AN OPTIMAL TRANSPORT VIEW OF ACTIVATION STEERING IN MASKED DIFFUSION MODELS

Gert Lek[1] Chaoyi Zhu[2] Pin-Yu Chen[3] Robert Birke[4] Lydia Y. Chen[1,2]

[1]University of Neuchâtel
[2]Delft University of Technology
[3]IBM Research
[4]University of Turin
gert.lek@unine.ch

ABSTRACT

Diffusion Large Language Models (dLLMs) offer a non-autoregressive alternative to left-to-right decoding, but inference-time control in dLLMs remains underdeveloped relative to autoregressive LLMs. Prior activation-steering methods for masked diffusion models (MDMs), a prominent dLLM formulation, focus primarily on concept negation and employ heuristics that do not explicitly optimize the transport objective. We introduce an Optimal Transport (OT) view of activation steering for MDMs: given contrastive prompt distributions, we learn a lightweight affine map that transports pooled activation distributions from a source behavior to a target behavior. This perspective unifies common steering rules (activation addition, mean-shift, directional ablation) as special cases of an affine transport map, and motivates the use of the OT estimator that matches first- and second-order moments. Across three state-of-the-art dLLMs (LLaDA-Instruct, LLaDA 1.5, Dream-Instruct), affine OT steering improves instruction-following accuracy (e.g., +6.5 to +11.9 absolute points) with no inference-time overhead.

1 INTRODUCTION

Diffusion Large Language Models (dLLMs) replace left-to-right decoding with diffusion-style denoising (Ghazvininejad et al., 2019; Chang et al., 2022). The model starts from a fully corrupted sequence and denoises until the text becomes coherent. This paradigm supports parallel token unmasking, which can improve latency. This work focuses on Masked Diffusion Models (MDMs) which utilize a masked absorbing and discrete corruption and denoising process (Nie et al., 2025; Ye et al., 2025; Austin et al., 2021). Recent efforts suggest that this approach can be competitive with autoregressive LLMs (Nie et al., 2025; Ye et al., 2025).

A persistent challenge in generative modeling is inference-time control: shifting behavior without expensive backpropagation. In diffusion models, control is often implemented with sampling-time guidance (Dhariwal & Nichol, 2021; Ho & Salimans, 2022). While effective, guidance assumes an explicit conditioning mechanism or auxiliary training setup. A more mechanistic perspective on inference-time control is provided by activation steering, which shows promising results in autoregressive LLMs (Turner et al., 2023; Stolfo et al., 2025) and diffusion models (Rodriguez et al., 2024), but remains unexplored in dLLMs.

Prior work on autoregressive models shows that affine interventions reliably modulate high-level attributes. Activation Addition (Turner et al., 2023) computes a steering direction from paired prompts and adds it to activations at inference time. Instruction-vector methods (Stolfo et al., 2025) compute activation differences between inputs and use these to improve adherence to constraints without finetuning. For MDMs, the field is less mature. The most directly related work (Shnaidman et al., 2025) shows that such control is possible for LLaDA (Nie et al., 2025); however, it focuses on a single task (concept negation) and uses a suboptimal activation steering map (Stolfo et al., 2025), Figure 1 displays these methods through an OT framing. As a result, the prominent applications of concept induction and instruction-following are underexplored.

Building on the Optimal Transport (OT) perspective on activation steering by Rodriguez et al. (2024), we design a rigorous inference-time control framework in MDMs beyond concept nega-

tion (Shnaidman et al., 2025) that generalizes across state-of-the-art backbones. We formalize OT steering for MDMs: learning an affine map to steer the unmasking trajectory over diffusion steps.

