LEARN AND STEER: MECHANISTICALLY INFORMED INTERVENTIONS FOR DISCRETE DIFFUSION LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Discrete diffusion language models (DLMs) generate text by iteratively denoising all positions in parallel, offering an alternative to autoregressive models. Controlled generation methods for DLMs, imported from autoregressive models, apply uniform intervention at every denoising step. We show this uniform schedule is inefficient and degrades quality, and the damage compounds when multiple attributes are steered jointly. To diagnose the failure, we train sparse autoencoders on four DLMs (124M-8B parameters) and find that different attributes commit on distinct schedules, varying in timing, sharpness, and magnitude. For instance, topic commits within the first 2% of denoising on MDLM, whereas sentiment emerges gradually over 20% of the process. Motivated by these profiles, we propose an adaptive scheduling mechanism that concentrates intervention where each attribute is actively forming. An idealized allocation analysis predicts that attributes with more sharply concentrated emergence benefit more from adaptive scheduling, a prediction we confirm empirically. Across seven single- and multi-attribute steering tasks on four DLMs, adaptive steering consistently improves the control-quality balance over uniform and interval-restricted baselines.

1 INTRODUCTION

Discrete diffusion language models (DLMs) generate text by iteratively denoising corrupted token sequences and have recently become competitive with autoregressive models (Austin et al., 2021; Sahoo et al., 2024; Lou et al., 2024; Ye et al., 2025; Nie et al., 2026). Unlike autoregressive generation, where each token is produced once in fixed left-to-right order, DLMs refine all positions in parallel across hundreds of steps, creating a temporal trajectory where semantic content is progressively committed. This raises a fundamental question: how do DLMs organize semantic attributes across denoising, and can that temporal structure be exploited for controlled generation?

Sparse autoencoders (SAEs) decompose dense activations into sparse, interpretable features (Cunningham et al., 2023; Bricken et al., 2023; Templeton et al., 2024) and have been used to steer autoregressive models (Goyal et al., 2025; Yeon et al., 2025; Cho et al., 2025; Arad et al., 2025). Concurrent work explores applying SAEs to DLMs (Wang et al., 2026) and steering DLMs via reference-sequence alignment (Avrahami & Nachmani, 2026). Other approaches from autoregressive models, including contrastive vectors (Zou et al., 2023; Rimsky et al., 2024) and probes (Belinkov, 2022), can also be adapted to DLMs. All existing methods, however, apply the same intervention at every denoising step, treating the trajectory as temporally uniform. This degrades generation quality in both single- and multi-attribute control, inflating perplexity and collapsing diversity (Gu et al., 2022a), while increasing cross-attribute interference, the unintended shifting of non-target attributes (Gu et al., 2022b; Oozeer et al., 2025; Shafran et al., 2026), as we verify in §5. Yet how attributes form and commit across denoising steps in DLMs remains underexplored, leaving open whether temporally informed intervention can avoid this quality cost.

We take an interpretability-first approach to address this gap (Figure 1). We train SAEs on four DLMs spanning two training objectives, three architectures, and a $60\times$ scale range (MDLM (Sahoo et al., 2024), SEDD (Lou et al., 2024), DREAM (Ye et al., 2025), and LLaDA (Nie et al., 2026)) and use the resulting features to characterize how attributes evolve across the denoising trajectory.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

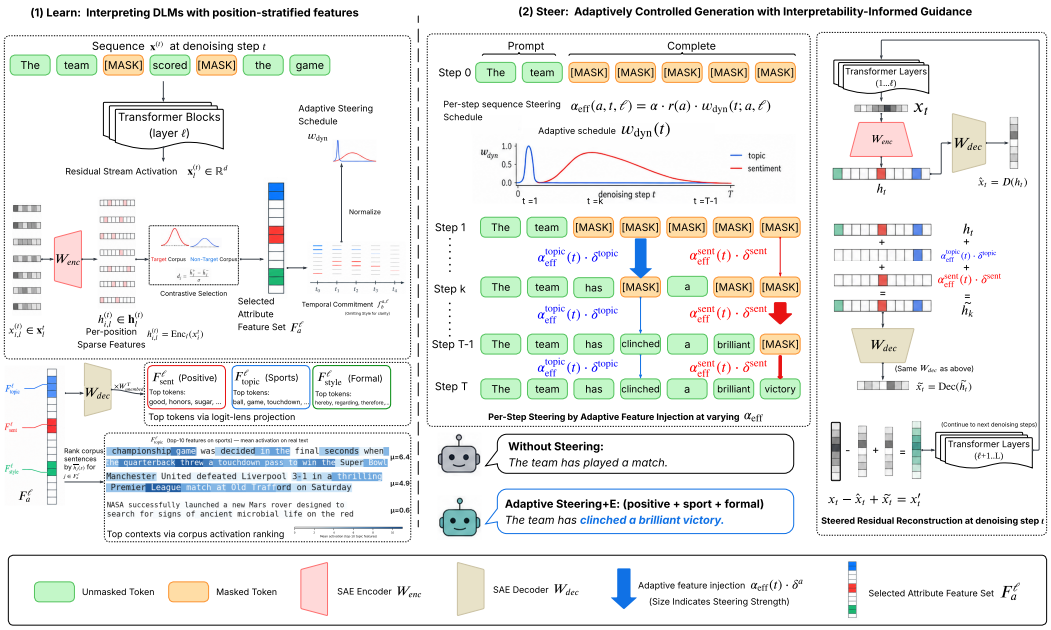


Figure 1: **Method overview.** (1) SAEs trained per-layer decompose the residual stream into interpretable features. Contrastive selection identifies attribute-relevant feature sets F_a^ℓ, whose temporal commitment across denoising steps defines an adaptive schedule $w_{\text{dyn}}(t)$. (2) At each step t, a sparse contrastive shift $\alpha_{\text{eff}}(a,t,\ell)\,\delta^{(a)}$ is applied to selected features and decoded with residual correction, enabling multi-attribute composition. All notations and terms introduced are summarized in Table 6.

Our analysis (§3) identifies three properties that directly inform steering: *commitment timing*, where attributes commit at different denoising steps; *commitment sharpness*, where emergence may be concentrated in a narrow window or spread across the trajectory; and *effect-size asymmetry*, where attributes carry different per-feature discriminative strength. These findings explain the quality cost of uniform steering: a constant intervention spends budget on steps where the target attribute has already committed or has not yet begun to emerge. From these empirical profiles, we derive *adaptive steering*, a schedule that concentrates intervention where each attribute is actively forming and leaves the rest of the trajectory undisturbed. An idealized budgeted-allocation analysis shows that the optimal schedule should concentrate intervention where emergence is sharpest, motivating our design (§4). For multi-attribute composition, we steer through disjoint SAE feature sets and rebalance per-attribute strengths via an effectiveness-ratio calibration so weaker attributes are not drowned out.

We conduct extensive experiments across four DLMs, seven single- and multi-attribute steering tasks, and six baselines spanning residual-stream, prompt-based, and SAE-based approaches. Our key findings are: (i) SAE-based steering provides the most consistent control profile, avoiding the model-dependent failure modes of residual-stream and prompt-based approaches; and (ii) adaptive scheduling consistently improves the control-quality balance over uniform and step-function Kynkäänniemi et al. (2024) steering, achieving lower quality cost at matched control and shifting the Pareto frontier toward higher quality (§5). In summary, we contribute: (i) a mechanistic characterization of how semantic attributes emerge during DLM denoising, identifying commitment timing, sharpness, and effect-size asymmetry as primary axes of variation; (ii) an adaptive steering framework that concentrates intervention where each attribute is actively forming, with per-attribute effect-size calibration, informed by a budgeted-allocation analysis; and (iii) a comprehensive evaluation showing that adaptive steering achieves the best quality-control balance across models.

2 Related Work

SAEs have become a standard tool for feature interpretability in autoregressive LMs, linking features to concepts via vocabulary projections (Cunningham et al., 2023; Bricken et al., 2023; Templeton et al., 2024), with applications to causal steering such as detoxification and safety control (Goyal

Table 1: Comparison of our work with concurrent efforts. ✓ = fully, ∼ = partially, ✗ = not addressed.

Work	Models	Approach	Feature Interpretability	Causal Steering	Temporal Dynamics	Cross-Objective	Compositional Control	Multi-Scale	Baselines
DLM-Scope (Wang et al., 2026)	Dream, LLaDA	SAE (Top-K, 16K)	∼	✓	✓	✗	✗	✗	∼
ILRR (Avrahami & Nachmani, 2026)	MDLM, LLaDA	Reference-based	✗	✓	∼	✗	✗	∼	✓
Skip-to-Good-Part (Goel et al., 2026)	LLaDA, Dream	Probing / RSA	✗	✗	∼	∼	✗	✗	✗
TDGNet (Hemmat et al., 2026)	LLaDA, Dream	Temporal graphs	✗	✗	✓	✗	✗	✗	✗
Ours	MDLM, SEDD, Dream, LLaDA	SAE (Top-K, 12–16K)	✓	✓	✓	✓	✓	✓	✓

et al., 2025; Yeon et al., 2025; Cho et al., 2025; Rodriguez et al., 2026; 2025; Arad et al., 2025), though effectiveness can be sensitive to feature selection and layer choice (Ronge et al., 2026; Basu et al., 2026). Optimal-transport methods (Rodriguez et al., 2025; 2026) fix operator strength within a single forward pass; our schedule modulates when to intervene across the T denoising steps, making the two approaches complementary. For DLMs, concurrent work has begun exploring interpretability and control. DLM-Scope (Wang et al., 2026) trains SAEs on DREAM and LLaDA and studies temporal decoding dynamics, but does not perform cross-objective comparison or demonstrate compositional control. ILRR (Avrahami & Nachmani, 2026) steers MDLM and LLaDA via reference-sequence alignment without feature-level interpretability. Other efforts study AR vs. DLM representations (Goel et al., 2026) and temporal attention for hallucination detection (Hemmat et al., 2026). Multi-attribute control in AR models requires specialized methods (Oozeer et al., 2025; Shafran et al., 2026); we show that SAE feature disjointness in DLMs enables additive composition. Table 1 summarizes how our work compares with concurrent efforts.

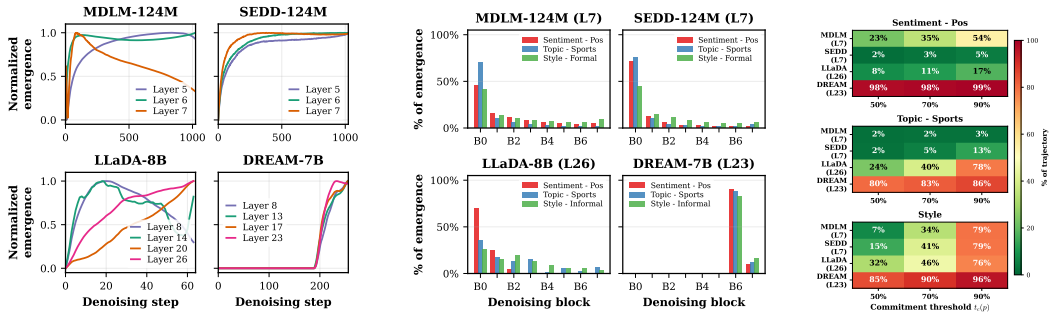
3 Interpreting Discrete Diffusion LMs via Sparse Autoencoders

DLMs offer a structural advantage for controlled generation: all positions are refined across a temporal trajectory, so different attributes may commit at different denoising steps, creating natural intervention points. Characterizing when and how sharply each attribute commits, and how attributes interact through temporal overlap or shared features, is essential for designing effective steering schedules. We use sparse autoencoders (SAEs), which decompose dense activations into interpretable feature dictionaries and have been applied to steer AR models (Goyal et al., 2025; Yeon et al., 2025; Cho et al., 2025; Arad et al., 2025), but remain largely unstudied for DLMs. We train SAEs on four DLMs and study how features encoding sentiment, topic, and style evolve and interact.

3.1 SAE Training and Attribute Feature Extraction

First, we train TopK SAEs (k=32) (Gao et al., 2025; Makhzani & Frey, 2013; Zhu et al., 2025) on residual-stream activations of four DLMs: MDLM (124M), SEDD (124M), LLaDA (8B), and DREAM (7B). They vary by architecture and diffusion objective: SEDD uses score-entropy loss while the other three use absorbing loss. We train at layers selected by our diffusion logit lens and layer study (§C.4, §C.5.2). DLM activations shift with the masking rate, so we sample training activations across uniformly distributed masking rates to cover the full denoising trajectory (see C.1 for details).

Next, we study three attributes: *sentiment*, *topic*, and *style*, a natural combination for multi-attribute study as they capture orthogonal aspects of text: subject matter, opinion polarity, and writing register. Following work on contrastive steering directions (Zou et al., 2023; Rimsky et al., 2024; Belinkov, 2022), we construct balanced binary corpora: Sports vs. Business from AG News (Zhang et al., 2015) for topic, positive vs. negative IMDB reviews (Maas et al., 2011) for sentiment, and formal vs. informal text for style, using the Pavlick corpus (Pavlick & Tetreault, 2016) for MDLM/SEDD and classifier-labeled IMDB reviews (Babakov et al., 2023) for LLaDA/DREAM to match their generation domain (§C.7.1). For style, we track the direction with greater room for control: formality on MDLM/SEDD (baseline 55–58%) and informality on LLaDA/DREAM (baseline 22–27%). To identify attribute-encoding features, we noise each text at multiple mask ratios, pass the noised inputs through the model, and ReLU-encode the hidden states through the SAE at each layer (§C.7.2). For each feature j, we compute the Cohen's d effect size (Cohen, 2013): $d_j = (\bar{h}_j^+ - \bar{h}_j^-)/\sqrt{(\text{Var}(h_j^+) + \text{Var}(h_j^-))/2}$, where $\bar{h}_j^+$ and $\bar{h}_j^-$ are the mean activations for the two contrastive classes. We retain the top 50 features per direction per layer that pass a Mann–Whitney U test (Mann



(a) Emergence trajectories (b) Block fractions (c) Commitment times

Figure 2: Temporal hierarchy of feature emergence: (a) Topic emergence (normalized $|a(t) - a(0)|$) across layers; (b) Block fractions (B=8) at deepest layers; (c) Commitment time $t_c(p)$: % of trajectory at which p% of emergence has occurred (deepest layer, all attributes). Green = early; Red = late.

& Whitney, 1947) ($p < 0.01$, Bonferroni-corrected (Bonferroni, 1936)) for the interpretability analysis below. Vocabulary grounding via $\mathbf{W}_{\text{unembed}}$ confirms that retained features are semantically coherent (§B).

3.2 Interpreting Temporal Dynamics of Features

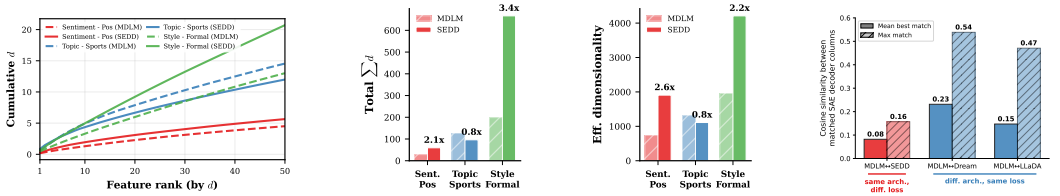
To analyze feature dynamics, we generate N samples per model with SAE hooks active at every step, record per-step feature activations, filter to classifier-confident samples using off-the-shelf attribute classifiers (since not all generated texts express the target attribute; §C.8), and partition the trajectory into B temporal blocks. We quantify emergence timing via *block fractions*, the share of total non-negative activation change in each block: $f_b = \max(0, \bar{h}_{e_b} - \bar{h}_{s_b}) / \sum_{b'} \max(0, \bar{h}_{e_{b'}} - \bar{h}_{s_{b'}})$, where s_b and e_b are the start and end steps of block b. We summarize our key observations below.

Commitment timing: attributes commit at different denoising steps. Attribute emergence timing varies by attribute, model, and layer. For topic (Figure 2a), MDLM and SEDD commit within the first few percent of denoising, with deeper layers faster; LLaDA shows gradual trajectories; DREAM defers emergence to the final quarter, likely due to its AR-based initialization which inherits late-trajectory token-finalization behavior. Figure 2c reports $t_c(p)$ (trajectory percentage at which p% of emergence has occurred) at the deepest layer per model. The patterns differ strikingly: on MDLM, topic reaches 50% emergence at t_c=22 (2% of trajectory) while sentiment reaches it at t_c=232 (23%), a $\sim$10$\times$ gap. Emergence trajectories for sentiment and style are in §D.1.

Commitment sharpness: models differ in how concentrated emergence is. Figure 2b quantifies emergence concentration via block fractions. MDLM and SEDD show consistent early concentration across all layers; at the deepest layer, block 0 captures 71% of topic emergence on MDLM and 75% on SEDD. LLaDA shows a different pattern: sentiment concentrates in block 0 ($\sim$72%) while topic and style are more spread (27–35%), and non-deepest layers show milder concentration overall (§D.2). DREAM concentrates *all* emergence in the final two blocks at every layer, with blocks 0–5 contributing zero. Commitment timing and sharpness together define each model's temporal profile; the diversity of these profiles across models, layers, and attributes means that *no single fixed schedule can be optimal for controllability*, motivating adaptive steering (§4.1).

Effect-size asymmetry: attributes differ in per-feature discriminative strength. Topic features carry $2-3\times$ larger per-feature Cohen's $|d|$ than sentiment across all models (Figure 3a), meaning equal-strength interventions over-steer topic and under-steer sentiment, motivating effectiveness-ratio calibration (§4.1). Comparing MDLM and SEDD (same architecture and data, different loss) at layer 7 reveals further structure (Figure 3): SEDD carries $2.1\times$ (sentiment) and $3.4\times$ (formality) more total signal ($\sum d$, Figure 3b), while topic is comparable ($0.8\times$). The effective dimensionality ($(\sum d)^2 / \sum d^2$; Figure 3c) shows SEDD spreads sentiment across $2.6\times$ and formality across $2.2\times$ more features; topic is more concentrated. We examine the effect on steering in §5.

Training objective shapes feature geometry. Cross-model feature alignment (details in § D.5) confirms the training objective as the dominant factor (Figure 3d): MDLM and SEDD yield near-



(a) Cumulative effect size (b) Total effect size (c) Effect spread (d) Cross-model alignment

Figure 3: Training objective shapes feature geometry. **(a–c)** MDLM vs. SEDD (layer 7, Cohen's $d > 0$ features): **(a)** cumulative Cohen's d vs. rank; **(b)** total $\sum d$; **(c)** effective dimensionality. **(d)** Cross-model alignment: sharing a loss family yields higher similarity than sharing architecture.

orthogonal features (cosine sim. 0.08), while same-loss-family pairs show substantially higher alignment: MDLM and DREAM (different architecture, $56\times$ scale gap) achieve 0.23 mean and 0.54 max similarity, and MDLM and LLaDA ($64\times$ scale gap) achieve 0.15 mean and 0.47 max.

Together, these findings motivate a steering framework that adapts to each model's temporal structure.

4 Interpretability Informed Steering

The analysis in §3 reveals that attributes commit at different steps with different sharpness and carry different per-feature discriminative strengths. We exploit this by steering in SAE feature space, modulating intervention strength across denoising steps and attributes.

Setup. At denoising step t and layer ℓ, we encode $h_t = \mathrm{Enc}(x_t)$, reconstruct $\hat{x}_t = \mathrm{Dec}(h_t)$, and modify features on the attribute-specific set F_a^ℓ:

$$\tilde{h}_{t,j} \;=\; h_{t,j} + \alpha_{\text{eff}}(a,t,\ell)\,\delta_j^{(a)} \qquad (j \in F_a^\ell), \tag{1}$$

where $\delta_j^{(a)} = \bar{h}_{j,+}^{(a)} - \bar{h}_{j,-}^{(a)}$ is the contrastive shift, signed so that positive α_{eff} steers toward the target pole. The forwarded residual is $x_t' = x_t - \hat{x}_t + \tilde{x}_t$ with $\tilde{x}_t = \mathrm{Dec}(\tilde{h}_t)$; the off-basis term $x_t - \hat{x}_t$ preserves SAE-residual information, and the affine decoder gives $x_t' - x_t = W_{F_a^\ell}(\alpha_{\text{eff}}\,\delta^{(a)})$, so the intervention acts only through the selected coordinates.

4.1 Steering Configurations

We decompose the effective steering strength into three components:

$$\alpha_{\text{eff}}(a,t,\ell) = \underbrace{\alpha}_{\text{overall scale}} \cdot \underbrace{r(a)}_{\text{attribute weight}} \cdot \underbrace{w_{\text{dyn}}(t;a,\ell)}_{\text{temporal weight}}. \tag{2}$$

The scalar α controls the overall steering magnitude (swept during evaluation); $r(a) \in (0,1]$ rebalances strength across attributes; and $w_{\text{dyn}}(t;a,\ell) \in [0,1]$ modulates strength across denoising steps according to each attribute's empirical commitment profile.

For multiple attributes $\mathcal{A} = \{a_1, \ldots, a_m\}$, we compose interventions additively in feature space:

$$\tilde{h}_{t,j} \;=\; \mathrm{ReLU}\left(h_{t,j} + \sum_{i:\, j \in F_{a_i}^\ell} \alpha_{\text{eff}}(a_i,t,\ell)\,\delta_j^{(a_i)}\right), \tag{3}$$

with steered residual $x_t' = x_t - \hat{x}_t + \tilde{x}_t$. The ReLU clamp enforces the non-negative activation domain of the TopK SAE (a single $\delta^{(a)}$ at moderate α_{eff} does not drive coordinates negative, but a sum of updates can). This formulation is effective because the contrastive selection procedure (§C.2) yields near-disjoint feature sets across attributes. Different choices of w_{dyn} and r yield four steering configurations of increasing sophistication, each incorporating a successive interpretability finding.

Uniform (r=1, w_{dyn}=1). Constant α at every step. This ignores the temporal structure uncovered in §3.2: different attributes commit at different denoising steps (Figure 2c), so a uniform schedule over-steers already-committed attributes and under-steers those still emerging.

Adaptive (r=1). Directly operationalizes commitment timing and sharpness (Figures 2a–2b): intervention is concentrated during active emergence and relaxed elsewhere. We partition the T denoising steps into K blocks and measure the *commitment fraction*, the share of total feature activation growth in each block: $f_b^{(a,\ell)} = \frac{\sum_{j\in F_a^\ell} \max\left(0, \bar{h}_j^{(s_b^{\text{end}})} - \bar{h}_j^{(s_b^{\text{start}})}\right)}{\sum_{b'} \sum_{j\in F_a^\ell} \max\left(0, \bar{h}_j^{(s_{b'}^{\text{end}})} - \bar{h}_j^{(s_{b'}^{\text{start}})}\right)}$, where $\bar{h}_j^{(t)}$ is the mean activation of feature j at step t (averaged over the contrastive corpus) and negative differences are clamped to zero, capturing net emergence only. The adaptive weight for all steps t within block b is

$$w_{\text{dyn}}(t; a, \ell) = \frac{f_b^{(a,\ell)}}{\max_{b'} f_{b'}^{(a,\ell)}}, \tag{4}$$

with the peak-emergence block receiving weight 1. When emergence is sharply concentrated (e.g., topic on MDLM, where block 0 captures 71%; §3.2), $w_{\text{dyn}} \approx 1$ early and ≈ 0 thereafter; when emergence is perfectly uniform ($\eta = 1$), $w_{\text{dyn}} \approx 1$ everywhere and Adaptive reduces exactly to Uniform. In practice, no condition in our set reaches this limit (η ranges from 1.44 to 5.01; §A.5), so Adaptive retains a quality advantage across all conditions.

Uniform+E and Adaptive+E. Addresses effect-size asymmetry (Figure 3a): attributes carry different per-feature Cohen's d magnitudes, so the same α produces disproportionate perturbations (e.g., topic features carry $2{-}3\times$ larger $|d|$ than sentiment; §3.2), causing stronger attributes to saturate before weaker ones reach target. We calibrate from single-attribute dose-response sweeps: for each attribute a, let α_a^* be the minimum strength reaching a reference control threshold. The effectiveness ratio is $r(a) = \alpha_a^* / \max_{a'} \alpha_{a'}^*$, so the hardest-to-steer attribute receives r=1 and easier attributes are scaled down. We calibrate empirically rather than deriving $r(a)$ directly from Cohen's d ratios because the relationship between feature-level effect size and classifier response is nonlinear and model-dependent. This normalizes the effect of α across attributes, allowing a single global budget without attribute-specific imbalances. Uniform+E sets w_{dyn}=1; Adaptive+E uses both components.

4.2 Analysis

To build intuition for why temporal concentration matters, we analyze an idealized allocation problem and show how its solution motivates our proxy schedule.

Trajectory-level cost decomposition. The steered and unsteered reverse processes are Markov chains with common initial distribution, so the chain rule gives $\text{KL}(\tilde{p}_\theta \| p_\theta) = \sum_t \mathbb{E}_{x_t}[\text{KL}(\tilde{p}_\theta(x_{t-1}|x_t) \| p_\theta(x_{t-1}|x_t))]$. Writing the per-step gain as $\Delta g_t = \alpha_t s_t + o(\alpha_t)$ ($s_t \geq 0$) with $\Delta q_t(\alpha_t) = \frac{1}{2} c_t \alpha_t^2 + o(\alpha_t^2)$, the allocation problem is then

$$\min_{\alpha_t \geq 0} \ \frac{1}{2} \sum_{t=0}^{T-1} c_t \alpha_t^2 \quad \text{s.t.} \quad \sum_{t=0}^{T-1} \alpha_t s_t \ \geq \ E_{\text{target}}. \tag{5}$$

Adaptive vs. uniform scheduling. The KKT conditions for equation 5 yield $\alpha_t^\star \propto s_t / c_t$ (§A.2). We compare against the uniform allocation $\alpha_t \equiv \alpha$.

The following result characterizes the idealized setting where per-step gains and costs are known exactly. While these quantities are not directly observable in practice (we use an SAE-derived proxy, as described below), the result yields a testable ordinal prediction.

Theorem 1 (Adaptive-vs-uniform efficiency ratio)**.** *Assume $s_t \geq 0$, $c_t > 0$, and $E_{\text{target}} > 0$. Among schedules with $\frac{1}{2}\sum_t c_t \alpha_t^2 = B$, let $E^\star$ and E_{unif} denote the maximum attribute shifts under the optimal and uniform schedules. Then*

$$\rho^2 := \left(\frac{E^\star}{E_{\text{unif}}}\right)^2 = \frac{\left(\sum_t s_t^2/c_t\right)\left(\sum_t c_t\right)}{\left(\sum_t s_t\right)^2} = 1 + \text{CV}_c^2(s/c), \tag{6}$$

where CV_c is the coefficient of variation of s_t/c_t under the cost-weighted measure $c_t / \sum_{t'} c_{t'}$, with equality $\rho = 1$ iff s_t/c_t is constant in t. (See §A.2 for assumptions and full proof.)

Interpretation. The adaptive gain is governed by cost-weighted dispersion of s_t/c_t: large when the attribute commits in a few steps, near one when emergence is spread evenly. Within each

model, attributes with more concentrated emergence profiles should benefit more from adaptive scheduling. We test this prediction via the schedule concentration statistic $\eta = N \sum_i w_i^2$ and confirm it empirically in §5.2 (full details in §A.5).

Proxy fidelity. Neither s_t nor c_t is directly observable. The proxy w_{dyn} approximates $\alpha_t^\star \propto s_t/c_t$ under two conditions (§A.2.6): (i) active-set growth tracks the marginal sensitivity s_t, and (ii) c_t varies slowly relative to s_t. These conditions are plausible but not formally verified; the proxy is justified primarily by the empirical quality gains reported in §5.

Cross-attribute coupling. For additive composition equation 3, we select top-20 features per attribute in §5.1. The resulting sets are near-disjoint but not fully so, since some attributes are inherently correlated (e.g., positive sentiment and formality). A formal analysis of the idealized fully-disjoint case, including a decoder-Gram bound on cross-attribute interference, is in §A.3; realized cosine coupling values and feature-overlap effects are in §A.4. Empirical interference is reported in §5.2.

5 Experiments

5.1 Experimental Setup

Models and attributes: We evaluate the four SAE steering modes from §4.1 (Uniform, Adaptive, Uniform+E, and Adaptive+E) on all four DLMs across 7 attribute combinations: 3 single-attribute, sentiment (S), topic (T), and style (St), and 4 multi-attribute (S+T, S+St, St+T, S+T+St). Sentiment is steered toward positive, topic toward sports, and style toward formality (MDLM/SEDD) or informality (LLaDA/DREAM). Each attribute is steered using the top 20 features per direction based on a feature count ablation (§E.4) that shows: lower-ranked features increasingly encode correlated attributes (e.g., sentiment features that also shift formality), increasing cross-attribute interference. MDLM and SEDD generate unconditionally over 1024 denoising steps; LLaDA and DREAM use 10 domain-neutral prompts (e.g., "I think this is") with temperature 0.8/0.7 over 64/256 steps (§C.9).

Baselines: We compare against three residual-stream baselines without SAE decomposition: *Contrastive Vectors* (mean class difference (Zou et al., 2023; Rimsky et al., 2024)), *Probe* (logistic regression weight (Belinkov, 2022)), and *PCA* (top principal component of contrastive activations). We additionally implement a *Step Function* baseline following Kynkäänniemi et al. (2024) that applies uniform steering within a commitment-derived interval (§C.10.3), isolating the contribution of adaptive scheduling from interval restriction alone. For LLaDA and DREAM, we additionally evaluate *Prompt*, prepending an attribute instruction to the prompt (§C.10). We note that other baselines such as ILRR (Avrahami & Nachmani, 2026) require a reference sequence (a different problem setup), and classifier-based guidance methods (Dhariwal & Nichol, 2021) require differentiable classifiers integrated into the denoising loop and are not directly applicable to masked discrete diffusion.

