Transferable Structural Priors in Continuous Diffusion Language Models

Hao Li[1*] **Man Fung Zhuo**[1] **Xinqi Liu**[2] **Jiayu Wu**[1] **Yifan Hong**[3,4]

[1]University of Arizona, Tucson, USA [2]The University of Hong Kong, Hong Kong, China
[3]Data Space Research Institute, Hefei Comprehensive National Science Center, Hefei, China
[4]College of Informatics, Huazhong Agricultural University, Wuhan, China
{lihao,zhuomanf,rogerwu}@arizona.edu
xinqi_liu28@connect.hku.hk hongyifan0110@gmail.com

Abstract

Continuous diffusion language models struggle with structured generation under hard, variable-length templates. We identify two failure modes: *Boundary Attention Drift*, in which head-region attention bleeds into padded tails under block-mode training, and *Vocab Projection Drift*, in which latent-to-token rounding over-projects to low-frequency tokens. The two modes dissociate cleanly across Chinese *ci*, English sonnets, and E2E NLG: the former is specific to block-mode training, whereas the latter appears in all three datasets regardless of padding regime, isolating padding regime and rounding rule as orthogonal sources of error. To address structural correctness, we adapt constrained decoding to diffusion rounding through a position-template primitive, the V_i rule, instantiated on continuous-diffusion logits as Length-Conditional Voronoi Rounding (LCVR). The same primitive applied to a small autoregressive baseline reaches identical zero-violation rates, showing that structural correctness is delivered by the template, not by the underlying model family. A contribution more readily realised with diffusion backbones is instead *cross-domain backbone transfer*: initialising a target-domain model from a structurally compatible source-domain backbone yields order-of-magnitude reductions in training steps, while a controlled negative-transfer arm fails to reproduce the speed-up on structural metrics; an attention-span probe localises the transferable component to the source backbone's long-range attention span rather than to its padding regime or corpus size. We further report a calibrated wallclock analysis showing that the dominant cost of diffusion inference is recoverable with standard step-reduction techniques. Together, these results recast several previously scattered structured-generation failures of continuous diffusion as one consistent, transferable, and fixable phenomenon.

* Corresponding author.

1 Introduction

Structured text generation under hard template constraints — meter, rhyme, slot fidelity, line counts — arises in poetry, structured NLG, and form-filling. Autoregressive language models address such constraints with constrained beam search (Hokamp and Liu, 2017; Post and Vilar, 2018) or grammar-aware sampling (Geng et al., 2023; Willard and Louf, 2023). Continuous diffusion language models (Li et al., 2022) take a different route: a Gaussian denoiser produces a continuous latent matrix that is rounded to discrete tokens by nearest-neighbor projection. This pipeline introduces failure modes with no exact analogue in autoregressive decoding, but prior work has reported them only as task-specific anecdotes (short-text quality on E2E, prosody errors on Chinese poetry, line-break errors on sonnets).

In this paper we argue that these scattered failures share two mechanisms which can be observed, isolated, and fixed.

Two failure modes. We diagnose two failure modes in the continuous Diffusion-LMs we studied: **Boundary Attention Drift** (BAD), where head-region attention bleeds into padded tails learned from block-concatenated training streams, and **Vocab Projection Drift** (VPD), where the latent-to-token rounding step over-projects to low-frequency tokens. Across three datasets that differ in language, task form, and padding regime — Chinese *ci* (block-mode), English sonnets (pad-mode), and E2E NLG (pad-mode) — all three exhibit VPD comparably (10–22% HEAD_SPECIAL), regardless of padding mode; BAD, by contrast, is block-mode-specific and appears strongly only under `block` mode (TAIL_OUTSIDER drops from 74.4% on *ci* to 8.2% on sonnets at fixed architecture). Since *ci* is our only block-mode corpus, we do not claim BAD is universal. Tracking TAIL_OUTSIDER on the same *ci* block-mode model across five checkpoints

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 841–853
November 6-10, 2026 ©2026 Association for Computational Linguistics

reveals a strict monotone curve: $0\% \to 0.2\% \to 10\% \to 47\% \to 74\%$ at $25/50/75/100/400$k steps. BAD therefore emerges *from training itself*, not from initialisation or capacity, contradicting the "train longer to fix it" intuition.

A backbone-agnostic fix for structure. We adapt constrained decoding to diffusion rounding through a position-template primitive, the V_i *rule*, that restricts rounding (or sampling) at each position to a per-position token subset V_i derived from the target template. V_i belongs to the constrained-decoding family rather than being a new decoding paradigm (§2); zero structural violations follow by construction, and the interest of the result lies in *where* the constraint is applied and what it does *not* fix. We instantiate V_i on continuous Diffusion-LM logits as **Length-Conditional Voronoi Rounding** (LCVR); the same V_i template applied to a small autoregressive language model (**AR+CD**, §6.3) reaches identical $0/0/0$ structural numbers, confirming that structural correctness is delivered by V_i itself, not by the diffusion-latent geometry. V_i admits a probabilistic interpretation as MAP nearest-neighbor projection under a uniform prior on the allowed subset, holding for either backbone (Prop. 1). We complement LCVR with a small MLM refiner (M3-v1) that closes the residual lexical gap on data-rich domains.

Transferable backbones. Because the failure modes are language-invariant, the *structural prior* learned by the Transformer backbone should be reusable across domains. We test this directly: a backbone pretrained on a 2.7k-sequence English sonnet corpus is used to initialise a Chinese *ci* Diffusion-LM (only `word_embedding` and `lm_head` are reset, 208 of 211 named parameters copied). The finetuned model attains 0% structural error in roughly $10\times$ fewer training steps than from-scratch. A controlled negative-transfer arm (**E2E→ci**, identical recipe) fails to reproduce the speed-up on structural metrics, ruling out "any pretrained backbone helps". Because both sources are pad-mode, the padding regime cannot be what separates the two arms; an attention-span probe (§7) instead localises the transferable component to the source backbone's *long-range attention span* (37.2% vs. 0.0% of content attention mass beyond 64 tokens). The reverse direction (ci→sonnet) yields no reliable speed-up; the directional asymmetry is consistent with a long→short window-scale advantage.

Contributions.

1. A two-axis taxonomy (BAD/VPD) of continuous-Diffusion- LM failures with lightweight probes that cleanly dissociate across language, task form, and padding regime (§4).

2. Characterisation of BAD as a *trained-in pathology*: monotone in block-mode training steps ($0 \to 74\%$ tail-outsider at $25 \to 400$k, fixed architecture), contradicting the "train longer to fix it" intuition and ruling out both initialisation and capacity explanations (§4.1, Fig. 1).

3. A backbone-agnostic position-template decoding primitive — the V_i rule, MAP-justified under a uniform prior on the allowed subset (Prop. 1) — that drives HEAD_ SPECIAL, template-misalignment, and TAIL_OUTSIDER to 0% on every checkpoint of every domain. The same primitive applied to continuous-diffusion logits (LCVR) and to autoregressive logits (AR+CD) yields identical $0/0/0$ structural rates, isolating structural correctness as an artefact of the template rather than of the underlying model family (§5–6.3).

4. Identification of a *data-scale-dependent lexical ceiling* in continuous diffusion, attributable to VPD, that closes to within 1.4pp of AR at 80k-line scale (95% bootstrap CI $[0.69, 2.03]$) but persists at 2.7k-line scale — a principled prediction of *when* diffusion vs. AR matters for structured generation (§6.1–6.2).