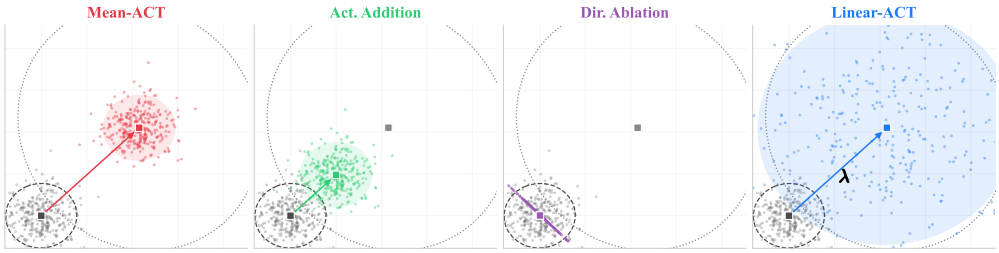

Figure 1: Affine optimal-transport view of activation steering: for source/target centroids m_a, m_b, Mean-ACT (Rodriguez et al., 2024) applies a mean shift $\Delta m = m_b - m_a$, Activation Addition (Stolfo et al., 2025) applies a scaled shift, Directional Ablation (Shnaidman et al., 2025) removes the component along Δm, and Linear-ACT matches second-order structure.

This paper contains the following contributions:

1. **First OT-guided steering formulation for dLLMs:** we introduce an Optimal Transport (OT) framing of activation steering specialized to MDM language models, defining steering as transporting pooled activation distributions between contrastive prompts over diffusion steps.
2. **First non-mitigation use of activation steering in dLLMs:** beyond concept negation/mitigation interventions, we demonstrate activation steering for positive behavior induction in dLLMs, targeting instruction constraint-following on IFEval (Zhou et al., 2023).
3. **Improvement over existing dLLM steering:** across multiple dLLM backbones, OT steering yields the largest and most consistent instruction-following gains relative to (i) no steering, (ii) the prior MDM-steering formulation of Shnaidman et al. (2025), and (iii) autoregressive baselines.

2 Optimal Transport in MDMs

We represent the diffusion sequence at time t as $x_t = [p_0; r_t]$, where p_0 denotes the prompt tokens and r_t denotes the response tokens at diffusion time t (which may be partially or fully masked, depending on t). We map x_t to a pooled activation representation $Z(x_t) \in \mathbb{R}^{M \times L}$ via ϕ-pooling, i.e., we apply an aggregation operator ϕ over the token dimension (in our experiments, we use mean-pooling [1]). Specifically, we compute steering parameters using activations at $t = 1.0$ (fully masked). This serves as a direct extension of using 0-th step activations in AR steering (Rodriguez et al., 2024; Shnaidman et al., 2025; Stolfo et al., 2025): both use the alignment direction at the initialization of the generative process and apply that at every decoding step. Unlike autoregressive steering, where information accumulates at the final position (favoring last-token pooling), diffusion distributes semantic information across the full sequence length (K tokens), so by pooling over K, our map steers the entire window rather than biasing only the next-token prediction.

Consider the probability distributions of sentences p and q on the pooled activation matrices: $\mu := Z\sharp p$ and $\nu := Z\sharp q$. We have contrastive samples $x_t^1, ..., x_t^n \sim p$ and $y_t^1, ..., y_t^n \sim q$ (differing along a specific attribute of interest), yielding activation matrices $a_t^i := Z(x_t^i)$ and $b_t^i := Z(y_t^i)$ from μ, ν respectively. The objective is to learn a map $T : \mathbb{R}^{M \times L} \to \mathbb{R}^{M \times L}$ pushing μ and ν: $T\sharp\mu \approx \nu$.

We restrict us to affine maps, which are computationally cheap and, subsume prior steering rules as special cases. We want to find affine maps pushing $A := a_t^i$ to $B := b_t^i$ for $t \in [0, 1], i \in 1, ..., n$. Note that t ranges over a discretized grid due to the discrete nature of MDMs. We represent this affine map as $T(a; A, B) := \omega a + \beta$ (for simplicity, we ignore the neuron subscripts).

OT view of common steering rules. Next, we show that the baselines can be naturally expressed within the OT framework. For MDM-steer (Shnaidman et al., 2025), this yields the following form

[1] Mean-pooling is a natural choice in MDMs because semantic information is distributed across the entire masked window during denoising, rather than concentrating at a single "last" token as in autoregressive decoding; empirically, it is also the most stable choice in our ablations (Appendix Table 3).

