DASH: STEERING ATTENTION HEADS FOR FINE-GRAINED CONCEPT CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Attention heads are widely used as units of analysis and intervention in transformer models, but treating an entire head as concept-specific can make interventions too coarse. We study concept control at the level of head–direction pairs and introduce **DASH**, **D**irection-**A**ware **S**teering of **H**eads. Given a concept dictionary and a selected set of attention heads, DASH identifies concept-related directions within each head and selectively modifies their activation components without retraining the model. Experiments on toxicity suppression, image captioning, image classification, and factual question answering show improved control–preservation trade-offs over whole-head rescaling across a broad range of settings. We further demonstrate simultaneous editing of multiple attributes, including when their selected heads overlap, while largely preserving an untargeted attribute. Together, these results demonstrate the value of within-head directional editing for selective and composable concept control.

1 INTRODUCTION

Language and vision–language models support diverse generative tasks, from question answering to image captioning. Their usefulness also depends on controlling which concepts appear in their outputs and how strongly they are expressed (Dathathri et al., 2020). Applications may require suppressing toxic language (Liu et al., 2021), emphasizing visual attributes (Basile et al., 2025), or adjusting several aspects of a response simultaneously. Such control must balance the desired behavioral change with preserving fluency, relevance, and information unrelated to the intervention.

Interventions on internal representations offer a direct way to influence generation without updating model parameters (Turner et al., 2023; Rimsky et al., 2024). Localizing these interventions connects control with interpretability: identifying where a concept is represented provides candidate sites for modifying its expression. Attention heads are useful intervention sites because their contributions can be inspected and edited before entering the residual stream (Li et al., 2023; Adila et al., 2024). Head Pursuit (Basile et al., 2025) uses Simultaneous Orthogonal Matching Pursuit (SOMP) Tropp et al. (2006) to reconstruct each head's output from concept-related token directions. It ranks heads by the fraction of output variance explained, then rescales the highest-ranked heads to suppress or amplify the target concept.

However, selecting relevant heads does not by itself ensure selective control. A head can carry information about multiple aspects of generation, so rescaling its entire activation may alter unrelated content or degrade general performance. When concepts share selected heads, a single scalar also cannot suppress one concept while amplifying another. These limitations motivate combining sparse head selection with finer-grained control over the concept-related components within each head.

We introduce **DASH**, *Direction-Aware Steering of Heads*, a training-free framework that retains Head Pursuit's concept dictionaries and SOMP selection procedure while editing directions within selected heads. DASH maps unembedding vectors into each head's activation space through the transpose of its output projection. It then edits either a mean concept direction ($\text{DASH}_{\text{mean}}$) or a subspace constructed from multiple dictionary-derived directions ($\text{DASH}_{\text{multi}}$). A continuous strength parameter controls suppression or amplification, leaving orthogonal activation components unchanged at the intervention site.

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.

DASH also composes interventions by summing concept-specific updates over the union of their selected head sets. Each concept has its own strength and sign, supporting joint suppression, joint amplification, and mixed interventions even when selected heads overlap.

We evaluate DASH on toxicity suppression, caption-attribute steering, image-classification suppression, and factual question answering. We combine fixed strengths shared across datasets and head budgets with intervention-strength sweeps to assess target control alongside preservation. DASH improves this trade-off across a broad range of settings: the mean direction often favors preservation, whereas multiple directions enable stronger control. The benefits depend on the task and preservation metric, motivating comparison across operating points rather than by maximum target effect alone.

Our contributions are:

- We **introduce DASH**, a training-free framework for direction-aware head editing, with mean-direction and multi-direction constructions for selective concept control.
- We **enable composable concept control** through simultaneous suppression and amplification, demonstrated on captioning attributes whose selected head sets overlap.
- We **demonstrate improved control–preservation trade-offs** over whole-head rescaling across a broad range of language and vision–language settings. Component ablations establish the importance of concept-informed heads and directions, while geometric diagnostics characterize the resulting concept representations.

2 Related Work

Controllable generation and representation steering. Controlling pretrained generative models without retraining has been studied through both decoding-time guidance and interventions on internal representations. Approaches such as PPLM (Dathathri et al., 2020), GeDi (Krause et al., 2021), FUDGE (Yang & Klein, 2021), and DExperts (Liu et al., 2021) guide generation using auxiliary attribute models or modifications to token probabilities. A closely related line of work directly manipulates model activations. Latent steering vectors (Subramani et al., 2022), Activation Addition (Turner et al., 2023), Contrastive Activation Addition (Rimsky et al., 2024), and Representation Engineering (Zou et al., 2023) demonstrate control of semantic and behavioral properties through directions in activation space. Related work identifies low-dimensional representations associated with behaviors such as refusal (Arditi et al., 2024) and develops structured geometric formulations of steering (Gaintseva et al., 2026b).

Representation steering also supports control in visual generative models. Concept Sliders (Gandikota et al., 2024) learn composable directions for continuous control of diffusion models, while CASteer (Gaintseva et al., 2026a) performs training-free concept erasure by steering cross-attention representations. DASH studies selective and composable control in language and vision–language models through concept-specific edits within attention heads.

Attention-head specialization and control. Individual attention heads can exhibit specialized behavior, including syntactic and positional functions (Voita et al., 2019), induction (Olsson et al., 2022), and copy-suppression mechanisms (McDougall et al., 2024). These findings motivate localized interventions on head activations. Inference-Time Intervention (Li et al., 2023) derives truth-related directions from labeled examples and shifts activations in selected heads. SteerFair (Adila et al., 2024) extracts bias directions from unlabeled rule-following demonstrations and shifts head activations to improve robustness to prompt variations. DASH constructs editing subspaces from concept-token dictionaries and rescales each activation's projection onto them.

Head Pursuit (Basile et al., 2025) uses dictionary-based SOMP reconstruction to rank heads by explained variance, then rescales selected heads. DASH retains this selection and maps dictionary directions through the transpose of each head's output projection. It uses the full-dictionary mean or a subspace of SOMP-selected atoms, and sums concept-specific updates to support simultaneous suppression and amplification on shared heads.

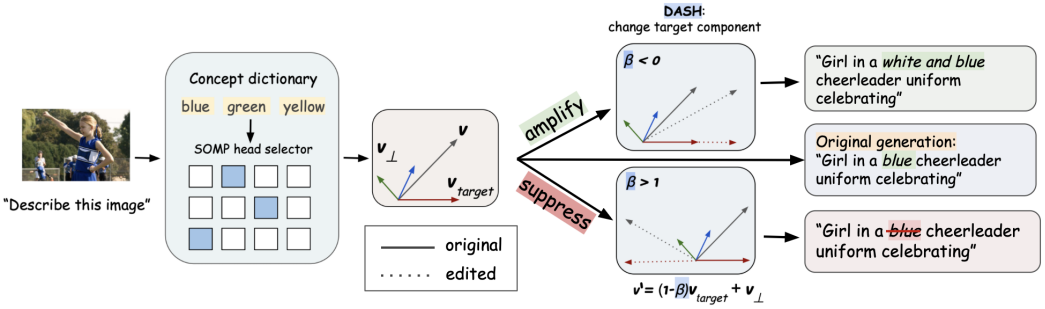


Figure 1: **Overview of DASH.** A concept dictionary guides SOMP-based selection of relevant attention heads. DASH edits the concept-related component of each selected head's activation to amplify or suppress an attribute, leaving its orthogonal component unchanged. On the right, the generated captions illustrate color steering: DASH amplifies or suppresses mentions of color while retaining the rest of the description.

3 METHOD

3.1 PRELIMINARIES: HEAD PURSUIT

Head Pursuit (Basile et al., 2025) controls semantic attributes through sparse head selection followed by whole-head rescaling. A concept c is specified by a dictionary $\mathcal{D}_c = \{t_1, \ldots, t_M\}$ of vocabulary token IDs. Each token corresponds to a residual-stream direction $W_U[t] \in \mathbb{R}^{d_{\text{model}}}$, given by a row of the unembedding matrix $W_U \in \mathbb{R}^{|\mathcal{V}| \times d_{\text{model}}}$.

Head selection. Using normalized concept-dictionary vectors, Simultaneous Orthogonal Matching Pursuit (SOMP) reconstructs each head's residual-stream contributions across calibration samples through a sparse selection of dictionary atoms. Heads are ranked by the fraction of variance explained, yielding the top-K set $\mathcal{H}^{(c)}_{\text{SOMP}} = \{(\ell_1, h_1), \ldots, (\ell_K, h_K)\}$, where (ℓ_i, h_i) identifies the i-th ranked head: ℓ_i is its layer index and h_i is its head index within that layer.

Whole-head rescaling. For each selected head, Head Pursuit rescales its activation $z_{\ell,h} \in \mathbb{R}^{d_h}$:

$$z_{\ell,h} \mapsto \alpha z_{\ell,h}.$$

Head Pursuit uses sign reversal ($\alpha = -1$) for suppression and amplification ($\alpha > 1$) to enhance the target attribute. DASH retains its concept dictionaries and head selection, replacing whole-head rescaling with edits to concept-specific activation components.

3.2 CONCEPT SUBSPACES IN HEAD SPACE

Let $z_{\ell,h}(x,t) \in \mathbb{R}^{d_h}$ denote the activation of head h in layer ℓ at token position t, before the output projection $W^O_{\ell,h} \in \mathbb{R}^{d_{\text{model}} \times d_h}$. The head contributes $W^O_{\ell,h} z_{\ell,h}(x,t)$ to the residual stream. DASH operates in this head space, mapping output-side concept directions through $(W^O_{\ell,h})^\top$.

For each token t in the concept dictionary $\mathcal{D}_c$, define the normalized unembedding atom

$$a_t = \frac{W_U[t]}{\|W_U[t]\|_2}.$$

We construct either a single mean direction or a rank-r subspace from these atoms.

Mean direction. We average all normalized atoms in the concept dictionary and normalize the resulting direction:

$$u^{(c)} = \frac{\sum_{t \in \mathcal{D}_c} a_t}{\left\| \sum_{t \in \mathcal{D}_c} a_t \right\|_2}.$$

For each selected head, we pull the mean direction into the head space and normalize the resulting vector:

$$v^{(c)}_{\ell,h} = (W^O_{\ell,h})^\top u^{(c)}, \qquad \hat{v}^{(c)}_{\ell,h} = \frac{v^{(c)}_{\ell,h}}{\|v^{(c)}_{\ell,h}\|_2}.$$

This pullback follows from

$$\left\langle u^{(c)}, W^O_{\ell,h} z_{\ell,h} \right\rangle = \left\langle (W^O_{\ell,h})^\top u^{(c)}, z_{\ell,h} \right\rangle .$$

Thus, $v^{(c)}_{\ell,h}$ represents the head-space linear functional measuring alignment of the head's residual-stream contribution with $u^{(c)}$. Its normalized form defines the rank-one subspace used by $\mathrm{DASH}_{\mathrm{mean}}$.

Rank-r subspace. For $\mathrm{DASH}_{\mathrm{multi}}$, we reuse the ordered dictionary atoms selected by SOMP for each head. SOMP selects atoms using residual correlations across calibration examples, refitting the reconstruction after each selection. The resulting ordering is therefore head-specific and calibration-dependent. Full details are provided in Appendix C.

For each selected head, let $s_1, \ldots, s_r$ index the first r atoms in its SOMP ordering. We pull these atoms into head space and apply a thin QR decomposition:

$$A^{(c)}_{\ell,h} = \left[(W^O_{\ell,h})^\top a_{s_1}, \ldots, (W^O_{\ell,h})^\top a_{s_r} \right], \qquad A^{(c)}_{\ell,h} = V^{(c)}_{\ell,h} R_{\ell,h}.$$

The orthonormal columns of $V^{(c)}_{\ell,h} \in \mathbb{R}^{d_h \times r}$ define the subspace used by our intervention. The requested editing rank is capped by the head dimension and the number of available selected atoms. We use $r = 10$ by default, independently of the larger SOMP atom budget used for head ranking.

3.3 Direction-aware operator

To express both DASH variants uniformly, let $\widetilde{V}^{(c)}_{\ell,h} \in \mathbb{R}^{d_h \times r}$ have orthonormal columns spanning the concept subspace: the unit mean direction $\hat{v}^{(c)}_{\ell,h}$ for $\mathrm{DASH}_{\mathrm{mean}}$ ($r = 1$), or the basis $V^{(c)}_{\ell,h}$ constructed above for $\mathrm{DASH}_{\mathrm{multi}}$ ($r \geq 1$). The corresponding orthogonal projector is

$$P^{(c)}_{\ell,h} = \widetilde{V}^{(c)}_{\ell,h} \big(\widetilde{V}^{(c)}_{\ell,h}\big)^\top .$$

DASH: Direction-Aware Steering of Heads. For every selected head, DASH modifies only the component of the activation lying in the target concept subspace:

$$z'_{\ell,h} = z_{\ell,h} - \beta\, P^{(c)}_{\ell,h} z_{\ell,h} = z_{\ell,h} - \beta\, \widetilde{V}^{(c)}_{\ell,h} \big(\widetilde{V}^{(c)}_{\ell,h}\big)^\top z_{\ell,h}.$$

Since $P^{(c)}_{\ell,h}$ is the Euclidean projector onto the target subspace, $0 < \beta < 1$ attenuates its contribution, $\beta = 1$ removes it, $\beta > 1$ moves the activation beyond removal in the opposite direction, and $\beta < 0$ amplifies it. Components orthogonal to the target concept subspace are left unchanged.