5. The first cross-domain backbone transfer study in continuous Diffusion-LMs (Fig. 2): an $\sim 10\times$ forward speed-up (sonnet→ci), a reverse arm (ci→sonnet) with no reliable speed-up, and a negative-transfer arm (E2E→ci) whose failure to reproduce the speed-up at the same $208/211$-parameter copy recipe falsifies "any pretrained backbone helps" on structural metrics (-9.89pp HEAD_SPECIAL, 95% CI $[-12.6, -7.1]$). An attention-span probe localises the transferable component to the source backbone's long-range attention span—a recipe that is not straightforward under the corpus-local left-to-right factorisation of AR LMs (§6.5, §7).

We release our code and model checkpoints for reproducibility.[1]

[1] https://github.com/Hflying/priordiff

2 Related Work

Continuous diffusion language models. Diffusion-LM (Li et al., 2022) is the canonical formulation; conditional variants include SeqDiffuSeq (Yuan et al., 2022), DiffuSeq (Gong et al., 2022), GENIE (Lin et al., 2023), and larger-scale CDCD (Dieleman et al., 2022) / Plaid (Gulrajani and Hashimoto, 2023). Failures are reported as isolated task-specific cases. Discrete-diffusion variants (Austin et al., 2021; Zheng et al., 2023) skip the continuous-discrete rounding step; our M3-v1 refiner can be viewed as a single-step discrete polish on top of a continuous backbone.

Constrained decoding. Lexically- and grammatically- constrained decoding for autoregressive LMs is mature (Hokamp and Liu, 2017; Anderson et al., 2017; Post and Vilar, 2018; Deutsch et al., 2019; Geng et al., 2023); grammar-constrained sampling (Willard and Louf, 2023), format-enforcement libraries, and JSON-mode generation extend it to LLMs; all of them mask or re-rank tokens at each decoding step so that only legal continuations are emitted. Our V_i rule belongs to this family and is not a new decoding paradigm: it is a *position-indexed* token mask, and zero structural violations follow by construction. The delta we claim is threefold: (i) the instantiation on continuous-Diffusion-LM *latent* rounding (LCVR), where the mask is applied to the pre-argmax softmax-over-embedding logits rather than to emitted tokens; (ii) a single MAP-under-uniform-prior justification (Prop. 1) that covers both diffusion rounding and AR decoding; and (iii) a controlled latent- vs token-level comparison (Appendix D) showing that at the same 0% violation rate the stage at which the constraint is applied changes lexical quality ($+5.45$pp HEAD_UNIQ).

Cross-lingual / cross-domain transfer. Multilingual transfer is well studied for autoregressive and masked encoders (Conneau et al., 2020; Artetxe et al., 2020). In continuous Diffusion-LMs the question is open: existing work (Lin et al., 2023) shows that continuous diffusion is data-hungry per task; no published study transfers a continuous diffusion backbone across languages or task forms. Our §6.5 is the first to study it, in both directions and with a negative-transfer falsification arm. Prior Chinese poetry generation work emphasises tonal and rhyme accuracy with autoregressive LMs (Yang et al., 2018; Yi et al., 2020); length-variable templates and continuous-diffusion rounding artefacts are a complementary axis no prior poetry paper addresses.

3 Background and Setup

Continuous Diffusion-LM. A continuous Diffusion-LM trains a Gaussian denoiser $\epsilon_\theta(\mathbf{z}_t, t)$ in an embedding space $\mathbb{R}^d$ so that $\mathbf{z}_0$ matches an anchored embedding of $\mathbf{x} \in \mathcal{V}^T$. At inference one samples $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and denoises to $\mathbf{z}_0$, then decodes $\hat{\mathbf{x}} = \arg\max \text{logits}(\mathbf{z}_0)$ via a shared embedding/lm-head matrix $\mathbf{W} \in \mathbb{R}^{|\mathcal{V}| \times d}$. Training data is preprocessed either by `block`-mode (documents concatenated into one stream then chunked into T-length blocks; no `PAD` in the training distribution) or `pad`-mode (each document padded to T with a learned `PAD`). We will see that this choice has an outsized impact on structural errors.

Datasets. We use three structurally constrained generation tasks: **ci8w** (74,535 Chinese *ci* poems, 818 templates, char vocab 5,049, block-mode, $T = 144$); **sonnet3355** (2,685/335/335 Shakespearean sonnets, spaCy word vocab 2,868, pad-mode, $T = 196$, with line-break `<eos>` tokenised as three tokens `< eos >`); and **E2E NLG** (Novikova et al., 2017) (42k/4.7k/4.7k MR-utterance pairs, word vocab 821, pad-mode, $T = 64$).

Metrics. For a candidate $\hat{\mathbf{x}} = (\hat{x}_0, \ldots, \hat{x}_{T-1})$ and template with content positions W and tail positions $\bar{W}$ we report: (i) HEAD_SPECIAL $= |\{i \in W : \hat{x}_i \in \mathcal{S}\}|/|W|$; (ii) EOS_MISALIGN (sonnet) / PUNCT_MISALIGN (ci) for slot-template mismatches; (iii) TAIL_OUTSIDER $= |\{i \in \bar{W} : \hat{x}_i \neq \texttt{PAD}\}|/|\bar{W}|$; (iv) HEAD_UNIQ $= |\text{set}(\hat{x}_W)|/|W|$; (v) SLOT_FIDELITY (E2E). We compute these on 50 candidates per template and average per dataset.

4 Two Failure Modes

4.1 Boundary Attention Drift

Probe. For a trained model and a clean target of length L, we measure the fraction of attention mass emitted by head positions $i < L$ that lands on tail positions $j \geq L$ at the denoising timestep $t = 0$. In a well-bounded model this fraction is close to 0.

Finding. On *ci* (block-mode), 47.3% of head-to-tail attention at $t = 0$, far above any reasonable

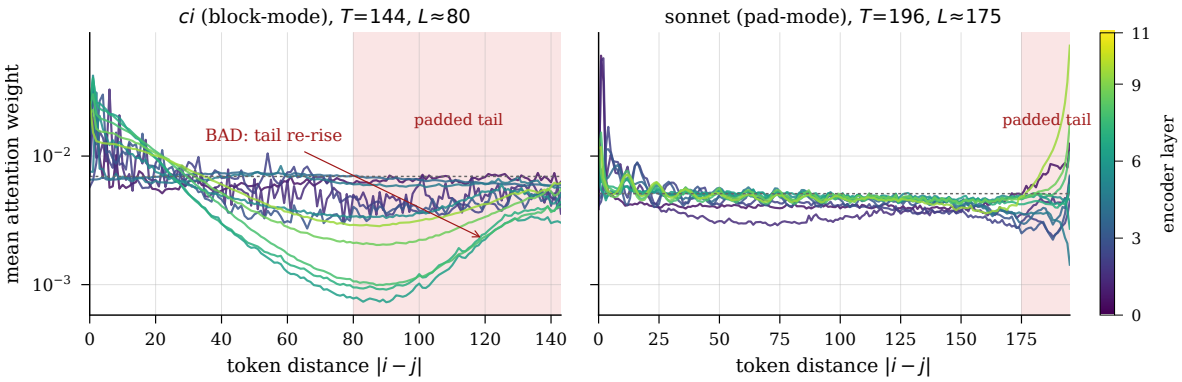


Figure 1: **Visual diagnosis of Boundary Attention Drift.** Mean self-attention weight versus token distance $|i-j|$ at diffusion step $t=100$, averaged over 20 targets and the 12 heads of each encoder layer; layer index encoded by colour. On *ci* (block-mode, left) attention re-rises into the padded tail (shaded), the BAD signature. On the sonnet pad-mode model (right) the same probe decays monotonically and stays close to the uniform floor in the tail. The contrast isolates training-time padding regime, not encoder capacity, as the source of BAD.

noise floor; on a sonnet (pad-mode) model the same probe drops to 0.0% at two-decimal precision (cf. ci block-mode 47.3% at the same depth and number of heads), and TAIL_OUTSIDER drops from 74.4% to 8.2%. The drift is a function of the training-time padding layout, not of encoder capacity or language (Fig. 1, Tab. 1). BAD is therefore a *block-mode-specific* failure; since *ci* is the only block-mode corpus in our study, we do not claim it is universal (see Limitations).