Metrics: Attribute control is measured by continuous classifier confidence (DistilBERT-SST2 (Sanh et al., 2019) for sentiment, BERT-AG News (Devlin et al., 2019) for topic, RoBERTa formality ranker (Babakov et al., 2023) for style; §C.11), which captures degree of attribute expression beyond binary thresholds. Quality is measured by Qwen2-7B perplexity (PPL) and LLM-judge fluency (LLMJ). We use PPL as the primary quality gate where reliable, following the established confidence-plus-PPL paradigm (Dathathri et al., 2019; Liu et al., 2021). For MDLM, SEDD, and DREAM, PPL and LLMJ correlate strongly ($|r_s| = 0.77$ to 0.88; §C.11). For LLaDA this correlation breaks down ($r_s = +0.21$) because aggressive steering produces repetitive text with artificially low PPL; we use LLMJ as the quality gate for this model. Human evaluation of fluency is prohibitively expensive at our scale (four models, seven conditions, multiple baselines); recent work has shown LLM judges to be a reliable proxy for human quality assessment in generation tasks (Goyal et al., 2025; Wu et al., 2025). We validate this for our setting via a human meta-evaluation across all four models covering fluency and all three attribute dimensions: the judge achieves 73% agreement with human consensus on a 5-point fluency scale, with disagreements concentrated in mid-range scores rather than the extremes that drive quality-gate decisions (full protocol and inter-annotator statistics in §C.11.2). For each method, we generate 200 samples per $\alpha \in \{1, \ldots, 20\}$ and select the operating point maximizing target confidence (single-attribute) or mean confidence (multi-attribute), subject to PPL < 100 and LLMJ fluency $\geq 50\%$ of baseline. MDLM/SEDD residual-stream baselines are exempt from the PPL gate, as they produce high PPL regardless of α.

Table 2: Steering performance comparison at best operating points (selection criteria in §5.1), averaged over single-attribute (sentiment (S), topic (T), style (St)) and multi-attribute (S+T, S+St, St+T, S+T+St) conditions. Conf = target attribute classifier confidence (%), PPL = Qwen2-7B perplexity, Flu = LLM-judge fluency (1–5 scale). Per-condition results with std in Tables 37–38.

		Single-attribute (mean of 3)								*Multi-attribute (mean of 4)*							
			SAE-based			Res. stream / prompt					SAE-based			Res. stream / prompt			
		Unsteered	Uniform	Step	Adaptive	Contr.V	Probe	PCA	Prompt	Unsteered	Uniform	Step	Adaptive	Contr.V	Probe	PCA	Prompt
MDLM	Conf↑	34.2	89.5	89.6	89.1	69.9	80.2	26.4	–	34.2	75.8	74.5	78.0	77.3	76.8	24.8	–
	PPL↓	46.0	86.3	88.5	59.7	309.3	462.6	300.0	–	46.0	85.5	82.6	70.7	353.5	441.4	281.1	–
	Flu↑	2.32	2.06	2.07	2.15	1.34	1.21	1.39	–	2.32	2.00	2.01	2.01	1.40	1.19	1.39	–
SEDD	Conf↑	33.4	97.7	96.5	89.1	97.1	92.5	41.0	–	33.3	89.7	88.8	88.6	92.2	84.6	45.3	–
	PPL↓	48.1	83.8	67.7	48.9	73.3	136.3	45.5	–	48.1	91.3	88.2	42.5	77.5	130.1	51.1	–
	Flu↑	2.23	2.00	2.10	2.17	2.12	1.86	2.25	–	2.23	1.99	1.99	2.11	2.02	1.83	2.25	–
LLaDA	Conf↑	35.1	95.0	91.6	84.8	52.8	44.9	41.7	63.6	35.1	84.1	81.2	76.5	48.1	44.0	39.6	57.0
	PPL↓	7.1	35.2	5.1	5.5	6.7	7.7	8.0	8.2	7.1	5.3	4.3	8.7	8.8	7.1	6.6	8.2
	Flu↑	3.56	2.56	2.64	2.97	3.41	3.55	3.54	3.15	3.56	2.20	2.30	2.75	3.48	3.60	3.57	3.19
DREAM	Conf↑	30.8	99.5	99.2	98.8	95.5	90.0	45.0	54.1	30.8	95.3	95.3	95.8	84.7	78.3	41.9	53.5
	PPL↓	8.6	35.8	26.7	20.8	12.2	13.2	26.9	7.7	8.6	63.6	58.6	64.2	11.7	15.7	21.7	8.8
	Flu↑	3.75	2.05	2.08	2.21	2.96	2.79	2.84	3.57	3.75	1.69	1.43	1.34	2.23	2.72	2.81	3.41

5.2 RESULTS

SAE steering vs. baselines: Table 2 reports mean steering performance averaged over conditions (per-condition results with selected α values and standard deviations in Tables 37–38; §E.1). Steering via latent-space intervention inherently incurs some quality cost, as observed in AR steering studies (Rimsky et al., 2024; Goyal et al., 2025). We observe the same in DLMs, but find that SAE-based methods generally degrade quality less than residual-stream baselines, with the exception of DREAM where residual-stream methods remain competitive. Residual-stream baselines exhibit model-dependent trade-offs: on MDLM, they achieve high confidence but at severe quality cost; on LLaDA, they frequently fail to steer beyond the unsteered rate for topic and style; on SEDD and DREAM, Contrastive Vectors performs competitively with SAE methods, while Probe steers effectively but with greater quality degradation, and PCA largely fails. Prompt steering (LLaDA, DREAM) preserves quality but achieves limited attribute control, particularly for style and multi-attribute combinations. SAE-based methods provide a more consistent profile across all four models by decomposing the residual stream into monosemantic features.

Adaptive balances quality and control: Among SAE-based methods, Adaptive scheduling achieves a better quality/control balance by concentrating perturbation on steps where attributes are actively forming. At best operating points (Tables 37–38), Adaptive achieves significantly better quality than the second-best SAE method in 22 of 28 conditions (Mann-Whitney U, $p < 0.05$), with 2 additional conditions showing non-significant improvement (§E.3.1); the 4 exceptions all involve DREAM, whose AR-based initialization compresses the temporal window (§E.2). Table 3 summarizes pairwise schedule comparisons across three matched-point analyses. Compared to Uniform: at matched confidence, Adaptive produces better quality in 82.2% of comparisons; at matched α, Adaptive achieves equal or better quality in 97.5% of comparisons (§E.3.2); and at matched quality, Adaptive achieves higher confidence in 46.7% of comparisons, confirming that Uniform's additional quality degradation does not translate into stronger steering. The full three-way ordering (Uniform $\leq$ Step Function $\leq$ Adaptive) holds in 88.2% of 560 (model, condition, α) triplets (§E.3.2). Figure 4 illustrates this on MDLM: the confidence/quality Pareto curves show that all schedules reach the same confidence plateau, but Adaptive achieves it at substantially lower perplexity, shifting the trade-off frontier toward higher quality. The same pattern holds across all four models.

Table 3: Schedule comparison: % of matched points where the first method is better (using model-specific quality metrics; see §E.3.2 and §E.3.4 for details).

	Matched conf. (quality $>$)	Matched α (quality $\geq$)	Matched qual. (conf. $>$)
Adaptive vs. Uniform	82.2%	97.5%	46.7%
Adaptive vs. Step	71.5%	95.0%	47.5%
Step vs. Uniform	66.5%	92.9%	44.6%

Table 4: α sensitivity: confidence drop (pp) when using a single global α^* across all 7 conditions vs. per-condition optimal α.

	Uniform		Step Fn.		Adaptive	
Model	mean	max	mean	max	mean	max
MDLM	9.1	50.9	5.6	32.2	0.6	3.2
SEDD	9.2	28.4	4.6	13.9	0.2	0.6
LLaDA	1.4	2.9	1.1	4.5	0.4	1.4
DREAM	3.1	9.1	3.2	10.8	1.2	2.1

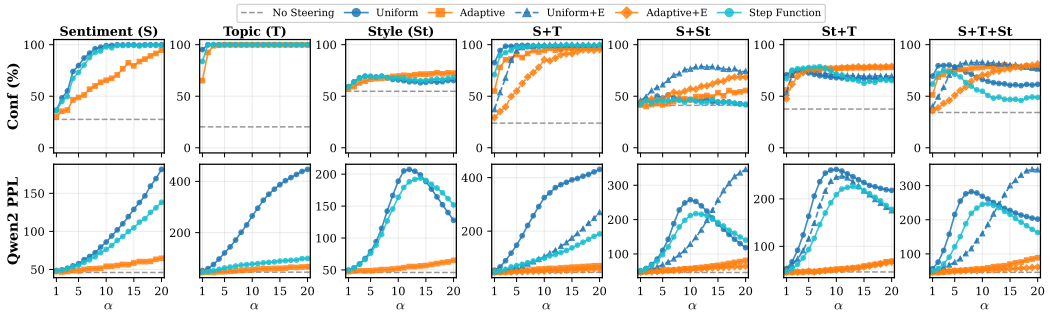


Figure 4: MDLM trade-off curves (SAE methods). Top: target confidence vs. α. Bottom: Qwen2-7B perplexity vs. α. Full curves with all methods and models are in §E.2.

α sensitivity: Although tuning a steering coefficient is standard practice (Rimsky et al., 2024; Goyal et al., 2025; Zou et al., 2023), Adaptive is robust to α selection. Fixing a single global α^* per (model, method) across all 7 conditions, Adaptive incurs only 0.2–1.2 pp mean confidence drop from the per-condition optimum, while Uniform suffers 1.4–9.2 pp because higher α values violate the quality gate, forcing a low global setting that under-steers most conditions (Table 4; details in §E.3.3).

E-ratio calibration: E-ratio re-balances per-attribute strength by effect size (§4). Table 5 reports the mean change from adding E-ratio calibration across multi-attribute conditions. Adaptive+E achieves better or comparable quality to Adaptive in 12/16 multi-attribute conditions, with the exception of SEDD, where distributed feature geometry (§3.2) increases cross-attribute overlap, reducing calibration reliability. Uniform+E shows more modest or inconsistent gains. Adaptive+E achieves significantly better quality than Uniform+E in all comparable multi-attribute conditions (§C.9.2).

Table 5: Effect of E-ratio scaling: mean Δ from adding E-ratio calibration, averaged over the 4 multi-attribute conditions. LLaDA PPL omitted as it is non-diagnostic. Flu = LLMJ fluency (1–5 scale).

	Uniform vs. Uniform+E			Adaptive vs. Adaptive+E		
Model	ΔConf	ΔPPL	ΔFlu	ΔConf	ΔPPL	ΔFlu
MDLM	+6.4	+2.5	+0.01	+3.0	−8.2	+0.05
SEDD	+1.1	+7.5	+0.01	−12.2	+10.6	+0.04
LLaDA	−6.7	–	+0.10	−10.2	–	+0.12
DREAM	+1.5	−5.0	+0.10	+0.2	−26.8	+0.70

Cross-attribute interference: We quantify interference as the mean absolute pp shift of non-target attributes from their unsteered baselines (Figure 14, §E.5). Some interference is unavoidable because attributes are naturally correlated (e.g., positive sentiment and formality). Adaptive consistently produces less interference than Uniform, with the largest gap on MDLM (up to 57pp for Uniform vs. 33pp for Adaptive), where targeting high-emergence steps avoids disrupting other attributes during their critical commitment.

Validating the efficiency prediction: Theorem 1 predicts that attributes with more concentrated emergence benefit more from adaptive scheduling. We measure this via the schedule concentration statistic $\eta = K \sum_i w_i^2$, where w_i are the normalized block weights ($\sum_i w_i = 1$, so $\eta = 1$ for a uniform schedule; §4.2). Across all 12 (model, attribute) conditions, using the appropriate quality metric for each model (§5.1), the within-model rank ordering of quality gains perfectly matches the η prediction for MDLM, SEDD, and LLaDA (Spearman $r_s = 1.0$ in each case; pooled exact permutation test across models, $p \approx 0.007$). DREAM shows $r_s = 0.5$ because its attributes share near-identical emergence profiles (η range 4.09–5.01), leaving little room for rank separation (§A.5).

6 Conclusion

We use SAEs to reveal how semantic features emerge during discrete diffusion denoising, discovering that attributes commit at different steps, with different sharpness, and at different magnitudes. These findings inform an adaptive schedule that concentrates intervention where each attribute is actively forming and improves the control–quality trade-off over uniform and interval-restricted schedules at matched control. Our analysis separates *when* concentration helps (headroom) from *how well* a schedule exploits it (alignment), turning interpretability into a practical lever for controlled generation.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript grammar correction and draft revision, and for experiment implementation. Generative AI tools were also used for literature review and all citations were manually collected and inserted to the relevant spans in the paper by the authors of this paper. All AI-assisted material included in the submission was reviewed and edited by the authors. The authors take responsibility for the final content of this work, including text, claims, analyses, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The paper and appendix document the details on experiment design and implementation, model and optimization settings, dataset evaluation splits, seeds, statistical procedures, and evaluation protocols used throughout this paper. Code will be made publicly available upon acceptance.

REFERENCES

Dana Arad, Aaron Mueller, and Yonatan Belinkov. Saes are good for steering–if you select the right features. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10252–10270, 2025.

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

Eden Avrahami and Eliya Nachmani. Ilrr: Inference-time steering method for masked diffusion language models. *arXiv preprint arXiv:2601.21647*, 2026.

Nikolay Babakov, David Dale, Ilya Gusev, Irina Krotova, and Alexander Panchenko. Don't lose the message while paraphrasing: A study on content preserving style transfer. In *International Conference on Applications of Natural Language to Information Systems*, pp. 47–61. Springer, 2023.

Sanjay Basu, Sadiq Y Patel, Parth Sheth, Bhairavi Muralidharan, Namrata Elamaran, Aakriti Kinra, John Morgan, and Rajaie Batniji. Interpretability without actionability: mechanistic methods cannot correct language model errors despite near-perfect internal representations. *arXiv preprint arXiv:2603.18353*, 2026.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R istituto superiore di scienze economiche e commericiali di firenze*, 8:3–62, 1936.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Seonglae Cho, Zekun Wu, and Adriano Koshiyama. Corrsteer: Generation-time llm steering via correlated sparse autoencoder features. *arXiv preprint arXiv:2508.12535*, 2025.

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization. *Proceedings of the National Academy of Sciences*, 100(5): 2197–2202, 2003.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, volume 2025, pp. 26721–26754, 2025.

Raghavv Goel, Risheek Garrepalli, Sudhanshu Agrawal, Chris Lott, Mingu Lee, and Fatih Porikli. Skip to the good part: Representation structure & inference-time layer skipping in diffusion vs. autoregressive llms. *arXiv preprint arXiv:2603.07475*, 2026.

Aaron Gokaslan and Vanya Cohen. Openwebtext corpus, 2019.

Agam Goyal, Vedant Rathi, William Yeh, Yian Wang, Yuen Chen, and Hari Sundaram. Breaking bad tokens: Detoxification of llms using sparse autoencoders. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12702–12720, 2025.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Jiaming Wu, Heng Gong, and Bing Qin. Improving controllable text generation with position-aware weighted decoding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 3449–3467, 2022a.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong, and Bing Qin. A distributional lens for multi-aspect controllable text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1023–1043, 2022b.

Arshia Hemmat, Philip Torr, Yongqiang Chen, and Junchi Yu. Tdgnet: Hallucination detection in diffusion language models via temporal dynamic graphs. *arXiv preprint arXiv:2602.08048*, 2026.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *Advances in Neural Information Processing Systems*, 37:122458–122483, 2024.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A Smith, and Yejin Choi. Dexperts: Decoding-time controlled text generation with experts and anti-experts. 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)*, pp. 6691–6706, 2021.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *International Conference on Machine Learning*, pp. 32819–32848. PMLR, 2024.

Jordan J Louviere, Terry N Flynn, and Anthony Alfred John Marley. *Best-worst scaling: Theory, methods and applications*. Cambridge University Press, 2015.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1015/.

Alireza Makhzani and Brendan Frey. K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.

Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pp. 50–60, 1947.

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

Narmeen Oozeer, Luke Marks, Fazl Barez, and Amirali Abdullah. Beyond linear steering: Unified multi-attribute control for language models. *Findings of the Association for Computational Linguistics: EMNLP*, pp. 23513–23557, 2025.

Ellie Pavlick and Joel Tetreault. An empirical analysis of formality in online communication. *Transactions of the association for computational linguistics*, 4:61–74, 2016.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Sudha Rao and Joel Tetreault. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: human language technologies, volume 1 (long papers)*, pp. 129–140, 2018.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Pau Rodriguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, Xavier Suau, et al. Controlling language and diffusion models by transporting activations. In *International Conference on Learning Representations*, volume 2025, pp. 89812–89855, 2025.

Pau Rodriguez, Michal Klein, Eleonora Gualdoni, Valentino Maiorca, Arno Blaas, Luca Zappella, Marco Cuturi, and Xavier Suau. Lineas: End-to-end learning of activation steering with a distributional loss. *Advances in Neural Information Processing Systems*, 38:82356–82398, 2026.

Raphael Ronge, Markus Maier, and Frederick Eberhardt. When the coffee feature activates on coffins: An analysis of feature extraction and steering for mechanistic interpretability. *arXiv preprint arXiv:2601.03047*, 2026.

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

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Or Shafran, Shaked Ronen, Omri Fahn, Shauli Ravfogel, Atticus Geiger, and Mor Geva. From directions to regions: Decomposing activations in language models via local geometry. *arXiv preprint arXiv:2602.02464*, 2026.

Adly Templeton, Tom Conerly, Jonathan Marcus, et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Xu Wang, Bingqing Jiang, Yu Wan, Baosong Yang, Lingpeng Kong, and Difan Zou. Dlm-scope: Mechanistic interpretability of diffusion language models via sparse autoencoders. *arXiv preprint arXiv:2602.05859*, 2026.

Barry Payne Welford. Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3):419–420, 1962.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Axbench: Steering llms? even simple baselines outperform sparse autoencoders. *arXiv preprint arXiv:2501.17148*, 2025.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

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

Jehyeok Yeon, Federico Cinus, Yifan Wu, and Luca Luceri. Gsae: Graph-regularized sparse autoencoders for robust llm safety steering. *arXiv preprint arXiv:2512.06655*, 2025.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

Xudong Zhu, Mohammad Mahdi Khalili, and Zhihui Zhu. Abstopk: Rethinking sparse autoencoders for bidirectional features. *arXiv preprint arXiv:2510.00404*, 2025.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Proofs and Extended Discussion for Steering Methods

We restate each result with a table explaining all notations used throughout the paper for convenience and provide full proofs.

A.1 Notation and Terms Explanation

Table 6: Notation used throughout the paper. Indices: t denotes a denoising step (of T total), ℓ a transformer layer, a an attribute, b a temporal block, j an SAE feature.

Symbol	Meaning
Model and denoising	
T	Total number of denoising steps
t	Denoising step index, $t \in \{1, \dots, T\}$ (or $0{:}T$)
ℓ	Transformer layer index
h_t^ℓ	Residual-stream activation at step t, layer ℓ
Sparse autoencoder	
Enc, Dec	SAE encoder / decoder
W_{dec}	SAE decoder matrix; column w_j^{dec} is feature j's direction
j	SAE feature index
F_a^ℓ	Attribute-specific feature set for attribute a at layer ℓ ($\lvert F_a^\ell \rvert = 20$ for steering; 50 for the interpretability analysis)
Attributes and emergence	
a	An attribute (sentiment, topic, or style)
b	Temporal block index, $b \in \{1, \dots, K\}$
K	Number of temporal blocks
$\bar{h}_j^{(t)}$	Mean activation of feature j at step t over the contrastive corpus
$f_b^{(a,\ell)}$	Commitment fraction: share of attribute a's total feature-activation growth occurring in block b
Steering	
$\delta_j^{(a)}$	Contrastive shift for feature j: $\bar{h}_{j,+}^{(a)} - \bar{h}_{j,-}^{(a)}$
α	Overall steering scale (swept during evaluation)
$r(a)$	Per-attribute effectiveness ratio, $r(a) \in (0, 1]$
$w_{\text{dyn}}(t; a, \ell)$	Adaptive temporal weight $\in [0, 1]$ from the commitment profile
$\alpha_{\text{eff}}(a, t, \ell)$	Effective strength $= \alpha \cdot r(a) \cdot w_{\text{dyn}}(t; a, \ell)$
Theory	
s_t	Realized per-step attribute gain (marginal sensitivity at step t)
c_t	Realized per-step KL cost of the intervention at step t
ρ^2	Adaptive-vs-uniform efficiency ratio, $= 1 + \text{CV}_c^2(s/c)$
CV_c	Coefficient of variation of s_t/c_t under the cost-weighted measure

A.2 Proof of Theorem 1 (Adaptive-vs-uniform efficiency ratio)

A.2.1 Trajectory-level KL: exact additive decomposition

Theorem 1 optimizes against the budget $\frac{1}{2}\sum_t c_t \alpha_t^2 \le B$, where each c_t is the second-order cost of perturbation at step t. This budget arises as a tractable additive decomposition of the trajectory-level KL between steered and unsteered reverse processes; we derive the decomposition here.

Let $p_\theta, \tilde{p}_\theta$ denote the joint distributions of the unsteered and steered reverse trajectories $(x_T, x_{T-1}, \dots, x_0)$. Both are Markov: the reverse transition at step t is $p_\theta(x_{t-1} \mid x_t)$ for the unsteered model and $\tilde{p}_\theta(x_{t-1} \mid x_t)$ for the steered model (which differs only in the residual stream perturbation at step t). Both processes share the same initial distribution at x_T (e.g., the fully-masked prior in absorbing-state diffusion), so $\text{KL}(\tilde{p}_\theta(x_T) \,\|\, p_\theta(x_T)) = 0$. The chain rule for KL on Markov chains applied to the reverse joint, then it gives

$$\text{KL}\big(\tilde{p}_\theta \,\big\|\, p_\theta\big) \;=\; \sum_{t=1}^{T} \mathbb{E}_{x_t \sim \tilde{p}_\theta}\Big[\text{KL}\big(\tilde{p}_\theta(x_{t-1} \mid x_t) \,\big\|\, p_\theta(x_{t-1} \mid x_t)\big)\Big]. \tag{7}$$

First-order change-of-measure approximation. The outer expectation in equation 7 is under the steered measure $\tilde{p}_\theta$, while x_t-marginal samples are easier to obtain under p_θ. Replacing $\tilde{p}_\theta$ by p_θ in the outer expectation incurs an error that is sub-leading in α: the Radon–Nikodym derivative $d\tilde{p}_\theta/dp_\theta$ equals 1 at $\alpha = 0$ and varies smoothly, so the difference $\mathbb{E}_{\tilde{p}_\theta}[\Phi(x_t)] - \mathbb{E}_{p_\theta}[\Phi(x_t)]$ is $O(\alpha)$ for any bounded Φ. Since the integrand $\Phi_t(x_t) := \mathrm{KL}(\tilde{p}_\theta(x_{t-1} \mid x_t) \,\|\, p_\theta(x_{t-1} \mid x_t))$ is itself $O(\alpha^2)$ in leading order (Lemma 1 below), the change-of-measure correction contributes $O(\alpha)\cdot O(\alpha^2) = O(\alpha^3)$, which is sub-leading relative to the $O(\alpha^2)$ leading cost. The replacement therefore preserves the leading-order behavior of the total cost. We use this approximation throughout.

Per-step quadratic expansion. For each summand, a second-order Taylor expansion of the KL around $\alpha_t = 0$ gives

$$\Delta q_t(\alpha_t) \;=\; \mathbb{E}_{x_t \sim p_\theta}\Big[\mathrm{KL}\big(\tilde{p}_\theta(x_{t-1} \mid x_t) \,\big\|\, p_\theta(x_{t-1} \mid x_t)\big)\Big] \;=\; \tfrac{1}{2} c_t \alpha_t^2 + o(\alpha_t^2),$$

where the linear term vanishes because $\alpha_t = 0$ globally minimizes each KL summand (KL is non-negative and zero at $\tilde{p}_\theta = p_\theta$, so the gradient at zero perturbation vanishes). The coefficient c_t is given by the Jacobian-pullback expression of Lemma 1. Summing across t and combining with the first-order attribute gain $\Delta g_t(\alpha_t) = \alpha_t s_t + o(\alpha_t)$ gives the budgeted allocation problem stated in the main text:

$$\min_{\alpha_t \geq 0} \; \tfrac{1}{2} \sum_{t=0}^{T-1} c_t \alpha_t^2 \quad \text{s.t.} \quad \sum_{t=0}^{T-1} \alpha_t s_t \;\geq\; E_{\text{target}}. \tag{8}$$

A.2.2 Proof of the efficiency ratio theorem

Assumption 1 (Finite-horizon Markov reverse process). The steered and unsteered reverse processes, denoted by $\tilde{p}_\theta$ and p_θ, respectively, are finite-horizon Markov chains with horizon $T < \infty$, sharing the same fully noised initial distribution at x_T.

Assumption 2 (Smoothness of the attribute score). The attribute score function g is twice continuously differentiable with respect to the steering strength at $\alpha_t = 0$. Specifically, g is instantiated as the softmax probability of a classifier evaluated on the realized output sequence, which is smooth with respect to the residual-stream activation at the intervention layer.

Assumption 3 (Local validity of Taylor expansions). There exists a neighborhood around $\alpha_t = 0$ in which

$$\Delta g_t(\alpha_t) = \alpha_t s_t + o(\alpha_t), \qquad \Delta q_t(\alpha_t) = \frac{1}{2} c_t \alpha_t^2 + o(\alpha_t^2),$$

and the operating regime satisfies $\alpha_t = O(\lambda)$ such that higher-order terms are negligible.

Assumption 4 (Nondegeneracy and feasibility). For all t, the coefficients satisfy $s_t \geq 0$ and $c_t > 0$, with $\sum_t s_t > 0$ and $E_{\text{target}} > 0$. These conditions ensure feasibility of the constrained optimization problem in Eq. equation 5.

Assumption A1 justifies the additive cost decomposition (App. §A.2.1). Assumptions A2 and A3 ensure that c_t is well defined as the second derivative of the output KL . Finally, Assumption A4 guarantees that the associated KKT system admits a unique nontrivial solution.

Theorem 1 (restated). *Adaptive-vs-uniform efficiency ratio.* Consider equation 8 with $s_t \geq 0$, $c_t > 0$, and $E_{\text{target}} > 0$. Fix a quality budget $B = \frac{1}{2}\sum_t c_t \alpha_t^2$ and let $E^\star$, E_{unif} denote the maximum attribute shifts achievable at this budget under the optimal and uniform schedules respectively. Then

$$\rho^2 \;:=\; \left(\frac{E^\star}{E_{\text{unif}}}\right)^2 \;=\; \frac{\big(\sum_t s_t^2/c_t\big)\big(\sum_t c_t\big)}{\big(\sum_t s_t\big)^2} \;=\; 1 \;+\; \mathrm{CV}_c^2(s/c), \tag{9}$$

where CV_c denotes the coefficient of variation of s_t/c_t under the cost-weighted measure $c_t/\sum_{t'} c_{t'}$. The optimum is achieved by $\alpha_t^\star = \lambda s_t/c_t$ with $\lambda = E_{\text{target}}/\sum_t s_t^2/c_t$. Equality $\rho = 1$ holds iff s_t/c_t is constant in t.

Proof. The proof has three parts: (i) derive the optimum location via KKT, (ii) compute E_{unif} for the constant-α schedule, and (iii) show the resulting ratio equals $1 + \text{CV}_c^2(s/c)$.

Part 1: optimum location. The Lagrangian for equation 8 is

$$\mathcal{L} \;=\; \tfrac{1}{2}\sum_t c_t\alpha_t^2 \;-\; \mu\left(\sum_t \alpha_t s_t - E_{\text{target}}\right) \;-\; \sum_t \nu_t\alpha_t,$$

with $\mu \geq 0$ and $\nu_t \geq 0$. Stationarity gives $c_t\alpha_t = \mu s_t + \nu_t$, i.e., $\alpha_t = (\mu s_t + \nu_t)/c_t$.

For t with $s_t > 0$: if $\alpha_t = 0$ then $\nu_t = -\mu s_t \leq 0$, which combined with $\nu_t \geq 0$ forces $\mu s_t = 0$. Since $s_t > 0$ this requires $\mu = 0$, but then all $\alpha_t = 0$, violating $\sum_t \alpha_t s_t \geq E_{\text{target}} > 0$. Therefore $\alpha_t > 0$ on $\{t : s_t > 0\}$, complementary slackness gives $\nu_t = 0$, and $\alpha_t^\star = \mu s_t/c_t$. For t with $s_t = 0$, $\alpha_t^\star = 0$ is consistent with all KKT conditions and contributes nothing to either objective or constraint.

The control constraint binds at the optimum (else scaling all α_t down by $1 - \epsilon$ would strictly decrease the objective while remaining feasible), so

$$\sum_t \mu\frac{s_t^2}{c_t} \;=\; E_{\text{target}} \quad\Longrightarrow\quad \mu^\star \;=\; \frac{E_{\text{target}}}{\sum_t s_t^2/c_t}.$$

Setting $\lambda := \mu^\star$ recovers the stated $\alpha_t^\star = \lambda s_t/c_t$. The optimal cost is

$$B^\star \;=\; \tfrac{1}{2}\sum_t c_t(\alpha_t^\star)^2 \;=\; \tfrac{1}{2}\lambda^2\sum_t \frac{s_t^2}{c_t} \;=\; \frac{E_{\text{target}}^2}{2\sum_t s_t^2/c_t},$$

and the maximum shift at any budget B is $E^\star = \sqrt{2B \cdot \sum_t s_t^2/c_t}$. Strict convexity of the objective (since $c_t > 0$) gives uniqueness.

Part 2: uniform schedule shift. Fix the uniform schedule $\alpha_t = \alpha$ for all t. The budget constraint becomes

$$\tfrac{1}{2}\sum_t c_t\alpha^2 \;=\; B \quad\Longrightarrow\quad \alpha \;=\; \sqrt{\frac{2B}{\sum_t c_t}}.$$

The corresponding attribute shift is

$$E_{\text{unif}} \;=\; \sum_t \alpha s_t \;=\; \alpha\sum_t s_t \;=\; \sqrt{\frac{2B}{\sum_t c_t}}\cdot\sum_t s_t.$$

Part 3: Adaptive-vs-uniform efficiency ratio and CV identity. Taking the ratio of $E^\star$ and E_{unif} at matched budget B,

$$\rho^2 \;=\; \frac{(E^\star)^2}{(E_{\text{unif}})^2} \;=\; \frac{2B\cdot\sum_t s_t^2/c_t}{(2B/\sum_t c_t)\cdot(\sum_t s_t)^2} \;=\; \frac{(\sum_t s_t^2/c_t)(\sum_t c_t)}{(\sum_t s_t)^2}.$$

This proves the first equality in equation 9.