Both variants intervene at all token positions during prompt processing and autoregressive decoding, including image-token positions in the VLM.

3.4 Composing multiple edits

Because DASH operates on concept-specific directions rather than entire heads, multiple edits can be applied within the same head. Let $\mathcal{C} = \{c_1, \ldots, c_m\}$ be a set of target concepts, where each c_j has its own SOMP-selected head set $\mathcal{H}^{(c_j)}$, concept frame $\widetilde{V}^{(c_j)}_{\ell,h}$, and intervention strength β_j. We intervene on the union

$$\mathcal{H}_{\mathcal{C}} = \bigcup_{j=1}^{m} \mathcal{H}^{(c_j)}.$$

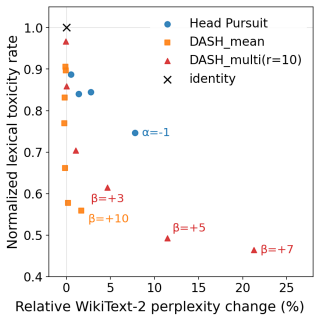
For each head in this union, the concept-specific updates are summed:

$$z'_{\ell,h} = z_{\ell,h} - \sum_{j:\,(\ell,h)\in\mathcal{H}^{(c_j)}} \beta_j\, \widetilde{V}^{(c_j)}_{\ell,h} \big(\widetilde{V}^{(c_j)}_{\ell,h}\big)^\top z_{\ell,h}.$$

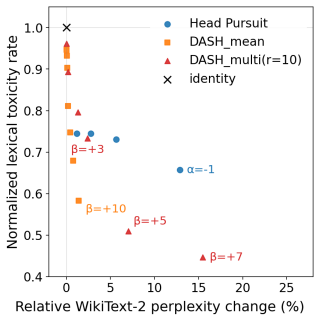
On shared heads, these additive updates support simultaneous suppression and amplification with concept-specific strengths, whereas whole-head rescaling applies a single scalar to every direction. We evaluate this setting in Section 4.2.3.

K	Operator	RTP (↓/↓)	TET (↓/↓)
8	HP	0.98 / +1.69	0.84 / +2.50
	$\text{DASH}_{\text{mean}}$	**0.87** / **+0.45**	**0.73** / **+0.55**
	$\text{DASH}_{\text{multi}}$	0.93 / +4.50	0.78 / +1.43
16	HP	0.82 / +3.65	0.76 / +10.68
	$\text{DASH}_{\text{mean}}$	0.67 / **+0.93**	**0.62** / **+0.91**
	$\text{DASH}_{\text{multi}}$	**0.66** / +7.77	0.69 / +4.46
32	HP	0.76 / +7.80	0.63 / +12.89
	$\text{DASH}_{\text{mean}}$	**0.50** / **+1.71**	0.56 / **+1.42**
	$\text{DASH}_{\text{multi}}$	0.55 / +11.46	**0.55** / +7.02

(a) Fixed-strength results.



(b) RTP trade-off.



(c) TET trade-off.

Figure 2: Toxicity suppression on Mistral-7B-Instruct-v0.2. **(a)** Fixed strengths: HP $\alpha = -1$, $\text{DASH}_{\text{mean}}$ $\beta = 10$, $\text{DASH}_{\text{multi}}$ $\beta = 5$. Entries report normalized lexical toxicity / relative WikiText-2 perplexity change (%), on 5,000 RTP and 1,273 TET prompts. **(b)–(c)** Strength sweeps at $K = 32$, varying HP's α and DASH's β on 1,000 prompts per dataset.

4 Experiments

4.1 Setup

We evaluate direction-aware head editing on the same broad task families studied by Head Pursuit: language generation, image captioning, image classification, and factual question answering. Language generation and image captioning are presented in the main text; image classification and question answering are reported in Appendix D. Across all tasks, we compare Head Pursuit, DASH_MEAN, and DASH_MULTI under matched head-selection budgets, sweeping the number of selected heads $K \in \{8, 16, 32\}$. Concept dictionaries follow Basile et al. (2025) (Appendix B). Holding dictionaries and selected heads fixed isolates directional editing from whole-head rescaling. We use $r = 10$ for DASH_MULTI by default. Ablations over r are provided in Appendix E.

We evaluate *target control* — concept suppression or amplification — alongside *preservation* of general capabilities and untargeted concepts, since strong target effects can accompany model degradation.

Intervention strength. Both Head Pursuit and DASH expose a continuous strength parameter: α for whole-head rescaling and β for directional editing. We evaluate the methods in two complementary ways. First, we compare fixed representative operating points across datasets and head budgets. For suppression, we use $\alpha = -1$ for Head Pursuit, following Basile et al. (2025), and $\beta = 10$ and $\beta = 5$ for $\text{DASH}_{\text{mean}}$ and $\text{DASH}_{\text{multi}}$, respectively. We chose these representative DASH strengths by inspecting RTP strength sweeps on 1,000 evaluation prompts (Figure 5, Appendix D.1), then kept one setting per variant across subsequent tasks and head budgets. Second, we sweep each method's intervention strength at $K = 32$ to characterize the control–preservation trade-off. Fixed-strength evaluations use task-specific sample sizes, detailed in Appendix D; strength sweeps use the same subset of 1,000 evaluation examples within each task across methods.

4.2 Results

4.2.1 Toxicity suppression

Setup. Following Basile et al. (2025), we evaluate toxicity suppression with Mistral-7B-Instruct-v0.2 on RealToxicityPrompts (RTP) (Gehman et al., 2020) and Thoroughly Engineered Toxicity (TET) (Luong et al., 2024). We measure target suppression using the normalized lexical toxicity rate, $\tau_{\text{lex}}/\tau_{\text{lex},0}$, where τ_{lex} is the fraction of generated continuations containing at least one match from a held-out toxic-word vocabulary disjoint from the head-selection dictionary. The denominator is the corresponding unedited-model rate. We additionally report mean toxic-word counts and RoBERTa-classified toxicity rates in the Appendix D.1. We measure language-modelling preservation by the

Table 1: Captioning interventions on LLaVA-NeXT-7B (Flickr30k, $n = 5{,}000$). Entries report normalized attribute count / normalized CIDEr. Suppression uses $\alpha = -1$ for HP and $\beta = 10, 5$ for $\mathrm{DASH}_{\mathrm{mean}}$, $\mathrm{DASH}_{\mathrm{multi}}$; amplification uses $\alpha = 5$ and $\beta = -10, -5$, respectively. $\mathrm{DASH}_{\mathrm{multi}}$ uses $r = 10$. The unedited baseline contains 26 sentiment matches.

		Suppression ↓ / ↑			Amplification ↑ / ↑		
K	Method	Colors	Sentiments	Quantity	Colors	Sentiments	Quantity
32	HP	0.13 / 0.75	0.04 / 0.95	0.08 / 0.79	2.29 / 0.95	5.54 / 0.88	1.62 / 0.91
	$\mathrm{DASH}_{\mathrm{mean}}$	0.03 / **0.94**	0.23 / **1.00**	0.71 / **0.95**	1.90 / **0.98**	4.65 / **0.97**	1.45 / **0.92**
	$\mathrm{DASH}_{\mathrm{multi}}$	**0.01** / 0.91	**0.00** / 0.96	**0.03** / 0.75	**2.99** / 0.96	**18.50** / 0.87	**1.93** / 0.88
16	HP	0.21 / 0.90	**0.00** / 0.97	0.37 / 0.89	1.39 / 0.95	5.54 / 0.92	1.29 / **1.04**
	$\mathrm{DASH}_{\mathrm{mean}}$	0.17 / **0.95**	0.27 / **1.00**	0.78 / **0.97**	1.68 / **0.99**	6.58 / **0.97**	1.10 / 0.99
	$\mathrm{DASH}_{\mathrm{multi}}$	**0.05** / 0.92	0.04 / 0.99	**0.08** / 0.82	**1.83** / 0.98	**14.08** / 0.90	**1.43** / 0.99

relative percentage change in WikiText-2 perplexity on 1,000 samples:

$$\Delta\mathrm{ppl}_{\mathrm{wt}} = 100 \left(\frac{\mathrm{ppl}_{\mathrm{edited}}}{\mathrm{ppl}_{\mathrm{unedited}}} - 1 \right).$$

Results. Figure 2 compares the methods at fixed representative strengths and across intervention-strength sweeps. At the fixed operating points, $\mathrm{DASH}_{\mathrm{mean}}$ achieves lower lexical toxicity rates and smaller perplexity increases than Head Pursuit across both datasets and all head budgets. The advantage generally grows with the head budget: on RTP at $K = 32$, normalized lexical toxicity falls from 0.76 to 0.50, while the relative WikiText perplexity increase falls from 7.8% to 1.7%.

$\mathrm{DASH}_{\mathrm{multi}}$ offers a complementary trade-off. On TET, it achieves stronger lexical suppression and better preservation than Head Pursuit at all three head budgets. On RTP, it also achieves stronger lexical suppression, but with larger perplexity increases. Thus, its additional control can come at a higher language modeling cost.

The strength sweeps show that favorable DASH operating points extend beyond the fixed choices. On both datasets, directional editing offers lower perplexity costs at comparable lexical toxicity rates and extends the observed range of suppression beyond whole-head rescaling. This supports an improved trade-off between reducing toxic-vocabulary mentions and preserving WikiText-2 language modelling. At the fixed RTP operating points with $K = 32$, the classifier ratios are 0.74 for $\mathrm{DASH}_{\mathrm{mean}}$ and 0.62 for HP (Appendix D.1).

4.2.2 Image captioning

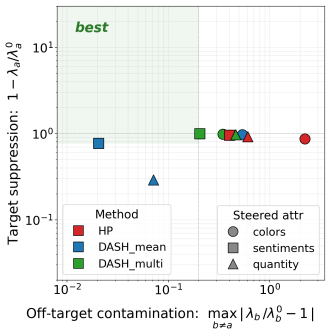
Setup. Following Basile et al. (2025), we evaluate LLaVA-NeXT-7B on Flickr30k. We suppress or amplify three caption attributes — *colors*, *sentiments*, and *quantity*. For each attribute a, we report the normalized property count λ_a/λ_a^0, where λ_a is the mean number of lexicon matches per generated caption and λ_a^0 is the unedited baseline. Caption preservation is measured by normalized CIDEr c/c_0 on the same $n = 5000$ images.

Results. Table 1 shows complementary control–preservation trade-offs for the two DASH variants. $\mathrm{DASH}_{\mathrm{mean}}$ generally favors preservation. For color suppression at $K = 32$, it reduces the normalized attribute count to 0.03 while retaining normalized CIDEr of 0.94, compared with 0.13 and 0.75, respectively, for Head Pursuit. $\mathrm{DASH}_{\mathrm{multi}}$ often achieves stronger control, with larger preservation costs on some attributes. For sentiment amplification at the same head budget, it reaches a normalized count of 18.50, compared with 5.54 for Head Pursuit, at similar CIDEr scores (0.87 versus 0.88).

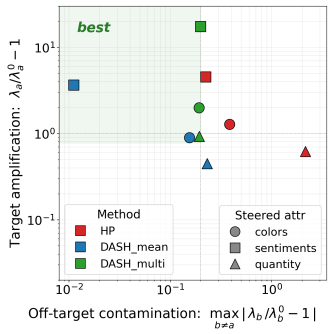
Intervention-strength sweeps at $K = 32$ in Appendix D.3 show strong color and sentiment suppression with near-baseline CIDEr and a wider amplification range for $\mathrm{DASH}_{\mathrm{multi}}$. At the sentiment operating point above, attribute presence rises 18.88-fold, showing that amplification reaches more captions (Appendix D.3).

Table 2: Qualitative examples of concept editing in Flickr30k captioning. Left: color suppression. Right: quantity amplification. HP denotes Head Pursuit. Highlighted words denote the target attribute; underlined text indicates non-target semantic changes introduced by the intervention.

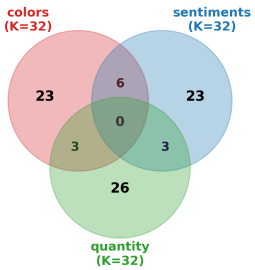
Color suppression	**Quantity** amplification
	
Original: Girl in blue cheerleader uniform celebrating.	**Original:** A group of people on a beach with a boat.
HP: Girl in a school uniform celebrating.	**HP:** A group of people enjoying a day at the beach.
DASH$_{multi}$**:** Girl in cheerleading uniform celebrating.	**DASH**$_{multi}$**:** Four men on a beach with a boat.