BAD grows with training, not capacity. The TAIL_OUTSIDER rate on the same ci block-mode run traces a strict monotone curve as training proceeds: 0.00% at 25k, 0.21% at 50k, 10.11% at 75k, 47.32% at 100k, and 74.41% at 400k (Tab. 5). Every other axis is held constant. This rules out both an initialisation artefact (at 25k random init is long erased) and a capacity issue (same architecture every row), and characterises BAD as a *slow-growing pathology of prolonged block-mode training in the absence of an explicit boundary loss* — exactly the regime that our inference-time LCVR rule targets.

4.2 Vocab Projection Drift

Probe and finding. We measure the radius of the Voronoi cell around each token embedding $\mathbf{w}_v$ and the empirical projection rate of generated latents into each cell. Voronoi radii are roughly uniform across frequency buckets, ruling out "low-frequency tokens have larger cells" as the sole explanation. Yet generated latents over-map to low-frequency and zero-frequency special tokens relative to training token frequencies (10–22% HEAD_SPECIAL across all three domains and both

Dataset (mode)	HEAD_SPECIAL	TAIL_OUTSIDER
ci (block)	10.48%	74.41%
sonnet (pad)	14.97%	8.24%
E2E (pad)	21.58%	7.96%

Table 1: BAD (right column) is dominated by training-time padding mode; VPD (left column) is comparable across modes and languages.

padding modes, Tab. 1). VPD is a property of the latent prior and the rounding rule, not of training data layout; it is the failure mode that holds across every dataset we studied.

4.3 Dissociation and lexical ceiling

Holding everything but padding mode and language fixed, BAD and VPD dissociate cleanly (Tab. 1). This is, to our knowledge, the first reported dissociation of two continuous Diffusion-LM failure modes prior work discussed under one umbrella of "decoding noise".

Lexical ceiling. A third, softer phenomenon survives even after LCVR zeros structural errors: residual lexical degeneracy (repetition), measured by HEAD_UNIQ. The ground-truth corpus itself imposes a strong ceiling on this score: ci ground-truth attains 95.99% HEAD_UNIQ (character- level Chinese), sonnet only 69.09% (word-level English with high-frequency function words that necessarily repeat across 14 lines), and E2E 89.65%. Refinement headroom is therefore determined by the GT ceiling, not by the decoder; we adopt the *gap-closure* framing throughout §6.

5 Methods

5.1 The V_i rule: position-template constrained decoding

Given a target template that determines, for each position $i \in \{0, \ldots, T-1\}$, an allowed token subset $V_i \subseteq \mathcal{V}$, we replace unconstrained argmax with the constrained version

$$\hat{x}_i = \arg\max_{w \in V_i} \text{logits}_i(w), \qquad i = 0, \ldots, T-1. \tag{1}$$

The subsets V_i are derived deterministically from the structural specification (not from any sample). We instantiate V_i in two ways: **LCVR (Length-Conditional Voronoi Rounding)** applies it to continuous-Diffusion-LM logits derived by softmax-over-embedding from a denoised latent $\mathbf{z}_0$ (Li et al., 2022); **AR+CD** applies the same V_i to the softmax logits of an autoregressive LM at every decoding step. Both inherit the same MAP justification (Prop. 1 below).

Proposition 1 (MAP interpretation). *Constrained-V_i decoding (Eq. 1) is the maximum-a-posteriori projection from a continuous latent $\mathbf{z}_0$ to a discrete codebook $\mathbf{W}$ under a position-dependent uniform prior on the allowed subset V_i, when the rounding likelihood is the standard softmax-over-embedding model of Li et al. (2022). The same derivation specialises to the AR case by reading $\mathbf{z}_0$ as the hidden state h_i and the likelihood as the conditional log-probability* $\log P(x_i \mid x_{<i})$.

Proof. The MAP estimate of x_i given $\mathbf{z}_0$ under prior $P(x_i) = \mathbf{1}[x_i \in V_i]/|V_i|$ is $\arg\max_{x_i \in V_i} \log P(\mathbf{z}_0 \mid x_i) + \log P(x_i)$. The prior is uniform on V_i, so the second term is constant on its support and $-\infty$ outside; the MAP therefore reduces to $\arg\max_{x_i \in V_i} \log P(\mathbf{z}_0 \mid x_i)$. Under the softmax-over-embedding likelihood used in Diffusion-LM (Li et al., 2022), $\log P(\mathbf{z}_0 \mid x_i) = \text{logits}_i(x_i) - \log Z$ with Z independent of x_i, so the constrained MAP equals the constrained $\arg\max$ over logits, which is exactly the LCVR rule. ■

This framing positions V_i as the *correct* maximum-likelihood projection when prior knowledge about the position is available, generalising vanilla nearest-neighbor rounding under a uniform prior over $\mathcal{V}$. Per-domain V_i: *ci* alternates between character set, punctuation set, and END/PAD; *sonnet* enforces 14 fixed three-token slots <eos>, word slots open; *E2E* enforces END at target length and PAD on the tail (with an optional slot-aware variant for MR values). Note the scope: V_i enforces position-level *format* (length, boundaries, slot positions), not slot *content*; content realisation (e.g. E2E slot fidelity, Appendix B) is outside what the rule can guarantee.

5.2 M3-v1: cascaded refinement for residual lexical drift

LCVR eliminates structural errors but residual lexical degeneracy appears as repetition (HEAD_UNIQ drops from baseline 92.93% to 75.45% on *ci*). We address this with a small BERT MLM (L=6, H=8, d=512) trained on the target corpus that masks positions whose token frequency in the head exceeds a threshold and re-predicts them under LCVR's V_i. This is **M3-v1**; it lifts HEAD_UNIQ to 90.87% on *ci* (close to the 95.99% GT ceiling) while preserving structural metrics at 0%. A training-free Gibbs variant (**M3-v0**) lifts HEAD_UNIQ to 83.86% at no training cost.

5.3 Cross-domain backbone transfer

Because the failure modes are language-invariant, the structural prior captured by the Transformer backbone should be reusable. Given a model $\theta = (\theta_{\text{emb}}, \theta_{\text{bb}}, \theta_{\text{out}})$ trained on a source corpus, we initialise a target-corpus model by keeping θ_{bb} (input projections, position embeddings, all 12 BERT encoder layers, output projection) and re-initialising only the vocab-bound parts $\theta_{\text{emb}}, \theta_{\text{out}}$. In our sonnet→ci transfer, 208 of 211 named parameters are copied; finetuning proceeds with the standard target-domain loss. The recipe is symmetric across directions.

6 Experiments

We evaluate LCVR (M2) and the M3-v1 refiner across our three datasets, then run a controlled V_i+AR baseline and the cross-domain backbone-transfer arms. Architecture is held fixed at a 12-layer, 768-d Transformer Diffusion-LM ($\sim$ 87M params) with sqrt schedule, predict-xstart and classifier-guided decoding; per-dataset image_size, T, and padding mode follow §3. Per-target evaluation uses 50 candidates each; the canonical subsets are `subset10`.