For the CV form, define the cost-weighted probability measure $\pi_t = c_t/C$ where $C = \sum_{t'} c_{t'}$, and write $X_t = s_t/c_t$. Under π,

$$\mathbb{E}_\pi[X] \;=\; \sum_t \pi_t X_t \;=\; \frac{1}{C}\sum_t s_t, \qquad \mathbb{E}_\pi[X^2] \;=\; \sum_t \pi_t X_t^2 \;=\; \frac{1}{C}\sum_t \frac{s_t^2}{c_t}.$$

Substituting these into the ratio derived above,

$$\rho^2 \;=\; \frac{(\sum_t s_t^2/c_t)(\sum_t c_t)}{(\sum_t s_t)^2} \;=\; \frac{C\cdot\mathbb{E}_\pi[X^2]\cdot C}{(C\cdot\mathbb{E}_\pi[X])^2} \;=\; \frac{\mathbb{E}_\pi[X^2]}{(\mathbb{E}_\pi[X])^2}.$$

By the elementary identity $\mathbb{E}[X^2]/\mathbb{E}[X]^2 = 1 + \text{Var}(X)/\mathbb{E}[X]^2 = 1 + \text{CV}^2(X)$,

$$\rho^2 \;=\; 1 + \text{CV}_c^2(s/c).$$

Equality condition. $\rho = 1$ iff $\mathrm{CV}_c(s/c) = 0$ iff s_t/c_t is constant on $\{t : c_t > 0\}$. Equivalently, by Cauchy–Schwarz applied to $\sqrt{c_t}$ and $s_t/\sqrt{c_t}$,

$$\Big(\sum_t s_t\Big)^2 \;=\; \Big(\sum_t \sqrt{c_t}\cdot\frac{s_t}{\sqrt{c_t}}\Big)^2 \;\le\; \Big(\sum_t c_t\Big)\Big(\sum_t \frac{s_t^2}{c_t}\Big),$$

with equality iff $\sqrt{c_t} \propto s_t/\sqrt{c_t}$, i.e., $s_t \propto c_t$. This recovers the same condition. □

Corollary 1 (Specialization under slowly-varying cost)**.** *When $c_t \approx c$ for all t, the efficiency ratio simplifies to*

$$\rho^2 \;\approx\; \frac{T\sum_t s_t^2}{(\sum_t s_t)^2} \;=\; 1 + \mathrm{CV}^2(s),$$

the squared coefficient of variation of s_t under the uniform measure. The optimum is $\alpha_t^\star \propto s_t$ and the achievable shift at fixed budget is governed by the dispersion of s_t alone. This is the regime in which the proxy of Appendix A.2.6 approximates the optimum most directly: when c_t is approximately constant, the proxy's tightness reduces to the alignment between activation growth and s_t.

A.2.3 Suboptimality of any feasible schedule

Theorem 1 compares optimal to uniform; the same Cauchy–Schwarz machinery characterizes the suboptimality of any feasible schedule, including the proxy used in our experiments.

Corollary 2 (Schedule efficiency)**.** *For any schedule $\{\alpha_t \ge 0\}$ that exhausts the budget ($\frac{1}{2}\sum_t c_t\alpha_t^2 = B$), the achieved attribute shift satisfies*

$$\frac{E}{E^\star} \;=\; \cos\theta, \qquad \theta \;=\; \angle\Big(\big(\alpha_t\sqrt{c_t}\big)_t,\; \big(s_t/\sqrt{c_t}\big)_t\Big),$$

with equality $E = E^\star$ iff $\alpha_t \propto s_t/c_t$. Equivalently, $B/B^\star = 1/\cos^2\theta$ at matched control.

Proof. Set $u_t = \alpha_t\sqrt{c_t}$ and $v_t = s_t/\sqrt{c_t}$. The budget constraint is $\|u\|^2 = 2B$ and the achieved shift is $E = \langle u, v\rangle$. From Part 1 of the proof of Theorem 1, $\|v\|^2 = \sum_t s_t^2/c_t$, and the maximum shift at budget B is $E^\star = \sqrt{2B\cdot\sum_t s_t^2/c_t} = \|u\|\cdot\|v\|$. By Cauchy–Schwarz,

$$E \;=\; \langle u, v\rangle \;\le\; \|u\|\,\|v\| \;=\; E^\star, \qquad \frac{E}{E^\star} \;=\; \frac{\langle u, v\rangle}{\|u\|\,\|v\|} \;=\; \cos\theta,$$

with equality iff $u \propto v$, i.e., $\alpha_t \propto s_t/c_t$.

For the cost-ratio form: at matched control level $\bar{E}$, suppose the heuristic schedule achieves $E = \bar{E}$ at cost B, while the optimal schedule achieves $\bar{E}$ at cost $B^\star$. From the $E^\star$ formula, $E^\star(B) = \sqrt{2B}\,\|v\|$ and $E^\star(B^\star) = \sqrt{2B^\star}\,\|v\| = \bar{E}$, so $B^\star = \bar{E}^2/(2\|v\|^2)$. From the cosine identity, $\bar{E} = E^\star(B)\cos\theta = \sqrt{2B}\,\|v\|\cos\theta$, so $B = \bar{E}^2/(2\|v\|^2\cos^2\theta)$. Dividing,

$$\frac{B}{B^\star} \;=\; \frac{1}{\cos^2\theta}.$$

□

Corollary 3 (Cost of using a heuristic schedule shape)**.** *For any schedule $\alpha_t = \kappa w_t$ with $w_t \ge 0$ and κ chosen to exhaust the budget,*

$$\frac{E}{E^\star} \;=\; \frac{\sum_t w_t s_t}{\sqrt{\sum_t w_t^2 c_t}\cdot\sqrt{\sum_t s_t^2/c_t}}.$$

The scalar κ cancels: efficiency depends only on the shape *of w_t, not its magnitude.*

Proof. The budget constraint $\frac{1}{2}\sum_t c_t(\kappa w_t)^2 = B$ gives $\kappa = \sqrt{2B/\sum_t w_t^2 c_t}$. Substituting into Corollary 2, the angle factor in the numerator becomes $\langle \kappa w\sqrt{c}, s/\sqrt{c}\rangle = \kappa\sum_t w_t s_t$, and $\|\kappa w\sqrt{c}\| = \kappa\sqrt{\sum_t w_t^2 c_t} = \sqrt{2B}$. The ratio $E/E^\star = \cos\theta$ from Corollary 2 expands as stated, and κ cancels between numerator and denominator. □

This corollary licenses the practice in the experimental section of fixing the shape via w_{dyn} and sweeping α separately: only the shape determines the proxy's distance from optimal in the Fisher-weighted geometry.

A.2.4 OPERATIONAL FORM OF c_t

The main text defines c_t as the second derivative of the per-step output-distribution KL at $\alpha_t = 0$. We give an explicit chain-rule expression and verify positivity.

Let $u_t \in \mathbb{R}^d$ denote the steering direction in residual-stream space at step t: under composition rule equation 1, $u_t = W_{F^\ell_a}\delta^{(a)}$. Let $f_t : \mathbb{R}^d \to \mathbb{R}^V$ be the map from the residual stream at the intervention layer to the output logits at step t, with Jacobian $J_t = \partial f_t / \partial x_t$. Let $F_t^{\text{out}} \in \mathbb{R}^{V\times V}$ denote the Fisher information of the output categorical distribution with respect to its logits, evaluated at the unsteered logits $f_t(x_t)$.

Lemma 1. *$c_t = u_t^\top J_t^\top F_t^{\text{out}} J_t u_t$, with $c_t \geq 0$. Strict positivity holds whenever $J_t u_t$ is not in the null space of F_t^{out}.*

Proof. Writing the perturbed logits as $\tilde\theta(\alpha_t) = f_t(x_t + \alpha_t u_t)$ and Taylor-expanding the per-step KL,

$$\mathrm{KL}\big(p_{f_t(x_t)} \,\big\|\, p_{\tilde\theta(\alpha_t)}\big) \;=\; \tfrac{1}{2}\alpha_t^2 \cdot (J_t u_t)^\top F_t^{\text{out}} (J_t u_t) + O(\alpha_t^3).$$

The linear term vanishes because $\alpha_t = 0$ globally minimizes the KL (non-negative everywhere, zero at the origin), forcing the gradient to vanish there. Matching against $\Delta q_t(\alpha_t) = \frac{1}{2} c_t \alpha_t^2 + o(\alpha_t^2)$ gives the stated form. Positivity follows from $F_t^{\text{out}} \succeq 0$, with strict inequality when $J_t u_t$ has a non-uniform component (the null space of F_t^{out} for a categorical distribution is the all-ones direction in logit space, corresponding to logit shifts that leave the softmax invariant). □

Remark. c_t is the pullback of the output Fisher through the layers above the intervention site. It is not the Fisher information of a parametric family in the standard sense; calling it "Fisher information" without specifying the family would be a misuse of the term. The operational characterization above is what enters the optimization; the chain-rule expression is included for completeness and is not used directly in our schedule. The global hypothesis $c_t > 0$ in Theorem 1 holds operationally because $J_t u_t$ is generically non-uniform (the steering direction $u_t = W_{F^\ell_a}\delta^{(a)}$ has support across many output dimensions, and the layers above the intervention site mix coordinates), placing $J_t u_t$ outside the all-ones null space of F_t^{out}.

A.2.5 POSITION-WISE DECOMPOSITION OF THE LOCAL GAIN

The local gain $s_t = \partial g / \partial \alpha_t$ admits a position-wise decomposition into direct and indirect pathways. The direct pathway provably localizes to masked positions; the indirect pathway is non-zero in general but is sub-leading in the regime supported by our measurements.

Setup. Fix a denoising trajectory and a step t. The steering operation applies a residual-stream perturbation $\alpha_t u_t$ at step t, where $u_t = W_{F^\ell_a}\delta^{(a)} \in \mathbb{R}^d$ is the same direction at every sequence position (the SAE injection is position-wise but uses the same $\delta^{(a)}$ across positions). Let $\mathcal{M}_t \subseteq \{1, \ldots, L\}$ denote the set of masked positions at step t; we condition on the masking schedule (which positions are unmasked when), so that the only stochasticity is in token content. For unmasked positions $p \notin \mathcal{M}_t$, the realized token token_p was determined at the step at which p was unmasked, prior to step t, and is fixed conditional on the trajectory up to that step.

Lemma 2 (Direct-effect localization)**.** *Under the setup above, the local gain decomposes as*

$$s_t \;=\; \underbrace{\sum_q \frac{\partial g}{\partial\, \mathrm{token}_q} \cdot \frac{\partial\, \mathrm{token}_q}{\partial x_{t,q}} \cdot \frac{\partial x_{t,q}}{\partial \alpha_t}}_{\textit{direct effect } s_t^{\text{dir}}} + \underbrace{\sum_q \frac{\partial g}{\partial\, \mathrm{token}_q} \sum_{p\neq q} \frac{\partial\, \mathrm{token}_q}{\partial x_{t,p}} \cdot \frac{\partial x_{t,p}}{\partial \alpha_t}}_{\textit{indirect effect } s_t^{\text{ind}}},$$

where the direct sum collects the $p = q$ diagonal of the position-by-position chain rule and the indirect sum collects the off-diagonal terms. The direct effect reduces to a sum over masked positions:

$$s_t^{\text{dir}} \;=\; \sum_{q\in\mathcal{M}_t} \frac{\partial g}{\partial\, \mathrm{token}_q} \cdot \frac{\partial\, \mathrm{token}_q}{\partial x_{t,q}} \cdot \frac{\partial x_{t,q}}{\partial \alpha_t}.$$

Proof. For $q \notin \mathcal{M}_t$, the realized token_q was determined at the step at which q was unmasked (prior to step t) and does not depend on $x_{t,q}$. By the chain rule, $\partial \mathrm{token}_q / \partial x_{t,q} = 0$, so the diagonal $p = q$ term vanishes. The direct-effect sum therefore restricts to $q \in \mathcal{M}_t$. □

Indirect pathway. The lemma controls only the direct effect. The indirect effect at unmasked position p is non-zero in general: although token_p is fixed, the residual-stream activation $x_{t,p}$ enters the key/value projections used by self-attention at masked positions $q \in \mathcal{M}_{t'}$ for $t' \geq t$, and so can influence token_q at those positions. Whether the indirect path is sub-leading relative to the direct path is a regime-dependent question that we cannot settle in closed form.

We treat direct-path dominance as a modeling assumption supported by the masked-fraction analysis of Appendix D.4: across our four models, attribute-relevant features fire preferentially on masked positions throughout the early-to-middle trajectory, with masked-fraction values ranging from 0.55 to 0.80 for anticipatory attributes. This is consistent with — though does not prove — direct-path dominance, since features that fire on masked positions are positioned to exert direct influence on the tokens decoded there. The within-model agreement between schedule concentration η and observed quality gains (5) is consistent with this assumption being operationally adequate, though it does not test the assumption directly.

A.2.6 Connecting the optimal schedule to the active-set proxy

Theorem 1 prescribes $\alpha_t^\star \propto s_t/c_t$, but neither s_t nor c_t is directly observable: s_t requires gradient information about a target attribute classifier on the unrolled denoising trajectory, and c_t requires the Jacobian-pullback of Lemma 1. Our schedule w_{dyn} uses positive activation growth as a proxy. We justify the proxy through representational commitment under the TopK gate and characterize when it is tight.

Active-set transitions as a measure of representational commitment. The TopK gate produces, at each step t and each input, a sparse active set $\mathcal{A}_t \subseteq \{1, \ldots, d_{\mathrm{SAE}}\}$ of features whose pre-gate scores exceed the top-k threshold. The steering operation itself does *not* depend on $\mathcal{A}_t$: the residual-stream perturbation $\alpha\, W_{F_a^\ell}\delta^{(a)}$ is added directly to the post-encode hidden state regardless of which features are currently active. The proxy is therefore not a statement about gradient localization through the gate, but about the model's *use* of attribute-relevant directions. When attribute features $j \in F_a^\ell$ enter $\mathcal{A}_t$ on average across the contrastive corpus, the model is, on average, beginning to compute on those directions; subsequent layers read out from the residual stream, and a perturbation along directions the model is starting to use propagates through more of the downstream computation than a perturbation along directions still effectively unused.

This rationale is intuitive rather than rigorous: we cannot prove that residual-space perturbations along high-growth directions produce larger $\partial g/\partial \alpha_t$ than perturbations along low-growth directions, only that this is plausible given how the model's forward pass distributes information. The framework's predictive accuracy is consistent with results from Section 5.

Why activation growth, not activation level. A natural alternative would be to weight the schedule by raw activation level $A_t^{(a,\ell)} = \sum_{j \in F_a^\ell} \bar{h}_j^{(t)}$. This over-weights steps where attribute features are already stably active across the corpus: the model has already committed to those features, and perturbing further along directions that are already saturated may have diminishing marginal effect. Activation growth $[\Delta A_t^{(a,\ell)}]_+$ targets steps where features are crossing into the active set on the corpus, where the model's representational state for that attribute is most malleable. The asymmetric clamping at zero is necessary: steps with negative growth correspond to features dropping out of the active set, which we do not want to amplify steering on.

Proxy fidelity and matched-shape efficiency. By Corollary 3, the efficiency of the proxy schedule relative to optimal is

$$\frac{E_{\mathrm{proxy}}}{E^\star} \;=\; \frac{\sum_t w_{\mathrm{dyn}}(t) s_t}{\sqrt{\sum_t w_{\mathrm{dyn}}(t)^2 c_t}\cdot\sqrt{\sum_t s_t^2/c_t}} \;=\; \cos\theta_{\mathrm{proxy}},$$

where θ_{proxy} is the angle between $w_{\mathrm{dyn}}(t)\sqrt{c_t}$ and $s_t/\sqrt{c_t}$. The proxy is exactly optimal when (i) active-set growth is proportional to s_t/c_t and (ii) c_t varies slowly relative to s_t, so the geometry collapses to the constant-cost case of Corollary 1.

When the proxy is loose, and what we observe. Two regimes follow from Theorem 1 and Corollary 3:

- *Sharp regime ($\mathrm{CV}_c(s/c)$ large).* The optimal schedule concentrates budget on a few steps, and any schedule whose shape matches that concentration recovers most of $E^\star$. On MDLM, where block 0 captures essentially all of topic emergence, the activation-growth profile is itself sharp, and Adaptive substantially outperforms Uniform at matched control — consistent with the proxy aligning with the sensitivity profile.
- *Flatter regime ($\mathrm{CV}_c(s/c)$ small).* By Theorem 1, $\rho \to 1$ as emergence flattens, and the potential gain from any non-uniform schedule shrinks. Conditions with lower η (e.g., MDLM sentiment, $\eta = 1.44$) are consistent with the prediction (§A.5). No condition in our experiments reaches the theoretical $\eta \approx 1$ limit where Adaptive and Uniform would be indistinguishable.

The proxy can therefore be loose without being harmful: in the flat regime, any schedule does about equally well; in the sharp regime, the proxy tracks the optimal shape closely enough to recover most of the available efficiency ratio. The regime where the proxy would actively underperform Uniform is one where activation growth and sensitivity are anti-correlated. We do not have direct evidence that this regime is avoided in our four models: our argument that Adaptive does not underperform Uniform is empirical (Table 37 and Table38) rather than structural.

A.3 Decoder-space interference: bound and analysis

We analyze cross-attribute interference under the additive composition equation 3. The bound below assumes disjoint attribute feature sets; §A.4 documents that this assumption holds in only a minority of deployed configurations (20/42 pairs at $|F_a^\ell| = 20$). Where it does not hold, we rely on empirical coupling measurements (also in §A.4), which are roughly an order of magnitude tighter than the bound. We therefore present the bound as a lemma, with the empirical observation as the operative claim, and analyze the gap between them.

Assumption 1 (Disjoint feature sets). For all $i \neq j$, the feature sets satisfy

$$F_{a_i}^\ell \cap F_{a_j}^\ell = \emptyset.$$

The small-overlap correction for the empirically near-disjoint regime (deployed top-20 sets share 0–5 features, mean 1.10; see §A.4) is discussed in App. §A.3.4.

Assumption 2 (Positive-definite intra-attribute Gram matrices). For each attribute a_i, the intra-attribute Gram matrix

$$G^{(ii)} = W_{F_{a_i}^\ell}^\top W_{F_{a_i}^\ell}$$

is positive definite on its support. Equivalently, the decoder columns indexed by $F_{a_i}^\ell$ are linearly independent. This condition generically holds when $|F_{a_i}^\ell| \ll d$ and decoder columns are unit normalized.

Assumption 3 (Nonzero perturbations). The perturbations satisfy

$$\Delta x_t^{(i)} \neq 0, \qquad \Delta x_t^{(j)} \neq 0,$$

ensuring that the cosine term in the bound (Lemma 4) is well defined. This condition holds whenever $\alpha_{\text{eff}}(a_i, t, \ell) > 0$, $\delta^{(a_i)}$ has a nonzero component, and the ReLU clamp does not zero the entire perturbation over $F_{a_i}^\ell$.

Assumption 1 provides the structural precondition of the proposition. Assumption 2 guarantees that the denominator of the bound is strictly positive. Finally, Assumption 3 ensures that the cosine similarity is well defined.

A.3.1 Subspace containment and the role of ReLU

The bound is stated for post-ReLU residual perturbations. We first establish that disjointness places each attribute's perturbation in its own decoder subspace and that ReLU preserves this structure.

Lemma 3 (Subspace containment under disjointness)**.** *Let $\Delta x_t^{(i)} = \tilde{x}_t - \tilde{x}_t^{[\setminus i]}$ denote the residual-space contribution of attribute i under composition rule equation 3, where $\tilde{x}_t^{[\setminus i]}$ is the steered reconstruction with attribute i omitted. If $F_{a_i}^\ell \cap F_{a_j}^\ell = \emptyset$ for all $j \neq i$, then $\Delta x_t^{(i)} \in \mathrm{span}(W_{F_{a_i}^\ell})$, and the ReLU clamp in equation 3 alters the magnitudes within $\Delta h_t^{(i)} := \tilde{h}_t - \tilde{h}_t^{[\setminus i]}$ but preserves its support on $F_{a_i}^\ell$.*

Proof. Write the pre-ReLU composed feature update as $\hat{h}_{t,j} = h_{t,j} + \sum_{i:\, j \in F_{a_i}^\ell} \alpha_{\text{eff}}(a_i, t, \ell)\delta_j^{(a_i)}$. Under disjointness, the inner sum has at most one term: for any j, either $j \in F_{a_i}^\ell$ for exactly one i (call it $i(j)$), or j lies in no attribute set. Consequently $\hat{h}_{t,j} - h_{t,j}$ depends on attribute $i(j)$ alone.

The ReLU clamp $\tilde{h}_{t,j} = \mathrm{ReLU}(\hat{h}_{t,j})$ acts coordinate-wise; its clipping decision at j depends only on $\hat{h}_{t,j}$, which depends on only one attribute. Therefore $\tilde{h}_{t,j} - \tilde{h}_{t,j}^{[\setminus i]}$ is non-zero only for $j \in F_{a_i}^\ell$, regardless of whether ReLU clipped the value. Applying W_{dec},

$$\Delta x_t^{(i)} \;=\; W_{\text{dec}} \Delta h_t^{(i)} \;=\; W_{F_{a_i}^\ell}\big(\Delta h_t^{(i)}\big)\big|_{F_{a_i}^\ell} \;\in\; \mathrm{span}\big(W_{F_{a_i}^\ell}\big).$$

ReLU may shrink magnitudes on $F_{a_i}^\ell$ but cannot move mass off the support. □

What disjointness buys. A weaker version of subspace containment — that disjoint feature *indices* produce non-overlapping coordinate updates — follows trivially from the structure of equation 3. Lemma 3 is stronger: it controls the post-ReLU residual perturbation, not just the feature update, by exploiting that ReLU is coordinate-wise and that disjointness localizes its decision per attribute. This is what makes the Gram bound apply to the actual perturbations rather than only to a first-order linearization.

A.3.2 Decoder-Gram bound

Lemma 4 (Decoder-Gram interference bound)**.** *Let $G = W_{\text{dec}}^\top W_{\text{dec}}$ and let $G^{(ij)}$ denote its submatrix indexed by $F_{a_i}^\ell \times F_{a_j}^\ell$. If $F_{a_i}^\ell \cap F_{a_j}^\ell = \emptyset$, then for any post-ReLU residual perturbations $\Delta x_t^{(i)}, \Delta x_t^{(j)}$ produced by equation 3,*

$$\big|\cos\big(\Delta x_t^{(i)},\, \Delta x_t^{(j)}\big)\big| \;\le\; \frac{\sigma_{\max}(G^{(ij)})}{\sqrt{\sigma_{\min}(G^{(ii)})\,\sigma_{\min}(G^{(jj)})}}. \tag{10}$$

Proof. By Lemma 3, $\Delta x_t^{(i)} = W_{F_{a_i}^\ell} \Delta h_t^{(i)}$ with $\Delta h_t^{(i)}$ supported on $F_{a_i}^\ell$, similarly for attribute j. Write $\delta^{(i)} = (\Delta h_t^{(i)})|_{F_{a_i}^\ell}$. Then

$$\big\langle \Delta x_t^{(i)},\, \Delta x_t^{(j)} \big\rangle \;=\; \big\langle W_{F_{a_i}^\ell}\, \delta^{(i)},\, W_{F_{a_j}^\ell}\, \delta^{(j)} \big\rangle \;=\; \delta^{(i)\top} G^{(ij)} \delta^{(j)}.$$

By the operator-norm definition of $\sigma_{\max}$,

$$\big|\delta^{(i)\top} G^{(ij)} \delta^{(j)}\big| \;\le\; \|\delta^{(i)}\| \cdot \|G^{(ij)}\delta^{(j)}\| \;\le\; \sigma_{\max}(G^{(ij)})\, \|\delta^{(i)}\|\, \|\delta^{(j)}\|.$$

For each attribute,

$$\|\Delta x_t^{(i)}\|^2 \;=\; \delta^{(i)\top} G^{(ii)} \delta^{(i)} \;\ge\; \sigma_{\min}(G^{(ii)})\, \|\delta^{(i)}\|^2,$$

where positive definiteness of $G^{(ii)}$ on the support of $\delta^{(i)}$ ensures $\sigma_{\min}(G^{(ii)}) > 0$. Combining,

$$\big|\cos\big(\Delta x_t^{(i)},\, \Delta x_t^{(j)}\big)\big| \;=\; \frac{|\delta^{(i)\top} G^{(ij)} \delta^{(j)}|}{\|\Delta x_t^{(i)}\|\, \|\Delta x_t^{(j)}\|} \;\le\; \frac{\sigma_{\max}(G^{(ij)})}{\sqrt{\sigma_{\min}(G^{(ii)})\,\sigma_{\min}(G^{(jj)})}}.$$

The bound is independent of $\delta^{(i)}, \delta^{(j)}$, so it holds uniformly over post-ReLU perturbations and is computable a priori from the SAE decoder. □

A.3.3 Why the bound is small in practice

Three structural properties of SAE-based steering favor a small bound.

Contrastive selection produces near-orthogonal feature sets. Features are selected by attribute-specific Cohen's d ($|d_j| \geq 0.2$, $p < 0.01$ via Mann–Whitney). A feature with high $|d|$ for sentiment typically has low $|d|$ for topic and vice versa, because sentiment and topic separate text along weakly correlated semantic axes. The realized overlaps in the deployed top-20 sets range from 0 to 5 shared features (mean 1.10; 20/42 pairs have zero overlap and 9/42 exceed 3 shared features), with the largest overlaps occurring at early layers of the smaller models. See §A.4 for the full distribution.

Overcompleteness and unit-norm decoder columns suppress off-diagonal correlation. Our SAEs use expansion factors $4\times$ to $16\times$ ($K \in \{12{,}288, 14{,}336, 16{,}384\}$ vs. $d \in \{768, 3{,}584, 4{,}096\}$). Decoder columns are normalized to unit norm after each training step. Trained overcomplete dictionaries with $K \gg d$ are observed to exhibit approximate incoherence, with $|\langle w_j, w_k \rangle|$ small for typical $j \neq k$, in line with what Donoho & Elad (2003) establish for optimal incoherent dictionaries (their result is for random/optimal dictionaries; for trained SAEs the same scaling is empirical, not guaranteed). The Frobenius bound

$$\sigma_{\max}(G^{(ij)}) \;\leq\; \|G^{(ij)}\|_F \;\leq\; \sqrt{|F^\ell_{a_i}| \cdot |F^\ell_{a_j}|} \cdot O(1/\sqrt{d})$$

is loose in the rank dimension but captures the qualitative scaling: the bound shrinks as d grows. The diagonal blocks $G^{(ii)}$ have unit diagonal and small off-diagonal entries, so by Gershgorin's theorem $\sigma_{\min}(G^{(ii)}) \geq 1 - (|F^\ell_{a_i}| - 1) \cdot O(1/\sqrt{d})$, bounded away from zero whenever $|F^\ell_{a_i}| \ll \sqrt{d}$.

Realized Gram values. Appendix A.4 reports both the realized Lemma 4 bound and the empirical cosine coupling at the deployed steering configuration ($|F^\ell_a| = 20$). The realized coupling is at most **0.15** across disjoint feature sets (mean **0.06**), roughly an order of magnitude tighter than the corresponding Frobenius/spectral bounds. A supplementary analysis at $|F^\ell_a| = 50$ (matching the interpretability convention used elsewhere in this appendix) gives similar magnitudes (max **0.25**, mean **0.08**).

Contrast with dense steering directions. Dense methods (PCA, linear probes, contrastive vectors) steer by adding a single direction $v_a \in \mathbb{R}^d$. For two attributes the cross-cosine is just $\cos(v_{a_1}, v_{a_2})$ — a single number, with no structural mechanism to keep it small. Empirically, these dense-method cross-cosines are substantially larger than the SAE coupling reported in §A.4; the corresponding cross-attribute interference for dense baselines is visible in Figure 14. SAE steering replaces this single coupling with the sparser $\sigma_{\max}(G^{(ij)})$ regime above.

A.3.4 Small-overlap correction

Strict disjointness $F^\ell_{a_i} \cap F^\ell_{a_j} = \emptyset$ is required for Lemma 3. Contrastive selection produces near-disjointness in most deployed pairs (20 of 42 have zero overlap, mean 1.10 shared features out of $|F^\ell_a| = 20$; §A.4). We characterize the resulting first-order correction.

Let $S_{ij} = F^\ell_{a_i} \cap F^\ell_{a_j}$. On S_{ij}, the pre-ReLU update at coordinate $j \in S_{ij}$ depends on both attributes: $\hat{h}_{t,j} = h_{t,j} + \alpha_{\text{eff}}(a_i, t, \ell)\delta_j^{(a_i)} + \alpha_{\text{eff}}(a_j, t, \ell)\delta_j^{(a_j)}$. ReLU's clipping decision at j depends on the sum, introducing genuine cross-attribute coupling. Decompose

$$\Delta x_t^{(i)} \;=\; \Delta x_t^{(i,\text{disjoint})} + \Delta x_t^{(i,\text{overlap})},$$

where the first term is the contribution from $F^\ell_{a_i} \setminus S_{ij}$ (covered by Lemma 3) and the second is from S_{ij}. The Gram bound applies to the disjoint part. The overlap part contributes

$$\big\|\Delta x_t^{(i,\text{overlap})}\big\| \;\leq\; \sigma_{\max}(W_{S_{ij}}) \cdot \big\|\Delta h_t^{(i)}\big|_{S_{ij}}\big\|,$$

which is bounded by the same $O(1/\sqrt{d})$ scaling as the off-diagonal Gram entries times the small factor $|S_{ij}|/|F^\ell_{a_i}|$. The overlap fraction $|S_{ij}|/|F^\ell_{a_i}|$ is at most $5/20 = 0.25$ and typically much smaller (average overlap fraction $1.10/20 = 0.055$), well within the correction regime for the disjoint-case bound. Realized coupling values are reported in Figure 14.

A.4 Decoder Gram structure and attribute coupling

Lemma 4 is a *conditional* statement: it assumes $F_{a_i}^{\ell} \cap F_{a_j}^{\ell} = \emptyset$ (disjoint feature sets), and under that assumption bounds cross-attribute coupling. It is not a claim that the deployed attribute feature sets are disjoint. Where the precondition holds, the bound is informative; where it does not, interference control is empirical (measured coupling, this section). This appendix reports both the bound and the empirical coupling, at both the deployed $|F_a^{\ell}| = 20$ (matching §4's steering recipe) and the interpretability-standard $|F_a^{\ell}| = 50$ used elsewhere in this appendix.