(see Appendix B): $\omega = I - \hat{v}\hat{v}^\top, \quad \beta = 0$ where $\hat{v} = \frac{m_b - m_a}{\|m_b - m_a\|_2}$ is the normalized difference-in-means vector derived from the contrastive centroids m_b and m_a. Here, ω acts as a projection matrix that ablates the variance along the direction $\hat{v}$, effectively collapsing the distribution's spread to zero along that axis, which represents the optimal transport map only under the assumption that the target concept is to be completely erased (where $\sigma_{target} = 0$), thus primarily effective for mitigation steering. Alternatively, Activation Addition steering (Stolfo et al., 2025) can be written as: $\omega = I, \quad \beta = \alpha\hat{v}$, where α is hyperparameter. Here, $\omega = I$ implies the distribution shape is preserved while β shifts the location. In both cases, $\hat{v} = \frac{m_b - m_a}{\|m_b - m_a\|_2}$ is the normalized difference-in-means vector derived from the concept centroids m_b and m_a. Here, using $\hat{v} = m_b - m_a$ gives Mean-ACT (Rodriguez et al., 2024), which is optimal under a Gaussian assumption and equal variances. For a Gaussian distribution but differing variances, $T(a) = \frac{\sigma_b}{\sigma_a}a + (m_b - \frac{\sigma_b}{\sigma_a}m_a)$ ("Gaussian OT") gives the optimal map (Peyré & Cuturi, 2020).

Optimal affine OT parameters. Moving to the optimal affine OT parameters, in Linear-ACT (Rodriguez et al., 2024), (ω, β) are obtained by solving a least-squares alignment between the samples:

$$(\omega, \beta) \in \arg\min_{\omega,\beta} \sum_{i=1}^{n} \big(b_{(i)} - (\omega a_{(i)} + \beta)\big)^2 \implies \quad \omega = \frac{\sum_{i=1}^{n} \tilde{a}_{(i)}\, \tilde{b}_{(i)}}{\sum_{i=1}^{n} (\tilde{b}_{(i)})^2}, \qquad \beta = m_b - \omega m_a$$

where $\tilde{a}_{(i)} = a_{(i)} - m_a$, $\tilde{b}_{(i)} = b_{(i)} - m_b$ are the centered and ordered activations.

Linear-ACT also introduces a strength parameter $\lambda \in [0, 1]$ via identity interpolation: $T_\lambda(a) = (1 - \lambda)a + \lambda(\omega a + \beta)$, so that $\lambda = 0$ applies no steering and $\lambda = 1$ applies the full learned transport. In practice, Linear-ACT may additionally constrain the transport to a support $Q_0 = [\min A, \max A]$ to avoid extrapolation, while using $Q_\infty = (-\infty, \infty)$ when extrapolation is desired (e.g., induction settings). For compactness, define $\Delta m := m_b - m_a$, then we can define our baselines as:

Method	$\boldsymbol{\omega}$	$\boldsymbol{\beta}$
Directional Ablation* [10]	$I - (\Delta m\, \Delta m^\top)/\|\Delta m\|_2^2$	$\mathbf{0}$
Activation Addition [11]	I	$\alpha\, \Delta m/\|\Delta m\|_2$
Mean-ACT [9]	I	Δm
Gaussian OT [7]	σ_b/σ_a	$m_b - (\sigma_b/\sigma_a)m_a$
Linear-ACT [9]	$\sum_{i=1}^{n} \tilde{a}_{(i)}\, \tilde{b}_{(i)}\ /\ \sum_{i=1}^{n} (\tilde{b}_{(i)})^2$	$m_b - \omega m_a$

Table 1: Affine transport parameters (ω, β) for activation-space steering methods.
*Directional Ablation is the only prior activation-steering method proposed specifically for MDMs.

Derivations of these affine forms are provided in Appendix B. Importantly, the affine assumption allows merging steering into activation functions, avoiding any inference-time overhead.