(a) Attribute suppression.



(b) Attribute amplification.



(c) Overlap of selected head sets.

Figure 3: Cross-attribute effects and head overlap on 5,000 Flickr30k images at $K = 32$. (a)–(b) Target effect versus off-target contamination at the strengths in Table 1. Target effect is the relative decrease (suppression) or increase (amplification) in target count; contamination is the largest absolute relative change among untargeted counts. Upper left indicates stronger control with less contamination. (c) Overlap among the 32 SOMP-selected heads for each attribute, showing that different concepts can share intervention sites.

Cross-attribute contamination. We next test whether editing one attribute inadvertently changes the others. For target attribute a, we define

$$\text{Effect}(a) = s\left(\frac{\lambda_a}{\lambda_a^0} - 1\right), \qquad \text{Contam}(a) = \max_{b\neq a}\left|\frac{\lambda_b}{\lambda_b^0} - 1\right|.$$

Here, $s = -1$ for suppression and $s = 1$ for amplification; larger effects and lower contamination are preferred. Figure 3 shows that DASH generally offers more favorable trade-offs between target modification and cross-attribute contamination.

One possible explanation is that the SOMP-selected head sets are not concept-exclusive. For $K = 32$, colors and sentiments share six selected heads, while color–quantity and sentiment–quantity each share three (Figure 3c). On such shared heads, scalar rescaling acts on all represented directions simultaneously, while DASH can target the concept-specific subspace.

Table 2 illustrates the same effect qualitatively. For color suppression, Head Pursuit removes the target color but also changes "cheerleader uniform" to "school uniform", while DASH removes the color reference without altering the remaining semantics. For quantity amplification, DASH adds the intended numerical attribute while retaining the beach and boat references.

Table 3: Multi-concept editing on LLaVA-NeXT-7B (Flickr30k, $n = 1{,}000$, $K = 32$ per concept). C/S/Q denote normalized mean color/sentiment/quantity vocabulary matches per caption; quantity is untargeted. The unedited baseline contains five sentiment matches across 1,000 captions, so each match corresponds to 0.2 in normalized S. CIDEr is also normalized to the unedited model. HP^{amp}/HP^{supp} prioritize amplification/suppression on shared heads.

(a) Suppression.

Method	C ↓	S ↓	Q	CIDEr
HP	0.72	0.60	1.13	0.68
$\text{DASH}_{\text{mean}}$	0.024	0.20	1.01	0.98
$\text{DASH}_{\text{multi}}$	0.034	0.000	0.94	0.89

(b) Amplification.

Method	C ↑	S ↑	Q	CIDEr
HP	1.87	3.00	0.70	0.90
$\text{DASH}_{\text{mean}}$	2.05	3.80	0.99	0.96
$\text{DASH}_{\text{multi}}$	2.69	19.00	1.06	0.86

(c) Mixed steering.

Method	C ↓	S ↑	Q	CIDEr
HP^{amp}	0.053	5.80	1.02	0.78
HP^{supp}	0.000	10.00	1.03	0.72
$\text{DASH}_{\text{mean}}$	0.019	6.20	0.99	0.95
$\text{DASH}_{\text{multi}}$	0.005	21.40	1.06	0.85

4.2.3 MULTI-TARGET INTERVENTION

We test composition through joint suppression, joint amplification, and mixed steering that suppresses one attribute while amplifying another.

Setup. We use the LLaVA-NeXT-7B Flickr30k captioning protocol from Section 4.2.2, evaluated on 1,000 images. We target colors and sentiments, the two attributes with the largest overlap between their SOMP-selected head sets. Each concept uses $K = 32$ heads, among which six are shared, giving a union of 58 distinct edited heads. The untargeted quantity attribute serves as a control for off-target effects.

We evaluate three compositions: jointly suppressing both attributes (*2-suppress*), jointly amplifying both (*2-amplify*), and a *mixed* setting that suppresses colors while amplifying sentiments. DASH composes the corresponding concept-specific projections as defined in Section 3.4, using the same strengths as in the single-concept experiments: $\beta = \pm 10$ for $\text{DASH}_{\text{mean}}$ and $\beta = \pm 5$ for $\text{DASH}_{\text{multi}}$ with $r = 10$.

For Head Pursuit, same-direction edits are applied to the union of the two head sets using $\alpha = -1$ for suppression and $\alpha = +5$ for amplification. In the mixed setting, the six shared heads cannot be both suppressed and amplified by a single scalar, so we report two variants: amplifying overlapping heads and suppressing them. We report normalized property counts λ_a/λ_a^0 for colors, sentiments, and the uninvolved quantity attribute, together with normalized CIDEr c/c_0 to measure caption preservation.

Results. Table 3 shows that both DASH variants improve target suppression, quantity preservation, and CIDEr under joint suppression. Under joint amplification, both achieve stronger target effects and better quantity preservation; $\text{DASH}_{\text{mean}}$ also improves CIDEr, whereas $\text{DASH}_{\text{multi}}$ trades some caption quality for stronger amplification. In mixed steering, both variants retain higher CIDEr than either HP variant. $\text{DASH}_{\text{mean}}$ keeps quantity mentions closest to baseline, while $\text{DASH}_{\text{multi}}$ achieves stronger sentiment amplification. Colors–quantity sweeps also show improved joint suppression and CIDEr at some DASH strengths (Appendix D.4).

4.2.4 ADDITIONAL TASKS

We also evaluate image-classification and TriviaQA country-answer suppression (Appendix D). DASH generally improves the control–preservation trade-off across six classification benchmarks. TriviaQA is more challenging: the mean variant provides limited suppression, while stronger multi-direction edits incur substantial QA and language-modelling costs, highlighting task-dependent selectivity.

Overall, the experiments demonstrate improved control–preservation trade-offs across a broad range of settings and support composition of multiple concept edits.

5 Ablations

We ablate the components underlying DASH's selective concept control. Full protocols and results can be found in Appendix E.

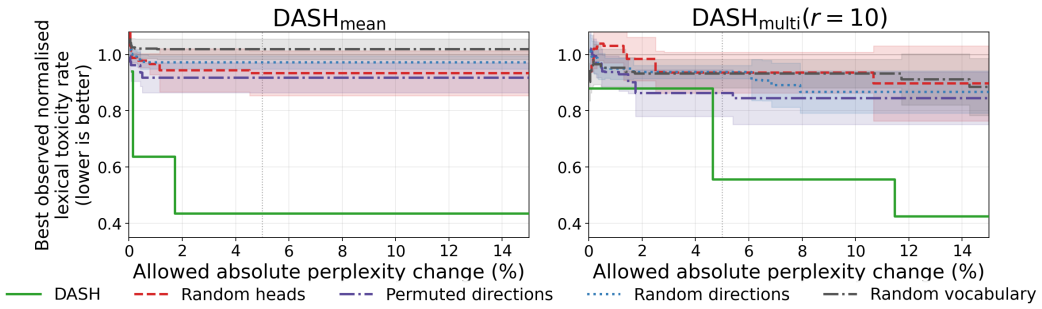


Figure 4: RTP component ablations at $K = 32$. Curves show the best observed normalized lexical toxicity rate over the tested strengths $\beta \in \{1, 3, 5, 10\}$, subject to the indicated absolute relative WikiText perplexity budget. Bands show standard deviations over five construction seeds; the vertical line marks a 5% budget.

Head and direction specificity. On RTP with $K = 32$, we compare four controls: (i) *random heads*: replace SOMP-selected heads with uniformly sampled heads, constructing concept directions or subspaces for the replacement heads using the same pipeline; (ii) *permuted directions*: reassign directions or subspaces across selected heads; (iii) *random directions*: replace concept directions with Gaussian unit vectors or random orthonormal subspaces of matching rank; and (iv) *random-vocabulary SVD*: keep selected heads fixed and use leading singular directions of a size-matched random dictionary projected into each head, changing both vocabulary and construction. Each randomized control uses five seeds, varying only the randomized component while holding calibration and evaluation prompts fixed.

Figure 4 compares the best observed normalized lexical toxicity rate within each WikiText perplexity budget, allowing each construction to choose among the tested strengths. Within a 5% budget, DASH$_{\text{mean}}$ and DASH$_{\text{multi}}$ reach approximately 0.43 and 0.56 of baseline, respectively, while randomized-control averages remain above 0.85. This separation persists across preservation budgets, showing that, at a fixed head budget, selective control depends on head selection, direction construction, and head–direction assignment.

Direction construction. Among rank-one constructions, the full-dictionary mean provides stronger lexical suppression than the SOMP-atom mean and leading dictionary singular direction within a 5% perplexity budget. At rank ten, SOMP gives stronger lexical suppression, while dictionary SVD yields lower classifier toxicity at comparable perplexity costs. These findings support our mean-direction default and identify dictionary SVD as a competitive subspace alternative (Appendix E).

Additional analyses. Appendix E examines rank, calibration size, activation energy, and dictionary geometry and stability. Scope sweeps show favorable RTP trade-offs for weak broad edits, while selected-head edits can better preserve captioning quality and untargeted attributes during sentiment steering.

6 Conclusion

We introduced DASH, a direction-aware approach to concept control that edits concept-specific directions within selected attention heads. Its mean-direction and multi-direction variants offer complementary control–preservation trade-offs, improving over whole-head rescaling across a broad range of language and vision–language settings. DASH also supports simultaneous suppression and amplification of multiple concepts, including when their selected heads overlap. Component ablations support the importance of concept-informed heads, directions, and their correspondence, while diagnostic analyses characterize the geometry of the edited subspaces. Together, these findings establish within-head directional editing as a flexible approach to selective concept control.

AI USE STATEMENT

We used generative AI tools to provide feedback on research methodology and experiments to help critique results. We also used them for brainstorming, analyzing and identifying relevant literature, formatting references, and editing the manuscript for clarity and readability. The authors reviewed the AI-assisted suggestions, checked the cited sources and reported results, and take responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

This paper uses publicly available pretrained models and datasets. We specify the concept dictionaries, prompts, evaluation metrics, intervention strengths and head-selection budgets, and report calibration and evaluation splits separately. We compare methods on shared evaluation examples, report strength sweeps alongside fixed operating points, and use repeated randomizations for the component controls. The method and experimental protocols are documented in Sections 3–4 and Appendices B–E. We will release the DASH code, experimental configurations and derived intervention parameters upon acceptance.

REFERENCES

Dyah Adila, Shuai Zhang, Boran Han, and Bernie Wang. Discovering bias in latent space: An unsupervised debiasing approach. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 246–261. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/adila24a.html.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://github.com/andyrdt/refusal_direction.

Lorenzo Basile, Valentino Maiorca, Diego Doimo, Francesco Locatello, and Alberto Cazzaniga. Head pursuit: Probing attention specialization in multimodal transformers. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/a30e060cde25f9c1e73924d2f996542f-Abstract-Conference.html.

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. In *International Conference on Learning Representations (ICLR)*, 2020.

Tatiana Gaintseva, Andreea-Maria Oncescu, Chengcheng Ma, Ziquan Liu, Martin Benning, Gregory Slabaugh, Jiankang Deng, and Ismail Elezi. CASteer: Cross-attention steering for controllable concept erasure. In *International Conference on Learning Representations*, 2026a. URL https://atmyre.github.io/casteer/.

Tatiana Gaintseva, Andrew Stepanov, Ziquan Liu, Martin Benning, Gregory Slabaugh, Jiankang Deng, and Ismail Elezi. Midsteer: Optimal affine framework for steering generative models. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026b.

Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: LoRA adaptors for precise control in diffusion models. In *Computer Vision – ECCV 2024*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/05660.pdf.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp.

1601–1611. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4929–4952, November 2021.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://neurips.cc/virtual/2023/poster/71200.

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. Association for Computational Linguistics, August 2021. URL https://aclanthology.org/2021.acl-long.522/.

Tinh Luong, Thanh-Thien Le, Linh Ngo, and Thien Nguyen. Realistic evaluation of toxicity in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, 2024.

Callum Stuart McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. Copy suppression: Comprehensively understanding a motif in language model attention heads. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 337–363. Association for Computational Linguistics, November 2024. URL https://aclanthology.org/2024.blackboxnlp-1.22/.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads, 2022. URL https://arxiv.org/abs/2209.11895.

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

s-nlp. RoBERTa toxicity classifier. Hugging Face model repository, n.d. URL https://huggingface.co/s-nlp/roberta_toxicity_classifier. Accessed September 25, 2026.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 566–581, Dublin, Ireland, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-acl.48/.

Joel A. Tropp, Anna C. Gilbert, and Martin J. Strauss. Algorithms for simultaneous sparse approximation. part i: Greedy pursuit. *Signal Processing*, 86(3):572–588, 2006. ISSN 0165-1684. Sparse Approximations in Signal and Image Processing.

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

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition, pp. 4566–4575, 2015. URL https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Vedantam_CIDEr_Consensus-Based_Image_2015_CVPR_paper.pdf.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808. Association for Computational Linguistics, July 2019. URL https://aclanthology.org/P19-1580/.