Setup. For each model–layer pair we extract the decoder submatrices W_{F_i} and W_{F_j} for the sentiment, topic, and style attributes used at that model, compute the inter-attribute Gram submatrix $G^{(ij)} = W_{F_i}^{\top} W_{F_j}$ and intra-attribute matrices $G^{(ii)}, G^{(jj)}$, and report:

(i) the *theoretical bound* $\sigma_{\max}(G^{(ij)}) / \sqrt{\sigma_{\min}(G^{(ii)})\, \sigma_{\min}(G^{(jj)})}$ from Lemma 4,

(ii) the *realized cosine coupling* $|\cos(\Delta x^{(i)}, \Delta x^{(j)})|$ where $\Delta x^{(i)} = \sum_k \delta_k^{(a_i)} \cdot W_{\text{dec}}[F_{a_i}[k]]$ uses the per-feature contrastive shifts $\delta^{(a)} = \bar{h}_+^{(a)} - \bar{h}_-^{(a)}$ actually applied at inference, and

(iii) the *disjoint variant* $|\cos|_{\text{disj}}$, which restricts each attribute to $F_{a_i} \setminus F_{a_j}$ so that the Lemma 4 precondition holds exactly.

Overlap and Lemma 4 status (deployed, $|F_a^{\ell}| = 20$). Across 42 model×layer×pair combinations, the top-20 sets share 0–5 features (mean 1.10; 20/42 pairs have zero overlap). The Lemma 4 bound is vacuous (RHS ≥ 1) in 36 of 42 deployed configurations; the 6 non-vacuous cases (RHS < 1) all have zero overlap. This is exactly what the conditional statement of Lemma 4 predicts: informative bounds appear precisely where the precondition (disjointness) is met (Spearman $\rho = 0.75$, $p \approx 10^{-8}$ between shared-feature count and RHS). In the majority of deployed configurations the precondition does not hold, and the interference claim rests on the empirical coupling measurements below, not on the analytic bound.

Empirical coupling is small (deployed). The realized cosine coupling is small across all models and layers:

- **With overlap (raw top-20):** $\max |\cos| = 0.734$ (MDLM L5 sentiment–formality, 3 shared features), mean $= 0.121$.
- **Disjoint feature sets:** $\max |\cos|_{\text{disj}} = 0.145$ (SEDD L6 sentiment–formality), mean $= 0.057$.

Removing shared features cuts the maximum coupling by $5\times$. Low geometric coupling does not by itself rule out behavioral interference: non-target attributes can still shift through downstream computation and through natural correlations between attributes (e.g., positive sentiment and formality). Behavioral interference is reported separately in Table 37 and Table 38.

Sensitivity to $|F_a^{\ell}| = 50$. For consistency with the interpretability characterizations in §C.6, we also report the same quantities at top-50. Overlap grows to 0–14 (mean 3.4), reflecting that the top-50 sets are more permissive. All 42 Lemma 4 bounds are vacuous at this scale. Empirical coupling is qualitatively similar: $\max |\cos|_{\text{disj}} = 0.252$ (MDLM L7 sentiment–topic), mean $= 0.075$. The deployed $|F_a^{\ell}| = 20$ view above is the operationally relevant measurement; the top-50 view characterizes the underlying SAE geometry at the scale used for other structural analyses.

Interpretation of the gap between bound and reality. Lemma 4 provides a valid but conservative bound: it guarantees that coupling cannot exceed a computable threshold under any post-ReLU perturbations consistent with disjoint feature sets. The empirical coupling is roughly an order of magnitude tighter. This gap is consistent with contrastive shifts $\delta^{(a)}$ avoiding the worst-case directions in the off-diagonal Gram structure i.e., contrastive selection finds attribute-aligned directions that are also geometrically near-orthogonal in the residual stream, beyond what disjointness alone guarantees. Lemma 4 characterizes *why* disjoint feature sets contain interference (subspace containment); it is not a precondition for the additive composition, which operates regardless of overlap. In deployment,

where sets are mostly not fully disjoint, interference is controlled empirically rather than by the analytic bound.

A.5 Empirical Validation of Theorem 1: η vs Quality Gain

Theorem 1 predicts that the adaptive-vs-uniform quality gain grows with the schedule concentration $\eta = N \sum_i w_i^2$, where w_i are normalized block fractions and N is the number of blocks. $\eta = 1$ corresponds to a perfectly uniform emergence profile (Adaptive $\approx$ Uniform), while $\eta = N$ indicates full concentration in a single block (maximum adaptive advantage). We test this prediction empirically by computing η from each model's emergence schedule and comparing it against the measured quality gain of Adaptive over Uniform steering.

Setup. For each of the 12 (model, attribute, primary-layer) conditions, we compute η from the B=8 block fractions used for steering, and measure the Adaptive-vs-Uniform quality gain as $(\text{Adaptive} - \text{Uniform})/\text{Uniform}$, averaged over $\alpha \in [1, 15]$. We report two quality metrics: LLMJ fluency (a 1–5 LLM-as-judge score) and Qwen2-7B perplexity (PPL).

Results. Table 7 reports all 12 conditions.

Table 7: Schedule concentration η and Adaptive-vs-Uniform quality gains for all 12 conditions (4 models $\times$ 3 attributes), averaged over $\alpha \in [1, 15]$.

Model	Attribute	η	LLMJ flu.	Qwen2 PPL
MDLM	sentiment	1.44	$+7.8\%$	$+29.0\%$
MDLM	style	1.57	$+28.7\%$	$+51.2\%$
MDLM	topic	2.58	$+10.9\%$	$+64.6\%$
SEDD	sentiment	3.18	$+8.0\%$	$+28.4\%$
SEDD	style	1.72	$+9.0\%$	$+26.4\%$
SEDD	topic	3.51	$+12.5\%$	$+53.7\%$
LLaDA	sentiment	1.68	$+0.5\%$	-8.3%
LLaDA	style	3.67	$+15.7\%$	-19.9%
LLaDA	topic	3.31	$+13.8\%$	-12.7%
DREAM	sentiment	4.72	$+1.6\%$	$+6.0\%$
DREAM	style	4.09	$+7.5\%$	$+27.5\%$
DREAM	topic	5.01	$+16.3\%$	$+29.8\%$

Within-model magnitude ordering. Theorem 1 predicts that, within a given model, attributes with higher η should show larger adaptive gains. We test this via Spearman rank correlation between η and each quality gain across the three attributes per model (n=3). On Qwen2 PPL, MDLM and SEDD both show perfect within-model rank agreement ($r_s = +1.0$): the attribute with the most concentrated schedule yields the largest PPL improvement from Adaptive steering. However, LLaDA's Qwen2 PPL gains are negative across all three attributes (-8% to -20%), because aggressive Adaptive steering on LLaDA induces repetitive text that achieves spuriously low Uniform PPL, inverting the metric (§5.1). For LLaDA, LLMJ fluency is the appropriate quality indicator, and under this metric η perfectly ranks all three attributes ($r_s = +1.000$). DREAM shows $r_s = +0.5$ on both metrics. With three attributes per model, no single within-model correlation can reach significance (the minimum one-sided p for a perfect ranking is 1/6). Pooling across models with an exact permutation test on the sum of r_s, the probability under random rankings of observing a sum ≥ 3.5 (three perfect orderings and one r_s=0.5) is 9/1296 $\approx$ 0.007.

B Vocabulary Grounding of Top Contrastive Features

To qualitatively verify that contrastive SAE features encode semantically coherent concepts, we project each feature's decoder column through the model's unembedding matrix and inspect the highest-scoring vocabulary tokens. Concretely, for a feature j with learned decoder vector $\mathbf{w}_j^{\text{dec}} \in \mathbb{R}^d$, we compute the logit-space projection $\mathbf{w}_j^{\text{dec}} \cdot \mathbf{W}_{\text{unembed}}^\top \in \mathbb{R}^{|\mathcal{V}|}$ and report the tokens with the largest scores. Intuitively, these are the vocabulary items that the feature most strongly promotes in the model's output distribution when active.

Table 8: Vocabulary grounding of top contrastive SAE features (via $\mathbf{W}_{\text{dec}} \cdot \mathbf{W}_{\text{unembed}}^{\top}$). For each model and attribute, the top-3 positive-d and top-3 negative-d features at the deepest SAE layer are shown.

Model	Attribute	Positive-d (target class)		Negative-d (contrast class)	
		Feature (d)	Top Tokens	Feature (d)	Top Tokens
MDLM (L7)	Topic	F6047 (+0.60)	midfield, Isles, winger, Trafford, Championship	F8014 (−0.67)	Securities, industrial, industry, regulatory, commercial
		F7923 (+0.58)	mph, Guinness, mph, Citation, maximum	F533 (−0.67)	2030, 2020, average, 2025, 2050
		F1000 (+0.57)	dyl, anooga, rongh, ickets, Sorceress	F6218 (−0.61)	Avg, [+, /$, DAQ, avg
	Sentiment	F6624 (+0.17)	solid, ocally, Scand, composure, ACTION	F11410 (−0.30)	!!!!!, ???, !!!!, ... , !!!
		F5123 (+0.16)	DJs, eaturing, magic, creat, aples	F10932 (−0.27)	my, idiot, NS, replies, CP
		F6921 (+0.15)	explan, overview, Learn, depth, detailing	F2643 (−0.27)	textures, texture, weird, Illusion, guts
	Formality	F1866 (+0.45)	reportedly, spokeswoman, watchdog, compensation, commissioner	F5013 (−0.56)	. . . , I, !, myself, ?!
		F8014 (+0.37)	Securities, industrial, industry, regulatory, commercial	F751 (−0.53)	. . . , ..., [...], . . . , [. . .]
		F7544 (+0.33)	quished, Instit, Aether, org, Presents	F7258 (−0.53)	GOODMAN, —, —
SEDD (L7)	Topic	F606 (+0.84)	matchups, rematch, matchup, challenger, clinch	F7260 (−0.73)	payable, cash, dollars, money, fees
		F5606 (+0.63)	championships, jerseys, championship, playoffs, playoff	F1914 (−0.66)	GDP, repairs, ppm, gallon, gallons
		F3948 (+0.45)	Featuring, teamed, eaturing, featuring, reunited	F9697 (−0.57)	liberal, government, neoliberal, privatization, atism
	Sentiment	F10956 (+0.27)	morrow, ,—, thence, tion, splend	F3169 (−0.21)	abusers, porn, sexist, moms, Kids
		F3948 (+0.25)	Featuring, teamed, eaturing, featuring, reunited	F8644 (−0.20)	nobody, nothing, nothing, Nothing, THING
		F2014 (+0.24)	dynam, affordable, unbeat, appreciated, versatile	F4439 (−0.20)	disgusting, vile, despicable, crap, disgrace
	Formality	F12001 (+0.63)	[CJK], Participant, ortium, ilaterally, ortment	F9490 (−0.67)	pie, hell, —, damn, . . .
		F4051 (+0.58)	Treaty, NATO, treaty, milit, Yugoslavia	F11298 (−0.61)	]], ín, tar, eq, ›
		F9697 (+0.56)	liberal, government, neoliberal, privatization, atism	F7306 (−0.55)	§§, Crossref, fmt, ÍÍ, nos
LLaDA (L26)	Topic	F9143 (+1.13)	Para, 增量 (increment), 晗, _event, events	F5553 (−0.80)	Oro, leness, experien, -war, Lah
		F1834 (+1.05)	Historic, atch, 奥斯 (Aus-), stacked	F6163 (−0.68)	INCLUDING, IllegalStateException, Subscriptions, -analytics
		F2897 (+0.63)	won, awarded, year, 提名 (nomination), win	F1518 (−0.66)	-producing, 资产负债 (balance sheet), .release, produce
	Sentiment	F12180 (+0.43)	(always, commended, throughout, PERF, -Sah	F5461 (−0.44)	Avoid, Avoid, avoided, avoid, AV
		F2833 (+0.38)	recommended, recommend, deserves, recommending, recomm	F10262 (−0.42)	throughout, whatsoever, anders, OEM, reative
		F7427 (+0.29)	aras, bec, 小小 (tiny), apical, 笃	F12851 (−0.42)	失败 (failure), 失败的 (failed (adj.)), amateur
	Formality	F8446 (+0.47)	monies, thru, angst, lite, youngster	F7446 (−1.07)	however, alot, barley, .D, ect
		F13972 (+0.44)	idency, "Don, cretion, ener, narc	F12552 (−0.66)	sublicense, fucking, empire, summer, fucked
		F12180 (+0.41)	(always, commended, throughout, PERF, -Sah	F9058 (−0.61)	}else, ",&, .const, ,免费, .第
DREAM (L23)	Topic	F6961 (+1.67)	zed, games, -sided, ided, 1	F14304 (−1.12)	company, angled, business, undergrad, ();
		F4856 (+1.28)	team, pall, Players, players, game	F13053 (−0.89)	computer, software, computers, Computer, data
		F7524 (+0.91)	team, Team, training, fre, trials	F6220 (−0.83)	union, union, unions, wright, SHA
	Sentiment	F10880 (+0.51)	simply, potency, consum, bullish, zen	F8145 (−0.51)	worse, S, literally, oa, Liter
		F5417 (+0.36)	and, all, each, is, even	F13133 (−0.34)	0, 1, prohibits, V, ement
		F3924 (+0.32)	in, also, that, really, is	F6642 (−0.33)	worse, worst, -bottom, Worst, buffalo
	Formality	F4302 (+0.59)	, ; … ... ~	F10886 (−1.14)	people, like, all, and, out
		F4221 (+0.58)	Parliamentary, anyhow, programmes, forb, practise	F11172 (−0.77)	abyte, oppos, imagin, Homeland, igy
		F5033 (+0.47)	cer, bloodstream, vintage, DNA, sap	F1918 (−0.70)	1, 2, 0, 4, ’

For each attribute (topic, formality/informality, sentiment), we select the top 3 features in each direction by Cohen's d: the 5 positive-direction features with the largest $d > 0$ (more active for the target class) and the 5 negative-direction features with the most negative $d < 0$ (more active for the contrast class). In Table 8, we report results at the deepest SAE layer for each model: layer 7 for MDLM-124M and SEDD-124M, layer 26 for LLaDA-8B, and layer 23 for DREAM-7B. MDLM and SEDD use the GPT-2 tokenizer (50K English tokens); LLaDA and DREAM use multilingual vocabularies (126K and 152K tokens, respectively), so some features project onto non-English tokens (Chinese glosses are provided where applicable).

C Detailed Experimental Setup

C.1 Training SAEs on Discrete Diffusion Models

We train TopK sparse autoencoders (Cunningham et al., 2023; Zhu et al., 2025) on residual-stream activations from MDLM (Sahoo et al., 2024), SEDD (Lou et al., 2024), LLaDA (Nie et al., 2026), and DREAM (Ye et al., 2025). We proposed a diffusion logits lens (§C.4) and conducted a layer probing study (§C.5.2), which together reveal when and where token identities crystallize during generation and identify the layers most suitable for SAE training in DLMs. Given a hidden state $\mathbf{x} \in \mathbb{R}^{d_{\text{model}}}$, the SAE encodes it into an overcomplete latent space ($d_{\text{SAE}} \gg d_{\text{model}}$), applies a TopK activation to retain only the k largest entries as a sparse code $\mathbf{h}$, and reconstructs via a learned decoder: $\mathbf{h} = \mathrm{TopK}\big((\mathbf{x} - \mathbf{b}_{\text{dec}})\,\mathbf{W}_{\text{enc}} + \mathbf{b}_{\text{enc}},\ k\big)$, $\hat{\mathbf{x}} = \mathbf{h}\,\mathbf{W}_{\text{dec}} + \mathbf{b}_{\text{dec}}$. Each nonzero entry h_j corresponds to a *feature* whose decoder direction $\mathbf{w}_j^{\text{dec}}$ (the j-th column of $\mathbf{W}_{\text{dec}}$) is the direction it adds to the residual stream. Training minimizes MSE reconstruction loss with an auxiliary dead-neuron loss (See AppendixC for details). Unlike AR models, DLM activations vary with masking rate, so we train on activations sampled across uniformly distributed masking rates to ensure generalization across the full denoising trajectory. Table 9 summarizes all models and SAE configurations.

Table 9: Model and SAE summary. All SAEs use TopK activation with k=32.

Model	Params	Arch.	Diffusion Loss	SAE Layers	d_{model}	d_{SAE}	$d_{\text{SAE}}/d_{\text{model}}$
MDLM (Sahoo et al., 2024)	124M	GPT-2 (Radford et al., 2019)	Absorbing	5, 6, 7	768	12,288	16×
SEDD (Lou et al., 2024)	124M	GPT-2	Score-entropy	5, 6, 7	768	12,288	16×
LLaDA (Nie et al., 2026)	8B	LLaMA (Touvron et al., 2023)	Absorbing	8, 14, 20, 26	4,096	16,384	4×
DREAM (Ye et al., 2025)	7B	Qwen2 (Yang et al., 2024)	Absorbing	8, 13, 17, 23	3,584	14,336	4×

C.2 Extracting Attribute Features from SAEs

We study three attributes: **sentiment**, **topic**, and **style**, common controllability targets and a natural combination for studying multi-attribute interplay, as they capture orthogonal aspects of text: subject matter, opinion polarity, and writing register. Following work on contrastive steering directions (Zou et al., 2023; Rimsky et al., 2024; Belinkov, 2022), we construct balanced binary corpora: Sports vs. Business from AG News (Zhang et al., 2015) for topic, positive vs. negative IMDB reviews (Maas et al., 2011) for sentiment, and formal vs. informal text for style, using the Pavlick corpus (Pavlick & Tetreault, 2016) for MDLM/SEDD and classifier-labeled IMDB reviews (Babakov et al., 2023) for LLaDA/DREAM to match their generation domain (§C.7.1). For style, we track formality features on MDLM/SEDD, whose baseline formality is near chance (∼55–58%), and informality features on LLaDA/DREAM, whose generations are already highly formal (∼73–78%; §5), so that the features we study correspond to the direction with greater room for control in later steering studies (§4). To identify attribute-encoding features, we noise each text at multiple mask ratios, pass the noised inputs through the model, and ReLU-encode the hidden states through the SAE at each layer (using ReLU rather than TopK to avoid biasing effect-size estimates; §C.7.2). For each feature j, we compute the Cohen's d effect size (Cohen, 2013): $d_j = (\bar{h}_j^+ - \bar{h}_j^-)/\sqrt{(\text{Var}(h_j^+) + \text{Var}(h_j^-))/2}$, where $\bar{h}_j^+$ and $\bar{h}_j^-$ are the mean activations for the two contrastive classes. We retain the top 50 features per direction per layer that pass a Mann–Whitney U test (Mann & Whitney, 1947) ($p < 0.01$, Bonferroni-corrected (Bonferroni, 1936)) for the interpretability analysis below. Vocabulary grounding (projecting decoder columns through $\mathbf{W}_{\text{unembed}}$) confirms that retained features are semantically coherent (see Appendix B for details).

C.3 Models and SAE Training

We study four DLMs that span two training objectives, two scales, and two provenance strategies: MDLM (Sahoo et al., 2024) and SEDD (Lou et al., 2024) are 124M-parameter GPT-2 architectures trained on OpenWebText with absorbing-state and score-entropy objectives, respectively—sharing architecture and data but differing only in loss function, forming a natural experiment for isolating the effect of training objective on learned representations. Dream (Ye et al., 2025) (7B) is fine-tuned from Qwen2.5 with masked diffusion, and LLaDA (Nie et al., 2026) (8B) is trained from scratch with a masked absorbing objective, providing a contrast in model provenance at the 7–8B scale.

For each model, we train TopK sparse autoencoders (Makhzani & Frey, 2013) on residual-stream activations extracted from middle-to-late transformer layers via forward hooks. The SAE encodes an input activation $\mathbf{x} \in \mathbb{R}^d$ into a sparse latent representation $\mathbf{h} \in \mathbb{R}^D$ ($D \gg d$) and reconstructs:

$$\mathbf{h} = \text{TopK}\big((\mathbf{x} - \mathbf{b}_{\text{dec}})\mathbf{W}_{\text{enc}} + \mathbf{b}_{\text{enc}},\ k\big), \tag{11}$$

$$\hat{\mathbf{x}} = \mathbf{h}\,\mathbf{W}_{\text{dec}} + \mathbf{b}_{\text{dec}}, \tag{12}$$

where TopK retains only the k largest activations and zeros the rest, providing an exact sparsity guarantee. We use k=32 for all models. Decoder columns are normalized to unit norm after each training step. We train with MSE reconstruction loss plus an auxiliary loss (λ=1/32) that encourages dead neurons (inactive for >10,000 steps) to reconstruct the current residual error. All SAEs are trained for 40-80k steps with AdamW (lr 10^{-4}, cosine schedule to 10^{-5}) on batches of 4,096 activations. All hyperparameters used for training SAE can be found in Table 15.

DLM-specific design choice: noise-level diversity. Unlike autoregressive models where each token has a single context, DLM activations depend on the noise level σ (or equivalently, the fraction of masked tokens). We extract activations across uniformly sampled noise levels during training, ensuring the SAE learns features that are robust across the full denoising trajectory rather than specialized to a single noise regime. This is validated by our sigma-stratified analysis (§C.6), which shows feature activations vary by less than 15% across noise levels.

C.4 Prediction Formation via Diffusion Logit Lens

We introduce the first *Diffusion Logit Lens*, adapting the logit lens technique (Belrose et al., 2023) to discrete diffusion language models. By projecting intermediate-layer hidden states through the unembedding matrix at each denoising step, we construct a two-dimensional *prediction formation surface* (layer $\times$ timestep) that reveals when and where token identities crystallize during generation.

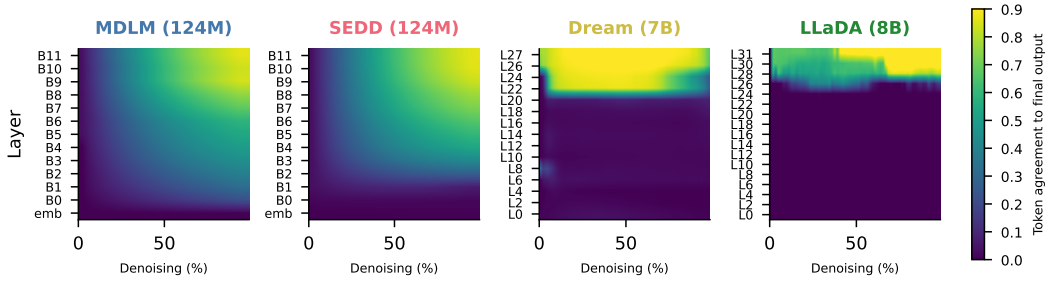


Figure 5: Prediction formation via logit lens across four diffusion language models. Each panel shows token agreement between intermediate-layer representations and the final output, as a function of layer depth and denoising progress. Small models (MDLM, SEDD) form predictions gradually across layers and time, while large models (DREAM, LLaDA) exhibit sharp layer boundaries where predictions crystallize abruptly.

C.4.1 Setup

For each model, we generate samples using the standard denoising schedule, hooking into every transformer block at each step. At layer ℓ and step t, we extract the hidden state $\mathbf{h}_t^\ell \in \mathbb{R}^{d_{\text{model}}}$ and project it into vocabulary space:

$$\hat{y}_t^\ell = \arg\max \ \mathbf{W}_{\text{unembed}}\, \mathbf{h}_t^\ell, \tag{13}$$

comparing $\hat{y}_t^\ell$ against the final generated token at each position. We report *agreement*: the fraction of positions where the intermediate prediction matches the final output. For models with explicit noise conditioning (MDLM, SEDD), we additionally evaluate a *conditioned* lens that applies the full output head including adaptive layer normalization (adaLN) with noise-level conditioning. Table 10 summarizes the experimental configuration.

Table 10: Logit lens experimental configuration per model.

Model	Samples	Steps	Layers	Lens Type	Notes
MDLM-124M	200	128	13	Raw + Conditioned	adaLN sigma conditioning
SEDD-124M	200	256	13	Raw + Conditioned	Score-entropy output head
DREAM-7B	100	64	15	Raw	AR-style logit shift
LLaDA-8B	64	64	17	Raw	No noise conditioning

C.4.2 Finding 1: Small Models Show Gradual Diagonal Crystallization

MDLM-124M's conditioned logit lens reveals a smooth diagonal gradient in the (layer, timestep) space. Agreement with the final output increases gradually across both dimensions simultaneously, reaching 86.9% at the final layer and step.

Table 11: MDLM-124M conditioned lens agreement (%) at selected layers and steps.

Layer	Step 0	Step 64	Step 127
Block 0	0.02	11.3	20.9
Block 5	3.8	37.9	54.9
Block 8	3.8	53.3	77.4
Block 11	3.6	60.1	**86.9**

Every layer contributes incrementally to prediction formation—there is no single "decision layer." The prediction surface is truly two-dimensional: both more denoising steps *and* deeper layers are jointly required for accurate predictions. This distributed processing is consistent with MDLM's relatively shallow architecture (12 blocks), which forces all layers to contribute.

Sigma conditioning is essential. The raw logit lens (projecting hidden states directly through $\mathbf{W}_{\text{unembed}}$ without noise-level conditioning) achieves a maximum of only $\sim$8% agreement on MDLM. The conditioned lens (applying the full output layer with adaLN sigma modulation) reaches 87%. This 10$\times$ gap proves that sigma conditioning is not merely a scaling factor—it fundamentally transforms the representation space. Without knowing the noise level, intermediate hidden states are essentially unreadable. The adaLN mechanism acts as a "decoder ring" that makes the residual stream interpretable at each noise level.

C.4.3 Finding 2: Masked vs. Unmasked Position Dynamics

Separating agreement by position type reveals distinct dynamics:

MDLM. Unmasked (already-revealed) positions are correctly predicted by block 5 onward ($>$80% agreement)—these are trivially predictable since the token is visible. Masked positions carry the interesting signal: their agreement rises from 3.6% (step 0) to 53.1% (final step) at block 11, indicating that the model progressively "plans" future tokens at still-masked positions. This anticipatory encoding is the mechanistic basis for why SAE features fire on masked positions before content commitment (§3.2).

DREAM. Unmasked agreement peaks at 97.6% (step 32) but *degrades* to 66.4% (step 63), mirroring the V-shaped entropy. Even the model's ability to predict already-visible tokens deteriorates at low mask ratios, confirming the out-of-distribution hypothesis.

C.4.4 Cross-Model Summary

Table 12: Cross-model comparison of prediction formation properties.

Property	MDLM	SEDD	DREAM	LLaDA
Crystallization	Gradual diagonal	Gradual diagonal	Sharp (L22)	Sharp (L26)
Layer contribution	All layers	All layers	L22+ only	L26+ only
Raw lens max	8%	72%	93%	100%
Entropy trajectory	Monotonic $\downarrow$	Monotonic $\downarrow$	V-shaped	Monotonic $\downarrow$
Sigma conditioning needed	Yes (essential)	Yes (but raw works)	N/A	N/A
Max agreement	87%	86%	93%	100%

The logit lens reveals a fundamental dichotomy in how DLMs form predictions: **small models crystallize gradually** across a 2D surface (requiring both depth and denoising progress), while **large models exhibit sharp phase transitions** at specific layers. The training objective further modulates readability—score-entropy (SEDD) produces representations that are 9$\times$ more readable without conditioning than absorbing-state (MDLM). These macroscopic patterns have direct implications for SAE-based steering: features in pre-crystallization layers (e.g., DREAM L8–L20) encode abstract semantic information without token commitment, making them natural intervention points for high-level attribute control without disrupting low-level token selection.

C.5 SAE Layer Selection

Selecting which transformer layers to train SAEs on is non-trivial: in autoregressive models, middle layers are known to balance representational richness with causal malleability (Zou et al., 2023), but this principle has never been validated for discrete diffusion transformers. We describe the systematic methodology used and summarize the results for each model.

C.5.1 Methodology

We apply a four-stage protocol uniformly across all models:

1. **Linear probing.** Logistic regression with 5-fold stratified cross-validation on mean-pooled activations from every transformer layer. We probe two attributes: sentiment (IMDB, 2-class) and topic (AG News, 4-class). To account for the stochastic nature of diffusion representations, each text is passed through the model with 3 independent noise samples (mask ratio 0.5), yielding $3\times$ the effective dataset size (6,000 data points per layer from 2,000 source texts).
2. **SAE training.** TopK SAEs are trained on candidate layers with identical hyperparameters per model (Table 15), ensuring fair comparison. We monitor reconstruction loss and dead neuron fraction as quality indicators.
3. **Contrastive feature identification.** Per-feature Cohen's d between attribute groups, retaining features with $|d| > 0.05$ and $p < 0.01$. This reveals which layers produce the strongest and most numerous attribute-discriminative features.
4. **Steering ablation.** The decisive test: we compare steering effectiveness (positive-rate swing across an α sweep) for different layer windows, isolating the causal contribution of each layer group.

C.5.2 Layer Probing Results

Table 13 reports sentiment probing accuracy at selected layers for all four models.

Table 13: Linear probing accuracy (sentiment, 5-fold CV) at selected layers. Layers in **bold** are those selected for SAE training. MDLM and SEDD share the GPT-2 architecture (12 layers); Dream has 28 layers (Qwen2); LLaDA has 32 layers (LLaMA).

Layer	MDLM (12L, 124M)	SEDD (12L, 124M)	Dream (28L, 7B)	LLaDA (32L, 8B)
0	80.9	81.5	88.5	86.4
4	82.2	77.8	90.7	88.1
5	**82.7**	**78.6**	—	—
6	**83.1**	**79.9**	—	—
7	**84.0**	**81.3**	—	—
8	84.7†	82.1	**95.0**	**91.4**
9	84.4	81.8	—	—
10	84.0	82.4	—	—
11	84.4	83.2†	—	—
13	—	—	**97.6**	—
14	—	—	—	**95.5**
17	—	—	**98.2**†	—
20	—	—	98.0	**97.3**
23	—	—	**98.0**	98.1†
26	—	—	—	**98.0**
27	—	—	98.0	—
31	—	—	—	97.9

†Peak probing accuracy for that model.

Three patterns emerge:

- **Monotonic rise then plateau.** All four models show increasing probing accuracy with depth, saturating in the upper layers. MDLM plateaus at L8–11 ($\sim$84%); Dream and LLaDA plateau earlier in relative terms ($\sim$L13/L17 and $\sim$L20/L23 respectively), with much higher absolute accuracy (>97%) reflecting the larger model capacity.
- **SEDD dips before rising.** Unlike MDLM, SEDD shows a characteristic dip at L1–5 (77–79%) below the L0 baseline (81.5%), before rising through L6–11. This U-shaped profile suggests that score-entropy diffusion initially disrupts the pretrained embeddings' linear separability before rebuilding it in later layers.
- **Scale raises the floor.** The 7–8B models (Dream, LLaDA) start at 86–89% even at L0, compared to 81% for the 124M models. Richer input embeddings (Qwen2.5 / LLaMA) provide a higher baseline that subsequent layers refine.

Topic probing follows a similar profile but peaks later: MDLM at L11 (79.1%), Dream at L23 (95.3%), LLaDA at L28 (95.7%). This offset is consistent with topic being a more global, multi-token property that requires deeper processing.