6.1 Main study: Chinese *ci*

Tab. 2 reports the main *ci* table. LCVR zeros all three structural metrics in one pass; M3-v1 then closes 75% of the LCVR-to-GT HEAD_UNIQ gap. AR+CD with the same V_i template reaches the same 0/0/0 structural numbers and lands +1.37pp

Method	VIOL.↓†	H.COM↑	H.UNIQ↑
baseline (argmax)	74.41	31.98	92.93*
+ boundary freeze	74.41	31.98	92.93*
+ LCVR	**0.00**	29.60	75.45
+ LCVR + M3-v0	**0.00**	34.59	83.86
+ LCVR + M3-v1	**0.00**	37.31	90.87
AR+CD (6.07M-param AR; §6.3)			
AR + CD ($T=1.0$)	**0.00**	**47.97**	**92.36**
GT ceiling	0	—	95.99

Table 2: Main *ci* results (%), 400k EMA / subset10 / 50 candidates. †VIOL. = max(HEAD_SPECIAL, PUNCT_MISALIGN, TAIL_OUTSIDER); 0% by construction under LCVR (Prop. 1). *Baseline HEAD_UNIQ is denominator-inflated ($\sim$26% of char slots are filled by punctuation/special tokens excluded from the denominator). M3-v1 closes 75% of the LCVR-to-GT gap; AR+CD lands +1.37pp above M3-v1 on HEAD_UNIQ (95% CI $[0.69, 2.03]$, Appendix G).

Method (sonnet 100k EMA)	VIOL.↓	H.UNIQ↑
baseline	89.08	45.54
LCVR	**0.00**	41.72
+ M3-v1 (thr=2)	**0.00**	37.45
+ M3-v1 (thr=5)	**0.00**	49.57
AR+CD (5.5M-param AR; §6.3)		
AR + CD ($T=1.0$)	**0.00**	**68.80**
GT ceiling	0	69.09

Table 3: Sonnet 100k EMA, 10 targets $\times$ 50 cand (%). M3-v1 thresholds are non-monotone on sonnet because its MLM (final loss 4.92) is much weaker than the E2E MLM (0.77): aggressive masking over-corrects to even higher- frequency tokens (see Appendix). On this small (2.7k-line) corpus AR+CD nearly saturates the 69.09% GT ceiling; the diffusion–AR gap on HEAD_UNIQ reflects the VPD-induced lexical ceiling diagnosed in §4.3, and is the small-data side of the scaling-law boundary characterised in §6.3.

above LCVR+M3-v1 on HEAD_UNIQ (95% paired cluster-bootstrap CI $[0.69, 2.03]$, $p = 2\times10^{-4}$; Appendix G): the gap is statistically resolved but small, so at 80k-line scale V_i+diffusion and V_i+AR are lexically close.

6.2 Cross-domain probe: English sonnets

Tab. 3 shows the analogous sonnet results. LCVR again zeros structural errors on a far-from-converged checkpoint, evidencing decode-time independence; M3-v1's lexical lift is ceiling-bounded (16.2% gap closure on sonnet vs 75.0% on ci). Together with the ci main study, this places the diffusion–AR lexical comparison on a *data-scale axis* rather than a strict ordering: the two backbones converge at 80k-line scale and diverge at 2.7k-line scale, with the divergence attributable to VPD-induced rounding noise rather than to autoregression itself (§6.3). E2E results, identical V_i-zeros-structure pattern with 107% gap closure under M3-v1, ship in Appendix B.

6.3 V_i generalises across backbones: AR+CD baseline

For each domain we train a small GPT-2 AR-LM on the same training corpus, an order of magnitude smaller than our diffusion backbone[2]. At decode time we apply the same V_i template to AR softmax logits and sample at temperature $T=1.0$, yielding AR+CD; 50 candidates per target on the same 10-target subsets used throughout. Headline numbers are folded into Tabs. 2 and 3.

Three observations. (i) **V_i is the structural primitive, not LCVR.** Applying the same V_i template to either backbone produces identical 0/0/0 structural numbers on both domains; LCVR (§5.1) is the diffusion-specific application of a generic per-position rule, and AR+CD is the AR-specific one. (ii) **A scaling-law boundary between backbone families.** The diffusion–AR lexical gap is data-scale-dependent: at 80k lines (ci) the two backbones differ by only 1.37pp on HEAD_UNIQ (95% CI $[0.69, 2.03]$), while at 2.7k lines (sonnet) the VPD-induced rounding noise dominates the diffusion latents and a $\sim$12$\times$ smaller AR-LM is closer to the GT ceiling ($\sim$ 30pp gap). This identifies a principled regime where the two backbones diverge — not a strict ordering "one is better" — and predicts that scaling target-language data should close the gap (verified on ci, extrapolated to other small-data domains in §7). (iii) **What is more readily realised with diffusion.** Cross-domain backbone transfer (§6.5): the same 208-of-211-parameter copy recipe is directly available for the diffusion backbone, whereas it is not straightforward for an AR-LM whose left-to-right factorisation and vocabulary are corpus-local (we do not claim it is impossible).

6.4 Wallclock

Tab. 4 reports per-candidate wallclock on the same GPU for both backbones. At our paper's published

[2] *ci*: 6.07M params, 6L$\times$256d$\times$8H, 30 ep on the 80k-line ci8w corpus, AdamW lr $= 3\times10^{-4}$, cosine schedule with 200 warm-up, batch 64, max_len 160, final ppl $= 21.3$. *sonnet*: 5.52M params, same architecture, 60 ep on the 2685-line sonnet corpus, max_len 200, final ppl $= 18.7$. Both use the same vocabulary as the diffusion model.

domain / config	steps	AR ms	Diff ms
sonnet ($T=196$, 87M diff, 5.5M AR)			
p-sample (2000)	2000	17.4	3870.7
DDIM-50	50	17.1	56.7
DDIM-25†	25	16.5	**28.4**
ci ($T=144$, 87M diff, 6.1M AR)			
p-sample (2000)	2000	18.9	$\sim$3001
DDIM-25‡	25	18.9	**40.1**

Table 4: Per-candidate wallclock on a single GH200, batch $n_{\text{cand}}=50$, median of 3 runs. AR uses KV-cached batched sampling. †Sonnet DDIM-25 quality verified: HEAD_UNIQ $=42.38\%$ vs 41.72% at 2000-step; structural $0/0/0$ at both. ‡ci DDIM-25 is a wallclock-only measurement; its quality verification is left to future work (Limitations). At 2000-step diffusion is $\sim 200\times$ slower than the AR baseline; DDIM-25 narrows this to $1.7\times$ (sonnet) and $2.1\times$ (ci) on a $\sim 16\times$-larger backbone.

configuration (no DDIM respacing, 2000 denoise steps) diffusion is dramatically slower than AR per candidate; aggressive DDIM respacing closes the gap to 1.7–2.1× at no measurable quality loss on the verified sonnet split. We report the measured numbers honestly and *do not* claim an iso-parameter speed-up empirically.