3 Experiments

We evaluate three MDM models: LLaDA-8B-Instruct (Nie et al., 2025), LLaDA 1.5 (Zhu et al., 2025), and Dream-7B Instruct (Ye et al., 2025). We report IFEval (Zhou et al., 2023) accuracy aggregated across constraint types, and also analyze per-constraint behavior to identify where affine transport helps most. All methods use recommended decoding settings.

3.1 Utility Evaluation

We evaluate utility using IFEval (Zhou et al., 2023) (Instruction Following Evaluation) benchmark, focusing on verifiable constraints such as forbidden words and capitalization frequency (see Appendix Table 2 for the full set of verifiable instruction categories). We construct contrastive calibration sets by collecting prompts (e.g., "Write a text without using the letter 'e'") and use 128 positive and 128 negative samples, selecting λ on a 10-sample validation split, and evaluated on a 64 sample test set. Unless otherwise noted, we use the best-performing pooling scope and intervention points from the Appendix ablation (Table 3): mean-pooling over the full diffusion sequence ($[p_0; r_t]$), forward hooks at the FFN projection and FFN output (Stolfo et al., 2025), and no quantile bounds. Compared to prior MDM steering (Shnaidman et al., 2025), which hooks only prompt activations (evaluated on a single dLLM/task), we intervene throughout generation. Directional

Ablation (Shnaidman et al., 2025) and Activation Addition (Stolfo et al., 2025) are steered on the residual layers, as recommended by the authors.

We compare methods in the following order: (i) **No Steering**; (ii) **Directional Ablation:** Projecting out a single contrastive direction (Shnaidman et al., 2025); (iii) **Mean-ACT:** difference-in-means transport (Rodriguez et al., 2024); (iv) **Activation Addition:** Adding a scaled mean vector (Stolfo et al., 2025); (v) **Gaussian OT:** Scale-and-shift transport under Gaussianity (Peyré & Cuturi, 2020); and (vi) **Linear-ACT:** Least-squares affine transport (Rodriguez et al., 2024).

3.2 Instruction-following results

Linear-ACT steering across backbones. Figure 2 compares the baseline and steered variants of three instruction-tuned backbones. For baseline performance across all models (including Qwen-Instruct (Qwen et al., 2025), an autoregressive LLM), see Appendix Figure 4. Overall, OT steering yields consistent gains on most format/control metrics (e.g., enforcing constraints such as forbidden-word avoidance and casing), while preserving comparable performance on the remaining dimensions. Tasks where no-steering performs well ($\geq 80\%$ accuracy) show marginal performance gains.

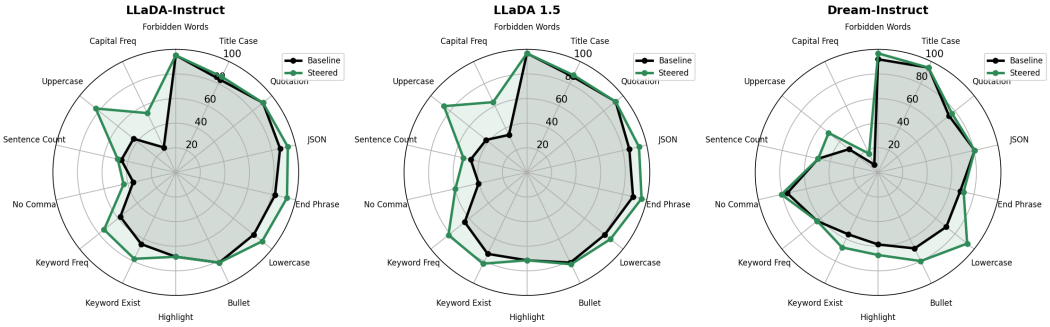

Figure 2: Baseline vs. Linear-ACT radar plots across dLLMs for different instruction-following IFEval tasks (LLaDA-Instruct, LLaDA 1.5, Dream-Instruct).