Kevin Yang and Dan Klein. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3511–3535, Online, June 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.naacl-main.276/.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency. In *ArXiV, 2310.01405*, 2023.

CONTENTS

1 Introduction **1**

2 Related Work **2**

3 Method **3**

3.1 Preliminaries: Head Pursuit 3

3.2 Concept subspaces in head space 3

3.3 Direction-aware operator 4

3.4 Composing multiple edits 4

4 Experiments **5**

4.1 Setup 5

4.2 Results 5

4.2.1 Toxicity suppression 5

4.2.2 Image captioning 6

4.2.3 Multi-target intervention 8

4.2.4 Additional tasks 8

5 Ablations **9**

6 Conclusion **9**

A Limitations **14**

B Concept dictionaries **15**

B.1 Toxicity (Mistral-7B-Instruct-v0.2) . . . 15
B.2 Factual question answering: TriviaQA (Mistral-7B-Instruct-v0.2) . . . 15
B.3 Image classification (LLaVA-NeXT-7B; six datasets) . . . 16
B.4 Image captioning (LLaVA-NeXT-7B, Flickr30k) . . . 16
B.5 Dictionary size after per-token expansion . . . 17

C SOMP-based head and atom selection **18**

D Evaluation details and additional results **20**
D.1 Toxicity suppression . . . 20
D.1.1 Setup . . . 20
D.1.2 Results . . . 21
D.2 Image classification . . . 23
D.2.1 Setup . . . 23
D.2.2 Results . . . 23
D.3 Captioning suppression/amplification . . . 27
D.3.1 Setup . . . 27
D.3.2 Results . . . 27
D.3.3 Additional qualitative examples . . . 31
D.4 Additional multi-attribute captioning results . . . 33
D.4.1 Setup . . . 33
D.4.2 Results . . . 33
D.5 Factual question answering: TriviaQA . . . 40
D.5.1 Setup . . . 40
D.5.2 Results . . . 40

E Ablations and diagnostic analyses **42**
E.1 Head and direction specificity . . . 43
E.2 Subspace rank . . . 46
E.3 Calibration size . . . 48
E.4 Activation-energy and geometric diagnostics . . . 50
E.4.1 Captured activation energy . . . 50
E.4.2 Mean–subspace alignment and dictionary geometry . . . 50
E.5 Alternative direction constructions . . . 53
E.6 Intervention scope: selected heads, selected layers, and all layers . . . 55

A LIMITATIONS

Concept specification and calibration. DASH relies on a token dictionary to represent each target concept and on calibration inputs to select heads and, for the multi-direction variant, dictionary atoms. Dictionary coverage, tokenization, and calibration representativeness therefore influence the resulting intervention. Extending control to context-dependent concepts or new domains may require revising these inputs. The shared strengths were chosen using preliminary RTP experiments; the preferred strength and subspace rank vary across tasks and preservation criteria.

Selectivity and composition. Orthogonal activation components remain unchanged at the intervention site, but subsequent model computation can propagate the edit to other behaviors. TriviaQA illustrates this limitation: stronger country-answer suppression is accompanied by substantial losses in non-country QA and language modelling (Appendix D.5). Composition likewise depends on concept geometry: overlapping subspaces can couple the effects of additive updates. Our composition experiments cover pairs of caption attributes; scaling to larger concept sets remains an open question.

Evaluation metrics. Lexical toxicity and attribute counts measure vocabulary use and can miss contextual meaning or changes expressed through alternative wording. Classifier-based toxicity, CIDEr, and WikiText-2 perplexity provide complementary but partial measures of control and preservation. Their rankings can differ, as the toxicity results demonstrate. Rare attributes also make normalized effects sensitive to small absolute changes: the unedited model produces only five sentiment matches in the 1,000-image composition evaluation. Broader human evaluation would help assess semantic fidelity and caption quality.

B CONCEPT DICTIONARIES

Across all tasks, we use the concept vocabularies from Head Pursuit (Basile et al., 2025).

For each concept c, the dictionary $\mathcal{D}_c$ contains the unique token IDs obtained from all subwords of its vocabulary entries. Each ID defines a normalized unembedding atom $a_t = W_U[t]/\|W_U[t]\|_2$. These atoms support SOMP head selection and DASH direction construction, as detailed in Appendix C.

Below, we describe the concept vocabularies used across the four task families and their roles in head selection and lexical evaluation. Vocabulary entries may be words, phrases, or class labels; dictionary atoms are unique token-level unembedding vectors.

B.1 TOXICITY (MISTRAL-7B-INSTRUCT-V0.2)

Each dataset uses disjoint validation and test vocabularies. The validation vocabulary is used for SOMP head selection and DASH dictionary construction; the test vocabulary is used only to compute held-out lexical toxicity rates $\tau_{\text{lex}}/\tau_{\text{lex},0}$ and toxic-token counts λ/λ_0 on generated continuations. We verified programmatically that the two vocabularies are disjoint for each dataset.

RealToxicityPrompts (RTP). 70 validation entries and 30 test entries.

Validation (94 dictionary atoms): cruelty, death, shame, terror, division, suffering, hate, gun, heinous, disturbing, twisted, chaos, crimes, false, adult, murder, inflammatory, atrocities, intimidation, crime, discrimination, divisive, blood, deadly, bruised, child, penis, danger, threatened, violated, violent, speech, explicit, threat, war, cruel, lies, offensive, broken, harm, anger, destruction, trauma, killed, brutal, killing, attacks, despair, hatred, traumatized, tragedy, brutality, sexual, white, terrorism, attack, deceit, horrific, extremist, inappropriate, derogatory, abuse, sex, terrorists, harmful, hateful, battered, racist, radical, threats.

Test (lexical evaluation only): violence, screams, racism, hell, threatening, prejudice, dangerous, victim, fuck, gruesome, violation, exploitation, kill, criminal, raped, assault, sexually, darkness, pain, menacing, fucking, rape, intolerance, harassment, misconduct, assaulted, horror, fear, terrorist, knife.

Thoroughly Engineered Toxicity (TET). 70 validation entries and 14 test entries.

Validation (106 dictionary atoms): slut, a, beast, taboo, baby, sensual, cock, graphic, cum, bodies, hardcore, breasts, body, seductive, creampie, failure, shit, steamy, hard, submissive, tits, forbidden, dick, nigga, sadistic, voluptuous, words, blowjob, rough, little, black, whore, dirty, semen, supple, slave, carnal, racist, skin, sinful, anal, nipples, passionate, twisted, filthy, small, pitiful, man, you, pleasure, thing, adult, white, climax, interracial, erect, naughty, passion, content, girl, desires, pussy, sex, her, moans, ass, fantasies, insignificant, men, desire.

Test (lexical evaluation only): sexy, cumshot, goddamn, plaything, fucking, perverted, flesh, orgasm, primal, worthless, fuck, explicit, lewd, pathetic.

B.2 FACTUAL QUESTION ANSWERING: TRIVIAQA (MISTRAL-7B-INSTRUCT-V0.2)

For country-answer questions in TriviaQA, we use the country-name vocabulary for both SOMP dictionary construction and lexical detection of country mentions in generated answers (`country_match_count`).

Country names (197 entries): Afghanistan; Albania; Algeria; Andorra; Angola; Antigua and Barbuda; Argentina; Armenia; Australia; Austria; Azerbaijan; Bahamas; Bahrain; Bangladesh; Barbados; Belarus; Belgium; Belize; Benin; Bhutan; Bolivia; Bosnia and Herzegovina; Botswana; Brazil; Brunei; Bulgaria; Burkina Faso; Burundi; Côte d'Ivoire; Cabo Verde; Cambodia; Cameroon; Canada; Central African Republic; Chad; Chile; China; Colombia; Comoros; Congo, Democratic Republic of the; Congo, Republic of the; Costa Rica; Croatia; Cuba; Cyprus; Czech Republic; Denmark; Djibouti; Dominica; Dominican Republic; Ecuador; Egypt; El Salvador; Equatorial Guinea; Eritrea; Estonia; Eswa-

tini; Ethiopia; Fiji; Finland; France; Gabon; Gambia; Georgia; Germany; Ghana; Greece; Grenada; Guatemala; Guinea; Guinea-Bissau; Guyana; Haiti; Honduras; Hungary; Iceland; India; Indonesia; Iran; Iraq; Ireland; Israel; Italy; Jamaica; Japan; Jordan; Kazakhstan; Kenya; Kiribati; Korea, North; Korea, South; Kosovo; Kuwait; Kyrgyzstan; Laos; Latvia; Lebanon; Lesotho; Liberia; Libya; Liechtenstein; Lithuania; Luxembourg; Madagascar; Malawi; Malaysia; Maldives; Mali; Malta; Marshall Islands; Mauritania; Mauritius; Mexico; Micronesia; Moldova; Monaco; Mongolia; Montenegro; Morocco; Mozambique; Myanmar (Burma); Namibia; Nauru; Nepal; Netherlands; New Zealand; Nicaragua; Niger; Nigeria; North Macedonia; Norway; Oman; Pakistan; Palau; Palestine; Panama; Papua New Guinea; Paraguay; Peru; Philippines; Poland; Portugal; Qatar; Romania; Russia; Rwanda; Saint Kitts and Nevis; Saint Lucia; Saint Vincent and the Grenadines; Samoa; San Marino; Sao Tome and Principe; Saudi Arabia; Senegal; Serbia; Seychelles; Sierra Leone; Singapore; Slovakia; Slovenia; Solomon Islands; Somalia; South Africa; South Sudan; Spain; Sri Lanka; Sudan; Suriname; Sweden; Switzerland; Syria; Taiwan; Tajikistan; Tanzania; Thailand; Timor-Leste; Togo; Tonga; Trinidad and Tobago; Tunisia; Turkey; Turkmenistan; Tuvalu; Uganda; Ukraine; United Arab Emirates; United Kingdom; United States; Uruguay; Uzbekistan; Vanuatu; Vatican City; Venezuela; Vietnam; Yemen; Zambia; Zimbabwe.

B.3 IMAGE CLASSIFICATION (LLAVA-NEXT-7B; SIX DATASETS)

For each dataset, the class labels form the concept vocabulary. SOMP scores heads by their alignment with the corresponding label unembedding vectors.

MNIST and SVHN (10 classes each). `0`, `1`, `2`, `3`, `4`, `5`, `6`, `7`, `8`, `9`. The shared digit vocabulary underlies Head Pursuit's cross-dataset transfer between MNIST and SVHN.

EuroSAT (10 classes). annual crop land, forest, brushland or shrubland, highway or road, industrial buildings or commercial buildings, pasture land, permanent crop land, residential buildings or homes or apartments, river, lake or sea.

RESISC45 (45 classes). airplane, airport, baseball_diamond, basketball_court, beach, bridge, chaparral, church, circular_farmland, cloud, commercial_area, dense_residential, desert, forest, freeway, golf_course, ground_track_field, harbor, industrial_area, intersection, island, lake, meadow, medium_residential, mobile_home_park, mountain, overpass, palace, parking_lot, railway, railway_station, rectangular_farmland, river, roundabout, runway, sea_ice, ship, snowberg, sparse_residential, stadium, storage_tank, tennis_court, terrace, thermal_power_station, wetland.

DTD (47 texture classes). banded, blotchy, braided, bubbly, bumpy, chequered, cobwebbed, cracked, crosshatched, crystalline, dotted, fibrous, flecked, freckled, frilly, gauzy, grid, grooved, honeycombed, interlaced, knitted, lacelike, lined, marbled, matted, meshed, paisley, perforated, pitted, pleated, polka-dotted, porous, potholed, scaly, smeared, spiralled, sprinkled, stained, stratified, striped, studded, swirly, veined, waffled, woven, wrinkled, zigzagged.

GTSRB (43 classes). GTSRB uses descriptive class labels, such as "red and white circle 20 kph speed limit", "blue circle with white right turn arrow mandatory", and "red circle with white horizontal stripe no entry". For space, we refer to `load_image_dataset('gtsrb')` in `hp_vision.py` for the complete list of 43 labels.

B.4 IMAGE CAPTIONING (LLAVA-NEXT-7B, FLICKR30K)

The *colors*, *sentiments*, and *quantity* vocabularies are used for both SOMP head selection and the normalized property count λ_a/λ_a^0. These vocabularies are disjoint at the entry level. We compute all three property counts from each generated caption, including in the composition experiments (Section D.4).

Task	Concept	Entries	Atoms $\lvert\mathcal{D}\rvert$
Toxicity RTP	toxicity (validation)	70	94
Toxicity TET	toxicity (validation)	70	106
TriviaQA	country	197	374
Captioning	colors	30	45
Captioning	sentiments	147	200
Captioning	quantity	68	67
Classification: MNIST/SVHN	digits (10)	10	10
Classification: EuroSAT	land use (10)	10	26
Classification: RESISC45	scene (45)	45	78
Classification: DTD	texture (47)	47	61
Classification: GTSRB	sign (43)	43	331

Table 4: Concept vocabulary sizes and the resulting numbers of unique token-level dictionary atoms under the corresponding model's tokenizer. Entry counts include repeated vocabulary entries. Dictionary atoms are unique token IDs; entries containing multiple tokens contribute multiple atoms.