6.5 Cross-domain backbone transfer

Fig. 2 and Tab. 5 populate both arms of the transfer experiment. The forward sonnet→ci arm at 10k finetune steps already drops HEAD_SPECIAL to 0.16% — below the from-scratch minimum of 5.71% at 50k — and lifts HEAD_UNIQ to 76.55%, an additional $+8.74$pp over from-scratch 25k. The structural prior that from-scratch training takes $\geq$ 25k steps to begin acquiring is delivered by the backbone copy directly. The negative-transfer arm (E2E→ci, identical $208/211$ parameter copy) on the same recipe peaks at 72.61% HEAD_UNIQ (-3.94pp vs sonnet→ci- 10k) and *declines* thereafter to 68.77% at 50k. A paired two-level cluster bootstrap (Appendix G) shows that this lexical gap is *not* significant (95% CI $[-0.54, +8.37]$, $p=0.085$), whereas the structural gap is: $208/211$ parameters transfer in both arms, yet HEAD_SPECIAL at 10k differs by 9.89pp (95% CI $[-12.6, -7.1]$, $p<10^{-4}$, $\sim 60\times$). We therefore anchor the falsification of "any pretrained backbone helps" on the structural metrics. Because both sources are pad-mode and both arms are pad→block, the padding regime is held constant and cannot be the discriminating variable; §7 localises it instead to the source backbone's effective attention span. LCVR still drives every E2E→ci checkpoint to $0/0$ one-shot, confirming source-independence. The reverse arm (ci→sonnet, block→pad) yields no reliable speed-up: the sonnet from-scratch reference is itself non-monotone, and its 25k checkpoint already beats ci→sonnet-50k (Appendix C). The direction asymmetry (a robust pad→block speed-up, none in block→pad) is consistent with a long→short window-scale advantage as well as with the boundary regime, and we do not attribute it to either alone.

init	step	H.SP↓	H.UNIQ↑	TAIL↓
from scratch (block-mode)				
ci	25k	8.42	67.81	**0.00**
ci	50k	5.71	82.17	**0.21**
ci	100k	7.00	90.70	47.32
ci	400k	10.48	92.93†	74.41
sonnet→ci (forward, pad→block)				
son→ci	10k	**0.16**	76.55	2.60
son→ci	30k	5.43	75.33	**0.00**
son→ci	50k	5.88	**82.51**	0.25
E2E→ci (negative transfer, pad→block)				
E2E→ci	10k	9.63	58.76	7.23
E2E→ci	30k	10.04	72.61	0.26
E2E→ci	50k	6.08	68.77	1.26

Table 5: Transfer curve (%), 10 targets × 50 cand. †The 92.93 raw ci-400k HEAD_UNIQ is denominator-inflated; once LCVR structurally cleans the output the true content-side value drops to 75.45% (§6.1). Forward sonnet→ci at 10k already beats from-scratch 50k on structural metrics ($\sim 10\times$ speed-up); E2E→ci on the identical recipe is 9.89pp worse on H.SP at 10k (95% CI $[-12.6, -7.1]$, Appendix G), a falsification of "any pretrained backbone helps". Its H.UNIQ deficit (peak 72.61 at 30k vs 76.55 at 10k for sonnet→ci) is *not* statistically significant ($p=0.085$). Reverse ci→sonnet yields no reliable speed-up (Appendix C).

6.6 Rejection-sampling baseline

A natural objection is the "free" alternative of running the unconstrained baseline N times until a candidate happens to pass all structural checks. Across 980 unconstrained candidates spanning 5 ci checkpoints, 2 sonnet checkpoints, and 1 E2E checkpoint, *not a single* candidate satisfies HEAD_SPECIAL $=0 \wedge$ TEMPLATE-MISALIGN $=0 \wedge$ TAIL_OUTSIDER $=0$ jointly, while V_i decoding achieves 100% structural acceptance one-shot (per-checkpoint breakdown in Appendix E). The asymmetry is sharp because the failure modes are independent: with the empirical $\sim 90\% \to 10\%$ per-constraint rates we measure, the joint clean rate is far below any usable threshold. This converts

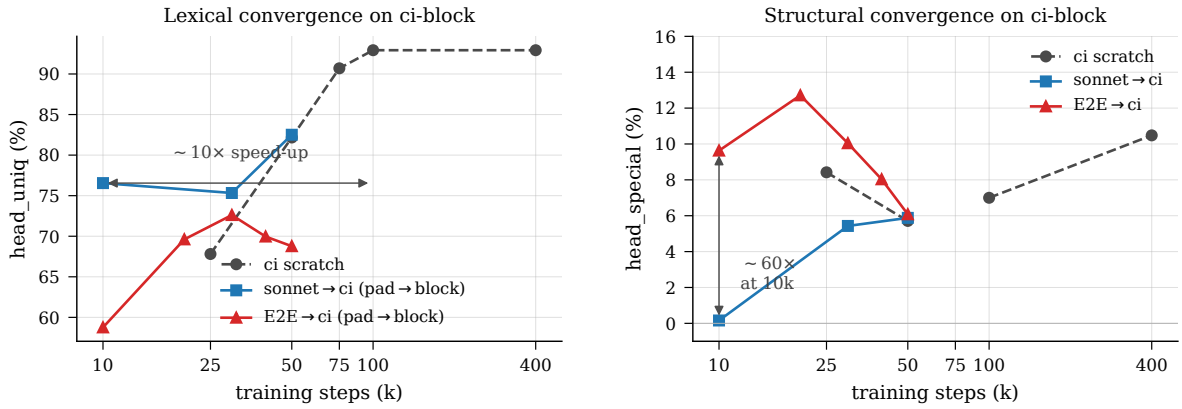


Figure 2: **Cross-domain backbone transfer on ci-block.** Each marker is one finetune checkpoint, 10 targets × 50 candidates. Left: lexical convergence. Initialising from the sonnet (pad-mode) backbone reaches head_uniq = 76.55% in 10k steps, which from-scratch training does not match until ∼100k (∼10× speed-up). Right: structural convergence. At 10k finetune steps, sonnet→ci has head_special = 0.16% vs E2E→ci's 9.63% (∼60× difference) on the same 208/211-parameter copy recipe, falsifying the "any pretrained backbone helps" hypothesis. Both sources are pad-mode, so the padding regime is held constant across the two arms; the attention-span probe of §7 localises the difference to the source backbone's long-range attention span.

a non-coverable region of the output space into a guaranteed- coverable one without any training-time cost.

7 Analysis

7.1 Why does sonnet→ci transfer work?

We attribute the benefit to three properties the *structural* part of the model captures and that are language-/ form-invariant: (i) **attention-span bias** — sonnet pad-mode training on $T = 196$ sequences encourages multi-head attention to attend across the whole realised content window (long-range, but not into the padded tail), a prior that survives the embedding swap; (ii) **LayerNorm / FFN statistics** — both adapt to the latent geometry induced by Gaussian denoising, which is independent of language; and (iii) **position-embedding semantics** — position embeddings encode "head" vs "tail" rather than word-class meaning. What does *not* transfer are tonal/rhyme regularities specific to ci (which live downstream of the diffusion backbone as classifier-guidance features) and vocabulary-frequency-dependent generation preferences (erased by the embedding swap by construction).

The negative E2E→ci arm in Tab. 5 adds the necessary falsifier: 208/211 parameters transfer in both arms, yet HEAD_SPECIAL at 10k differs by 9.89pp (∼ 60×; 95% CI $[-12.6, -7.1]$). This rules out "any pretrained backbone helps" as the explanation. It does *not*, however, implicate the padding regime: both sources are pad-mode and both arms are pad→block, so the boundary regime is identical in the arm that works and the arm that fails.

Attention-span probe. To find the discriminating variable we measured the effective attention span of each source backbone on 20 in-domain sequences at diffusion step $t = 100$ (Tab. 14). The sonnet source ($T = 196$) has a mean attention distance of 55.4 tokens and places 37.2% of its content-region attention mass beyond position 64; the E2E source ($T = 64$) has a mean distance of 17.2 tokens and places 0.0% beyond position 64 — 64 is its training length, an architectural ceiling that makes long-range attention impossible regardless of the probe sequence. The ci target ($T = 144$) has an effective content window of ≈80 tokens, which only the sonnet backbone can cover. This isolates the confounds a reader might raise: (1) *sequence length / padding*: the discriminator is attention-span scale, not pad/block matching; (2) *corpus size*: transfer is anti-correlated with source size (sonnet 2.7k vs. E2E 42k lines, ∼16×), ruling out "a larger corpus helps"; (3) *optimisation*: every E2E→ci checkpoint reaches 0/0 one-shot under LCVR. We therefore relabel the transferable prior as a *long-range attention bias*, and invoke the pad/block distinction only for the direction asymmetry. The probe is a content-agnostic attention-geometry measurement and thus argues against a poetry-specific explanation, but it does not by itself separate win-

dow scale from domain; window-scale-controlled sources (a long-window non-poetic source and a short-window poetic source) are left to future work.