C.5.3 Steering Ablation: MDLM

The decisive experiment compares four 3-layer windows for sentiment steering with identical settings (50 features per layer, $\alpha \in [-3, +3]$, 512 samples per α):

Table 14: MDLM sentiment steering ablation across layer windows. Swing = positive rate at $\alpha{=}-3$ minus $\alpha{=}+3$.

Layer window	Swing (pp)	Max pos. rate	Min pos. rate	PPL range
[5, 6, 7]	**74.2**	76.0%	1.8%	42–51
[7, 8, 9]	70.3	70.5%	0.2%	40–49
[8, 9, 10]	62.1	63.3%	1.2%	40–46
[9, 10, 11]	59.8	62.7%	2.9%	40–43

Steering swing decreases monotonically with depth (-3.6 pp per layer shift on average), revealing a stark **probing–steering dissociation**: the layers with the highest probing accuracy (L8–11) are the least causally effective for steering. This mirrors findings from autoregressive representation engineering (Zou et al., 2023) and extends the principle to discrete diffusion transformers.

The mechanism is *causal distance to output*: perturbations at L5–7 traverse 5–7 subsequent transformer blocks, amplifying the steering signal through residual-stream propagation. Later-layer perturbations have fewer blocks to propagate through and may be partially undone by the final layer norm and unembedding.

C.5.4 Layer Selection per Model

MDLM: layers [5, 6, 7]. Selected by the steering ablation above. Despite probing peaking at L8 (84.7%), L5–7 achieve the strongest steering (74.2pp swing). Topic steering on the same layers yields 100pp swing (100% Sports at $\alpha{=}-3$), independently validating the choice.

SEDD: layers [5, 6, 7]. SEDD shows *no* probing–steering dissociation: at L8–10 the swing reaches 90.8pp, far exceeding L5–7 (38.8pp). Score-entropy training appears to co-locate probing accuracy and causal malleability in the same layers which is the opposite of MDLM's pattern under absorbing-state training, suggesting the dissociation stems from the loss objective rather than the architecture. To preserve a clean MDLM–SEDD contrast (same architecture, different loss), we train SAEs at the same layers (L5–7) on both models throughout; per-model layer optimization would conflate the loss comparison with layer choice and obscure the controlled comparison in §3.2.

Dream-7B: layers [8, 13, 17, 23]. Selected to span the transition zone (L8, pre-plateau) through the probing plateau (L13–23). The plateau makes layer discrimination via probing alone impossible; direct steering comparisons confirm that L17 (sentiment) and L23 (topic) produce the strongest features. Even spacing across the 28-layer model ensures coverage of early-, mid-, and late-stage representations.

LLaDA-8B: layers [8, 14, 20, 26]. Follows the same even-spacing strategy across 32 layers, covering transition (L8), mid-rise (L14), near-peak (L20), and plateau (L26). Feature identification reveals the strongest contrastive effects at L14 (max $|d|{=}0.63$), while the dissociation experiment shows L26 dominates steering (67pp alone). This suggests a mild dissociation similar to MDLM: the layer with the strongest individual features is not the layer with the greatest causal impact.

C.5.5 SAE Training Hyperparameters

C.5.6 Summary

The layer selection study yields two findings of independent interest:

Table 15: SAE training configuration per model. All SAEs use TopK activation with auxiliary dead-neuron loss.

Parameter	MDLM	SEDD	Dream	LLaDA
d_{model}	768	768	3,584	4,096
d_{sae}	12,288	12,288	14,336	16,384
Expansion ratio	16×	16×	4×	4×
k (TopK)	32	32	32	32
Training steps	50K	80K	80K	40K
Batch size	4,096	4,096	4,096	4,096
Learning rate	10^{-4}	10^{-4}	10^{-4}	10^{-4}
LR schedule	Cosine	Cosine	Cosine	Cosine
Aux loss coef	1/32	1/32	1/32	1/32
EMA decay	0.999	0.999	0.999	0.999
Training data	OWT	OWT	OWT	OWT

1. **Probing–steering dissociation is loss-dependent.** MDLM (absorbing loss) shows clear dissociation — middle layers steer best despite later layers probing best. SEDD (score-entropy loss) shows no dissociation — later layers both probe and steer best. LLaDA shows suggestive dissociation similar to MDLM. This implicates the training objective, not the architecture, as the key determinant of where causal malleability resides.
2. **The causal malleability principle extends to diffusion transformers.** For absorbing-loss DLMs, middle layers provide the best balance of representational richness and propagation distance, consistent with findings from autoregressive representation engineering. This principle appears robust across model scales (124M to 8B) and architectures (GPT-2, Qwen2, LLaMA).

C.6 Sigma-Stratified Feature Analysis

Unlike autoregressive models where each token is conditioned on a fixed left context, DLM activations depend on the noise level σ (equivalently, the fraction of masked tokens). During SAE training, we extract activations across uniformly sampled noise levels, ensuring the SAE learns features that span the full denoising trajectory. A natural question is whether the learned features are *robust* across noise levels or *specialized* to particular noise regimes. We answer this with a sigma-stratified analysis of feature discriminability.

C.6.1 Methodology

We partition the noise schedule into five equal-frequency bins based on the percentiles of σ observed during activation extraction (Table 16). For each bin, we recompute the Cohen's d effect size for the top-20 most discriminative features per layer. We report the *coefficient of variation* (CV) of $|d|$ across bins as the primary measure of noise-level sensitivity:

$$\text{CV} = \frac{\text{std}(|d_1|, \ldots, |d_5|)}{\text{mean}(|d_1|, \ldots, |d_5|)} \times 100\% \tag{14}$$

where d_i is the Cohen's d of a given feature computed within sigma bin i.

Table 16: Sigma bin boundaries (equal-frequency, MDLM). Low σ corresponds to near-clean text; high σ corresponds to heavily masked text. Each bin contains $\sim$1.79M token positions (sentiment) or $\sim$334K (topic).

Bin	σ_{low}	σ_{high}	Interpretation
1	8.3×10^{-5}	0.221	Near-clean
2	0.221	0.514	Low noise
3	0.514	0.915	Medium noise
4	0.915	1.61	High noise
5	1.61	6.89	Near-fully-masked

C.6.2 RESULTS

Table 17 summarizes the distribution of CV values across the top-20 features for each layer and attribute. We observe moderate but heterogeneous variation across noise levels, with median CV values ranging from 14% to 22% depending on the layer and attribute.

Table 17: Coefficient of variation (CV) of $|d|$ across sigma bins for the top-20 features per layer. Lower CV indicates more noise-robust discriminability.

Attribute	Layer	Median CV	Mean CV	% with CV$<15\%$	% with CV$<25\%$
Sentiment	L5	19.8%	24.6%	25%	65%
	L6	19.9%	22.6%	25%	70%
	L7	14.5%	19.0%	50%	80%
Topic	L5	13.9%	16.6%	60%	80%
	L6	21.7%	23.3%	20%	70%
	L7	14.1%	15.4%	60%	90%

Several patterns emerge. First, a substantial fraction of features (20–60%, depending on the layer) exhibit CV $< 15\%$, indicating that their discriminability is largely preserved across the full noise schedule. Second, the majority of features (65–90%) have CV $< 25\%$, confirming that SAE features trained on mixed-noise activations successfully generalize across denoising stages rather than collapsing to noise-regime-specific detectors. Third, the remaining features with higher CV reveal a structured dependence on σ: some features strengthen with increasing noise (capturing distributional or structural cues available even in heavily corrupted text) while others weaken (encoding fine-grained lexical or semantic cues that emerge only as text nears its final form).

C.6.3 PER-FEATURE SIGMA PROFILES

While the CV analysis above uses $|d|$ to assess magnitude stability, the per-feature profiles below report signed d to preserve polarity information, which is relevant for the directional analysis that follows. Table 18 shows representative per-bin Cohen's d profiles for selected positive-d sentiment features at Layer 7, illustrating the diversity of sigma-dependence patterns.

Table 18: Per-bin signed Cohen's d for selected positive-d sentiment features (Layer 7, top-50 by overall d). Features are categorized by their sigma-dependence pattern.

Feature	Overall d	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Pattern
Noise-robust (CV $< 15\%$)							
F1561	+0.068	0.065	0.073	0.072	0.073	0.072	Stable
F6047	+0.091	0.090	0.082	0.092	0.099	0.093	Stable
F6360	+0.092	0.089	0.087	0.094	0.107	0.097	Stable
Noise-sensitive (CV $> 25\%$)							
F3121	+0.074	0.007	0.034	0.058	0.086	0.170	Strong increase
F3099	+0.117	0.052	0.079	0.103	0.152	0.186	Strong increase
F5486	+0.065	0.028	0.036	0.056	0.093	0.125	Strong increase

C.6.4 DIRECTIONAL ANALYSIS

We additionally examine whether features tend to become more or less discriminative at higher noise levels. For sentiment, the split is approximately balanced: 54/46 at L5, 48/52 at L6, 44/56 at L7 (fraction of top-50 features with increasing vs. decreasing d across σ bins). Topic features, by contrast, show a clear tendency to weaken at higher noise levels: only 28/72 at L5, 26/74 at L6, and 26/74 at L7 strengthen under noise.

This bidirectional structure suggests that the SAE decomposes DLM representations into complementary feature populations:

- **Low-σ features** encode fine-grained semantic content (e.g., sentiment-bearing lexical items) that is most discriminative when text is nearly resolved.

- **High-σ features** capture coarser distributional properties (e.g., topic or register) that the model infers even from heavily corrupted input.

This functional decomposition is unique to the diffusion setting and has no direct analog in autoregressive models, where all tokens share a single (fully observed) conditioning context.

C.6.5 Implications for Steering

The sigma-stratified analysis has practical implications for feature-based steering of DLMs. Since feature discriminability varies across the denoising trajectory, a sigma-adaptive steering strategy—applying different feature weights at different noise levels—could yield more targeted interventions than uniform steering across all timesteps. We leave exploration of sigma-adaptive steering to future work, noting that the infrastructure for step-conditional steering is already demonstrated in our denoising dynamics experiments (§3.2).

C.7 Contrastive Feature Extraction

C.7.1 Datasets

We study three attributes in this paper–Sentiment, Topic, and Style. The datasets used to model these attributes are explained below.

Sentiment: We use the IMDB movie review dataset proposed by Maas et al. (2011) for sentiment. We sample $5,000$ positive reviews (label $= 1$) and $5,000$ negative reviews (label $= 0$) from this dataset and use them for learning sentiment features. All texts are truncated to 512 tokens.

Topic: We use the AG News dataset Zhang et al. (2015), which contains news articles across four topics: World, Sports, Business, and Sci/Tech. We study the topic Sports in this paper. Two potential contrastive approaches exist for feature extraction: (1) contrasting Sports against all other topics combined (one-vs-all), and (2) contrasting Sports against a single other topic (one-vs-one). Prior work on contrastive representation learning suggests that narrower, more targeted contrasts yield cleaner feature directions: Zou et al. (2023) showed that contrastive pair specificity affects steering direction quality, Rimsky et al. (2024) demonstrated that carefully curated contrastive pairs outperform generic ones for activation steering, and Belinkov (2022) established that the choice of control task determines what linguistic properties a probe captures. Consistent with these findings, we empirically observe that one-vs-one contrasts produce features that are more attribute-specific and less likely to interfere with other attributes during multi-attribute steering. Hence, we sample $5{,}000$ Sports articles (label $= 1$) and $5{,}000$ Business articles (label $= 0$), totaling $10{,}000$ samples, and use them for learning sports topic related features. Texts are truncated to 512 tokens.

Style: We study formality as our style attribute. For MDLM and SEDD, which are unconditional models trained on OpenWebText (Gokaslan & Cohen, 2019), we use the formality corpus of Pavlick & Tetreault (2016). This dataset contains $9{,}274$ sentences from news articles and reader comments, each annotated with a continuous formality score on a $[-3, +3]$ scale by crowdworkers using Best-Worst Scaling (Louviere et al., 2015). We threshold these scores to create a binary contrast: sentences with average score ≥ 1.0 are labeled formal (class A, $2{,}597$ sentences) and sentences with average score ≤ -1.0 are labeled informal (class B, $2{,}600$ sentences). Sentences with scores between -1.0 and 1.0 are excluded as ambiguous. We balance to $2{,}500$ samples per class, totaling $5{,}000$ samples. Positive-d features correspond to formal features; negative-d to informal. LLaDA and DREAM are conditional models that we prompt with domain-neutral opinion starters (e.g., "I think this is", "The experience was") to generate review-style text. The corpus by Pavlick & Tetreault (2016) is domain-mismatched for these models: its formality features are learned from news articles and government proceedings, encoding vocabulary patterns (e.g., legislative terminology, formal news register) that do not appear in review-style generations. When applied to LLaDA, these features caused text degradation–steering toward formality produced repetitive "president of the Senate" loops rather than genuinely formal reviews, because the features activated domain-specific vocabulary instead of domain-general register shifts. Hence, we instead label $20{,}000$ IMDB reviews from (Maas et al., 2011) with an off-the-shelf formality classifier[1] (Babakov et al., 2023) that was trained on the GYAFC corpus (Rao & Tetreault, 2018) and the online formality corpus of Pavlick & Tetreault

[1] https://huggingface.co/s-nlp/roberta-base-formality-ranker

(2016). This ensures the formality features are learned from the same domain (movie reviews) that the prompted models generate. Reviews with formality score > 0.9 form the formal class (6,587 reviews) and reviews with score < 0.1 form the informal class (263 reviews). We balance to 263 samples per class, limited by the informal count. As with MDLM and SEDD, positive-d features correspond to formal features and negative-d to informal.

C.7.2 Feature Extraction Procedure

The feature extraction procedure works as follows. For each dataset pair (class A vs. class B):

1. **Tokenization:** We tokenize each text with the model's tokenizer, truncating or padding to the model's sequence length.
2. **Apply noise:** Then for each text, we create 3 noised versions at mask ratios $\{0.3, 0.5, 0.7\}$ by randomly replacing tokens with `[MASK]`. This ensures features are identified across noise levels rather than at a single denoising stage.
3. **Forward pass:** Next, we run the noised input through the model to obtain hidden states at each SAE layer.
4. **ReLU SAE encoding (per-token):** for each token position at each SAE layer, we compute:
$$\text{activations} = \text{ReLU}\big((\mathbf{h} - \mathbf{b}_{\text{dec}})\,\mathbf{W}_{\text{enc}} + \mathbf{b}_{\text{enc}}\big), \tag{15}$$
This produces a $(\text{seq_len} \times d_{\text{SAE}})$ activation matrix per sample per layer. Although the SAEs use TopK activation during training and normal inference, we use full ReLU encoding for feature extraction. TopK retains only the K most active features per token, so a contrastive feature outside the top K would appear inactive, biasing its Cohen's d toward zero. ReLU returns the true activation for all d_{SAE} features, enabling accurate effect-size estimation.
5. **Streaming statistics:** We accumulate per-feature mean and variance for class A and class B using Welford's online algorithm (Welford, 1962), avoiding the need to store all activations in memory.
6. **Cohen's d (Cohen, 2013):** for each of the d_{SAE} features at each layer, we compute:
$$d_j = \frac{\bar{h}_j^A - \bar{h}_j^B}{\sqrt{(\text{Var}(h_j^A) + \text{Var}(h_j^B))/2}}. \tag{16}$$
Positive d indicates a feature more active for class A; negative d for class B.
7. **Statistical validation:** Next, we rank features by $|d|$, take the top 300 candidates, and run Mann-Whitney U tests (Mann & Whitney, 1947) with $p < 0.01$ (Bonferroni-corrected (Bonferroni, 1936) across 300 tests). We retain the top 50 positive-d and top 50 negative-d features per layer (indices and statistics) for our study. The top 20 per direction are used for steering.
8. **Per-feature mean shifts:** For each retained feature j, we store the calibrated shift
$$\Delta_j = \bar{h}_j^A - \bar{h}_j^B, \tag{17}$$
Here, A and B are the two contrastive classes. At steering time, the hidden state is modified as $h_j \leftarrow h_j + \alpha \cdot \Delta_j$. This is self-calibrating: features with large class differences receive large shifts, while features with small differences receive proportionally small shifts.

Table 19 shows the feature extraction hyperparameters for all models, and Table 20 shows per-model configurations.

C.8 Temporal Dynamics Tracking

C.8.1 Tracking Procedure

To construct the dynamics-weighted (adaptive) steering schedule, we track how contrastive feature activations evolve across the denoising trajectory. Table 23 lists the per-model configurations.

1. We generate N samples with SAE hooks active at all denoising steps.

Table 19: Feature extraction hyperparameters for all models.

Parameter	Value
Batch size	48
Noise samples per text	3
Mask ratios	$\{0.3, 0.5, 0.7\}$
Top candidates for U-test	300
Mann-Whitney p-value	0.01 (Bonferroni)
Features saved per direction	50
Features used for steering	20
Random seed	42

Table 20: Per-model extraction configurations.

	MDLM	SEDD	LLaDA	DREAM
SAE layers	5, 6, 7	5, 6, 7	8, 14, 20, 26	8, 13, 17, 23
d_{SAE}	12,288	12,288	16,384	14,336
d_{model}	768	768	4,096	3,584
Sentiment source	IMDB (Maas et al., 2011)	IMDB (Maas et al., 2011)	IMDB (Maas et al., 2011)	IMDB (Maas et al., 2011)
Topic source	AG News (Zhang et al., 2015)	AG News (Zhang et al., 2015)	AG News (Zhang et al., 2015)	AG News (Zhang et al., 2015)
Formality source	Pavlick & Tetreault (2016)	Pavlick & Tetreault (2016)	IMDB+RoBERTa (Babakov et al., 2023)	IMDB+RoBERTa (Babakov et al., 2023)

2. At each step, we ReLU-encode the hidden states through the SAE and record activations for the top-20 contrastive features per attribute.
3. We store per-sample, per-step activations as an $(N \times T \times F)$ tensor per layer, where T is the number of denoising steps and F is the number of tracked features.
4. Since not all generated samples express the target attribute, we classify all N texts with attribute-specific classifiers (Table 21) and retain only confident samples (softmax probability above a threshold; see Table 22) to avoid diluting the emergence signal. Informality uses a lower threshold (0.6) because baseline formality is high (73–78% for LLaDA/DREAM), so very few generated texts are confidently informal at 0.8.
5. Then we average dynamics over the filtered samples to obtain per-step mean activation curves.
6. We compute block fractions by dividing the trajectory into B temporal blocks, computing the activation change in each block, clamping negatives to zero, and normalizing to sum to one.

Table 21: Evaluation classifiers. These classifiers are used for evaluation and for filtering dynamics samples. They are not used during contrastive feature extraction for sentiment or topic, which rely on human-labeled datasets. For formality, the RoBERTa ranker is additionally used to label IMDB reviews for LLaDA and DREAM extraction (§C.7.1); however, the labeled texts (IMDB training set) and the evaluated texts (model-generated) are disjoint populations, so there is no direct circularity.

Attribute	Model	HuggingFace ID
Sentiment	DistilBERT (Sanh et al., 2019) fine-tuned on SST-2	`distilbert/distilbert-base-uncased-finetuned-sst-2-english`
Topic	BERT (Devlin et al., 2019) fine-tuned on AG News	`fabriceyhc/bert-base-uncased-ag_news`
Formality	RoBERTa (Babakov et al., 2023) formality ranker	`s-nlp/roberta-base-formality-ranker`
Perplexity	Qwen2-7B (Yang et al., 2024)	`Qwen/Qwen2-7B`

C.8.2 Schedule Robustness

The adaptive steering schedule is derived from empirical emergence profiles, raising two practical questions: (1) how many filtered samples are needed for the schedule to stabilize, and (2) how sensitive is the schedule to the number of temporal blocks B?

Minimum sample size (bootstrap subsampling). We bootstrap-subsample $N \in \{25, 50, 100, 200, 500, 1000, 2000, 3000, 5000\}$ filtered samples without replacement (200 repetitions) from each model's per-sample dynamics, recompute the B-block emergence-fraction schedule for each subset, and measure cosine similarity to the full-data schedule. We report the smallest N for which the 5th-percentile cosine ≥ 0.95.

Table 22: Classifier confidence thresholds for dynamics filtering.

Attribute	Threshold	% samples retained
Sentiment	0.8	22–64%
Topic	0.8	7–18%
Formality	0.8	8–12%
Informality	0.6	11%

Table 23: Per-model dynamics tracking configurations.

	MDLM	SEDD	LLaDA	DREAM
Samples (N)	5,000	5,000	5,000	5,000
Denoising steps (T)	1,024	1,024	64	256
Temporal blocks (B)	16	16	8	8
Steps per block	64	64	8	32
Batch size	16	4	8	8

Models with concentrated emergence profiles (MDLM, SEDD, DREAM) stabilize at 25–100 filtered samples, while LLaDA's distributed profile requires 100–500 samples. One condition (LLaDA topic at L14, n_{filtered}=346) never reached the 0.95 threshold within the sweep range, reflecting the sparse, weakly-committed topic features at that layer. These results confirm that the dynamics data collected per Table 23 (5,000 samples per model) is sufficient by a wide margin.

Block-count B sensitivity. We recompute block fractions at $B \in \{4, 8, 16\}$ from the same full-data mean emergence curves, upsample each to a common length of 16, and compute pairwise cosine similarity between the resulting schedules across all (model, layer, attribute) conditions.

For sentiment, topic, and formality, the schedule is stable across B values, with cosine ≥ 0.73 in all comparisons and ≥ 0.88 between the adjacent B=8 and B=16. The low cosines for informality ($\sim$0.47) reflect degenerate profiles on LLaDA and DREAM where emergence curves are flat or decreasing at some layers, producing near-zero schedules where cosine similarity is dominated by noise. For these conditions, the schedule is effectively uniform regardless of B, so the sensitivity is moot. On MDLM specifically, formality matches exactly across B values (cosine 0.88/0.78/0.88), and sentiment is highly stable (0.94/0.91/0.97).

C.9 SAE Steering Experiments

C.9.1 Steering Mechanism

At each denoising step, for each SAE layer, a forward hook performs the following:

1. **Encode**: $\mathbf{h}_{\text{enc}} = \text{ReLU}(\mathbf{h}\,\mathbf{W}_{\text{enc}} + \mathbf{b}_{\text{enc}})$
2. **Modify**: $\mathbf{h}_{\text{enc}}[j] \leftarrow \mathbf{h}_{\text{enc}}[j] + \alpha \cdot \Delta_j$ for each target feature j
3. **Decode**: $\mathbf{h}_{\text{modified}} = \mathbf{h}_{\text{enc}}\,\mathbf{W}_{\text{dec}} + \mathbf{b}_{\text{dec}}$
4. **Replace**: substitute the original hidden state with $\mathbf{h}_{\text{modified}}$

The shift Δ_j for each feature is the calibrated mean difference computed during extraction (§C.7.2).

C.9.2 Steering Modes

We evaluate two schedule types with two alpha-scaling strategies, yielding four modes.

Uniform schedule. A constant α is applied at every layer and every denoising step.

Adaptive schedule. The denoising trajectory is divided into B temporal blocks (Table 23). Each block has a *block fraction* computed from the dynamics tracking (§C.8), reflecting what fraction of total feature emergence occurs in that block. At each step, the steering strength is scaled by the block fraction of the step's temporal block. Steps in high-activity blocks receive proportionally more

Table 24: Minimum filtered samples for schedule stability (5th-percentile cosine ≥ 0.95 to full-data schedule). Median, min, and max are taken across all (attribute, layer) conditions per model.

	Median min-N	**Best case**	**Worst case**
MDLM	25	25	100
SEDD	25	25	200
DREAM	50	25	100
LLaDA	200	100	500

Table 25: Schedule stability across block counts B. Mean cosine similarity (across models and layers) between schedules at different B values.

Attribute	$\cos(B{=}4, B{=}8)$	$\cos(B{=}4, B{=}16)$	$\cos(B{=}8, B{=}16)$	N
Sentiment	0.83	0.79	0.90	14
Topic	0.81	0.73	0.89	14
Formality	0.88	0.77	0.88	6
Informality	0.47	0.45	0.48	8

steering; steps in low-activity blocks receive less. This concentrates steering effort where the model is actively making semantic decisions.

E-ratio scaling (for multi-attribute steering). Different attributes require different steering strengths to achieve comparable classifier rates. We normalize this asymmetry with E ratios: (1) run single-attribute steering across $\alpha = 1$–20; (2) identify α^* at which each attribute reaches approximately 80% classifier rate; (3) compute $E_{\text{attr}} = \alpha^*_{\text{attr}} / \max_{\text{attrs}}(\alpha^*)$. The hardest attribute receives $E{=}1.0$; easier attributes receive $E{<}1.0$. During multi-attribute steering, $\alpha_{\text{eff}} = \alpha_{\text{global}} \times E_{\text{attr}}$.

Crossing the two schedules (uniform, adaptive) with two scaling strategies (raw α, E-scaled α) gives four modes: **Uniform**, **Adaptive**, **Uniform+E**, and **Adaptive+E**. Table 26 reports the E ratios.

Table 26: E ratios for multi-attribute steering. The hardest attribute (ratio $= 1.0$) is in bold.

Model	**Mode**	**Sentiment**	**Topic**	**Style**
MDLM	Uniform	**1.000**	0.167	0.333
	Adaptive	**1.000**	0.143	0.214
SEDD	Uniform	**1.000**	0.333	0.167
	Adaptive	**1.000**	0.100	0.100
LLaDA	Uniform	0.417	0.500	**1.000**
	Adaptive	0.900	0.450	**1.000**
DREAM	Uniform	**1.000**	0.400	0.800
	Adaptive	**1.000**	0.400	0.800

C.9.3 Attribute Combinations

We evaluate 7 attribute combinations per model: 3 single-attribute (sentiment, topic, style) and 4 multi-attribute (sentiment+topic, sentiment+style, style+topic, sentiment+topic+style). Sentiment is steered toward positive and topic toward Sports in all combinations. For the style attribute, MDLM and SEDD steer toward *formality* (positive-d features), since their baseline formality is near chance (55% and 58%, respectively). LLaDA and DREAM steer toward *informality* (negative-d features, with inverted shift direction), since their baselines are already highly formal (78% and 73%).

C.9.4 Steering Hyperparameters

Table 27 shows the per-model steering parameters. We generate with MDLM and SEDD unconditionally (no prompt). For LLaDA and DREAM, we use 10 domain-neutral prompts (e.g., "I think this is", "In my opinion") with 20 samples per prompt, yielding 200 total per condition. The 10 prompts are listed in Table 28. All models use the same α range of 1–20. Note that, MDLM and SEDD are unconditional diffusion models that do not use temperature-based sampling. Diversity arises from

the stochastic denoising process itself: MDLM samples a random initial sequence from the prior and applies stochastic DDPM updates at each step, while SEDD uses an analytic predictor with Gumbel-max sampling. Each generation starts from a different random noise realization, producing diverse outputs without an explicit temperature parameter.

Table 27: Per-model steering parameters.

	MDLM	SEDD	LLaDA	DREAM
Samples per condition	200	200	200	200
Denoising steps	1,024	1,024	64	256
Sequence length	1,024	1,024	64	64
Batch size	10	12–20	4	8
Temperature	N/A	N/A	0.8	0.7
Remasking strategy	N/A	N/A	low-confidence	entropy
Prompts	None	None	10 neutral	10 neutral
α range	1–20	1–20	1–20	1–20
Seed	42	42	42	42

Table 28: Prompts used for LLaDA and DREAM generation.

#	Prompt
1	"I think this is"
2	"In my opinion, this was"
3	"I would say this is"
4	"Overall, I found this to be"
5	"To be honest, this was"
6	"Many people found this to be"
7	"It turns out that this was"
8	"The experience was"
9	"This was definitely"
10	"I have to say, this is"

C.10 Baseline Methods

We compare SAE steering against three activation-space baselines and one prompt-based baseline.

C.10.1 Activation-Space Baselines

These baselines add $\alpha \cdot \mathbf{v}$ to hidden states at each SAE layer at each denoising step, using the same forward-hook mechanism as SAE steering but without the encode–modify–decode cycle.

Table 29: Activation-space baseline methods.

Method	Uses SAE?	Direction computation
Contrastive vector	No	$\mathbf{v} = \overline{\mathbf{h}}^A - \overline{\mathbf{h}}^B$, unit-normalized. Mean difference in raw activation space.
Probe	No	Weight vector from logistic regression on class labels.
PCA	No	First principal component of contrastive activations, oriented positive→negative.

These baselines use the same attribute combinations and alpha range as SAE steering, with 200 samples per condition. Activations for direction computation are loaded from the Phase 1 extraction cache.

C.10.2 Prompt Instruction Baseline

For LLaDA and DREAM, we additionally evaluate a prompt-based baseline that prepends an explicit attribute instruction to each generation prompt. No activation-space intervention is applied. The instruction prefix is concatenated with each of the 10 standard prompts (Table 28), and 20 samples

are generated per prefixed prompt (200 total), matching the SAE steering setup. Table 30 lists the prefixes for each attribute combination.

This baseline is not applicable to MDLM and SEDD, which are unconditional models that generate by denoising from pure noise with no prompt interface.

Table 30: Instruction prefixes for the prompt baseline (LLaDA and DREAM). Each prefix is prepended to the 10 standard prompts from Table 28.

Combination	**Prefix**
Sentiment	"Write in positive sentiment: "
Topic	"Write about sports: "
Style	"Write informally: "
Sentiment + Topic	"Write in positive sentiment about sports: "
Sentiment + Style	"Write in positive sentiment informally: "
Style + Topic	"Write about sports informally: "
Sentiment + Topic + Style	"Write in positive sentiment about sports informally: "

This baseline tests whether the base model can follow explicit attribute instructions through prompting alone, providing a reference for how much control is achievable without any activation-level intervention.

C.10.3 Step Function Baseline

To isolate the contribution of adaptive scheduling from interval restriction alone (Kynkäänniemi et al., 2024), we implement a *Step Function* baseline that applies uniform-strength steering within a commitment-derived temporal window and zero steering outside it.

Deriving the steering window. For each (layer, attribute) pair, we compute per-block emergence fractions (§D.2), yielding 8 block fractions that sum to 1.0 per layer/attribute. We then find the shortest contiguous block range that captures at least a fraction τ of total commitment mass: for each candidate window length $L = 1, \ldots, 8$ and start position $s = 0, \ldots, 8 - L$, we check whether $\sum_{i=s}^{s+L-1} f_i \geq \tau \cdot \sum_{i=0}^{7} f_i$, and return the shortest window satisfying this constraint. Block indices are converted to timestep bounds: start $= s \times \lfloor T/8 \rfloor$ and end $= (e+1) \times \lfloor T/8 \rfloor - 1$, where T is the total number of denoising steps.