Average accuracy across steering methods. Figure 3 summarizes the mean IFEval accuracy for each steering method across the three dLLM backbones. Linear-ACT achieves the highest accuracy, improving by 11.9, 10.1 and 6.5 points over LLaDA-Instruct, LLaDA 1.5 and Dream-Instruct, respectively. Gaussian OT is the second best method, which confirms that second order moments are vital in the steering map. Directional Ablation, which is the prior steering method, can underperform no-steering, confirming that such variance-collapsing map is not suited for non-mitigation steering. The consistent improvement from Mean-ACT to Gaussian OT to Linear-ACT supports that richer moment-matching yields more reliable steering.

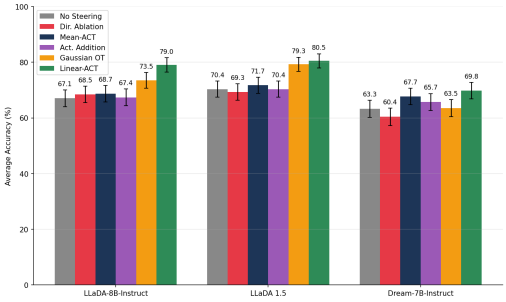

Figure 3: Average accuracy (%) for no steering, Directional Ablation, Mean-ACT, Activation Addition, Gaussian OT, and Linear-ACT across model backbones with a 95% sample-based confidence interval.

4 Conclusion

We introduced an optimal-transport framework for activation steering in masked diffusion language models, providing a notion of optimal inference-time control in dLLMs. The resulting transport map (Rodriguez et al., 2024) provides an activation steering estimator which, through diffusion-tailored representation and design choices, outperforms existing MDM steering baselines on IFEval (Zhou et al., 2023) across a range of dLLMs. Future work includes timestep-adaptive transport maps, richer transport parameterizations, incorporating utility-based evaluation of the responses, and broader evaluations beyond verifiable constraints such as concept induction.

ACKNOWLEDGMENTS

This research is part of the Priv-GSyn project, 200021E_229204 of Swiss National Science Foundation and the DEPMAT project, P20-22 / N21022, of the research programme Perspectief which is partly financed by the Dutch Research Council (NWO). This work was partly supported by the Spoke 1 "FutureHPC & BigData" of ICSC - Centro Nazionale di Ricerca in High-Performance Computing, Big Data and Quantum Computing, funded by the European Union - NextGenerationEU.

REFERENCES

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. 2021. doi: 10.48550/arXiv.2107.03006.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. 2022. doi: 10.48550/arXiv.2202.04200.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6112–6121, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1633. URL https://aclanthology.org/D19-1633/.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. 2022. doi: 10.48550/arXiv.2207.12598.

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. 2025. doi: 10.48550/arXiv.2502.09992.

Gabriel Peyré and Marco Cuturi. Computational optimal transport, 2020. URL https://arxiv.org/abs/1803.00567.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Pau Rodriguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, Marco Cuturi, and Xavier Suau. Controlling language and diffusion models by transporting activations. *arXiv preprint arXiv:2410.23054*, 2024.

Adi Shnaidman, Erin Feiglin, Osher Yaari, Efrat Mentel, Amit Levi, and Raz Lapid. Activation steering for masked diffusion language models, 2025. URL https://arxiv.org/abs/2512.24143.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering, 2025. URL https://arxiv.org/abs/2410.12877.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. 2023. doi: 10.48550/arXiv.2308.10248.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models. 2025. doi: 10.48550/arXiv.2508.15487.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

Fengqi Zhu, Rongzhen Wang, Shen Nie, Xiaolu Zhang, Chunwei Wu, Jun Hu, Jun Zhou, Jianfei Chen, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Llada 1.5: Variance-reduced preference optimization for large language diffusion models, 2025. URL https://arxiv.org/abs/2505.19223.

A Appendix

Table 2: IFEval (Zhou et al., 2023) verifiable instruction categories used in our evaluation. Each constraint is automatically verified by a deterministic condition.