7.2 Token-level vs latent-level V_i

At fixed V_i, using diffusion logits for the in-category argmax dominates a token-level modal-character fallback by $+5.45$pp HEAD_UNIQ on ci-400k (full table in Appendix D); both reach 0% structural error, isolating the value of operating in latent space. Removing the position-specific part of V_i (allowing the tail set $\{\texttt{END}, \texttt{PAD}\}$ everywhere) catastrophically inflates TAIL_OUTSIDER to 100%, confirming that the load-bearing contribution of V_i is the per-position template, not the "drop specials" restriction alone.

7.3 Speed, lexical ceiling, and what diffusion adds

The wallclock measurement (Tab. 4) and the controlled AR+CD comparison (Tabs. 2, 3) together yield an honest characterisation of the diffusion contribution. At our published 2000-step configuration diffusion is $\sim 200\times$ slower per candidate than the $\sim 16\times$-smaller AR baseline; DDIM-25 respacing narrows this to $1.7\times$ on sonnet at HEAD_UNIQ $= 42.4\%$ vs 41.7% at 2000 steps (no measurable quality loss). We do not claim an iso-parameter speed-up empirically. Lexically the two backbones nearly converge at 80k-line scale (ci, $\Delta = 1.37$pp) and diverge at 2.7k-line scale (sonnet), with the divergence attributable to the VPD lexical ceiling rather than to autoregression itself; we therefore frame the comparison as a *data-scale axis* along which the choice of backbone family matters, not as a strict ranking. After structural correctness is delegated to the backbone-agnostic V_i primitive and lexical quality is delegated to M3-v1 on data-rich domains, the residual value-add that is more readily realised with diffusion is **cross-domain backbone transfer** (Tab. 5): the same $208/211$-parameter copy is directly available for the diffusion backbone, whereas it is not straightforward under the corpus-local left-to-right factorisation and vocabulary of the AR baseline. This is a narrow but defensible contribution; we do not claim it is impossible for AR models.

The dissociation also clarifies engineering recommendations: practitioners using block-mode pretraining (most existing Diffusion-LMs) should treat BAD as a first-order failure mode and apply LCVR; practitioners using pad-mode face only VPD, which LCVR alone resolves; either way VPD appears regardless of padding mode and language in our three datasets, and the V_i fix applies in both regimes.

8 Conclusion

We diagnosed two failure modes of continuous Diffusion-LMs — BAD, specific to block-mode training, and VPD, present in all three datasets we studied — and adapted constrained decoding to diffusion rounding through the V_i rule, a MAP-justified position-template primitive that drives structural error to zero on both diffusion and AR backbones. A matched V_i+AR baseline exposes a data-scale boundary between the two backbone families; what diffusion more readily affords is cross-domain backbone transfer, with a negative-transfer falsifier anchored on structural metrics and an attention-span probe that localises the transferable component to the source backbone's long-range attention span, a recipe that is not straightforward under the corpus-local left-to-right factorisation of the AR baseline.

Limitations

The V_i rule enforces position-level *format*, not *content*: inter-position constraints (rhyme, tonal alternation) are not captured, and E2E SLOT_FIDELITY reaches only 11.04% even with M3-v1 (Appendix B). BAD is evidenced on a single block-mode corpus (*ci*), so we do not claim its universality; BMT remains an unvalidated recipe on which no conclusion depends. Our wallclock comparison is iso-implementation, not iso-parameter (~ 87M diffusion vs. ~ 5.5M AR), and DDIM-25 quality parity is verified on sonnet only. M3-v1 gap closure is domain-dependent (75% on ci, 16% on the 2.7k-line sonnet corpus). The transfer study uses one source–target pair per direction and a single seed; its arms do not isolate padding regime, the attention-span probe does not separate window scale from domain, and the negative-transfer falsification rests on structural metrics only (Appendix G). Finally, all metrics are automatic proxies and no human evaluation was run, so our claims are scoped to structural compliance and lexical diversity, not general text quality.

Ethical considerations

This work uses three publicly available text corpora (Chinese *ci* poetry, Shakespearean sonnets, and the

E2E NLG dataset), all of which are either public-domain literary text or research-release benchmarks. No personally identifying information is involved. The trained models generate structured text in the literary or restaurant-description style of their training data; we do not see plausible misuse vectors beyond those already present in small-scale open-source language models. Our code, sampling scripts, and evaluation harnesses are released under a research-friendly license to support reproducibility (`https://github.com/Hflying/priordiff`); model checkpoints are made available from the same repository.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2017. Guided open vocabulary image captioning with constrained beam search. In *Proceedings of EMNLP*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of ACL*.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. In *Proceedings of NeurIPS*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Daniel Deutsch, Shyam Upadhyay, and Dan Roth. 2019. A general-purpose algorithm for constrained sequential inference. In *Proceedings of CoNLL*.

Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, and 1 others. 2022. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. 2023. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of EMNLP*.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. 2022. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*.

Ishaan Gulrajani and Tatsunori B Hashimoto. 2023. Likelihood-based diffusion language models. *Advances in Neural Information Processing Systems*, 36:16693–16715.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of ACL*.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. 2022. Diffusion-LM improves controllable text generation. In *Proceedings of NeurIPS*.

Zhenghao Lin, Yeyun Gong, Yelong Shen, Tong Wu, Zhihao Fan, Chen Lin, Nan Duan, and Weizhu Chen. 2023. Text generation with diffusion language models: A pre-training approach with continuous paragraph denoise. In *Proceedings of ICML*, pages 21051–21064.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of SIGDIAL*.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of NAACL*.

Brandon T. Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *Preprint*, arXiv:2307.09702.

Cheng Yang, Maosong Sun, Xiaoyuan Yi, and Wenhao Li. 2018. Stylistic chinese poetry generation via unsupervised style disentanglement. In *Proceedings of EMNLP*.

Xiaoyuan Yi, Ruoyu Li, Cheng Yang, Wenhao Li, and Maosong Sun. 2020. MixPoet: Diverse poetry generation via learning controllable mixed latent space. In *Proceedings of AAAI*, pages 9450–9457.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Fei Huang, and Songfang Huang. 2022. Seqdiffuseq: Text diffusion with encoder-decoder transformers. *arXiv preprint arXiv:2212.10325*.

Lin Zheng, Jianbo Yuan, Lei Yu, and Lingpeng Kong. 2023. A reparameterized discrete diffusion model for text generation. *arXiv preprint arXiv:2302.05737*.

A Hyperparameters

B E2E NLG cross-domain probe (full table)

Same architecture as the ci main run, except image_size=8, $T = 64$, `padding_mode='pad'`. Evaluation: 30 random valid utterances drawn from `src1_valid.txt`; 50 candidates each. In addition to the metrics shared with ci/sonnet we report: BOUND_ALIGN $= \mathbf{1}\{\hat{x}_L = \texttt{END}\}$, ideal 1.0; and SLOT_FIDELITY, the fraction of MR slot values appearing as contiguous subsequences of the generated utterance.