Colors (30 entries). red, blue, green, yellow, orange, purple, pink, brown, black, white, gray, cyan, magenta, turquoise, indigo, maroon, beige, gold, silver, olive, lavender, navy, teal, peach, violet, ivory, charcoal, amber, emerald, coral.

Sentiments (147 entries; 145 unique). emotion, smile, happy, sad, facial expression, laugh, tears, frown, grin, joy, crying, anger, surprised, content, excited, worried, disappointed, cheerful, upset, blushing, grief, serious, delighted, anxious, fear, confusion, disgust, smirk, love, hate, relaxed, nervous, bored, concentration, amused, grumpy, pensive, thoughtful, frustrated, ecstatic, mournful, proud, relief, resentful, contentment, admiration, determined, gloomy, bashful, intense, awkward, reluctant, insecure, curious, playful, skeptical, sympathetic, bewildered, elated, optimistic, disheartened, triumphant, indifferent, jealous, melancholy, pleased, ashamed, grateful, loneliness, lonely, ambition, ambitious, despair, despairing, hope, hopeful, solitude, solitary, desire, desirous, freedom, free, conflict, conflicted, tranquility, tranquil, chaotic, powerful, weakness, trust, trustworthy, betrayal, betrayed, mystery, mysterious, identity, identifiable, memorable, imagination, imaginative, dreamlike, nightmare, nightmarish, inspiration, inspirational, faith, faithful, doubt, doubtful, regret, regretful, perseverance, persevering, illusion, illusory, destiny, destined, karma, karmic, purposeful, fated, realistic, transformation, transformative, fearlessness, fearless, emptiness, serenity, serene, wisdom, wise, vulnerability, vulnerable, alienation, alienated, acceptance, accepting, yearning, yearning, ambivalence, ambivalent, hopefulness, hopeful, liberation, liberated, introspection, introspective.

The vocabulary contains two repeated entries, *yearning* and *hopeful*; repeated token IDs are deduplicated when constructing the concept dictionary.

Quantity (68 entries). zero, one, two, three, four, five, six, seven, eight, nine, ten, eleven, twelve, twenty, hundred, thousand, million, billion, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 100, half, quarter, third, some, few, several, many, numerous, multiple, various, couple, handful, dozens, hundreds, thousands, plenty, a lot, lots, countless, more, less, fewer, most, least, enough, too many, too few, no, none, nothing, pair, set, bunch, group, amount, number, quantity, total.

B.5 Dictionary size after per-token expansion

The dictionary size $\lvert\mathcal{D}\rvert$ counts unique token-level unembedding vectors, rather than vocabulary entries. Each entry contributes one atom per token, with repeated tokens deduplicated across the vocabulary. Table 4 reports the vocabulary and dictionary sizes used in the main experiments.

C SOMP-BASED HEAD AND ATOM SELECTION

In this section, we describe the Simultaneous Orthogonal Matching Pursuit (SOMP) procedure used to select concept-relevant heads and construct $\mathrm{DASH}_{\mathrm{multi}}$ subspaces. Following the head-selection framework of Basile et al. (2025), we score heads using calibration activations and a concept dictionary. SOMP is applied separately to each head: it greedily selects *dictionary atoms* that reconstruct that head's calibration encodings. The resulting reconstruction scores are then used to rank heads. The atom budget k_{SOMP}, selected-head budget K, and DASH editing rank r are distinct quantities. Below, we specify the calibration and numerical details of our implementation, followed by DASH's use of the selected atoms for within-head editing.

Concept dictionary. For concept c, we tokenize each dictionary word without special tokens and retain all distinct subword token IDs. Each retained ID t_m defines a unit-norm unembedding direction

$$a_m = \frac{W_U[t_m]}{\|W_U[t_m]\|_2} \in \mathbb{R}^{d_{\mathrm{model}}}, \qquad m = 1, \dots, M.$$

We treat a_m as a column vector and write $D_c = [a_1, \dots, a_M]$. The atoms are normalized individually before selection, so their unembedding norms do not determine their correlation scores.

Calibration encodings. For each calibration example x_n, we collect head activations before the attention output projection and average over the calibration token positions T_n:

$$\bar{z}_{\ell,h,n} = \frac{1}{|T_n|} \sum_{t \in T_n} z_{\ell,h}(x_n, t) \in \mathbb{R}^{d_h}.$$

Text calibration pools over the content-token span of the user prompt. Vision–language calibration uses an image-only conversation and pools over image-token positions. These encodings are obtained from forward passes of the unedited model; calibration does not require generated continuations.

To compare head contributions with unembedding atoms, we map $\bar{z}_{\ell,h,n}$ through the head's output projection and apply the pooled final-RMSNorm scaling used by the encoding pipeline. Let $\bar{r}_n$ be the final pre-RMSNorm residual vector averaged over the same positions, and let γ and ϵ be the final RMSNorm parameters. Define

$$s_n = \frac{\|\bar{r}_n\|_2^2}{d_{\mathrm{model}}}, \qquad x_{\ell,h,n} = \frac{\gamma \odot (W_{\ell,h}^O \bar{z}_{\ell,h,n})}{\sqrt{s_n + \epsilon}}.$$

The normalization denominator is shared across heads for a given example and comes from the pooled final residual, rather than from the head's own contribution. This pooled encoding approximates the tokenwise normalization; it is used for selection, not as an additional intervention. Stacking the row vectors $x_{\ell,h,n}^\top$ gives $X_{\ell,h} \in \mathbb{R}^{N_{\mathrm{calib}} \times d_{\mathrm{model}}}$.

Greedy atom selection. For a fixed head, we suppress the (ℓ, h) indices. The procedure goes as follows: First, center its encodings across calibration examples:

$$\widetilde{X} = X - \mathbf{1}\bar{x}^\top, \qquad \bar{x} = \frac{1}{N_{\mathrm{calib}}} \sum_n x_n.$$

Next, initialize the ordered atom list $S_0 = ()$ and residual $R_0 = \widetilde{X}$. At iteration j, select the unused atom with the largest sum of absolute correlations across examples:

$$s_j = \operatorname*{arg\,max}_{m \notin S_{j-1}} \|R_{j-1} a_m\|_1, \qquad S_j = (s_1, \dots, s_j).$$

Thus, all calibration examples share the selected atom set, while their reconstruction coefficients may differ. Writing $B_j = [a_{s_1}, \dots, a_{s_j}]$, we refit the coefficients against the original centered encodings and recompute the residual:

$$C_j = \operatorname*{arg\,min}_{C \in \mathbb{R}^{j \times N_{\mathrm{calib}}}} \|\widetilde{X}^\top - B_j C\|_F^2, \tag{1}$$

$$\widehat{X}_j = (B_j C_j)^\top, \qquad R_j = \widetilde{X} - \widehat{X}_j. \tag{2}$$

At each iteration, coefficients are refit over all selected atoms, and the residual is recomputed from the original centered encodings. Our text implementation uses a least-squares solver, the vision implementation solves the normal equations with a $10^{-9}I$ diagonal regularizer.

Head ranking. For each head, we measure the fraction of calibration variance captured by the reconstruction after the SOMP atom budget:

$$e_{\ell,h}^{(c)} = \frac{\sum_{d=1}^{d_{\text{model}}} \text{Var}_n[\widehat{X}_{\ell,h,nd}]}{\sum_{d=1}^{d_{\text{model}}} \text{Var}_n[X_{\ell,h,nd}] + \varepsilon_{\text{num}}}, \qquad \varepsilon_{\text{num}} = 10^{-12}.$$

Variances use the population convention. For an exact orthogonal least-squares projection and nonzero input variance, this score is equivalent, up to the numerical stabilizer, to $1 - \|R\|_F^2/\|\widetilde{X}\|_F^2$. We select the K heads with the highest scores:

$$\mathcal{H}^{(c)} = \text{TopK}_{(\ell,h)}\, e_{\ell,h}^{(c)}.$$

The configured atom budget is $k_{\text{SOMP}} = 50$, capped by the dictionary size when it contains fewer atoms. Head ranking is performed after the independent per-head pursuits; there is no residual update across different heads. The score measures dictionary-based reconstruction of calibration activations and does not by itself establish semantic exclusivity.

From selected atoms to DASH subspaces. Head Pursuit uses the scores to select heads for whole-head rescaling. DASH additionally reuses the atom ordering to construct a concept subspace within each selected head. We retain the first r atoms and pull them into the pre-output-projection head space:

$$A_{\ell,h}^{(c)} = \left[(W_{\ell,h}^O)^\top a_{s_1}, \ldots, (W_{\ell,h}^O)^\top a_{s_r}\right], \qquad A_{\ell,h}^{(c)} = V_{\ell,h}^{(c)} R_{\ell,h}^{\text{QR}}.$$

A thin QR decomposition produces the orthonormal basis used by $\text{DASH}_{\text{multi}}$:

$$z'_{\ell,h} = z_{\ell,h} - \beta V_{\ell,h}^{(c)} (V_{\ell,h}^{(c)})^\top z_{\ell,h}.$$

The requested editing rank is capped by the atom budget and head dimension; when fewer atoms are available, only the valid selected prefix is used. Unless otherwise specified, $r = 10$. Consequently, the 50-atom scoring budget does not imply a rank-50 intervention. In the rank ablation, we vary the retained prefix while keeping the head-selection procedure fixed.

$\text{DASH}_{\text{mean}}$ uses the same selected heads but constructs its direction from the mean of the *full* normalized concept dictionary, rather than the SOMP-selected prefix. SOMP therefore determines the intervention sites for both variants, and additionally determines the atom ordering used by the multi-direction variant.

D EVALUATION DETAILS AND ADDITIONAL RESULTS

D.1 TOXICITY SUPPRESSION

We provide additional details on the toxicity suppression experiments in Section 4.2.1 and report results across intervention strengths and head budgets, including a classifier-based measure of toxicity.

D.1.1 SETUP

Our evaluation protocol follows Basile et al. (2025).

Datasets and splits. We partition each dataset into disjoint calibration and evaluation pools using seed 42:

- **RealToxicityPrompts (RTP).** We split the `train` set of `allenai/real-toxicity-prompts` into 20% calibration ($\sim$19,888 prompts) and 80% evaluation ($\sim$ 79,553 prompts). For calibration, we use 5,000 prompts from the calibration set. For evaluation, both the main-text toxicity table and the expanded results in Table 5 use 5,000 prompts from the evaluation set. Intervention-strength sweeps (Pareto plots) use 1,000 prompts from the evaluation set.
- **Thoroughly Engineered Toxicity (TET).** We divide the `train` set of `convoicon/Thoroughly_Engineered_Toxicity` equally into calibration and evaluation pools of 1,273 prompts each. We use the full calibration pool for calibration. For evaluation, both the main-text toxicity table and Table 5 use all 1,273 prompts, while intervention-strength sweeps use 1,000 prompts.

Calibration sizes remain fixed across strength sweeps. Head selection uses forward-pass calibration encodings as described in Appendix C. Ablation-specific sample sizes and generation settings are given in Appendix E.

Generation. For each evaluation condition, we format prompts with Mistral's built-in chat template and generate continuations using greedy decoding and `max_new_tokens=50`. We measure target suppression and preservation relative to the unedited model.

Metrics. We report four complementary metrics:

- **Normalized lexical toxicity rate, $\tau_{\text{lex}}/\tau_{\text{lex},0}$.** Here, τ_{lex} is the fraction of continuations containing at least one held-out toxic-vocabulary match, using the matching rule below; $\tau_{\text{lex},0}$ is the unedited baseline. This is the primary toxicity metric in the main text.
- **Normalized classifier toxic-generation rate, $\tau_{\text{rob}}/\tau_{\text{rob},0}$.** The `s-nlp/roberta_toxicity_classifier` model (s-nlp, n.d.) labels a continuation toxic when its toxic-class probability exceeds 0.5, using a maximum classifier input length of 512 tokens. Here, τ_{rob} is the fraction labelled toxic and $\tau_{\text{rob},0}$ is its unedited baseline.
- **Normalized held-out toxic-token count, λ/λ_0.** We count test-vocabulary word occurrences in each lowercased continuation using word-boundary regular expressions. The mean count is λ, and λ_0 is its unedited baseline. The test vocabulary is disjoint from the vocabulary used to construct the concept dictionary $\mathcal{D}$.
- **Relative WikiText-2 perplexity change, $\Delta\text{ppl}_{\text{wt}}$.** We measure perplexity on 1,000 sentences from the `wikitext-2-raw-v1` test split of `Salesforce/wikitext` and report its percentage change relative to the unedited model:

$$\Delta\text{ppl}_{\text{wt}} = 100\left(\frac{\text{ppl}_{\text{intervention}}}{\text{ppl}_{\text{unedited}}} - 1\right).$$

  Negative values indicate improved perplexity relative to the unedited model.