Steering application. Within the derived window $[\text{start}, \text{end}]$, steering is applied at constant strength α at every timestep (identical to Uniform within this range). Outside the window, no steering is applied ($\alpha = 0$). Each layer may have a different window, since block fractions vary across layers. For multi-attribute conditions, each attribute's window is computed independently from that attribute's dynamics; where windows overlap, steering effects compose additively at the residual stream.

Comparison with other methods. The Step Function baseline uses the same commitment-window information as Adaptive but discretizes it to a binary on/off signal rather than continuously weighting by emergence intensity. This isolates whether Adaptive's quality advantage comes from the *shape* of the per-step weights or merely from not steering everywhere.

Threshold selection. We experiment with two threshold values: $\tau = 0.90$ (capturing $\geq 90\%$ of commitment mass) and $\tau = 0.80$ ($\geq 80\%$). Table 31 compares their performance at the best operating point α^* (subject to the same quality gates as Table 37 and Table 38. The 80% threshold yields a tighter window that steers during a more concentrated interval, generally achieving better quality (lower PPL) at comparable or slightly lower confidence. The 90% threshold produces a wider window closer to Uniform, with correspondingly similar quality-control trade-offs. We select $\tau = 0.80$ for the main results as it better isolates the effect of temporal restriction from Adaptive's continuous weighting.

C.11 Evaluation Metrics

For each steering condition, we report the *mean classifier confidence* (mean softmax probability for the target class across 200 samples) using the attribute classifiers listed in Table 21. We report continuous confidence rather than binary classification rate because confidence captures the *degree* of

Table 31: Step Function threshold comparison: $\tau = 0.90$ vs. $\tau = 0.80$. Best operating point α^* under quality gates (PPL < 100 for MDLM/SEDD/DREAM; fluency $\geq 50\%$ baseline for LLaDA). Conf = mean target classifier confidence (%). PPL = Qwen2-7B perplexity.

		$\tau = 0.90$				$\tau = 0.80$			
Model	Combo	α^*	Conf	PPL	Flu	α^*	Conf	PPL	Flu
MDLM	S	12	99.7	94.3	2.04	14	99.5	100.0	2.06
MDLM	T	6	99.9	89.9	2.00	20	100.0	99.3	2.00
MDLM	St	1	41.8	49.6	2.33	1	42.2	49.6	2.33
MDLM	S+T	5	98.3	94.1	2.00	8	99.1	87.9	2.00
MDLM	S+St	4	47.2	82.9	2.06	4	46.0	79.1	2.06
MDLM	St+T	3	73.0	72.3	2.00	2	68.9	52.7	2.05
MDLM	S+T+St	3	77.7	83.5	2.00	3	75.2	70.3	2.00
SEDD	S	18	97.4	99.5	2.04	18	93.4	71.5	2.23
SEDD	T	10	100.0	99.8	1.99	16	100.0	79.7	2.01
SEDD	St	15	67.6	65.8	2.01	18	66.7	52.6	2.04
SEDD	S+T	7	93.5	97.5	2.00	15	92.5	99.3	2.00
SEDD	S+St	14	88.0	93.0	2.00	20	82.1	63.3	2.01
SEDD	St+T	6	85.9	76.5	2.00	9	81.9	72.8	2.00
SEDD	S+T+St	6	89.4	99.9	2.00	10	85.5	91.6	1.97
LLaDA	S	20	93.2	6.3	3.46	19	89.5	6.5	3.44
LLaDA	T	12	99.9	5.8	2.71	13	99.4	5.6	2.70
LLaDA	St	12	50.4	5.9	2.46	12	49.5	5.4	2.62
LLaDA	S+T	12	96.8	5.7	2.77	11	92.8	5.9	2.85
LLaDA	S+St	12	65.6	7.0	2.62	12	63.5	5.0	2.73
LLaDA	St+T	9	66.3	7.3	2.73	9	61.8	5.7	2.83
LLaDA	S+T+St	8	70.6	6.5	2.79	9	70.1	5.1	2.79
DREAM	S	18	100.0	11.0	3.60	18	100.0	11.0	3.59
DREAM	T	11	100.0	48.2	1.01	11	100.0	43.1	1.00
DREAM	St	16	98.5	39.5	1.64	20	98.2	25.4	1.61
DREAM	S+T	5	99.2	22.0	2.46	7	99.8	22.6	2.00
DREAM	S+St	20	98.7	57.4	1.68	20	97.5	42.5	1.51
DREAM	St+T	6	88.5	72.8	1.43	6	88.3	43.0	1.68
DREAM	S+T+St	6	90.4	70.4	1.60	10	91.9	85.5	1.02

steering: a method that shifts sentiment confidence from 30% to 85% is meaningfully stronger than one reaching 55%, even though both may exceed a binary threshold. We measure text quality with Qwen2-7B perplexity (PPL; closer to baseline is better) and LLM-judge fluency (higher is better). We also compute additional surface-level metrics (distinct bigram ratio, within-sample repetition rate) but do not report them in the main results, as they capture only shallow lexical properties and do not fully reflect generation quality. For each steering method, when reporting results, we select the best operating-point α that maximizes the mean confidence for the target attribute (single-attribute) or mean confidence across steered attributes (multi-attribute), subject to two quality gates: LLM-judge fluency $\geq 50\%$ of baseline and $\mathrm{PPL} < 100$.

C.11.1 LLM-AS-JUDGE EVALUATION

Following recent SAE steering work (Goyal et al., 2025; Wu et al., 2025), we use an LLM-as-judge to evaluate both text quality (fluency) and target attribute presence, providing a scalable complement to automatic metrics. All judgments are produced by Claude Sonnet 4.6.

Fluency scoring. Each generated sample is scored on a 5-point Likert scale using the following prompt:

> *Rate the fluency of the following text on a scale of 1 to 5.*
>
> *Scoring guide:*
> *1 = Incoherent gibberish, random word sequences, no grammatical structure, OR empty / near-empty / whitespace-only / punctuation-only output*
> *2 = Mostly incoherent, occasional recognizable phrases but no logical flow*
> *3 = Partially fluent, some grammatical sentences but awkward or repetitive*
> *4 = Mostly fluent, natural-sounding with minor issues*
> *5 = Fully fluent, reads like natural human-written text*

Attribute classification. Each sample is also classified on three attributes (sentiment, topic, formality) using a single prompt designed to handle degenerate text (repetitive or broken):

> *Classify the following text on three attributes. The text may be repetitive, grammatically broken, or partially incoherent—that is FINE. If a repeated or broken phrase still clearly carries an attribute, classify it by that attribute. For example: "match match match" is*

about sports; a repeated casual/slang phrase is informal; "terrible terrible terrible" is negative sentiment. Judge from the content and word choices present, even if the text is not well-formed.

Answer "none" for an attribute ONLY when the text is so degenerate—empty, only punctuation/symbols, or random unconnected tokens with no interpretable content—that the attribute genuinely cannot be determined. Do NOT use "none" just because the text is repetitive or broken.

1. SENTIMENT: positive sentiment (good, great, love, enjoy, excellent, happy) or negative sentiment (bad, terrible, hate, awful, boring, worst)?
Answer: "positive", "negative", or "none"

2. TOPIC: does the text relate to sports (games, teams, players, scores, competitions, leagues, championships)?
Answer: "sports", "not_sports", or "none"

3. FORMALITY: classify the register. Use a STRICT bar for informal.
"formal" = standard written English: complete, grammatical sentences with standard spelling and neutral/professional vocabulary. Conversational or first-person prose (e.g. "That is a great question. I have been thinking about it...") is still FORMAL as long as it is standard, grammatical English—do NOT call it informal merely for being conversational or opinionated.
"informal" = ONLY when the text has overt informal markers: slang or texting style (e.g. "gonna", "wanna", "lol", "omg", "u", "ur"), emoticons/emoji, non-standard spelling/grammar, profanity, or heavy casual interjections ("yeah", "haha", "dude", "kinda").
Answer: "formal", "informal", or "none"

C.11.2 LLM-JUDGE META EVALUATION

We validate the LLM judge (Claude Sonnet 4.6) against human annotations on a stratified sample of 100 generations.

Sample selection. From the pool of LLM-judge annotations across all four models and steering configurations, we drew a stratified sample of 100 generations, balanced by model (MDLM/SEDD/LLaDA/DREAM) and LLM-judge-predicted fluency bin (1–5) to give coverage across models and quality levels.

Annotation procedure. Two authors independently annotated all 100 samples on the same four dimensions the LLM judge scores: sentiment, topic (sports vs. not-sports), formality, and fluency (1–5). For each (sample, dimension) pair, the annotator marked whether they agreed with the LLM-judge prediction.

Inter-annotator agreement. Table 32 reports agreement between the two human annotators. Agreement is high on objective categorical dimensions (sentiment, topic) and lower on subjective ones (formality, fluency). The sentiment κ is near zero despite 98% raw agreement due to the κ paradox: when both annotators mark the same label on nearly every sample, chance agreement is also near 100%, driving κ toward zero; raw agreement is more informative in this regime.

Table 32: Inter-annotator agreement (two authors, 100 samples).

Dimension	Agreement (%)	Cohen's κ	Strength
Sentiment	98.0	$-0.01^{\dagger}$	Near-perfect (raw)
Topic	95.0	0.43	Moderate
Formality	87.0	0.31	Fair
Fluency	66.0	0.24	Fair

$^{\dagger}\kappa$ paradox: near-perfect raw agreement with extreme base rate yields $\kappa \approx 0$.

Tie-break resolution. For the 47 samples on which annotators disagreed on at least one dimension, each disagreement was resolved by discussion, producing a per-(sample, dimension) human consensus verdict.

LLM-judge agreement with human panel. Table 33 reports the LLM judge's agreement with the post-tie-break human consensus. The LLM judge shows strong agreement on objective categorical di-

mensions (sentiment 99%, topic 97%), good agreement on formality (89%), and moderate agreement on the more subjective ordinal fluency dimension (73%).

Table 33: LLM-judge agreement with human consensus (100 samples, post tie-break).

Dimension	**Agreement (%)**	**Strength**
Sentiment	99.0	Strong
Topic	97.0	Strong
Formality	89.0	Good
Fluency	73.0	Moderate

Error analysis. Where the LLM judge disagrees with humans, errors concentrate in specific label bins: formality "informal" (55% agreement vs. 99% for "formal") and fluency middle scores (score 3: 55% agreement), while extreme fluency scores and formal formality are reliable ($\geq$78%). These results justify using LLM-judge scores throughout our evaluation, with fluency treated as a coarser signal than the categorical dimensions.

C.11.3 Metric Correlations

Table 34 reports Spearman rank correlations between quality metrics across all operating points (each datapoint is the mean metric value for one condition–method–α combination). For MDLM, SEDD, and DREAM, Qwen2-7B PPL correlates strongly with LLM-judge fluency ($\rho = -0.77$ to -0.88, all $p < 0.001$), confirming that PPL rankings are reliable for these models. For LLaDA, PPL shows near-zero correlation with fluency ($\rho = +0.21$), confirming that PPL is unreliable for this model; distinct bigram ratio (d2) instead correlates very strongly with fluency ($\rho = 0.93$), validating its use as a secondary quality indicator.

Table 34: Quality metric correlations (Spearman ρ). Each datapoint is the mean metric value for one (condition, method, α) operating point ($n = 1{,}140$ per model). $^{***}p < 0.001$.

Model	Qwen2 PPL vs. LLMJ	d2 vs. LLMJ
MDLM	-0.775^{***}	-0.058 (ns)
SEDD	-0.774^{***}	$+0.367^{***}$
LLaDA	$+0.210^{***}$	$+0.927^{***}$
DREAM	-0.875^{***}	$+0.447^{***}$

Table 35 reports Spearman correlations between classifier confidence and LLM-judge attribute classification rate, aggregated across all four models. All three attributes show strong agreement ($\rho \geq 0.83$), with topic showing the highest correlation ($\rho = 0.92$) and style the lowest ($\rho = 0.83$), reflecting the inherent subjectivity of formality judgments.

Table 35: Target attribute correlations: classifier confidence vs. LLM-judge attribute rate (Spearman ρ, aggregated across all 4 models, $n = 4{,}560$). All $p < 0.001$.

Attribute	ρ
Sentiment	0.869
Topic	0.918
Style	0.830

D Extended Interpretability Analysis

This section provides the full set of interpretability analyses summarized in Section 3 of the main paper. For each analysis type, we describe the generation methodology, present the complete figures spanning all models and layers, and discuss patterns not covered in the main text. Throughout, the three attributes studied are positive sentiment (IMDB), sports topic (AG News), and style (formality for MDLM/SEDD; informality for LLaDA/DREAM), as detailed in §C.7.1.

D.1 Emergence Trajectories for All Attributes

Methodology. The main paper (Figure 2a) shows emergence trajectories for *topic* features only. Here we present the full 4×3 grid covering all three attributes across all four models. For each model, attribute, and layer, the emergence trajectory is computed as follows:

1. We generate N samples (Table 23) with SAE hooks active at every denoising step, recording per-step activations for the top-20 contrastive features per attribute per layer.
2. We filter to classifier-confident samples (Table 22) and average across retained samples to obtain a per-step mean activation curve $\bar{h}(t)$.
3. We compute absolute deviation from baseline: $|\bar{h}(t) - \bar{h}(0)|$, which captures emergence regardless of whether the mean activation increases or decreases during denoising.
4. We apply moving-average smoothing with model-specific window sizes (MDLM/SEDD: no smoothing for 1024-step trajectories since they are sufficiently smooth; LLaDA: window $= 5$ for 64-step trajectories; DREAM: window $= 9$ for 256-step trajectories) to reduce per-step noise while preserving emergence timing.
5. We normalize each curve to $[0, 1]$ via min-max scaling, so that 0 corresponds to baseline (fully masked) and 1 to full emergence.

Each panel in Figure 6 shows a single model–attribute combination, with one line per SAE layer. This reveals both the *timing* of emergence (how early or late features activate) and the *depth gradient* (whether shallow or deep layers commit first).

D.2 Block Fractions Across All Layers

Methodology. The main paper (Figure 2b) shows block fractions at the deepest layer per model only. Here we present block fractions at *every* SAE layer, revealing how the temporal distribution of emergence shifts with depth. Block fractions are computed as follows:

1. The denoising trajectory is divided into $K{=}8$ equal temporal blocks. For MDLM/SEDD (1024 steps), each block spans 128 steps; for LLaDA (64 steps), 8 steps; for DREAM (256 steps), 32 steps.
2. For each block b, we compute the activation increase from the block's start to its end: $\delta_b = \bar{h}(e_b) - \bar{h}(s_b)$, where s_b and e_b are the first and last steps of block b.
3. Negative deltas are clamped to zero: $\delta_b^+ = \max(0, \delta_b)$. This prevents blocks where activation temporarily decreases (e.g., due to non-monotonic trajectories) from receiving negative weight.
4. Block fractions are computed by normalizing: $f_b = \delta_b^+ / \sum_{b'} \delta_{b'}^+$, so they sum to one.

Figure 7 presents the full results, organized as a 4-row grid (one row per model, one column per layer). The deepest-layer panels correspond to the main paper's Figure 2b.

D.3 Cumulative Effect Size Across Models and Layers

Methodology. The main paper (Figure 3) compares cumulative effect size between MDLM and SEDD at layer 7 only. Here we present cumulative $|d|$ curves across all layers and all four models. For each model, attribute, and layer, the procedure is:

1. We load the full set of d_{SAE} Cohen's d values computed during contrastive feature extraction (§C.7.2).
2. We take the absolute value $|d_j|$ and sort features in descending order.
3. We plot the cumulative sum $\sum_{i=1}^{r} |d_{(i)}|$ versus feature rank r for the top 50 features.

The cumulative $|d|$ curve captures how discriminative signal is distributed across features: a steep initial rise indicates that a few features carry most of the signal (concentrated representation), while a gradual rise indicates distributed encoding across many features. Figure 8 presents the full $4 \times (3\text{–}4)$ grid.

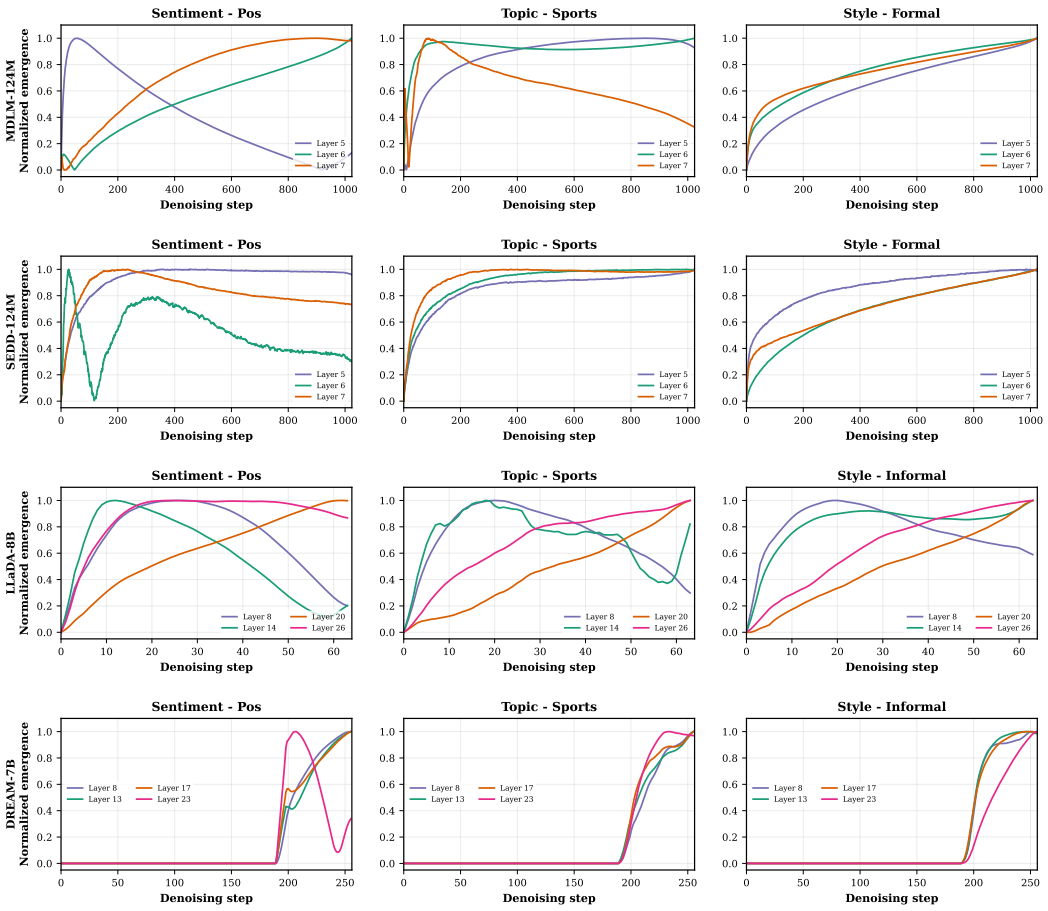


Figure 6: Emergence trajectories for all four models and all three attributes (positive sentiment, sports topic, formal/informal style). Each panel shows normalized absolute deviation $|\bar{h}(t) - \bar{h}(0)|$ from baseline across the full denoising trajectory, with one line per SAE layer. MDLM and SEDD (top two rows) use 1024 denoising steps; LLaDA uses 64 steps; DREAM uses 256 steps. The main paper Figure 2a shows the topic column only.

D.4 Anticipatory vs. Reactive Encoding

Methodology. To understand whether contrastive SAE features encode attribute information *before* a token is revealed (anticipatory) or *after* (reactive), we partition feature activations at each denoising step by mask status. For each step t, each token position is either masked (not yet revealed) or unmasked (already decoded). We compute separate mean activations over the top-20 contrastive features for each pool:

- $\bar{h}^{\text{mask}}(t)$: mean activation across masked positions.
- $\bar{h}^{\text{unmask}}(t)$: mean activation across unmasked positions.

The *masked fraction* is then $\bar{h}^{\text{mask}}(t)/(\bar{h}^{\text{mask}}(t) + \bar{h}^{\text{unmask}}(t))$. Values above 0.5 indicate *anticipatory* encoding: features fire preferentially on positions where no token has been revealed yet, suggesting the model is "planning ahead" for what will appear there. Values below 0.5 indicate *reactive* encoding: features respond to tokens that have already been placed. Smoothing is applied with the same model-specific window sizes as for emergence curves (MDLM/SEDD: window $= 31$; LLaDA: window $= 5$; DREAM: window $= 9$).

Figure 9 shows the masked fraction across denoising progress (0–100%) for each attribute and layer on all four models.

Discussion.

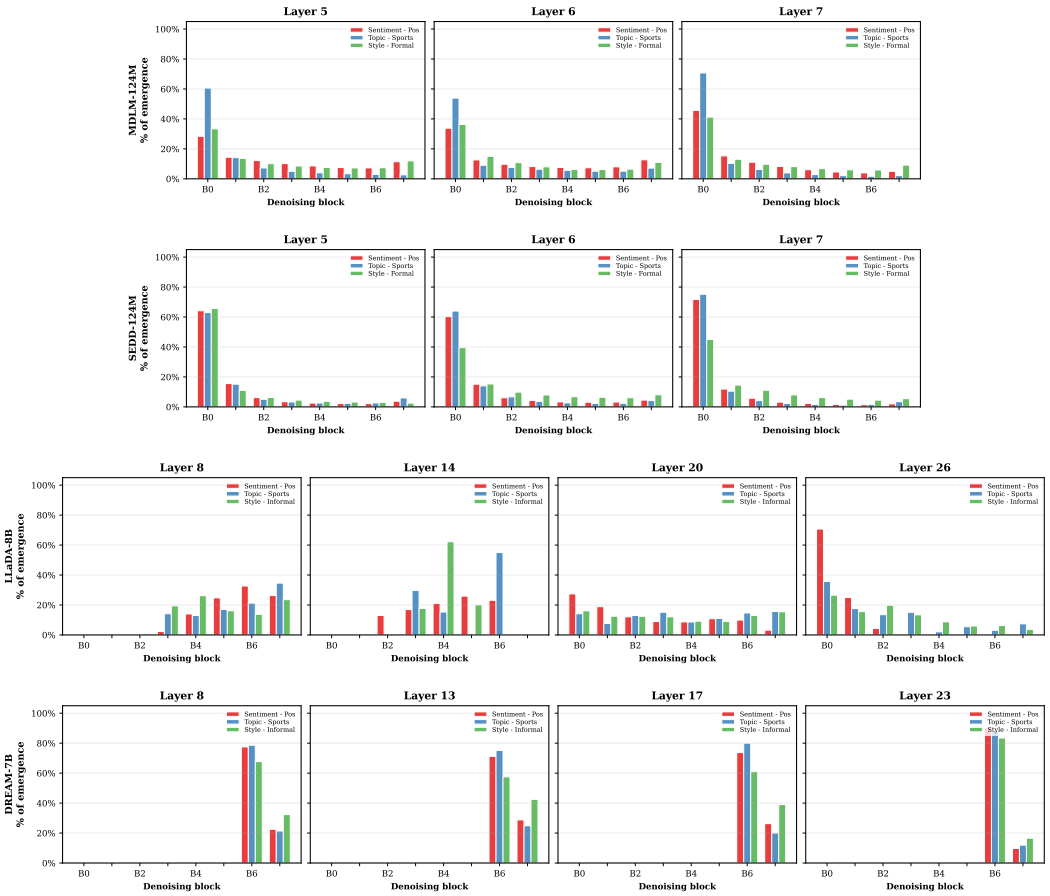


Figure 7: Block fractions across all SAE layers for all four models. Each row corresponds to a model; each column to a layer. Comparing columns within a row reveals how the temporal distribution of emergence shifts with layer depth. MDLM and SEDD (rows 1–2, 3 layers each) show clear depth gradients; LLaDA (row 3, 4 layers) shows mild variation; DREAM (row 4, 4 layers) is uniformly late across all layers.

- **MDLM.** Topic is consistently anticipatory across all layers ($\sim$0.55–0.80 early on), with the strongest signal at layer 7. Sentiment is strongly reactive at layer 5 (starting at $\sim$0.10) and progressively less so at deeper layers; layer 7 shows anticipatory sentiment early ($\sim$0.65). Formality exhibits a similar depth gradient: reactive at layer 5, mixed at deeper layers.
- **SEDD.** Qualitatively similar to MDLM—topic is anticipatory, sentiment is reactive at shallow layers—but with noisier trajectories and compressed dynamic range. This reflects SEDD's score-entropy training objective, which produces smaller absolute SAE activations: the masked fraction ratio remains meaningful (topic is still above 0.5, sentiment below), but the signal-to-noise ratio is lower because both masked and unmasked activations are closer to the ReLU threshold.
- **LLaDA.** Topic is anticipatory at layers 8–14, then transitions to reactive at deeper layers. Sentiment and informality are mildly reactive across most layers. Layer 26 shows anticipatory sentiment.
- **DREAM.** All attributes remain near 0.5 for the first $\sim$75% of denoising, consistent with the absence of feature emergence during this phase (§3.2). Late-stage differentiation is minimal, with slight reactive shifts for topic at layers 8 and 23.
- **Convergence toward equilibrium.** All anticipatory signals decay toward 0.5 as denoising progresses and the mask ratio drops. This is expected: as most positions become unmasked, the distinction between masked and unmasked pools diminishes.

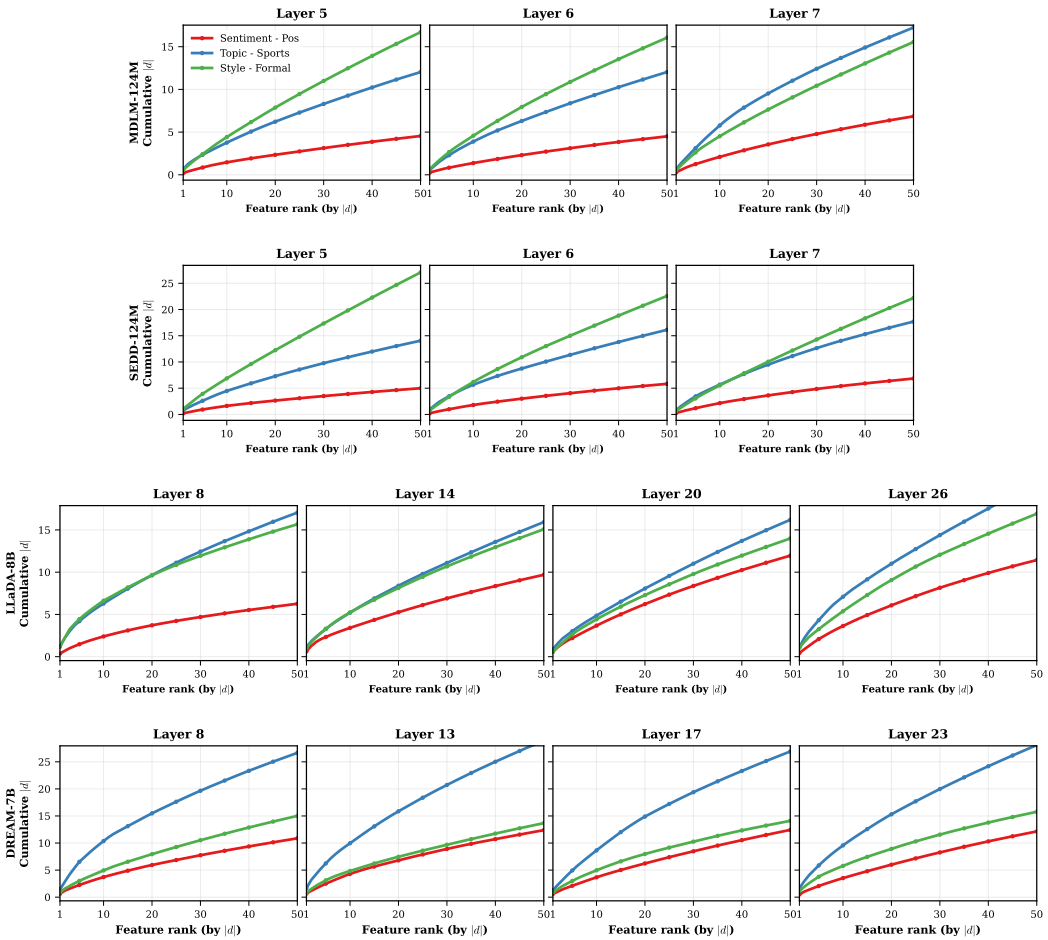


Figure 8: Cumulative |Cohen's d| vs. feature rank for all four models and all SAE layers. Each row is a model (MDLM-124M, SEDD-124M, LLaDA-8B, DREAM-7B); each column is a layer. Features are sorted by $|d|$ in descending order and the top 50 are shown. Topic features (blue) consistently dominate, accumulating the most signal per feature rank. Attributes: positive sentiment (red), sports topic (blue), and formal/informal style (green).

The anticipatory–reactive distinction has implications for steering: anticipatory features are active early and on positions that will be filled later, suggesting that intervening on these features during early denoising steps can influence the content of tokens not yet decoded. Reactive features, by contrast, are most active on already-visible tokens, making them suitable for reinforcing or modifying existing content in later denoising stages.

D.5 Cross-Model Feature Cosine Similarity

This section details the methodology and full results behind the cross-model feature alignment analysis. The goal is to test whether models that share a training objective learn geometrically aligned SAE features, even when they differ in architecture and scale.

Comparison pairs. We compare three model pairs that isolate the effect of training objective from architecture and scale:

(a) **MDLM** $\leftrightarrow$ **SEDD** (124M $\leftrightarrow$ 124M): same GPT-2 architecture, same OpenWebText training data, *different* training objectives (absorbing diffusion vs. score-entropy diffusion).

(b) **MDLM** $\leftrightarrow$ **Dream** (124M $\leftrightarrow$ 7B): *same* absorbing-diffusion loss family, different architectures (GPT-2 vs. Qwen2.5), $56\times$ scale difference.