Component	Setting
Diffusion (ci)	image_size=12, T=144, in_channel=16, padding_mode=block, BERT-base 12L/12H/768d, sqrt schedule, predict_xstart, EMA 0.9999
Diffusion (sonnet)	image_size=14, T=196, padding_mode=pad, otherwise identical
Diffusion (E2E)	image_size=8, T=64, padding_mode=pad, otherwise identical
LCVR	optional drop-low-zero frequency filter; mask logit indices $\geq \lvert\text{vocab}\rvert$ as forbidden
M3-v0	$\alpha \in \{2, 4, 8\}$, ≤ 10 Gibbs iters, repetition window over candidate's char slots
M3-v1 MLM	L=6, H=8, d=512, MLM mask prob 0.15, AdamW lr 1×10^{-4}, warmup 2000, steps 50k (default), batch 32
Transfer (sonnet→ci)	copy 208/211 params, re-init word_embedding + lm_head, finetune 50k steps lr 1×10^{-4}

Table 6: Per-component hyperparameters.

Method	VIOL.↓	BOUND↑	H.UNIQ↑	S.FID.↑
baseline	21.58	3.73	92.52‡	9.40
LCVR	**0.00**	**100.00**	80.20	8.58
+ M3-v1 (thr=3)	**0.00**	**100.00**	87.75	9.99
+ M3-v1 (thr=2)	**0.00**	**100.00**	**90.33**	**11.04**
GT ceiling	0	100	89.65	100

Table 7: E2E NLG (25k EMA; %). ‡Baseline denominator-inflated as in Tab. 2. M3-v1 uses a BERT MLM trained 10k steps on the 42k-utterance E2E train split (final loss 0.77). VIOL. = max(HEAD_SPECIAL, TAIL_OUTSIDER); 0% by construction under LCVR.

The dominant qualitative finding is the boundary signal: the baseline produces END at position L in only 3.73% of candidates, LCVR in 100% by construction (+96.27pp). M3-v1 (thr=2) reaches 107% LCVR-to-GT gap closure on HEAD_UNIQ by substituting some of the high-frequency repeats in the GT itself with slot-aligned content tokens; this also bumps SLOT_FIDELITY from 8.58 to 11.04. The latter number is a limitation of the method rather than a success: V_i enforces position-level format (length, boundaries, slot positions) but not slot *content*, so LCVR solves the formatting half of structured generation and leaves content realisation open (see Limitations). The optional slot-aware V_i variant (§5.1) restricts MR-value positions to the value's tokens and is the natural next step, but we do not evaluate it here. VPD shows up here too: 21.58% HEAD_SPECIAL is the highest of the three datasets despite E2E having the smallest vocabulary, attributable to the very short content windows ($L \in [8, 30]$) which give Gaussian latent noise more leverage to project to special tokens; LCVR sets it to 0% by per-position constraint.

C Full cross-domain transfer tables

We report the complete transfer curves, including the E2E→ci falsification arm at five finetune checkpoints (10/20/30/40/50k) and the ci→sonnet reverse-direction arm (E4) at five finetune checkpoints. The body uses the abbreviated version (Tab. 5).

C.1 E3: E2E→ci as a negative-transfer stress test

Both transfer arms (sonnet→ci, E2E→ci) use the same 208-of-211-parameter copy recipe and both sources are pad-mode; only the source corpus (and hence its training length T and attention span, §7) differs. Under the hypothesis that *any* pretrained backbone helps, E2E→ci should show *comparable* early-training structural speed-up to sonnet→ci; under the hypothesis that the benefit depends on the source backbone's long-range attention span (or on an English poetry-specific prior), E2E→ci should show minimal or negative transfer.

source	step	H.SP↓	H.UNIQ↑	H.COM↑	TAIL↓
scratch	25k	8.42	67.81	24.70	**0.00**
scratch	50k	5.71	**82.17**	—	0.21
son→ci	10k	**0.16**	76.55	28.94	2.60
son→ci	30k	5.43	75.33	37.12	**0.00**
son→ci	50k	5.88	**82.51**	39.57	0.25
E2E→ci	10k	9.63	58.76	39.10	7.23
E2E→ci	20k	12.71	69.61	42.96	0.16
E2E→ci	30k	10.04	*72.61*	44.75	0.26
E2E→ci	40k	8.03	69.96	45.51	1.42
E2E→ci	50k	6.08	68.77	44.95	1.26
E2E→ci 10k + LCVR	—	**0.00**	53.70	46.21	**0.00**
E2E→ci 30k + LCVR	—	**0.00**	65.59	51.72	**0.00**
E2E→ci 50k + LCVR	—	**0.00**	65.99	49.36	**0.00**

Table 8: E3 negative-transfer stress test (%; 10 targets × 50 candidates). E2E→ci's HEAD_UNIQ peaks at 72.61% at 30k and *declines* to 68.77% at 50k; the peak is 3.94pp below sonnet→ci-10k (76.55%), a gap that is *not* statistically significant (95% CI $[-0.54, +8.37]$, $p = 0.085$; Appendix G). At 10k, sonnet→ci's HEAD_SPECIAL is $\sim 60\times$ below E2E→ci's (0.16 vs 9.63; −9.89pp, 95% CI $[-12.6, -7.1]$, $p < 10^{-4}$), which is the gap the falsification rests on. Crucially, LCVR drives every E2E→ci checkpoint to H.SP/TAIL = 0 one- shot, confirming the LCVR claim is source-independent.

C.2 E4: ci→sonnet as a reverse-direction arm

The reverse arm (E4) initialises sonnet pad-mode from the ci-400k block-mode backbone with the same 208-of-211-parameter copy recipe (target vocab 22,970) and fine-tunes for 50k steps on the

2.7k sonnet train split. The source here is block-mode and the target pad-mode, reversing both the boundary regime and the language of the forward arm.

init	step	H.SP↓	EOS↓	TAIL↓	H.UNIQ↑	+LCVR↑
sonnet scratch references						
scratch	25k	14.97	90.28	8.24	53.25	47.42
scratch	100k	15.02	89.08	10.84	45.54	41.72
ci→son	10k	*4.25*†	*93.94*†	*94.35*†	*74.18*†	*71.92*†
ci→son	20k	11.38	90.65	13.82	55.29	50.42
ci→son	30k	22.53	90.67	10.46	46.55	37.69
ci→son	40k	19.98	90.70	8.96	49.21	41.24
ci→son	50k	17.72	90.43	9.50	**51.61**	**44.68**

Table 9: E4 reverse-direction backbone transfer (ci→sonnet) (%; 10 targets × 50 cand). †The 10k row is in italic as a caveat: with TAIL = 94.35% the model has not yet re-learned pad-mode boundary behaviour and its apparent high HEAD_UNIQ is inflated by random-looking tail tokens. At 50k, ci→sonnet matches sonnet scratch at 100k on structural metrics with a +6.07pp HEAD_UNIQ gain (+2.96pp under LCVR); however, sonnet scratch at 25k (53.25 H.UNIQ, 14.97 H.SP) already beats ci→sonnet-50k (51.61, 17.72) on both metrics, and the scratch reference is itself non-monotone (53.25 at 25k → 45.54 at 100k), so no single-checkpoint reverse speed-up claim is reliable and we make none. By contrast the forward arm is baseline-independent (H.SP 0.16% at 10k, below the best scratch value 5.71% at 50k). The direction asymmetry is consistent with both the boundary regime and a long→short window-scale advantage.