Category	**Constraint**	**Description**
Keywords	Forbidden Words	Response must not contain specified words
Keywords	Keyword Existence	Response must include specified keywords
Keywords	Keyword Frequency	A keyword must appear at least k times
Keywords	Letter Frequency	A given letter must appear $\geq n$ or $\leq n$ times
Change Case	Capital Frequency	At least n words must be fully capitalized
Change Case	All Uppercase	Entire response in capital letters
Change Case	All Lowercase	Entire response in lowercase letters
Det. Format	Title	Response must include a title in markdown
Det. Format	JSON Format	Entire response must be valid JSON
Det. Format	Bullet Lists	Response must contain $\geq n$ bullet points
Det. Format	Highlighted Sections	$\geq n$ sections wrapped in markdown highlights
Length	Sentence Count	Response must have $\{<, >, =\}$ n sentences
Punctuation	No Comma	Response must not contain any commas
Start/End	End Phrase	Response must end with an exact phrase

Table 3: Ablations over steering configurations. Reported numbers are mean improvement in IFEval accuracy (in percentage points) relative to no steering, aggregated across tasks and configurations. We use the best-performing configuration in subsequent experiments.

Category	**Setting**	**Mean Δ (pp)**
Hook point	**Both (FFN proj + FFN out)**	**+17.23**
Hook point	FFN projection	+14.52
Hook point	FFN output	+8.25
Quantile bounds	**None (no bounds)**	**+17.22**
Quantile bounds	$[0, 100]$	+9.45
Pooling scope	**Full sequence**	**+13.49**
Pooling scope	Prompt tokens	+13.39
Pooling scope	Response tokens	+13.14

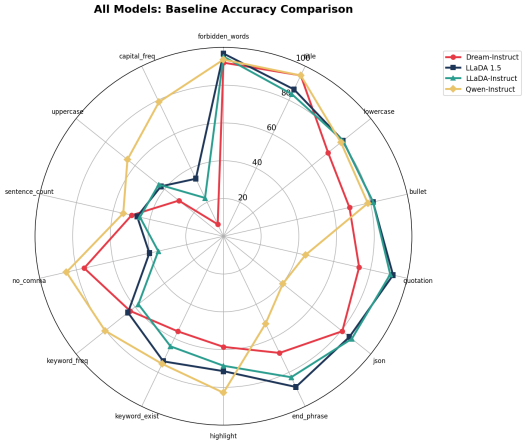

Figure 4: Baseline IFEval (Zhou et al., 2023) accuracy breakdown across instruction-following constraint types for all evaluated dLLMs, also including Qwen-Instruct (Qwen et al., 2025) (a similarly sized autoregressive LLM baseline).

B Optimal Transport Derivations

Directional Ablation

The Directional Ablation (Shnaidman et al., 2025) update rule is defined as removing the projection of the activation a onto the steering direction $\hat{v}$:

$$T(a) = a - \langle a, \hat{v} \rangle \hat{v}$$

Using the property that $\langle a, \hat{v} \rangle = \hat{v}^\top a$ and associativity:

$$\begin{aligned} T(a) &= a - \hat{v}(\hat{v}^\top a) \\ &= a - (\hat{v}\hat{v}^\top)a \\ &= (I - \hat{v}\hat{v}^\top)a \end{aligned}$$

Comparing this to the affine map form $T(a) = \omega a + \beta$, we identify: $\omega = I - \hat{v}\hat{v}^\top, \quad \beta = 0$ Thus, ω is the projection matrix onto the orthogonal complement of $\hat{v}$, and the shift β is zero.

Activation Addition (Mean Shift)

The standard Activation Addition (Stolfo et al., 2025) update rule consists of adding a scaled steering vector to the activation a:

$$T(a) = a + \alpha\hat{v}$$

We can factorize a explicitly using the identity matrix I:

$$T(a) = Ia + \alpha\hat{v}$$

Comparing this to the affine map form $T(a) = \omega a + \beta$, we identify: $\omega = I, \quad \beta = \alpha\hat{v}$ This corresponds to a pure translation of the activation space, where the covariance is assumed to be equal between source and target ($\omega = I$).