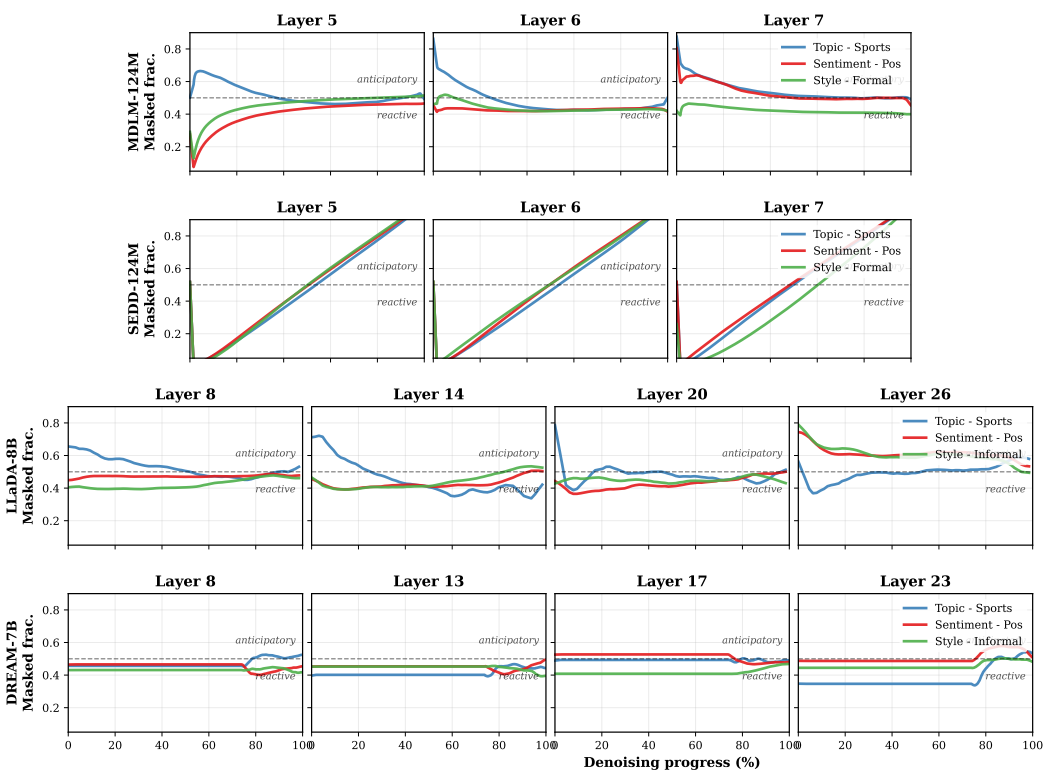


Figure 9: Masked fraction of SAE feature activation vs. denoising progress for all four models and layers. Values above the dashed line (0.5) indicate *anticipatory* encoding; values below indicate *reactive* encoding. Each row is a model; each column is a layer. Attributes studied: positive sentiment (red), sports topic (blue), and formal/informal style (green). SEDD shows qualitatively similar patterns to MDLM but with compressed dynamic range due to smaller absolute SAE activations from score-entropy training (see text).

(c) **MDLM ↔ LLaDA** (124M ↔ 8B): *same* absorbing-diffusion loss family, different architectures (GPT-2 vs. LLaMA-based), $64\times$ scale difference.

For each pair, we select the top-50 features by |Cohen's d| for sentiment and topic attributes from each model's SAE.

Method 1: Direct decoder-space comparison (same hidden dimension). When both models share the same hidden dimension d (MDLM and SEDD, both d=768), we compare decoder directions directly:

$$\text{sim}(f_i^A, f_j^B) = \cos\big(\mathbf{w}_{\text{dec},i}^A,\ \mathbf{w}_{\text{dec},j}^B\big), \tag{18}$$

where $\mathbf{w}_{\text{dec},i} \in \mathbb{R}^d$ is the i-th row of $\mathbf{W}_{\text{dec}}$, ℓ_2-normalized. We compute the full 50×50 similarity matrix across each model's top features and extract, for each feature in model A, its *best-match* similarity $\max_j \text{sim}(f_i^A, f_j^B)$. The "mean best-match similarity" averages these 50 best-match values.

This comparison is valid because both SAEs operate in the same residual-stream vector space: both models share the GPT-2 architecture and the SAEs are trained on activations at corresponding layers (L5, L6, L7).

Method 2: Vocabulary-projected comparison (different hidden dimensions). When hidden dimensions differ (MDLM d=768 vs. Dream d=3584 or LLaDA d=4096), direct decoder comparison is not possible. Instead, we project each decoder direction through the model's unembedding matrix

$\mathbf{W}_{\text{unembed}}$ into a shared vocabulary space:

$$\mathbf{v}_i^A = \frac{\hat{\mathbf{w}}_{\text{dec},i}^A \cdot \mathbf{W}_{\text{unembed}}^{A\top}}{\left\| \hat{\mathbf{w}}_{\text{dec},i}^A \cdot \mathbf{W}_{\text{unembed}}^{A\top} \right\|}, \qquad \text{sim}(f_i^A, f_j^B) = \mathbf{v}_i^A \cdot \mathbf{v}_j^B, \tag{19}$$

where $\hat{\mathbf{w}}_{\text{dec},i}$ denotes the ℓ_2-normalized decoder row and $\mathbf{v}_i^A \in \mathbb{R}^{|V_{\text{shared}}|}$ is the projected, renormalized vector restricted to shared vocabulary tokens. The shared vocabulary consists of all tokens present in both models' tokenizers ($|V_{\text{shared}}| = 42{,}284$ for MDLM–Dream; $42{,}450$ for MDLM–LLaDA).

This comparison is inherently *asymmetric*: the best match for MDLM feature f_i among Dream's features ($A \to B$ direction) need not coincide with the best match in the reverse direction ($B \to A$). We report both directions separately and average them for the summary statistics.

Layer alignment. For each model and attribute, we select the layer with the highest mean |Cohen's d| among the top-50 features:

- Sentiment: MDLM L7 ↔ Dream L17 ↔ LLaDA L20 ↔ SEDD L7.
- Topic: MDLM L7 ↔ Dream L13 ↔ LLaDA L26 ↔ SEDD L7.
- Formality: MDLM L5 ↔ Dream L23 ↔ LLaDA L26 ↔ SEDD L5.

For MDLM ↔ SEDD, both models share the same architecture, so we compare at matched layers directly.

Full results. Results are shown in Table 36.

Table 36: Cross-model feature alignment. *Best-match sim.* is the mean cosine similarity between each feature in model A and its nearest neighbour in model B. For cross-architecture pairs, both search directions ($A \to B$ and $B \to A$) are shown. *Overall mean* is the mean of the full 50×50 similarity matrix (a random-alignment baseline; expected ≈ 0 for unrelated feature sets).

Pair	Attribute	Best-match ($A \to B$)	Best-match ($B \to A$)	Max sim.	Overall mean
MDLM ↔ SEDD (diff. loss, same arch.)	Sentiment (L5)	0.079		0.122	—
	Topic (L6)	0.085		0.122	—
	Formality (L5)	0.084		0.125	—
MDLM ↔ Dream (same loss, diff. arch.)	Sentiment (L7↔L17)	0.170	0.229	0.410	+0.007
	Topic (L7↔L13)	0.166	0.259	0.414	−0.003
	Formality (L5↔L23)	0.191	0.373	0.539	+0.005
MDLM ↔ LLaDA (same loss, diff. arch.)	Sentiment (L7↔L20)	0.133	0.146	0.335	−0.005
	Topic (L7↔L26)	0.135	0.149	0.369	+0.009
	Formality (L5↔L26)	0.142	0.178	0.471	−0.008

Key observations.

- **Different loss ⇒ orthogonal features.** MDLM and SEDD features show mean best-match cosine of 0.08–0.085 across all attributes—indistinguishable from random alignment in 768-dimensional space (expected random cosine $\approx 1/\sqrt{d} \approx 0.036$, with the slightly higher observed values attributable to both models learning *some* shared structure from the same training data). The maximum pairwise similarity is only 0.125, confirming that no individual feature pair is aligned.
- **Same loss ⇒ substantial alignment despite scale gap.** Same-loss-family pairs show $\sim 2\times$ higher alignment. MDLM–Dream mean best-match similarity averages 0.23 (across both directions and all three attributes), with individual feature pairs reaching 0.54 cosine. MDLM–LLaDA averages 0.15 mean / 0.47 max—lower than Dream, likely reflecting the greater architectural distance (LLaMA-based vs. GPT-2).
- **Directional asymmetry.** Dream→MDLM best-match similarity (0.23, 0.26, 0.37) consistently exceeds MDLM→Dream (0.17, 0.17, 0.19) across sentiment, topic, and formality respectively. This indicates that Dream's larger SAE (K=14,336, d=3,584) learns more specific features—each Dream feature maps cleanly to a single MDLM counterpart—while

Table 37: Single-attribute steering results with standard deviations. Format: mean±std. Superscripts on Conf denote the selected α. Same quality gate and selection criteria as Table 2

	Sentiment (S)			Topic (T)			Style (St)		
	Conf↑	PPL↓	Flu↑	Conf↑	PPL↓	Flu↑	Conf↑	PPL↓	Flu↑
MDLM-124M									
No Steering	27.5±39.2	46.0±10.6	2.32±0.48	20.3±39.9	46.0±10.6	2.32±0.48	54.8±17.6	46.0±10.6	2.32±0.48
Uniform	99.3 ± 6.6^{11}	94.0±24.3	2.07±0.27	99.9 ± 0.0^{4}	97.2±16.3	2.00±0.00	69.4 ± 8.8^{4}	67.8±10.9	2.12±0.32
Step Function	99.5 ± 2.0^{14}	100.0±20.1	2.06±0.23	100.0 ± 0.0^{20}	99.3±18.1	2.00±0.00	69.3 ± 8.8^{4}	66.2±10.5	2.16±0.37
Adaptive	94.8 ± 18.9^{20}	64.6±13.3	2.19±0.39	99.9 ± 0.0^{9}	54.5±15.3	2.03±0.17	72.7 ± 8.5^{17}	60.0±8.7	2.24±0.43
Contrastive Vec.	97.4 ± 11.8^{20}	348.1±113.8	1.52±0.50	68.9 ± 45.7^{11}	296.4±88.8	1.07±0.26	43.3 ± 16.2^{4}	283.3±85.7	1.44±0.50
Probe	74.4 ± 36.2^{17}	474.2±131.6	1.21±0.41	99.9 ± 0.1^{19}	166.0±30.7	1.27±0.45	66.4 ± 11.3^{13}	747.5±212.8	1.16±0.37
PCA	20.7 ± 36.3^{14}	345.0±114.7	1.45±0.50	15.7 ± 36.1^{4}	268.0±75.0	1.34±0.47	42.8 ± 16.6^{5}	287.1±82.1	1.37±0.48
SEDD-124M									
No Steering	23.4±37.6	48.1±22.2	2.23±0.42	19.1±39.2	48.1±22.2	2.23±0.42	57.6±17.1	48.1±22.2	2.23±0.42
Uniform	97.2 ± 13.7^{11}	94.1±31.4	2.02±0.14	100.0 ± 0.0^{4}	65.8±19.2	1.98±0.16	96.0 ± 3.9^{19}	91.4±14.8	1.99±0.12
Step Function	93.4 ± 18.4^{18}	71.5±25.2	2.23±0.42	100.0 ± 0.0^{16}	79.7±15.2	2.01±0.10	96.2 ± 3.6^{20}	51.8±7.8	2.06±0.23
Adaptive	76.7 ± 37.6^{19}	59.3±24.1	2.27±0.45	100.0 ± 0.0^{9}	44.7±12.8	2.02±0.12	90.5 ± 7.2^{19}	42.6±7.0	2.22±0.41
Contrastive Vec.	99.8 ± 0.3^{19}	83.4±22.5	2.02±0.16	99.9 ± 0.0^{19}	59.4±15.2	2.08±0.31	91.7 ± 8.5^{20}	77.2±12.7	2.27±0.45
Probe	95.9 ± 16.0^{9}	166.5±49.9	1.88±0.32	87.8 ± 31.9^{7}	78.6±27.9	1.93±0.32	93.8 ± 6.9^{15}	163.9±38.5	1.76±0.43
PCA	33.3 ± 43.1^{5}	51.9±12.8	2.13±0.36	32.9 ± 46.5^{20}	41.8±10.8	2.29±0.48	56.7 ± 18.4^{1}	42.9±11.5	2.33±0.48
LLaDA-8B									
No Steering	69.1±44.9	7.1±5.5	3.56±0.86	13.9±31.4	7.1±5.5	3.56±0.86	22.2±23.0	7.1±5.5	3.56±0.86
Uniform	97.7 ± 13.7^{20}	6.4±3.4	3.39±0.66	99.9 ± 0.1^{15}	5.8±2.1	2.52±0.50	87.4 ± 6.1^{19}	93.3±1256.0	1.78±0.47
Step Function	89.5 ± 29.2^{19}	6.5±5.2	3.44±0.73	99.9 ± 0.4^{17}	4.7±1.6	2.54±0.52	85.4 ± 8.7^{19}	4.0±5.8	1.94±0.47
Adaptive	85.4 ± 34.2^{20}	6.3±3.3	3.52±0.81	99.3 ± 7.1^{18}	6.0±3.8	2.94±0.55	69.8 ± 25.3^{20}	4.3±2.6	2.46±0.70
Contrastive Vec.	91.0 ± 27.2^{19}	6.1±3.2	3.56±0.80	28.6 ± 41.9^{19}	6.9±6.2	3.42±0.76	38.8 ± 27.1^{19}	7.1±6.9	3.25±0.81
Probe	81.9 ± 36.7^{19}	7.5±7.7	3.62±0.84	24.6 ± 39.4^{20}	7.2±6.4	3.45±0.85	28.3 ± 25.7^{19}	8.5±23.4	3.59±0.74
PCA	73.2 ± 42.7^{8}	7.5±8.1	3.67±0.93	15.3 ± 31.9^{18}	6.6±4.3	3.62±0.73	36.5 ± 26.8^{19}	10.0±24.3	3.34±0.88
Prompt	80.9±36.5	7.2±4.5	3.33±0.80	64.9±45.6	6.0±3.9	3.40±0.78	45.1±24.9	11.5±33.2	2.71±0.82
DREAM-7B									
No Steering	52.2±48.0	8.6±3.4	3.75±0.68	13.3±29.1	8.6±3.4	3.75±0.68	26.9±22.7	8.6±3.4	3.75±0.68
Uniform	100.0 ± 0.0^{19}	12.9±4.4	3.44±0.65	100.0 ± 0.0^{11}	49.2±16.6	1.01±0.07	98.5 ± 1.3^{19}	45.3±18.5	1.70±0.46
Step Function	100.0 ± 0.0^{18}	11.0±3.7	3.59±0.66	100.0 ± 0.0^{11}	43.1±15.1	1.00±0.00	97.7 ± 2.2^{18}	25.9±11.3	1.66±0.47
Adaptive	99.5 ± 6.8^{16}	10.0±3.7	3.67±0.67	100.0 ± 0.0^{12}	27.0±30.4	1.19±0.39	97.0 ± 3.0^{17}	25.4±12.8	1.76±0.42
Contrastive Vec.	98.0 ± 12.3^{17}	10.3±3.9	3.63±0.66	95.7 ± 18.2^{19}	15.2±5.8	2.95±0.61	92.8 ± 7.1^{9}	11.0±8.5	2.31±0.77
Probe	90.8 ± 26.7^{19}	10.0±3.6	3.37±0.66	80.7 ± 36.1^{18}	10.8±4.7	3.17±0.67	98.5 ± 3.1^{14}	18.9±16.4	1.84±0.72
PCA	65.6 ± 45.1^{16}	7.2±2.9	3.79±0.64	38.1 ± 42.4^{20}	64.6±37.6	1.08±0.29	31.3 ± 23.5^{20}	9.0±3.2	3.66±0.67
Prompt	80.3±36.6	7.1±2.8	3.50±0.66	61.5±47.1	8.2±2.6	3.73±0.59	20.6±21.5	7.9±3.2	3.48±0.75

MDLM's smaller SAE (K=12,288, d=768) produces more diffuse features that partially match multiple Dream features. For MDLM–LLaDA, the asymmetry is smaller ($\Delta \leq 0.04$), suggesting that LLaDA's features, while in a higher-dimensional space, are not as sharply specialized as Dream's.

- **Overall mean ≈ 0 confirms no systematic bias.** The overall mean similarity (averaging all 2,500 entries in the 50×50 matrix) is effectively zero for all cross-architecture pairs ($|\text{mean}| < 0.01$), confirming that the high best-match values reflect genuine feature-level correspondence rather than a global bias in the projection space.

- **Formality > topic > sentiment in alignment strength.** Across both cross-architecture pairs, formality features show the highest best-match similarity (MDLM–Dream: 0.28; MDLM–LLaDA: 0.16), followed by topic (0.21; 0.14) and sentiment (0.20; 0.14). This ordering is consistent with the degree of lexical grounding: formality features project onto tight vocabulary clusters of register markers (e.g., contractions, punctuation), topic features onto domain-specific terms (e.g., "touchdowns", "quarterback"), while sentiment features rely on subtler evaluative vocabulary that is less likely to be shared across models.

E Steering Analysis and Ablations

E.1 Steering Results with Standard Deviations

The steering results with standard deviations for single attributes are shown in Table 37 and results with standard deviations for multi-attributes are shown in Table 38. The best operating-point α for each method is indicated as a superscript.

Table 38: Multi-attribute steering results with per-attribute confidence. S: sentiment confidence; T: topic confidence; St: style confidence (formal for MDLM/SEDD, informal for LLaDA/DREAM). Superscripts on the first attribute denote selected α. Same quality gate and selection criteria as Table 2.

	S+T				S+St				St+T				S+T+St				
	S↑	T↑	PPL↓	Flu↑	S↑	St↑	PPL↓	Flu↑	St↑	T↑	PPL↓	Flu↑	S↑	T↑	St↑	PPL↓	Flu↑
MDLM-124M																	
No Steering	27.5±39.2	20.3±39.9	46.0±10.6	2.32±0.48	27.5±39.2	54.8±17.6	46.0±10.6	2.32±0.48	54.8±17.6	20.3±39.9	46.0±10.6	2.32±0.48	27.5±39.2	20.3±39.9	54.8±17.6	46.0±10.6	2.32±0.48
Uniform	97.1±10.7^{3}	99.9±0.0	89.3±16.6	1.99±0.10	32.3±39.1^{4}	68.1±8.5	84.9±13.0	2.02±0.16	49.5±5.3^{3}	99.9±0.0	91.8±14.7	2.00±0.07	89.3±24.4^{2}	99.9±0.0	49.6±5.7	75.9±14.4	1.99±0.10
Step Function	98.3±6.1^{8}	100.0±0.0	87.9±15.0	2.00±0.00	24.2±36.5^{4}	67.9±8.7	79.1±12.3	2.06±0.24	55.8±6.2^{5}	99.9±0.6	93.2±15.4	1.98±0.14	73.0±37.7^{3}	99.5±7.0	53.2±7.2	70.3±12.6	2.00±0.07
Adaptive	96.1±14.0^{18}	100.0±0.0	68.8±15.2	2.01±0.10	37.1±42.6^{20}	74.0±10.1	81.2±12.1	2.04±0.21	56.2±6.8^{19}	99.9±0.0	67.3±11.9	2.00±0.10	85.1±29.7^{14}	99.9±0.0	56.5±6.6	65.5±11.8	2.00±0.10
Uniform+E	99.5±1.9^{9}	100.0±0.0	94.9±18.9	2.01±0.10	78.4±35.6^{8}	65.4±11.4	97.8±15.9	2.01±0.10	51.7±6.9^{3}	99.7±3.1	73.5±13.5	2.00±0.07	93.1±19.5^{6}	98.4±12.2	51.7±7.2	85.8±16.3	2.00±0.00
Adaptive+E	94.2±18.9^{20}	96.5±18.4	58.8±11.8	2.06±0.25	74.6±37.8^{18}	62.7±12.8	62.8±11.7	2.15±0.36	58.0±6.8^{20}	99.9±0.0	67.3±12.5	2.00±0.00	91.2±23.3^{20}	97.9±12.9	54.6±9.0	61.2±11.5	2.03±0.17
Contrastive Vec.	82.6±31.0^{20}	76.5±41.6	336.6±88.3	1.08±0.28	97.0±13.6^{20}	72.9±16.4	427.1±134.5	1.72±0.45	28.4±8.5^{17}	99.9±0.1	289.2±68.3	1.47±0.50	96.5±12.9^{20}	90.0±29.7	55.7±14.2	360.9±99.7	1.33±0.47
Probe	75.3±36.1^{16}	99.9±0.1	251.0±47.5	1.30±0.46	70.0±38.3^{19}	73.0±13.1	600.3±211.0	1.26±0.44	53.0±11.5^{18}	99.9±0.0	593.6±165.9	1.07±0.26	63.2±40.6^{19}	100.0±0.0	52.3±8.9	320.6±147.5	1.14±0.34
PCA	20.1±34.2^{12}	11.2±30.8	257.8±78.0	1.29±0.46	19.9±34.3^{7}	40.7±14.7	300.5±90.9	1.44±0.50	44.3±16.4^{4}	12.8±33.2	270.6±78.2	1.41±0.49	18.6±34.7^{6}	15.4±35.4	39.6±14.5	295.3±93.0	1.43±0.49
SEDD-124M																	
No Steering	23.4±37.6	19.1±39.2	48.1±22.2	2.23±0.42	23.4±37.6	57.6±17.1	48.1±22.2	2.23±0.42	57.6±17.1	19.1±39.2	48.1±22.2	2.23±0.42	23.4±37.6	19.1±39.2	57.6±17.1	48.1±22.2	2.23±0.42
Uniform	93.0±19.3^{4}	100.0±0.0	82.4±26.9	2.00±0.00	79.6±31.1^{9}	96.5±3.9	98.9±17.3	2.00±0.00	71.0±8.6^{5}	100.0±0.0	89.4±20.7	1.99±0.12	93.2±19.5^{4}	100.0±0.0	73.1±8.2	94.4±24.5	1.97±0.17
Step Function	85.0±28.4^{15}	100.0±0.0	99.3±19.3	2.00±0.12	65.7±37.4^{20}	98.5±2.1	63.3±10.2	2.01±0.10	91.0±5.8^{13}	99.3±7.2	98.8±19.6	2.00±0.07	69.2±37.9^{10}	98.5±11.8	88.9±6.5	91.6±21.0	1.97±0.17
Adaptive	90.6±23.3^{20}	100.0±0.0	51.2±12.1	2.01±0.10	60.5±42.1^{19}	95.2±5.1	45.2±7.1	2.41±0.49	79.9±7.7^{20}	100.0±0.0	34.7±7.5	2.00±0.10	91.0±22.0^{19}	100.0±0.0	83.0±7.6	38.8±8.6	2.01±0.10
Uniform+E	95.7±13.8^{8}	100.0±0.0	99.8±45.7	2.00±0.10	94.8±17.9^{11}	93.0±8.6	98.1±39.7	2.04±0.18	66.3±7.3^{6}	100.0±0.0	98.6±21.1	1.98±0.14	96.4±13.6^{8}	100.0±0.0	68.3±9.0	98.5±25.5	1.99±0.10
Adaptive+E	78.6±35.4^{19}	62.6±48.2	59.6±25.3	2.17±0.37	73.9±38.4^{19}	85.2±14.1	60.5±38.6	2.27±0.45	79.9±7.7^{20}	100.0±0.0	34.7±7.5	2.00±0.12	76.9±37.1^{20}	41.8±49.1	77.3±15.4	57.5±22.4	2.14±0.35
Contrastive Vec.	99.0±4.0^{20}	100.0±0.0	78.1±19.7	2.00±0.07	99.7±0.6^{20}	96.4±5.1	74.7±13.5	2.08±0.28	68.2±13.3^{20}	99.5±6.8	79.8±20.5	2.00±0.07	97.2±12.8^{20}	85.5±34.9	79.8±11.9	77.4±17.6	2.02±0.14
Probe	91.6±25.9^{7}	78.6±39.8	89.1±27.0	1.89±0.31	94.7±17.0^{10}	91.5±11.7	177.4±55.6	1.94±0.25	57.2±18.9^{10}	93.4±24.5	122.6±37.2	1.84±0.38	93.5±20.7^{11}	83.1±36.8	77.9±18.6	131.3±44.7	1.64±0.48
PCA	28.3±40.5^{5}	21.9±40.7	49.3±12.7	2.19±0.40	59.7±43.2^{19}	80.4±14.5	70.9±18.5	2.21±0.40	54.5±18.5^{6}	27.2±44.0	40.0±10.2	2.33±0.49	41.4±44.1^{17}	33.3±46.6	61.7±16.3	44.2±11.3	2.29±0.47
LLaDA-8B																	
No Steering	69.1±44.9	13.9±31.4	7.1±5.5	3.56±0.86	69.1±44.9	22.2±23.0	7.1±5.5	3.56±0.86	13.9±31.4	22.2±23.0	7.1±5.5	3.56±0.86	69.1±44.9	13.9±31.4	22.2±23.0	7.1±5.5	3.56±0.86
Uniform	94.7±20.8^{12}	99.9±0.3	5.7±2.5	2.78±0.41	74.0±42.5^{17}	78.1±15.1	4.2±1.5	2.29±0.54	99.0±9.9^{20}	65.8±17.4	5.6±2.3	1.85±0.39	78.8±35.8^{20}	98.4±11.9	65.2±17.3	5.8±2.2	1.88±0.36
Step Function	87.2±31.3^{11}	98.4±10.7	5.9±5.6	2.85±0.40	72.2±43.8^{17}	77.4±16.4	4.1±1.5	2.26±0.51	99.5±4.4^{19}	60.6±19.5	3.6±1.0	2.06±0.34	73.5±38.4^{20}	98.6±11.2	59.3±20.8	3.8±1.1	2.03±0.31
Adaptive	87.0±30.9^{15}	97.0±15.9	6.8±4.3	3.10±0.61	72.0±43.5^{20}	58.7±29.9	17.8±131.9	2.71±0.64	94.4±20.6^{20}	55.1±24.6	4.8±3.6	2.48±0.56	77.1±39.7^{18}	93.8±22.6	50.6±28.0	5.5±5.0	2.69±0.54
Uniform+E	93.5±22.8^{15}	99.4±7.1	6.0±2.6	2.81±0.43	48.4±47.8^{17}	82.5±13.3	100.3±1259.0	2.06±0.40	80.5±34.2^{19}	71.5±18.5	4.6±1.8	2.00±0.38	70.1±43.6^{15}	82.9±35.5	61.9±25.6	4.3±1.6	2.31±0.51
Adaptive+E	88.9±30.1^{20}	83.8±35.5	6.8±7.1	3.40±0.71	67.6±45.2^{20}	59.7±30.1	18.0±131.9	2.62±0.62	53.0±47.6^{20}	54.1±29.7	9.1±34.7	2.71±0.61	76.8±40.2^{18}	56.7±47.4	51.4±26.4	6.3±9.3	2.75±0.56
Contrastive Vec.	88.1±30.5^{17}	27.2±41.1	6.3±5.0	3.62±0.77	83.9±34.0^{17}	33.2±26.5	8.1±19.4	3.40±0.90	22.5±36.7^{20}	35.9±26.6	13.3±59.8	3.43±0.78	87.7±31.4^{20}	22.6±37.9	30.7±26.5	7.5±11.8	3.48±0.75
Probe	78.5±39.6^{20}	22.9±37.7	6.7±4.8	3.58±0.82	82.6±35.6^{14}	31.5±24.8	6.7±3.3	3.71±0.76	20.3±36.2^{20}	30.8±27.1	7.1±4.9	3.58±0.83	81.4±37.2^{19}	19.5±36.8	26.8±25.1	7.9±16.5	3.52±0.81
PCA	71.4±42.8^{19}	15.4±30.0	6.4±4.0	3.71±0.80	68.7±44.5^{20}	33.5±23.8	6.8±5.4	3.53±0.87	14.9±29.7^{20}	33.7±26.3	6.5±3.9	3.46±0.76	69.8±43.7^{20}	15.6±29.9	33.2±25.9	6.7±5.8	3.59±0.80
Prompt	88.7±29.2	50.5±48.6	8.3±6.9	3.23±0.75	77.6±40.3	19.3±23.9	9.5±8.3	2.83±0.85	74.5±42.0	37.3±24.0	6.7±4.0	3.43±0.87	92.4±25.2	48.2±49.3	21.2±22.8	8.4±6.2	3.25±0.96
DREAM-7B																	
No Steering	52.2±48.0	13.3±29.1	8.6±3.4	3.75±0.68	52.2±48.0	26.9±22.7	8.6±3.4	3.75±0.68	13.3±29.1	26.9±22.7	8.6±3.4	3.75±0.68	52.2±48.0	13.3±29.1	26.9±22.7	8.6±3.4	3.75±0.68
Uniform	99.8±1.1^{5}	99.9±0.0	21.8±8.9	2.50±0.50	100.0±0.0^{20}	96.5±2.9	63.3±20.6	1.55±0.50	100.0±0.0^{7}	84.1±4.6	96.6±28.1	1.12±0.32	95.8±17.4^{6}	100.0±0.0	77.4±9.1	72.7±20.9	1.60±0.49
Step Function	99.7±1.6^{7}	99.9±0.0	22.6±8.7	2.00±0.19	100.0±0.0^{20}	94.9±4.5	42.5±13.1	1.51±0.50	99.5±7.1^{8}	84.3±6.0	83.9±26.2	1.18±0.38	91.3±22.5^{10}	100.0±0.0	84.4±6.8	85.5±22.5	1.02±0.12
Adaptive	98.1±9.6^{14}	100.0±0.0	48.5±18.5	1.24±0.43	100.0±0.0^{20}	93.9±4.3	33.3±12.3	1.92±0.27	100.0±0.0^{15}	88.3±4.2	92.6±30.4	1.07±0.26	95.7±15.6^{16}	100.0±0.0	83.2±5.9	82.5±27.5	1.13±0.34
Uniform+E	100.0±0.0^{11}	100.0±0.0	17.9±6.7	2.81±0.42	100.0±0.0^{20}	93.2±3.9	52.0±19.3	1.79±0.41	99.5±7.1^{15}	94.4±3.9	99.4±30.7	1.04±0.20	99.9±0.2^{19}	100.0±0.0	81.7±8.3	65.1±18.3	1.53±0.50
Adaptive+E	99.9±0.6^{12}	99.9±0.1	13.3±5.3	3.08±0.45	100.0±0.0^{20}	90.4±5.1	28.3±10.2	1.98±0.16	97.5±14.4^{17}	93.2±4.9	62.8±24.6	1.53±0.50	99.4±7.1^{20}	100.0±0.0	81.1±9.9	45.3±15.1	1.58±0.49
Contrastive Vec.	99.9±0.3^{17}	85.1±34.2	13.4±5.4	3.23±0.68	92.7±23.2^{13}	89.1±9.8	9.7±7.4	2.06±0.74	53.6±44.9^{14}	93.0±6.6	10.8±8.0	1.98±0.64	93.1±19.8^{18}	63.0±44.7	89.7±9.0	12.8±9.3	1.66±0.64
Probe	87.6±31.3^{20}	59.2±45.5	12.5±5.2	3.17±0.74	69.7±42.5^{12}	93.5±9.5	16.3±10.4	2.73±0.87	69.6±41.6^{14}	94.1±7.4	17.4±11.2	2.29±0.84	77.0±38.9^{16}	56.7±43.5	95.4±6.5	16.4±10.9	2.67±0.88
PCA	63.2±45.2^{20}	44.3±45.8	7.6±3.8	2.82±0.65	54.0±47.9^{4}	27.1±21.2	8.5±3.4	3.94±0.66	22.8±37.7^{18}	32.7±26.0	7.8±3.5	3.37±0.68	32.0±42.9^{20}	43.3±43.8	60.9±24.1	62.9±46.2	1.11±0.32
Prompt	80.3±37.1	63.2±47.1	8.5±3.6	3.47±0.68	59.6±46.2	22.7±23.3	7.9±3.8	3.21±0.67	62.2±46.6	24.0±23.3	9.3±3.4	3.65±0.62	76.2±40.3	70.6±44.4	26.7±25.1	9.4±4.4	3.30±0.70

E.2 Steering Trade-off Curves

Figures 10–13 present the full steering trade-off curves for all four models with all methods. Each figure shows target classifier confidence (top) and LLM-judge fluency (bottom) across the α sweep for all 7 attribute conditions. The main paper (Figure 4) shows SAE methods only on MDLM.