D Token-level vs latent-level V_i: full ablation

A natural follow-up to the rejection-sampling baseline (§6.6) is whether LCVR's *latent-level* operation is materially different from *token-level* constrained decoding along the lines of Hokamp and Liu (2017); Willard and Louf (2023) or the widely deployed LM Format Enforcer.[3] A token-level enforcer applied *post-hoc* to the unconstrained-argmax baseline only has access to the emitted token sequence: when it encounters a disallowed token it has no logit information and the natural fallback is to replace it with a *corpus-modal* in-category token. LCVR has access to the pre-argmax latent logits and can pick *argmax* over V_i for every position.

[3] https://github.com/noamgat/lm-format-enforcer

method	H.SP↓	H.UNIQ↑	H.COM↑	TAIL↓
Unconstrained baseline	10.48	92.93⋆	31.98	74.41
Token-level post-hoc (modal-char fallback)				
shared V_i (no_pos)	**0.00**	70.00	49.93‡	**0.00**
tpl, specials-only	**0.00**	81.22	41.96	**0.00**
full template	**0.00**	70.00	49.93‡	**0.00**
Latent-level LCVR ablations				
shared V_i (no_pos)	**0.00**	75.45	29.60	100.00†
lax vocab (puncts allowed)	**0.00**	83.81	34.33	**0.00**
LCVR full [ours]	**0.00**	**75.45**	29.60	**0.00**

Table 10: Latent-level vs token-level constrained decoding on ci-400k (%). ⋆Inflated by the 10.48% special leak and 74.41% tail-outsider of the baseline. †Without the tail{END, PAD} position-rule, all $T-L$ tail positions receive character tokens, inflating TAIL to 100%: this ablation confirms that the position-specific part of V_i is load-bearing, not just "drop specials". ‡Token-level post-hoc's high H.COM is misleading: it reflects repetitions of the single modal-char fallback, not content diversity. The crucial comparison is LCVR full vs token-level full: both reach 0% structural error, but LCVR's H.UNIQ is +5.45pp higher, the value of operating in latent space.

E Rejection-sampling per-checkpoint breakdown

Per-checkpoint detail of the rejection-sampling baseline of §6.6. For each domain we count how many of the $N \in \{1, 5, 50\}$ baseline candidates per target satisfy HEAD_SPECIAL = 0 ∧ TAIL_OUTSIDER = 0 ∧ TEMPLATE-MISALIGN = 0 simultaneously.

domain (ckpt)	cand/target	cand-clean	target-acc
ci-block (25k)	50	0/500 = 0%	0/10 = 0%
ci-block (50k)	50	0/500 = 0%	0/10 = 0%
ci-block (75k)	50	0/500 = 0%	0/10 = 0%
ci-block (100k)	50	0/500 = 0%	0/10 = 0%
ci-block (400k)	50	0/500 = 0%	0/10 = 0%
sonnet-pad (25k)	5	0/50 = 0%	0/10 = 0%
sonnet-pad (100k)	5	0/50 = 0%	0/10 = 0%
E2E-pad (25k)	1	0/30 = 0%	0/30 = 0%
LCVR	1	**500/500**	**10/10**

Table 11: Rejection sampling cannot serve as a free baseline: *not a single* unconstrained candidate, across 5 ci checkpoints, 2 sonnet checkpoints, and 1 E2E checkpoint (980 candidates total), satisfies all three structural constraints simultaneously. LCVR achieves 100% structural acceptance one-shot.

F Automated fluency proxy (n-gram coverage)

A widely asked question is whether LCVR's structural compliance is purchased at the price of *fluency*. Because human evaluation at scale is out of scope for this submission, we report a cheap and reproducible automated proxy: the fraction of bi/trigrams in the generated content slots that occur in the corresponding training corpus, plus a Laplace-smoothed trigram cross-entropy.

domain	method	bg cov.↑	tg cov.↑	tg-xent↓
ci-block (400k)	baseline†	57.46	4.26	12.45
	+ LCVR	67.91	4.00	12.47
	+ LCVR + M3-v0	67.23	3.58	12.49
	+ LCVR + M3-v1	**70.93**	**10.70**	**12.36**
sonnet-pad (100k)	baseline†	57.01	11.64	14.17
	+ LCVR	33.16	5.43	14.26
	+ LCVR + M3-v1$_{\text{THR5}}$	**63.91**	**7.05**	**14.23**
E2E-pad (25k)	baseline†	**93.72**	**81.22**	**5.66**
	+ LCVR	69.68	58.75	7.29
	+ LCVR + M3-v1$_{\text{THR2}}$	86.82	72.99	6.49
	+ LCVR + M3-v1$_{\text{THR3}}$	85.04	69.42	6.70

Table 12: Fluency proxy across three domains (%). †Baseline rows have HEAD_SPECIAL/template- misalign $\geq 10\%$ and (in ci) TAIL_OUTSIDER=74%; their n-gram coverage is therefore inflated by structurally broken outputs. The direction of LCVR's n-gram change depends on whether the baseline is structurally clean: on ci-block (severe BAD), LCVR simultaneously fixes structure *and* raises bigram coverage by $+10$pp; on sonnet/E2E (clean pad-mode), LCVR sacrifices $\sim$24pp bigram coverage for guaranteed structure. M3-v1 recovers most of the lost coverage in 2/3 domains.

G Variance estimates and attention-span probe

All headline numbers in the body are computed over 10 targets $\times$ 50 candidates from a single training run per system. To quantify the uncertainty of the small gaps the argument depends on, we run a paired two-level cluster bootstrap: resample the 10 targets with replacement, then the 50 candidates within each resampled target, recompute both systems' metrics on the same resample, and take the difference ($B = 20{,}000$). We report percentile 95% confidence intervals and two-sided p-values. The pipeline reproduces the paper's point estimates (90.87 / 76.55 / 0.16 / 72.61 exactly; AR+CD is 92.24 in the bootstrap pipeline vs. 92.36 in Tab. 2, which is why the gap below reads $+1.37$pp rather than $+1.49$pp).

comparison	metric	Δ (pp)	95% CI	p
AR+CD − (LCVR+M3-v1), ci	H.UNIQ	$+1.37$	$[+0.69, +2.03]$	2×10^{-4}
son→ci-10k − E2E→ci-30k	H.UNIQ	$+3.94$	$[-0.54, +8.37]$	0.085
son→ci-10k − E2E→ci-10k	H.SP	-9.89	$[-12.6, -7.1]$	$<10^{-4}$

Table 13: Paired two-level cluster bootstrap ($B = 20{,}000$) for the three gaps the argument depends on. The diffusion–AR convergence gap on ci is small but statistically resolved; the lexical negative-transfer gap is *not* significant, whereas the structural negative-transfer gap is. We therefore anchor the "any pretrained backbone helps" falsification on HEAD_SPECIAL rather than on HEAD_UNIQ.

Attention-span probe. For each source backbone we take 20 in-domain sequences, run the denoiser at diffusion step $t = 100$, and measure (i) the mean attention distance $|i - j|$ weighted by attention mass over content positions, and (ii) the fraction of content-region attention mass that lands beyond position 64 (the E2E training length). Both source backbones are trained in pad-mode, so the probe isolates window scale from the padding regime.

backbone	role	T	mean dist. (tok)	mass >64
sonnet (pad)	source, works	196	55.4	37.2%
E2E (pad)	source, fails	64	17.2	0.0%
ci (block)	target	144	—	content window ≈ 80

Table 14: Effective attention span of the two transfer sources (20 in-domain sequences, $t = 100$). The E2E source cannot place any attention mass beyond its training length 64, so it cannot cover the ci target's ≈ 80-token content window; the sonnet source routinely does. Corpus size is anti-correlated with transfer success (2.7k vs. 42k lines), ruling out "a larger corpus helps".