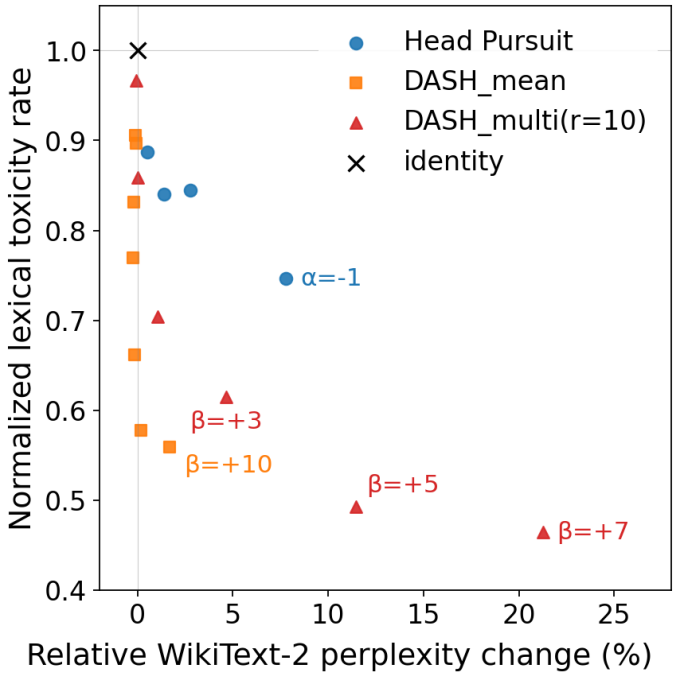


Figure 5: RTP strength sweep on 1,000 prompts at $K = 32$, varying α for HP and β for DASH. The vertical axis is normalized lexical toxicity and the horizontal axis relative WikiText-2 perplexity change (%). Selected points are annotated with their strengths.

D.1.2 Results

Table 5 expands the main toxicity results in Figure 2 to include both DASH variants at $\beta \in \{3, 5, 10\}$, alongside Head Pursuit at $\alpha = -1$, for $K \in \{8, 16, 32\}$. These results use 5,000 evaluation prompts for RTP and 1,273 for TET, as specified above.

RTP intervention-strength sweep. Figure 5 reports the RTP strength sweep on 1,000 evaluation prompts at $K = 32$, with 5,000 calibration prompts. The shared suppression strengths $\beta = 10$ for $\text{DASH}_{\text{mean}}$ and $\beta = 5$ for $\text{DASH}_{\text{multi}}$ were chosen by inspecting the preliminary RTP strength curves as representative control–preservation operating points, then retained across subsequent tasks and head budgets. The sweep shows their positions within the observed lexical-toxicity–perplexity trade-off.

Table 5: Toxicity suppression with Mistral-7B-Instruct-v0.2 and SOMP. Entries report $(\tau_{\text{lex}}/\tau_{\text{lex},0})$ / $(\tau_{\text{rob}}/\tau_{\text{rob},0})$ / (λ/λ_0) / $\Delta\text{ppl}_{\text{wt}}$ (%). DASH_MULTI uses $r = 10$.

RTP ($n_{\text{eval}} = 5{,}000$)			
Method and strength	$K = 8$	$K = 16$	$K = 32$
HP $\alpha = -1$	0.98 / 0.78 / 0.98 / +1.7	0.82 / 0.60 / 0.76 / +3.6	0.76 / 0.62 / 0.69 / +7.8
$\text{DASH}_{\text{mean}}$ $\beta = 3$	0.96 / 0.97 / 0.92 / -0.2	0.84 / 0.92 / 0.75 / -0.2	0.79 / 0.83 / 0.70 / -0.2
$\text{DASH}_{\text{mean}}$ $\beta = 5$	0.93 / 0.94 / 0.90 / -0.2	0.76 / 0.83 / 0.67 / -0.2	0.67 / 0.74 / 0.58 / -0.1
$\text{DASH}_{\text{mean}}$ $\beta = 10$	0.87 / 0.88 / 0.79 / +0.5	0.67 / 0.76 / 0.55 / +0.9	0.50 / 0.74 / 0.42 / +1.7
$\text{DASH}_{\text{multi}}$ $\beta = 3$	0.98 / 0.86 / 0.97 / +2.8	0.75 / 0.80 / 0.67 / +3.6	0.67 / 0.80 / 0.58 / +4.6
$\text{DASH}_{\text{multi}}$ $\beta = 5$	0.93 / 1.00 / 0.89 / +4.5	0.66 / 0.84 / 0.54 / +7.8	0.55 / 0.70 / 0.48 / +11.5
$\text{DASH}_{\text{multi}}$ $\beta = 10$	0.71 / 1.03 / 0.63 / +22.2	0.35 / 1.07 / 0.27 / +39.9	0.28 / 1.86 / 0.38 / +64.8
TET ($n_{\text{eval}} = 1{,}273$)			
Method and strength	$K = 8$	$K = 16$	$K = 32$
HP $\alpha = -1$	0.84 / 0.83 / 0.82 / +2.5	0.76 / 0.69 / 0.66 / +10.7	0.63 / 0.48 / 0.57 / +12.9
$\text{DASH}_{\text{mean}}$ $\beta = 3$	0.93 / 0.91 / 0.89 / +0.0	0.82 / 0.86 / 0.76 / +0.1	0.80 / 0.80 / 0.74 / +0.2
$\text{DASH}_{\text{mean}}$ $\beta = 5$	0.81 / 0.86 / 0.81 / +0.1	0.74 / 0.77 / 0.67 / +0.2	0.74 / 0.73 / 0.66 / +0.4
$\text{DASH}_{\text{mean}}$ $\beta = 10$	0.73 / 0.78 / 0.68 / +0.5	0.62 / 0.67 / 0.55 / +0.9	0.56 / 0.62 / 0.50 / +1.4
$\text{DASH}_{\text{multi}}$ $\beta = 3$	0.81 / 0.84 / 0.78 / +0.8	0.73 / 0.73 / 0.64 / +1.8	0.74 / 0.62 / 0.61 / +2.4
$\text{DASH}_{\text{multi}}$ $\beta = 5$	0.78 / 0.73 / 0.69 / +1.4	0.69 / 0.62 / 0.60 / +4.5	0.55 / 0.43 / 0.54 / +7.0
$\text{DASH}_{\text{multi}}$ $\beta = 10$	0.56 / 0.67 / 0.43 / +5.3	0.39 / 0.46 / 0.29 / +19.9	0.21 / 0.63 / 0.15 / +39.0

D.2 IMAGE CLASSIFICATION

In this section, we provide results on the image classification task. We evaluate suppression of image-classification capability while measuring preservation of captioning and language modelling. Following Basile et al. (2025), we use LLaVA-NeXT-7B on six datasets: MNIST, SVHN, GTSRB, EuroSAT, RESISC45, and DTD.

D.2.1 SETUP

Datasets and splits. We evaluate on the following datasets:

- **MNIST** (`ylecun/mnist`, 10 classes).
- **SVHN** (`ufldl-stanford/svhn`, `cropped_digits`, 10 classes).
- **GTSRB** (`tanganke/gtsrb`, 43 classes).
- **EuroSAT** (`tanganke/eurosat`, 10 land-use classes).
- **RESISC45** (`timm/resisc45`, 45 scene classes).
- **DTD** (`tanganke/dtd`, 47 texture classes).

For each dataset, both fixed-strength results and intervention-strength sweeps use 1,000 classification evaluation images, sampled using seed 42. Head selection uses the first 200 images from this evaluation set for calibration (`calib_split=test`). The class-label vocabularies used to construct the concept dictionaries are listed in Appendix B.

Prompting and generation. Following the prompt template of Head Pursuit, we ask:

> What {object} is depicted in this image? Answer EXACTLY with one of the following class names only:
> $\{\text{class}_1, \text{class}_2, \dots\}$ Answer:

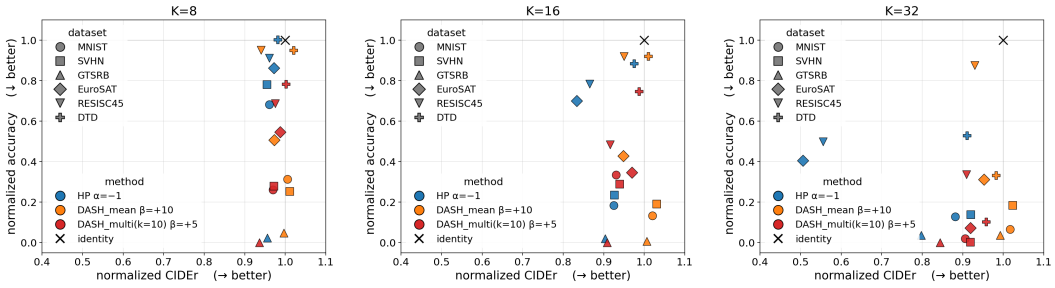
Here, {object} is *digit* for MNIST and SVHN, *traffic sign* for GTSRB, *texture* for DTD, and *remote sensing scene* for EuroSAT and RESISC45. The class-name list contains the dataset's full label vocabulary. We use greedy decoding with `max_new_tokens=50`. Following Head Pursuit, a prediction is counted as correct when the ground-truth class name appears as a substring of the lowercased continuation and no other class name appears; class names are also lowercased before matching. We also record F1 overlap between the prediction and ground truth for supplementary analysis.

Metrics. We measure target suppression using classification accuracy and preservation using captioning CIDEr and text-only perplexity:

- **Normalized classification accuracy, a/a_0.** Here, a is the fraction of evaluation images scored as correct under the label-inclusion rule above, and a_0 is the unedited-model accuracy. Lower ratios indicate stronger suppression. Unedited accuracies range from 0.30 to 0.85 across datasets (Table 6).
- **Normalized captioning CIDEr, c/c_0.** We generate captions for 200 held-out Flickr30k test images, score them against the reference captions, and normalise by the unedited-model score c_0. Values near 1 indicate preservation of baseline captioning performance on this evaluation set.
- **Relative WikiText perplexity change, $\Delta\text{ppl}_{\text{wt}}$.** We measure the language backbone's perplexity on 200 WikiText-2 test sentences using the relative percentage-change metric defined in Section D.1.1. This measures preservation of text-only language modelling alongside captioning performance.

D.2.2 RESULTS

Figure 6 compares the methods at canonical intervention strengths across $K \in \{8, 16, 32\}$, while Figure 7 sweeps intervention strength at $K = 32$. Table 6 provides numerical results, including WikiText perplexity changes. Together, they show that direction-aware editing generally improves the suppression–preservation trade-off.



(a) $K = 8$ (b) $K = 16$ (c) $K = 32$

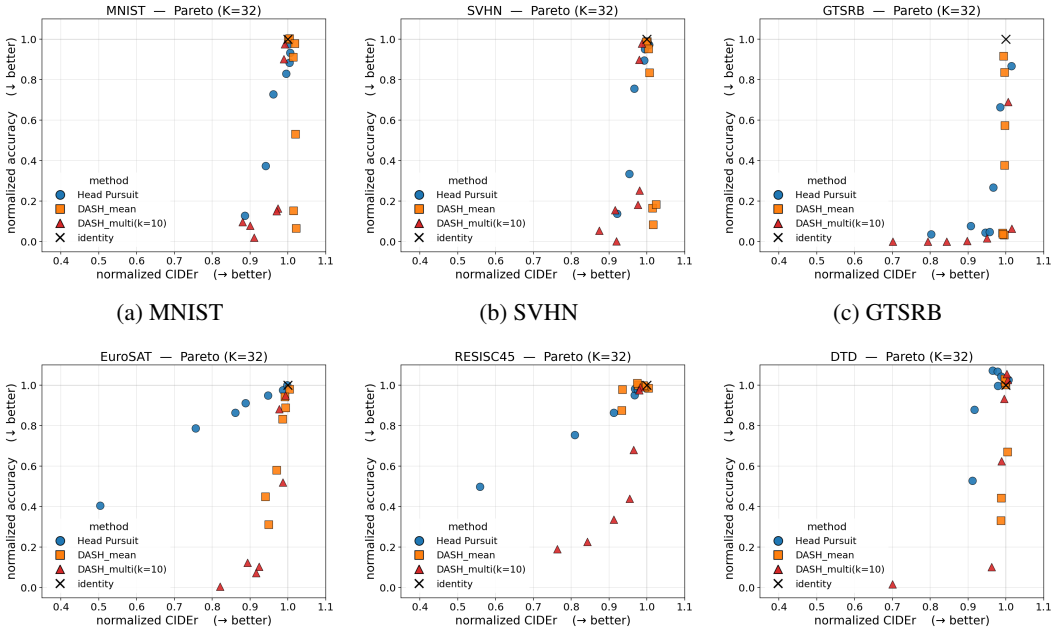
Figure 6: Image-classification suppression and captioning preservation at the canonical strengths: $\alpha = -1$ for Head Pursuit, $\beta = 10$ for DASH_MEAN, and $\beta = 5$ for $\text{DASH}_{\text{multi}}$ ($r = 10$). Panels vary the head budget $K \in \{8, 16, 32\}$. Lower normalized accuracy and higher normalized CIDEr are preferred. The cross marks the unedited model.

