; paired bootstrap and Wilcoxon tests are below 0.05 throughout. This is a robustness check under a fixed update budget, not a comparison of scaling with data size.

the smallest nonzero |ΔCOMET|, and 25 random cases.

Rating dimensions. *Adequacy* measures how completely and accurately the translation conveys the source meaning, on a 1–5 scale: 5 = all source information conveyed accurately with no omissions, additions, or errors; 4 = nearly all informa-

	A	B	C	Agg.
Panel A: En→Zh (N=100)				
Δ Adequacy (EV − Ratio), mean	$+0.19$	$+0.17$	$+0.19$	$+0.18$
Δ Fluency (EV − Ratio), mean	$+0.18$	$+0.20$	$+0.16$	$+0.18$
Preference EV / Ratio / Tie	28/19/53	28/16/56	30/27/43	28/19/53
Wilcoxon p (Adequacy)	0.038^{*}	0.070	0.065	—
Wilcoxon p (Fluency)	0.056	0.035^{*}	0.135	—
Panel B: Zh→En (N=100)				
Δ Adequacy (EV − Ratio), mean	$\mathbf{+0.54}$	$\mathbf{+0.52}$	$\mathbf{+0.43}$	$\mathbf{+0.50}$
Δ Fluency (EV − Ratio), mean	$+0.12$	$+0.09$	$+0.12$	$+0.11$
Preference EV / Ratio / Tie	43/27/30	38/23/39	52/32/16	45/25/30
Wilcoxon p (Adequacy)	0.0001^{***}	0.0007^{***}	0.003^{**}	0.001^{**}
Wilcoxon p (Fluency)	0.52	0.79	0.44	0.95
Agreement and metric correlation:				
Fleiss' κ (pref., En→Zh / Zh→En)	0.574 / 0.522			
Kendall's W (score ranks, range)	En→Zh: 0.68–0.74; Zh→En: 0.81–0.90			
Majority-pref. sign-test p	En→Zh: 0.24; Zh→En: 0.022^{*}			
$\rho(\Delta\mathrm{COMET}, \Delta\mathrm{Adeq})$ En→Zh / Zh→En	0.637 / 0.689			
$\rho(\Delta\mathrm{COMET}, \Delta\mathrm{Flu})$ En→Zh / Zh→En	0.461 / 0.455			

Table 20 Detailed human-evaluation statistics. The first three columns report annotator-level results. The aggregate preference uses per-sentence majority vote; aggregate Zh→En adequacy uses the paired Wilcoxon test on the three-expert mean.

tion, with only minor omissions or imprecision; 3 = main idea preserved but visible information loss; 2 = only some information correct, with major mistranslations or omissions; 1 = unrelated or unintelligible. *Fluency* measures how natural and grammatical the target text is, on a 1–5 scale: 5 = reads as native; 4 = mostly natural with occasional unnaturalness; 3 = understandable but with multiple unnatural expressions; 2 = frequent grammatical errors, hard to parse; 1 = severe grammar errors or incoherent text. Adequacy and fluency are scored independently.

Preference rule. `Preference` is a single A / B / Tie choice that combines both dimensions. If one system dominates on both adequacy and fluency, pick that system; if the two trade off, pick the more competent overall translation; if the difference is tiny (e.g. $4/4$ vs $4/5$), pick `Tie`.

Blinding and independence. Per-row slot randomisation prevents annotators from inferring system identity from row order. The three annotators worked independently on the same 100 sentences per direction with no cross-annotator discussion during rating. No annotator was shown the other direction's ratings before completing their own. The study used public WMT22 sentences and did not collect sensitive personal data. Annotators consented to the use of their anonymised ratings and aggregate statistics for research reporting.

Special-case handling. Empty outputs (no translation produced) are scored $1/1$ on both adequacy and fluency. Outputs with repeated fragments (the same phrase repeated) receive a fluency penalty. Unusual observations can be recorded in the free-text `Notes` column and do not affect the numeric ratings.

Compensation and pacing. Annotators were compensated at local professional-translator hourly rates. Annotation averaged 1–2 minutes per sentence; each 100-sentence batch was completed in 2–3 sittings, with the directions rated on separate days.