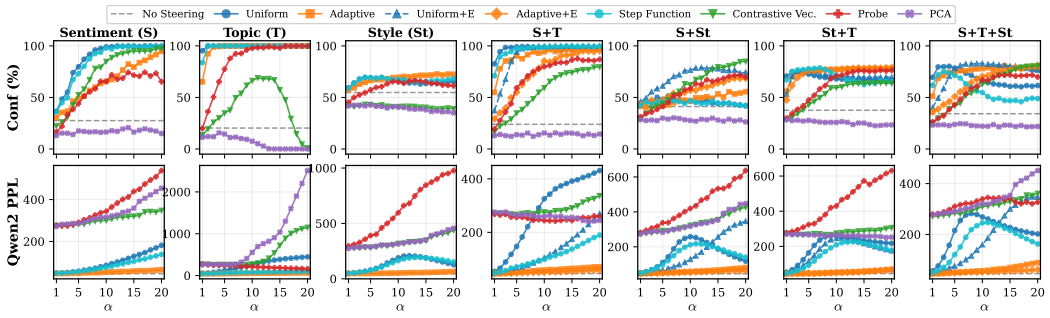


Figure 10: MDLM-124M steering trade-off curves (all methods). Top: target confidence vs. α. Bottom: Qwen2-7B perplexity vs. α. MDLM shows the clearest Adaptive–Uniform separation: Adaptive matches or exceeds Uniform in steering while preserving lower perplexity.

E.3 Statistical Analysis of Steering Quality

We conduct four complementary analyses to quantify the quality advantage of Adaptive steering over Uniform and Step Function baselines. Throughout, we compare the three non-augmented SAE methods: **Uniform**, **Step Function**, and **Adaptive**. We exclude the $+E$ variants (Uniform+E, Adaptive+E) to isolate the effect of the steering schedule. All numbers are computed from per-sample data (200 samples per condition) in the consolidated evaluation files.

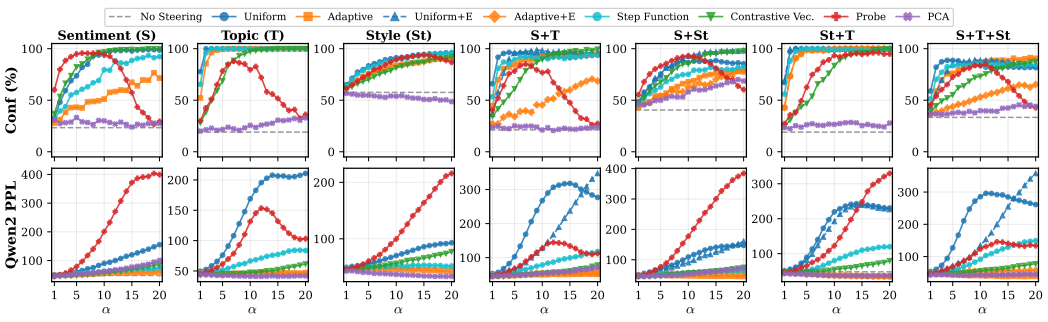


Figure 11: SEDD-124M steering trade-off curves (all methods). Top: target confidence vs. α. Bottom: Qwen2-7B perplexity vs. α. Contrastive Vectors matches SAE methods on sentiment and topic, consistent with SEDD's distributed feature geometry (§3.2).

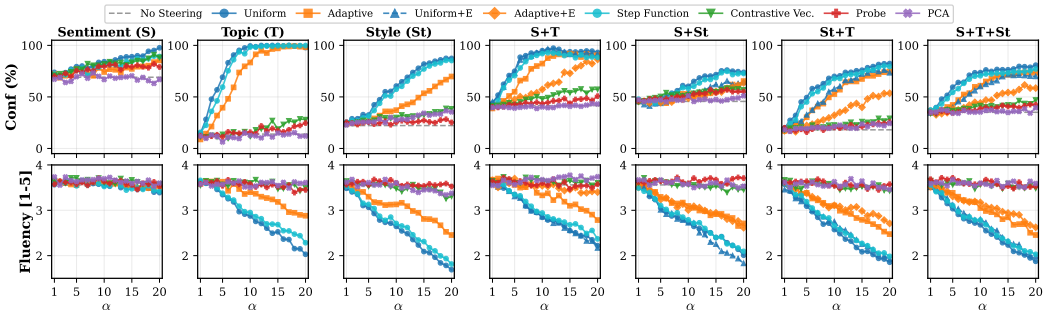


Figure 12: LLaDA-8B steering trade-off curves (all methods). Top: target confidence vs. α. Bottom: LLM-judge fluency vs. α. SAE methods consistently outperform baselines, with the gap largest for topic and multi-attribute conditions.

E.3.1 Significance of Quality Advantage

For each of the 28 conditions (4 models $\times$ 7 attribute combinations), each method selects its best operating point $\alpha^* \in \{1, \ldots, 20\}$ that maximizes target classifier confidence subject to the quality gate described in §5.1 (Qwen2-7B PPL < 100 for MDLM, SEDD, and DREAM; LLM-judge fluency $\geq 50\%$ of baseline for LLaDA). We then compare Adaptive's per-sample quality distribution against the second-best method using a one-sided Mann–Whitney U test ($p < 0.05$). The quality metric is Qwen2-7B perplexity (lower is better) for MDLM, SEDD, and DREAM, and LLM-judge fluency (higher is better) for LLaDA.

Table 39 reports the results. Adaptive achieves significantly better quality than the second-best method in **22 of 28 conditions (78.6%)**: 6/7 for MDLM (86%), 7/7 for SEDD (100%), 6/7 for LLaDA (86%), and 3/7 for DREAM (43%). The non-significant cases are MDLM S+St (where Step Function's PPL is marginally lower), LLaDA Sentiment (where both methods produce high fluency, $p = 0.088$), and four DREAM conditions where aggressive steering at high α degrades PPL.

When using LLM-judge fluency as the quality metric for all models, significance drops to 14/28 (50%) because fluency on small models (MDLM, SEDD) is heavily quantized around integer values, compressing meaningful PPL differences into near-ties.

E.3.2 Matched-α Quality Ordering

We test whether the quality ordering Adaptive $\geq$ Step Function $\geq$ Uniform holds at matched steering strength. For every (model, combo, α) triplet where all three methods have data ($n = 560$ points: 4 models $\times$ 7 combos $\times$ 20 α values), we check whether the ordering holds on both quality metrics independently.

Table 40 reports the results. On LLM-judge fluency, the full ordering holds in **86.4%** of cases, with Adaptive $\geq$ Uniform in 98.0%. Using the model-specific quality metric (Qwen2-7B PPL for

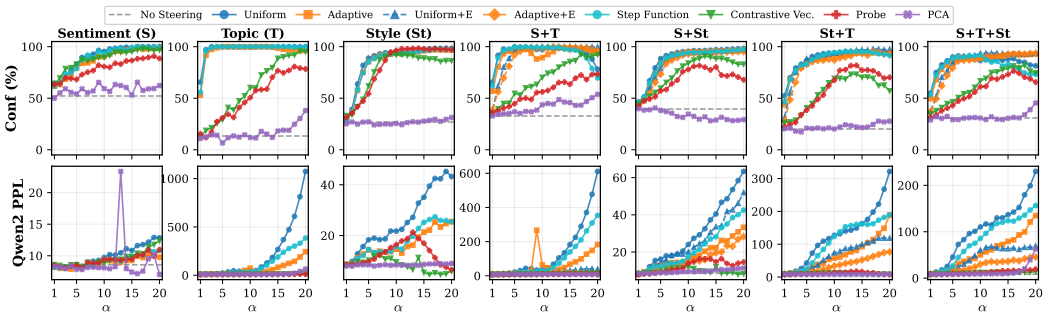


Figure 13: DREAM-7B steering trade-off curves (all methods). Top: target confidence vs. α. Bottom: Qwen2-7B perplexity vs. α. DREAM achieves near-perfect steering at moderate α but exhibits the sharpest quality degradation, consistent with its concentrated late-block emergence.

Table 39: Significance of Adaptive quality advantage at best operating points. For each condition, we compare Adaptive vs. the second-best method (2nd) using a one-sided Mann–Whitney U test. Quality metric: Qwen2-7B PPL for MDLM/SEDD/DREAM (P, lower better), LLM-judge fluency for LLaDA (F, higher better). $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Model	Combo	α_A	Q_A	2nd Best	α_{2nd}	Q_{2nd}	p-value	
MDLM	Sent	20	64.6^P	Uniform	11	94.0^P	<0.001	***
MDLM	Topic	9	54.5^P	Uniform	4	97.2^P	<0.001	***
MDLM	Style	17	60.0^P	Step Fn	4	66.2^P	<0.001	***
MDLM	S+T	18	68.8^P	Step Fn	8	87.9^P	<0.001	***
MDLM	S+St	20	81.2^P	Step Fn	4	79.1^P	0.988	
MDLM	St+T	19	67.3^P	Uniform	3	91.8^P	<0.001	***
MDLM	S+T+St	14	65.5^P	Step Fn	3	70.3^P	<0.001	***
SEDD	Sent	19	59.3^P	Step Fn	18	71.5^P	<0.001	***
SEDD	Topic	9	44.7^P	Uniform	4	65.8^P	<0.001	***
SEDD	Style	19	42.6^P	Step Fn	20	51.8^P	<0.001	***
SEDD	S+T	20	51.2^P	Uniform	4	82.4^P	<0.001	***
SEDD	S+St	19	45.2^P	Step Fn	20	63.3^P	<0.001	***
SEDD	St+T	20	34.7^P	Uniform	5	89.5^P	<0.001	***
SEDD	S+T+St	19	38.8^P	Step Fn	10	91.6^P	<0.001	***
LLaDA	Sent	20	3.52^F	Step Fn	19	3.44^F	0.088	
LLaDA	Topic	18	2.94^F	Step Fn	17	2.54^F	<0.001	***
LLaDA	Style	20	2.46^F	Step Fn	19	1.94^F	<0.001	***
LLaDA	S+T	15	3.10^F	Step Fn	11	2.85^F	<0.001	***
LLaDA	S+St	20	2.71^F	Uniform	17	2.29^F	<0.001	***
LLaDA	St+T	20	2.48^F	Step Fn	19	2.06^F	<0.001	***
LLaDA	S+T+St	18	2.69^F	Step Fn	20	2.03^F	<0.001	***
DREAM	Sent	16	10.0^P	Step Fn	18	11.0^P	<0.001	***
DREAM	Topic	12	27.0^P	Step Fn	11	43.2^P	<0.001	***
DREAM	Style	17	25.4^P	Step Fn	18	25.9^P	0.173	
DREAM	S+T	14	48.5^P	Uniform	5	21.8^P	1.000	
DREAM	S+St	20	33.3^P	Step Fn	20	42.5^P	<0.001	***
DREAM	St+T	15	92.6^P	Step Fn	8	83.9^P	0.999	
DREAM	S+T+St	16	82.5^P	Uniform	6	72.7^P	1.000	

MDLM/SEDD/DREAM, fluency for LLaDA), the full ordering holds in **88.2%** of cases. The ordering is remarkably consistent across models: even the weakest model (DREAM, 79–80%) shows the predicted trend in the vast majority of conditions.

E.3.3 α Sensitivity

We test α sensitivity by fixing one global α^* per (model, method) pair across all 7 conditions, specifically, the α that maximizes mean confidence while passing the quality gate described in §5.1 for all conditions simultaneously, then measuring the per-condition confidence drop relative to the per-condition optimum.

Table 41 reports the mean and maximum drop in percentage points. Adaptive is robust: its global α^* sits at 15–20 and incurs only 0.2–1.2 pp mean drop (maximum 3.2 pp). Uniform is forced to low global α values (2–7 for MDLM/SEDD/DREAM) because higher values violate the quality gate for at least one condition; this low setting under-steers other conditions, producing mean drops of

Table 40: Matched-α quality ordering. Percentage of (model, combo, α) points where the ordering holds. $n = 560$ total (4 models $\times$ 7 combos $\times$ 20 α).

Ordering	LLMJ Fluency		Model-specific	
	Count	%	Count	%
Adaptive $\geq$ Step Fn $\geq$ Uniform	484/560	86.4	494/560	88.2
Adaptive $\geq$ Step Fn	518/560	92.5	532/560	95.0
Step Fn $\geq$ Uniform	524/560	93.6	520/560	92.9
Adaptive $\geq$ Uniform	549/560	98.0	546/560	97.5
Per-model breakdown (LLMJ fluency, full ordering):				
MDLM	134/140	95.7	125/140	89.3
SEDD	122/140	87.1	139/140	99.3
LLaDA	118/140	84.3	118/140	84.3
DREAM	110/140	78.6	112/140	80.0

3.1–9.2 pp and worst-case drops up to 50.9 pp. Step Function falls in between (mean 1.1–5.6 pp, max 4.5–32.2 pp). The asymmetry arises because Adaptive concentrates its intervention where attributes form, so increasing α does not degrade generation quality the way uniform-weight intervention does.

Table 41: α sensitivity: confidence drop (pp) when using a single global α^* across all 7 conditions vs. per-condition optimal α. α^* is the value that maximizes mean confidence subject to the quality gate passing for all conditions.

	Uniform			Step Function			Adaptive		
Model	mean	max	α^*	mean	max	α^*	mean	max	α^*
MDLM	9.1	50.9	2	5.6	32.2	4	0.6	3.2	20
SEDD	9.2	28.4	4	4.6	13.9	10	0.2	0.6	19
LLaDA	1.4	2.9	19	1.1	4.5	20	0.4	1.4	20
DREAM	3.1	9.1	7	3.2	10.8	8	1.2	2.1	15

The rank ordering from Analysis E.3.2 is also stable across α ranges. Table 42 shows that the full quality ordering (Adaptive $\geq$ Step Function $\geq$ Uniform) holds in 75–80% of points even at the lowest α range [1,3] and strengthens to 86–88% at the full [1,20] range, confirming that the Adaptive advantage is not an artifact of specific α values.

Table 42: Rank stability across α ranges. Percentage of matched (model, combo, α) points where the full quality ordering (Adaptive $\geq$ Step $\geq$ Uniform) holds, computed separately for LLMJ fluency and model-specific quality.

α range	LLMJ Fluency		Model-specific	
	Count	%	Count	%
$[1, 3]$	63/84	75.0	67/84	79.8
$[1, 5]$	110/140	78.6	116/140	82.9
$[1, 10]$	235/280	83.9	240/280	85.7
$[1, 15]$	357/420	85.0	369/420	87.9
$[1, 20]$	484/560	86.4	494/560	88.2

E.3.4 Matched-Confidence and Matched-Quality Comparisons

The previous analyses compare methods at the same α, but different methods may achieve different confidence levels at the same α. To control for steerability, we compare quality at *matched target confidence* using an interpolation-based approach.

For each (model, combo) pair, we compute each method's confidence–quality curve across $\alpha \in \{1, \ldots, 20\}$. We identify the overlapping confidence range where all compared methods can operate, then probe at intervals of 2 percentage points within that range. At each probe, we linearly interpolate each method's quality (between adjacent α values) and compare.

Table 43 reports the results. Using LLM-judge fluency as the quality metric, Adaptive produces better quality than Uniform in **84.0%** of matched-confidence probes and better than Step Function in **72.8%**. Using the model-specific metric, the rates are 82.2% and 71.5% respectively. The advantage is consistent across models, with SEDD showing the strongest results (93–100% Adaptive $>$ Uniform).

Table 43: Matched-confidence quality comparison. At each confidence level within the overlap range (probed every 2 pp), the percentage of probes where one method produces better quality than the other.

Comparison	LLMJ Fluency		Model-specific	
	Count	%	Count	%
Adaptive $>$ Uniform	405/482	84.0	396/482	82.2
Adaptive $>$ Step Fn	389/534	72.8	382/534	71.5
Step Fn $>$ Uniform	332/516	64.3	343/516	66.5
Per-model breakdown (Adaptive $>$ Uniform, LLMJ fluency):				
MDLM	62/67	92.5	60/67	89.6
SEDD	104/104	100.0	89/104	85.6
LLaDA	115/157	73.2	115/157	73.2
DREAM	124/154	80.5	132/154	85.7

Matched-quality confidence comparison. We perform the complementary analysis: at matched quality, does one schedule achieve higher confidence? The procedure mirrors the matched-confidence analysis above but along the quality axis. For each (model, condition) pair we sort each method's points by quality, identify the overlapping quality range, and probe at evenly spaced intervals (step size 1.0 for PPL-based models, 0.02 for LLM-judge fluency on LLaDA). At each probed quality level we linearly interpolate each method's confidence and compare.

Using the model-specific quality metric, Adaptive achieves higher confidence than Uniform in **46.7%**, higher than Step Function in **47.5%**, and Step Function achieves higher confidence than Uniform in **44.6%** of matched-quality probes. All values are near 50%, confirming that no method consistently achieves stronger steering at the same quality level. This is expected: Adaptive's advantage lies in quality preservation—at matched confidence it produces better text, but at matched quality it does not reach higher confidence. The approximately even split demonstrates that Uniform's additional quality degradation does not translate into stronger attribute control.

E.4 Feature Count Ablation

We steer sentiment toward positive using $N \in \{1, 2, 5, 10, 20, 50\}$ features per direction with Uniform steering at α=5 and report the positive-direction classifier confidences for all three attributes. Table 44 shows that increasing N raises target sentiment confidence but also increases interference on non-target attributes (topic and style). At N=20, all models achieve strong sentiment control (71–84%) with modest interference; at N=50, SEDD formality shifts from 57.7% baseline to 94.5% and MDLM formality rises to 67%. We therefore use N=20 features throughout.

Table 44: Feature count ablation: positive-direction sentiment steering at α=5, varying the number of features N. Pos = positive sentiment confidence (%), Spt = sports topic confidence (%), Frm = formality confidence (%).

	MDLM-124M			**SEDD-124M**			**DREAM-7B**			**LLaDA-8B**		
N	Pos	Spt	Frm	Pos	Spt	Frm	Pos	Spt	Frm	Pos	Spt	Frm
No Steering	25.5	18.6	53.3	29.0	20.3	57.7	57.7	10.5	73.5	69.1	13.9	77.8
1	41.5	20.0	51.0	29.0	23.0	61.5	69.5	7.5	82.5	73.5	7.0	81.5
2	50.5	23.5	50.0	37.5	26.0	66.5	70.0	8.0	82.0	72.5	9.0	80.0
5	65.0	22.0	46.0	53.5	24.5	77.5	78.0	6.5	85.0	75.0	8.0	83.5
10	75.0	26.5	51.5	65.0	20.5	80.0	80.5	5.0	90.5	79.0	7.0	83.5
20	84.0	25.0	55.5	71.5	20.0	90.0	82.5	4.0	92.5	80.0	3.5	82.0
50	92.0	18.0	67.0	88.5	17.5	94.5	89.5	7.5	91.5	78.0	6.0	84.5

E.5 Cross-Attribute Interference

For each method at its best operating point, we measure cross-attribute interference as the mean absolute percentage-point deviation of non-target attribute classifier confidences from their unsteered baselines. Given m attributes steered simultaneously, the interference for method M is $\frac{1}{|\mathcal{A}\setminus T|}\sum_{a\notin T}|c_a^{(M)} - c_a^{(\text{unsteered})}|$, where c_a is the classifier confidence for attribute a. Figure 14 reports this metric across all models and steering conditions.

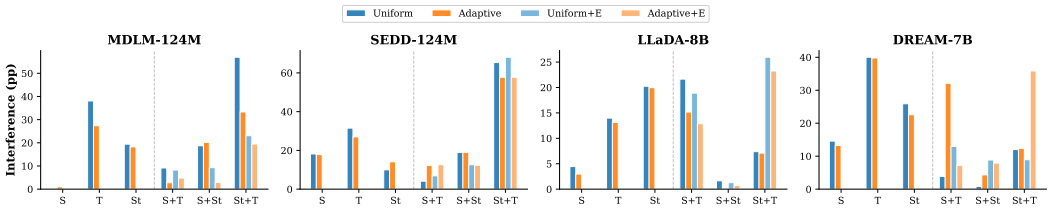


Figure 14: Cross-attribute interference at the best operating points: mean absolute pp deviation of non-target attributes from unsteered baselines. On MDLM, Adaptive (orange) produces substantially less interference than Uniform (blue); on other models the gap narrows or reverses, reflecting model-specific temporal dynamics. E-ratio variants (lighter shades) appear only for multi-attribute conditions.

E.6 Latency Overhead

Table 45 reports per-sample wall-clock times for unsteered generation, SAE steering (uniform schedule, α=5), and contrastive-vector steering, measured over 20 single-sample generations per method after a warm-up pass.

Table 45: Generation latency (seconds per sample, batch size 1). SAE overhead is the ratio of SAE steering time to baseline. CV = contrastive vector.

Model	Steps	Baseline (s)	SAE (s)	SAE Overhead	CV (s)	CV Overhead
MDLM (124M)	1024	25.66	30.07	1.17×	25.73	1.00×
SEDD (124M)	1024	21.49	22.55	1.05×	22.15	1.03×
DREAM (7B)	256	7.52	8.41	1.12×	7.53	1.00×
LLaDA (8B)	64	0.83	1.25	1.51×	0.81	0.97×

SAE steering adds 5–51% overhead depending on model size and step count. The overhead is lowest on SEDD (1.05×), where the 1024-step analytic predictor dominates runtime and the SAE encode–decode ($768 \to 12{,}288 \to 768$) is negligible by comparison. MDLM (1.17×) has a simpler per-step update, so the SAE cost is proportionally larger. For the 7B-scale models, DREAM (1.12×, 256 steps) amortizes the SAE cost across expensive transformer forward passes, while LLaDA (1.51×, 64 steps) has the highest overhead because each of its few steps is fast (∼13ms), making the per-step SAE cost ($4096 \to 16{,}384 \to 4096$ across 4 layers) a larger fraction. Contrastive-vector steering adds negligible overhead (∼1.00×) since it requires only a single vector addition per layer. MDLM was benchmarked on an A10G GPU; all other models on A100.

E.7 Qualitative Steering Examples

Tables 46–49 show representative generations under triple-attribute steering (S+T+St) for each model, all four SAE steering modes, and baselines (Contrastive Vectors and Prompt where applicable). MDLM and SEDD are steered toward positive sentiment, sports topic, and formality; LLaDA and DREAM toward positive sentiment, sports topic, and informality. Texts are truncated; classifier confidences are from the off-the-shelf classifiers described in §C.11.

On MDLM and SEDD, Uniform steering achieves high target confidence but produces garbled text (PPL >250), while Adaptive matches confidence at lower perplexity with coherent output. E-ratio calibration (Uniform+E, Adaptive+E) provides intermediate quality. On LLaDA and DREAM, all SAE modes successfully shift attributes from near-chance baselines while maintaining low perplexity;

Table 46: MDLM qualitative examples (triple-attribute: positive, sports, formal). S/T = sentiment/topic confidence (%); Frm = formality confidence (%).

Method	S	T	Frm	PPL	Generated Text (truncated)
Unsteered	1	0	62	43	Its immigration reform bill, making reference again to immigration law Tuesday night… The disagreements declared on cases on which judges are working are "a case of man's versus a judge, the issues that arise that have to be heard by one court."
Uniform	100	100	47	259	with also the former rivals side and the stars for the replacement and they was make a breakthrough at the end of 2011, moving to the series win which capped a five-match Premier career…
Adaptive	100	100	50	54	"I'm really happy, I'm thinking and looking forward to Southampton," Thomas this week told Sky Sport. "I am always hungry to go home." The Shearsborough stadium won't have any "New England" colours against Coventry…
Uniform+E	100	100	51	101	went back into form, making the Crystal Palace bench on Tuesday night and his hopes of winning the game on Sunday and defeat Manchester United. The former first-time winner and two-time ace culminated on the pre-match session…
Adaptive+E	99	100	40	51	went deeper into future, making it again to take news Tuesday night when his team will win the AFC championship Sunday at Dick Bowl in Spokane. "I'm going there at 1 p.m. tomorrow to try to see what happens," Bowl said.
Contrastive Vec.	100	0	37	329	with also the fan t; and feed through for the fans and you with the fan letters to some of her characters, and then they started doing me and fronted and played us those episodes…

Table 47: SEDD qualitative examples (triple-attribute: positive, sports, formal). S/T = sentiment/topic confidence (%); Frm = formality confidence (%).

Method	S	T	Frm	PPL	Generated Text (truncated)
Unsteered	80	0	30	29	everybody, that's what happens if you know people are really getting the pressure there is you look at no one else they're busy with that. I'm like things keep on happening every day is that it is thrusting into my life…
Uniform	99	100	67	312	complete the third and 2nd all Mexico and scored a one goal that had the European League victories final of ten and scored the final that turned for victory of World title against Brazil USA's Ners Joao…
Adaptive	100	100	61	49	Two games hard-won, Wade again in the fourth quarter and scored again with one tip of Jimmy Wingard's fade-away drive to draw a one minute tie with the Pacers. He scored twice…
Uniform+E	100	100	74	135	great nation is case of it all, although there is one of the most sophisticated players of the sport, Diego Brasil, who was restored for a new level to have with a strong analytical background…
Adaptive+E	99	100	77	102	of Jordi has emerged as the next president in each club and country, the extraordinary result is that Spain has re-established itself… In an elegant football football held to the centre of the Catalan club victory…
Contrastive Vec.	100	100	85	36	in the bottom of the 10th inning was the night when Babe Ruth tied the National League Division scoreless with a three-run solo third, the first time his career led, but fell foul. The Pirates were 2 ahead of the San Francisco Giants…

E-ratio variants achieve the highest simultaneous control across all three attributes. The DREAM unsteered example (formal, technical text with St=1%) contrasts sharply with steered outputs that exhibit clear informality markers (lowercase, casual phrasing, slang).

BROADER IMPACTS

Mechanistic interpretability is a foundational tool for AI safety: models whose internal computations can be traced and steered are easier to audit, correct, and align. Our work extends this agenda to discrete diffusion language models, a class that has received less interpretability attention than autoregressive LLMs but is becoming competitive in capability. The framework supports legiti-

Table 48: LLaDA qualitative examples (triple-attribute: positive, sports, informal). S/T = sentiment/topic confidence (%); Inf = informality confidence (%).

Method	S	T	Inf	PPL	Generated Text (truncated)
Unsteered	25	0	12	6	a great question. I have been thinking about it for a while and I am trying to come up with a solution that would work for me. I have a lot of money and I want to invest it in real estate.
Uniform	1	100	93	6	what you mean by "after the game" lol but not at the end of the game but not at the end of the game but not at the end of the game...
Adaptive	100	83	70	5	a great idea for the game. I think it would be a great way to make the game more fun and competitive by giving the players a big advantage at the start of the game.
Uniform+E	94	100	72	11	a great way to get a kick out of the game. you can get the ball at the start of the game and then after the game you can get it back and throw it in the end zone. i was a hater in the past.
Adaptive+E	100	99	75	3	a great way to get to know the game. I think this is a great way to make friends. I think this is a great way to get to know the game.
Contrastive Vec.	97	100	20	8	a good question. It's always difficult to predict the outcome of a match, but it's also difficult to predict the outcome of a game. I think the important thing is to be prepared for the match...
Prompt	100	2	0	9	a good book for people who are interested in sports. It has a lot of information about different sports and how to play them. I like the way it is written because it is easy to understand...

Table 49: DREAM qualitative examples (triple-attribute: positive, sports, informal). S/T = sentiment/topic confidence (%); Inf = informality confidence (%).

Method	S	T	Inf	PPL	Generated Text (truncated)
Unsteered	0	0	1	14	a good way for beginners to understand that if you have an array that is large enough, you can treat it as a queue. However, when you use an array as a queue, it is easy to reach the capacity of the queue...
Uniform	100	100	77	144	a match in 3 final the all match good there time time and they was great team up all good lost and win a match of the game and go won a win of the game they 2 game...
Adaptive	100	100	76	37	a great match. Its a great match i love it and the match was 20-20 and when the match started was 2-3 they played alot of people thought they could be in the back of the match go 4-5...
Uniform+E	100	100	87	22	a good game, but it needs to be more fun and not so boring it is more of a best ball game. 1st game ever. this game was really fun. playa 11. this game was so great...
Adaptive+E	100	100	84	21	a good game, but it needs to be more fun and not so boring it is more of a basket ball game. 1st game ever. this game was good but i liked the first game better. This game was awesome i liked it it was fun and i like it even though it was hard. Great game
Contrastive Vec.	100	1	67	28	a great way to get kids to think about how many days are in a week! I love the whole concept of how many days in a week and then i can ask them how many days in a week...
Prompt	100	99	65	11	a way to do that. It's kind of a good example of how to do it. It's not a good example of how to do it, but I think it's a great example of how to do it...

mate downstream uses including content moderation, domain adaptation without retraining, and accessibility-oriented control of register or sentiment. The interpretability tools themselves let practitioners diagnose failures and understand where in the denoising trajectory particular attributes commit—information not visible from external behavior alone.

However, the same steering capability that enables beneficial control also enables manipulation: an adversary with white-box access could shift outputs toward biased or deceptive content while preserving fluency, in ways simple prompting cannot. We release no models, datasets, or steering pipelines targeting harmful attributes; trained SAEs and feature sets are intended for research on benign attributes and would need retraining for deployment