The two DASH variants provide complementary operating points. $\text{DASH}_{\text{mean}}$ favors preservation: at $\beta = 10$, normalized CIDEr remains between 0.93 and 1.03 across all datasets and head budgets, alongside strong suppression on MNIST, SVHN, and GTSRB. Its effect is more limited on RESISC45, where $\text{DASH}_{\text{multi}}$ achieves stronger suppression, as it also does on DTD. The relative advantage depends on the dataset and head budget, and the mean direction remains competitive on the digit datasets.

Increasing the head budget often strengthens suppression but can increase collateral effects. At $K = 32$, Head Pursuit substantially degrades captioning on EuroSAT and RESISC45, while DASH retains much higher CIDEr. At this budget, $\text{DASH}_{\text{multi}}$ with $\beta = 5$ improves all three metrics over Head Pursuit on EuroSAT, RESISC45, and DTD. On EuroSAT, for example, normalized accuracy falls from 0.41 to 0.07 while normalized CIDEr rises from 0.51 to 0.92. These gains are not uniformly monotonic in K.

The strength sweeps confirm that the advantages extend beyond the canonical settings. On MNIST and SVHN, $\text{DASH}_{\text{mean}}$ achieves strong suppression with nearly unchanged CIDEr; on GTSRB, it matches the strongest whole-head suppression with better captioning preservation. Both DASH variants exceed the observed Head Pursuit suppression range on EuroSAT while retaining substantially higher CIDEr. On RESISC45, this advantage is primarily obtained by $\text{DASH}_{\text{multi}}$.

Captioning preservation alone, however, can conceal language-model degradation. On SVHN, $\text{DASH}_{\text{multi}}$ at $K = 32$ and $\beta = 10$ retains normalized CIDEr of 0.92 despite a 44.5% increase in WikiText perplexity. On DTD, the same setting nearly eliminates classification accuracy but severely degrades captioning. Selecting an operating point therefore requires assessing target suppression alongside both preservation metrics.



(a) MNIST (b) SVHN (c) GTSRB

(d) EuroSAT (e) RESISC45 (f) DTD

Figure 7: Image-classification suppression and captioning preservation at $K = 32$, sweeping α for Head Pursuit and β for both DASH variants. The cross marks the unedited model. Lower normalized accuracy and higher normalized CIDEr are preferred.

Table 6: Image-classification suppression on LLaVA-NeXT-7B. Entries report normalized accuracy (a/a_0) / normalized CIDEr c/c_0 / relative WikiText perplexity change $\Delta\mathrm{ppl}_{\mathrm{wt}}$ (%). DASH_MULTI uses $r = 10$. Unedited accuracies are 0.84 (MNIST), 0.63 (SVHN), 0.31 (GTSRB), 0.46 (EuroSAT), 0.51 (RESISC45), and 0.41 (DTD).

(a) $K = 8$.

Operator	MNIST	SVHN	GTSRB	EuroSAT	RESISC45	DTD
HP α=−1	0.68 / 0.96 / +3.2	0.78 / 0.95 / +3.6	0.02 / 0.96 / +1.2	0.86 / 0.97 / +2.2	0.91 / 0.96 / +3.1	1.00 / 0.98 / +0.3
DASH_MEAN β=+3	0.95 / 1.01 / +0.3	0.96 / 1.00 / +0.3	0.76 / 1.00 / +0.0	0.88 / 0.99 / +0.2	0.99 / 0.98 / +0.1	1.02 / 1.01 / -0.2
DASH_MEAN β=+5	0.74 / 1.01 / +0.8	0.77 / 1.00 / +0.8	0.47 / 1.01 / +0.0	0.80 / 1.01 / +0.4	0.98 / 0.96 / +0.2	1.02 / 1.01 / -0.3
DASH_MEAN β=+10	0.31 / 1.01 / +2.7	0.25 / 1.01 / +3.1	0.05 / 1.00 / +0.1	0.51 / 0.97 / +1.2	0.95 / 0.94 / +0.7	0.95 / 1.02 / -0.2
DASH_MULTI β=+3	0.29 / 0.96 / +2.1	0.44 / 0.97 / +2.2	0.01 / 0.95 / +0.9	0.79 / 0.97 / +1.0	0.86 / 1.00 / +1.5	0.96 / 1.01 / -0.2
DASH_MULTI β=+5	0.26 / 0.97 / +5.0	0.28 / 0.97 / +5.5	0.00 / 0.94 / +1.7	0.55 / 0.99 / +2.9	0.69 / 0.98 / +4.2	0.78 / 1.00 / -0.1
DASH_MULTI β=+10	0.23 / 0.95 / +16.2	0.37 / 0.94 / +17.2	0.00 / 0.86 / +4.3	0.45 / 0.93 / +9.0	0.55 / 0.90 / +11.1	0.54 / 0.98 / +2.4

(b) $K = 16$.

Operator	MNIST	SVHN	GTSRB	EuroSAT	RESISC45	DTD
HP α=−1	0.18 / 0.92 / +5.1	0.23 / 0.93 / +6.5	0.02 / 0.90 / +4.3	0.70 / 0.83 / +5.4	0.78 / 0.87 / +4.1	0.88 / 0.98 / +1.9
DASH_MEAN β=+3	0.90 / 1.01 / +0.3	0.89 / 1.02 / +0.3	0.55 / 0.99 / +0.0	0.82 / 1.00 / +0.2	0.99 / 0.98 / +0.1	1.02 / 1.00 / -0.2
DASH_MEAN β=+5	0.56 / 1.01 / +0.9	0.57 / 1.01 / +0.9	0.17 / 0.99 / +0.1	0.70 / 1.01 / +0.4	1.01 / 0.95 / +0.2	1.02 / 1.01 / -0.3
DASH_MEAN β=+10	0.13 / 1.02 / +3.4	0.19 / 1.03 / +3.3	0.01 / 1.01 / +0.5	0.43 / 0.95 / +1.3	0.92 / 0.95 / +0.7	0.92 / 1.01 / -0.0
DASH_MULTI β=+3	0.17 / 0.97 / +3.5	0.21 / 0.98 / +4.0	0.00 / 0.94 / +1.5	0.42 / 0.99 / +1.6	0.71 / 0.96 / +1.9	0.89 / 1.00 / +0.1
DASH_MULTI β=+5	0.34 / 0.93 / +6.9	0.29 / 0.94 / +7.3	0.00 / 0.91 / +3.4	0.35 / 0.97 / +4.6	0.48 / 0.92 / +5.3	0.75 / 0.99 / +0.9
DASH_MULTI β=+10	0.12 / 0.93 / +18.6	0.19 / 0.94 / +18.7	0.00 / 0.86 / +9.4	0.30 / 0.88 / +12.3	0.32 / 0.85 / +13.6	0.22 / 0.93 / +7.4

(c) $K = 32$.

Operator	MNIST	SVHN	GTSRB	EuroSAT	RESISC45	DTD
HP α=−1	0.13 / 0.88 / +20.9	0.14 / 0.92 / +22.3	0.04 / 0.80 / +8.9	0.41 / 0.51 / +28.1	0.50 / 0.56 / +23.0	0.53 / 0.91 / +6.6
DASH_MEAN β=+3	0.91 / 1.01 / +0.4	0.83 / 1.01 / +0.6	0.38 / 1.00 / +0.1	0.83 / 0.99 / +0.3	1.00 / 0.97 / +0.2	1.03 / 0.99 / -0.6
DASH_MEAN β=+5	0.53 / 1.02 / +1.3	0.17 / 1.01 / +2.2	0.04 / 0.99 / +0.3	0.58 / 0.97 / +0.9	1.01 / 0.97 / +0.5	0.67 / 1.00 / -0.8
DASH_MEAN β=+10	0.07 / 1.02 / +6.1	0.18 / 1.02 / +12.7	0.04 / 0.99 / +2.1	0.31 / 0.95 / +3.4	0.88 / 0.93 / +2.1	0.33 / 0.98 / +0.3
DASH_MULTI β=+3	0.15 / 0.97 / +5.0	0.18 / 0.98 / +5.2	0.00 / 0.90 / +2.9	0.10 / 0.93 / +5.4	0.44 / 0.95 / +5.5	0.62 / 0.98 / +0.1
DASH_MULTI β=+5	0.02 / 0.91 / +11.3	0.00 / 0.92 / +13.2	0.00 / 0.84 / +6.8	0.07 / 0.92 / +12.1	0.34 / 0.91 / +12.1	0.10 / 0.96 / +2.0
DASH_MULTI β=+10	0.10 / 0.88 / +56.5	0.16 / 0.92 / +44.5	0.00 / 0.70 / +22.8	0.00 / 0.82 / +37.8	0.19 / 0.76 / +31.7	0.00 / 0.25 / +19.0

D.3 CAPTIONING SUPPRESSION/AMPLIFICATION

We provide additional details on the captioning experiments in Section 4.2.2, covering attribute suppression and amplification across intervention strengths and head budgets.

D.3.1 SETUP

Dataset and splits. We use the Flickr30k dataset, and divide its built-in `test` split (∼31,000 image–caption groups) equally into disjoint calibration and evaluation pools using seed 42. Within each pool, we subsample using a fixed permutation with the same seed.

For each attribute, we use 1,000 calibration images. Fixed-strength results use 5,000 evaluation images; intervention-strength sweeps use 1,000, with the same calibration set. Ablation-specific sample sizes are given in Appendix E.

Attributes and head selection. We consider three target attributes: *colors*, *sentiments*, and *quantity*. Their vocabularies are listed in Appendix B and are used for both SOMP head selection and lexical evaluation. Calibration uses the image-only encoding protocol in Appendix C.

Prompting and generation. We use the prompt:

> Provide a one-sentence caption for the provided image.

The prompt and image are formatted using the model's chat template. For each evaluation image, we generate one caption with greedy decoding and `max_new_tokens=50`. Generated captions are lowercased before metric computation.

Metrics. We measure attribute steering and captioning preservation using three metrics, each normalized by its unedited-model baseline:

- **Normalized property count, λ_a/λ_a^0.** For attribute a, λ_a is the mean number of vocabulary matches per generated caption, computed using word-boundary regular expressions, and λ_a^0 is the unedited baseline. Values below 1 indicate suppression; values above 1 indicate amplification. This is our primary target-control metric.
- **Normalized attribute presence, p_a/p_a^0.** Here, p_a is the fraction of captions containing at least one match for attribute a, and p_a^0 is its unedited baseline. Together with the property count, this distinguishes changes in how often an attribute appears across captions from changes in the number of mentions within a caption.
- **Normalized CIDEr, c/c_0.** We compute CIDEr (Vedantam et al., 2015) metric against the Flickr30k reference captions. Here, c is the score under intervention and c_0 is the unedited-model score on the same images. Values near 1 indicate preservation of baseline captioning performance.

D.3.2 RESULTS

Table 7 extends the main-text results across head budgets and intervention strengths, while Figure 8 shows the control–preservation trade-offs at $K = 32$. The table uses 5,000 evaluation images; the sweeps use 1,000. The results reinforce the complementary roles of the two DASH variants: $\text{DASH}_{\text{mean}}$ generally favors preservation, while $\text{DASH}_{\text{multi}}$ enables stronger attribute control.

For suppression, the advantage is particularly clear for colors. At $K = 32$, $\text{DASH}_{\text{mean}}$ with $\beta = 10$ reduces the normalized property count to 0.03 while retaining CIDEr 0.94, compared with 0.13 and 0.75 for Head Pursuit. Sentiment mentions can be nearly eliminated by either Head Pursuit or the multi-direction variant with little captioning degradation, leaving less room for improvement. Quantity presents a sharper trade-off: the mean direction preserves captioning but provides limited suppression, while multi-direction editing achieves much stronger suppression at a larger CIDEr cost. The sweeps show the same attribute-dependent pattern, rather than a uniform advantage at every operating point.

For amplification, $\text{DASH}_{\text{multi}}$ substantially extends the attainable control range. At $K = 32$ and $\beta = -5$, it raises the normalized sentiment count to 18.50 with CIDEr 0.87, compared with 5.54

Table 7: Captioning attribute suppression/amplification on LLaVA-NeXT-7B, evaluated on 5,000 Flickr30k images. Entries report normalized property count (λ_a/λ_a^0) / attribute presence (p_a/p_a^0) / CIDEr (c/c_0). Head Pursuit uses $\alpha = -1$ for suppression and $\alpha = 5$ for amplification; $\text{DASH}_{\text{multi}}$ uses $r = 10$.

(a) **Suppression** (β>0 for DASH; α<1 for HP). Lower λ_a/λ_a^0 = stronger suppression.

Colors			
Method and strength	K=8	K=16	K=32
HP α=−1	0.26 / 0.27 / 0.95	0.21 / 0.18 / 0.90	0.13 / 0.09 / 0.75
$\text{DASH}_{\text{mean}}$ β=+1	0.97 / 0.97 / 1.00	0.91 / 0.93 / 1.00	0.84 / 0.86 / 1.00
$\text{DASH}_{\text{mean}}$ β=+3	0.88 / 0.90 / 0.99	0.68 / 0.72 / 0.99	0.48 / 0.51 / 0.98
$\text{DASH}_{\text{mean}}$ β=+5	0.78 / 0.81 / 0.99	0.48 / 0.51 / 0.98	0.20 / 0.23 / 0.96
$\text{DASH}_{\text{mean}}$ β=+10	0.56 / 0.60 / 0.98	0.17 / 0.19 / 0.95	0.03 / 0.03 / 0.94
$\text{DASH}_{\text{multi}}$ β=+1	0.61 / 0.64 / 0.99	0.44 / 0.49 / 0.97	0.19 / 0.22 / 0.96
$\text{DASH}_{\text{multi}}$ β=+3	0.19 / 0.18 / 0.95	0.07 / 0.07 / 0.94	0.02 / 0.02 / 0.93
$\text{DASH}_{\text{multi}}$ β=+5	0.31 / 0.26 / 0.94	0.05 / 0.05 / 0.92	0.01 / 0.02 / 0.91
$\text{DASH}_{\text{multi}}$ β=+10	0.88 / 0.66 / 0.90	0.21 / 0.18 / 0.87	0.16 / 0.14 / 0.83
Sentiments			
Method and strength	K=8	K=16	K=32
HP α=−1	0.15 / 0.16 / 0.98	0.00 / 0.00 / 0.97	0.04 / 0.04 / 0.95
$\text{DASH}_{\text{mean}}$ β=+1	0.65 / 0.64 / 1.00	0.65 / 0.64 / 1.00	0.69 / 0.68 / 1.00
$\text{DASH}_{\text{mean}}$ β=+3	0.62 / 0.60 / 1.00	0.50 / 0.48 / 1.00	0.54 / 0.52 / 1.00
$\text{DASH}_{\text{mean}}$ β=+5	0.46 / 0.44 / 1.00	0.46 / 0.44 / 1.00	0.46 / 0.44 / 1.00
$\text{DASH}_{\text{mean}}$ β=+10	0.27 / 0.28 / 1.01	0.27 / 0.28 / 1.00	0.23 / 0.24 / 1.00
$\text{DASH}_{\text{multi}}$ β=+1	0.50 / 0.48 / 1.00	0.35 / 0.32 / 1.00	0.50 / 0.48 / 1.00
$\text{DASH}_{\text{multi}}$ β=+3	0.00 / 0.00 / 1.00	0.04 / 0.04 / 1.01	0.00 / 0.00 / 0.99
$\text{DASH}_{\text{multi}}$ β=+5	0.08 / 0.08 / 0.99	0.04 / 0.04 / 0.99	0.00 / 0.00 / 0.96
$\text{DASH}_{\text{multi}}$ β=+10	0.04 / 0.04 / 0.96	0.04 / 0.04 / 0.93	0.15 / 0.16 / 0.87
Quantity			
Method and strength	K=8	K=16	K=32
HP α=−1	0.84 / 0.85 / 0.97	0.37 / 0.38 / 0.89	0.08 / 0.08 / 0.79
$\text{DASH}_{\text{mean}}$ β=+1	0.99 / 1.00 / 1.00	0.98 / 0.98 / 1.00	0.97 / 0.97 / 1.00
$\text{DASH}_{\text{mean}}$ β=+3	0.99 / 0.99 / 1.00	0.94 / 0.95 / 0.99	0.91 / 0.91 / 0.99
$\text{DASH}_{\text{mean}}$ β=+5	0.98 / 0.98 / 0.99	0.91 / 0.91 / 0.98	0.86 / 0.86 / 0.98
$\text{DASH}_{\text{mean}}$ β=+10	0.96 / 0.96 / 0.99	0.78 / 0.79 / 0.97	0.71 / 0.71 / 0.95
$\text{DASH}_{\text{multi}}$ β=+1	0.94 / 0.94 / 1.00	0.76 / 0.76 / 0.97	0.61 / 0.61 / 0.96
$\text{DASH}_{\text{multi}}$ β=+3	0.73 / 0.73 / 0.98	0.22 / 0.23 / 0.89	0.09 / 0.09 / 0.83
$\text{DASH}_{\text{multi}}$ β=+5	0.46 / 0.46 / 0.95	0.08 / 0.08 / 0.82	0.03 / 0.03 / 0.75
$\text{DASH}_{\text{multi}}$ β=+10	0.42 / 0.38 / 0.95	0.05 / 0.04 / 0.72	0.22 / 0.06 / 0.61

and 0.88 for Head Pursuit. The normalized attribute-presence ratios are 18.88 and 5.68, respectively, indicating that amplification reaches more captions as well as increasing the total number of mentions. Color amplification also improves, whereas stronger quantity amplification comes with a more pronounced preservation cost. The mean-direction variant typically produces smaller changes while keeping CIDEr closer to baseline.

Larger head budgets and intervention magnitudes often strengthen control, but can also degrade captioning or weaken suppression. At $K = 32$, increasing the multi-direction suppression strength from $\beta = 5$ to $\beta = 10$ worsens both the property count and CIDEr for all three attributes. Strong amplification can also change the nature of the effect: for colors at $\beta = -10$, the count ratio reaches 14.56 while the presence ratio is only 3.61, indicating substantially more mentions per attribute-containing caption. These results favor selecting strengths using both target-control metrics and CIDEr, rather than maximizing the attribute count alone.

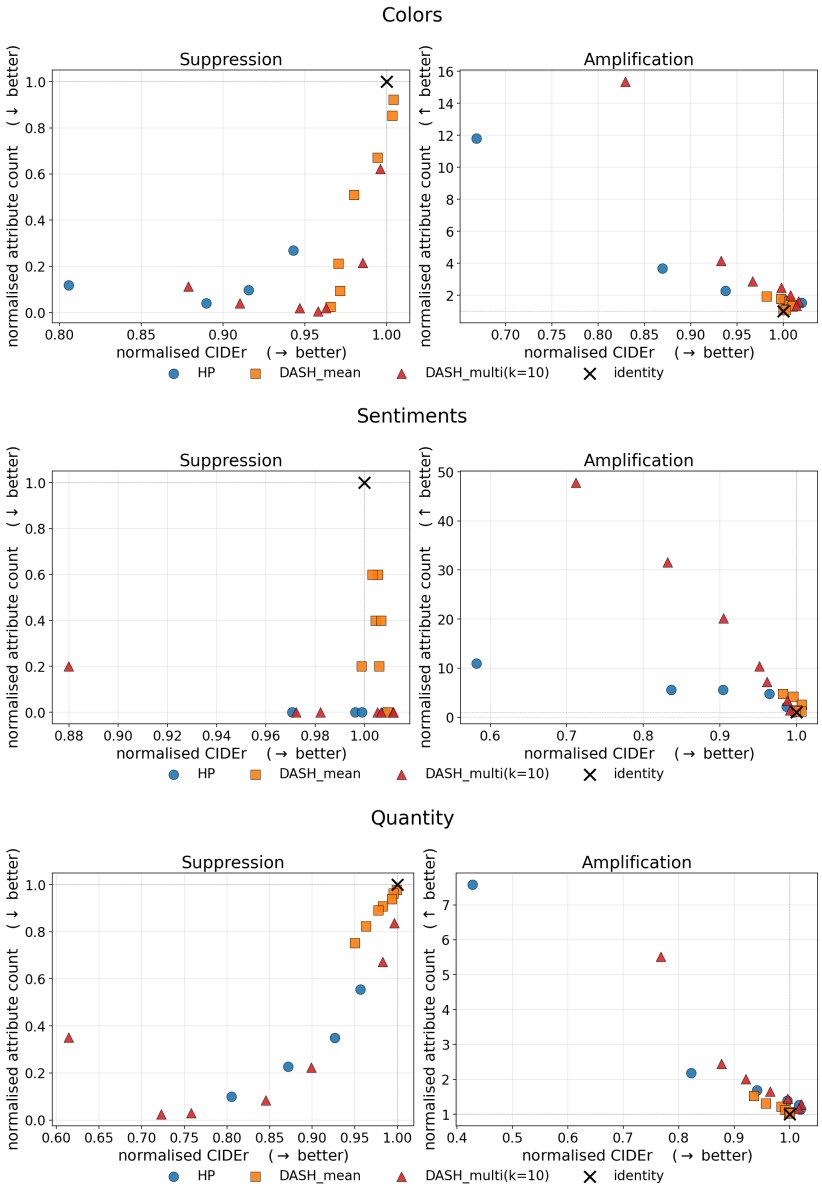


Figure 8: Captioning attribute steering at $K = 32$, sweeping intervention strength for Head Pursuit and both DASH variants ($r = 10$ for $\text{DASH}_{\text{multi}}$). Rows show colors, sentiments, and quantity; columns show suppression and amplification. Each point reports normalized property count versus normalized CIDEr on 1,000 Flickr30k images. Lower counts are better for suppression and higher counts for amplification. Higher CIDEr indicates better captioning preservation. Crosses mark the unedited model.

Table 7: Captioning attribute suppression/amplification on LLaVA-NeXT-7B (continued)

(b) **Amplification** ($\beta<0$ for DASH; $\alpha>1$ for HP). Higher λ_a/λ_a^0 = stronger amplification.

Colors			
Method and strength	$K=8$	$K=16$	$K=32$
HP $\alpha=+5$	1.43 / 1.32 / 0.99	1.39 / 1.28 / 0.95	2.29 / 1.99 / 0.95
$\text{DASH}_{\text{mean}}$ $\beta=-1$	1.04 / 1.03 / 1.00	1.11 / 1.10 / 1.00	1.16 / 1.12 / 1.00
$\text{DASH}_{\text{mean}}$ $\beta=-3$	1.12 / 1.09 / 1.00	1.28 / 1.22 / 1.01	1.38 / 1.32 / 1.00
$\text{DASH}_{\text{mean}}$ $\beta=-5$	1.18 / 1.15 / 1.00	1.42 / 1.35 / 1.00	1.55 / 1.46 / 0.99
$\text{DASH}_{\text{mean}}$ $\beta=-10$	1.31 / 1.24 / 1.01	1.68 / 1.52 / 0.99	1.90 / 1.72 / 0.98
$\text{DASH}_{\text{multi}}$ $\beta=-1$	1.26 / 1.21 / 1.01	1.40 / 1.32 / 1.01	1.60 / 1.49 / 1.01
$\text{DASH}_{\text{multi}}$ $\beta=-3$	1.53 / 1.41 / 1.01	1.63 / 1.49 / 0.99	2.36 / 2.09 / 0.99
$\text{DASH}_{\text{multi}}$ $\beta=-5$	1.57 / 1.44 / 1.00	1.83 / 1.61 / 0.98	2.99 / 2.49 / 0.96
$\text{DASH}_{\text{multi}}$ $\beta=-10$	1.79 / 1.54 / 0.98	3.44 / 2.04 / 0.93	14.56 / 3.61 / 0.83
Sentiments			
Method and strength	$K=8$	$K=16$	$K=32$
HP $\alpha=+5$	4.92 / 5.08 / 0.94	5.54 / 5.56 / 0.92	5.54 / 5.68 / 0.88
$\text{DASH}_{\text{mean}}$ $\beta=-1$	1.23 / 1.20 / 1.00	1.27 / 1.24 / 1.00	1.19 / 1.16 / 1.00
$\text{DASH}_{\text{mean}}$ $\beta=-3$	1.73 / 1.72 / 0.99	2.15 / 2.12 / 0.99	1.85 / 1.80 / 0.99
$\text{DASH}_{\text{mean}}$ $\beta=-5$	2.19 / 2.16 / 0.99	2.96 / 2.96 / 0.99	2.50 / 2.48 / 0.99
$\text{DASH}_{\text{mean}}$ $\beta=-10$	3.54 / 3.52 / 0.98	6.58 / 6.52 / 0.97	4.65 / 4.68 / 0.97
$\text{DASH}_{\text{multi}}$ $\beta=-1$	2.23 / 2.16 / 0.99	2.85 / 2.80 / 0.99	3.08 / 3.04 / 0.99
$\text{DASH}_{\text{multi}}$ $\beta=-3$	5.23 / 5.32 / 0.96	8.31 / 8.52 / 0.95	10.38 / 10.60 / 0.93
$\text{DASH}_{\text{multi}}$ $\beta=-5$	8.38 / 8.56 / 0.93	14.08 / 14.16 / 0.90	18.50 / 18.88 / 0.87
$\text{DASH}_{\text{multi}}$ $\beta=-10$	12.42 / 12.52 / 0.82	30.92 / 30.20 / 0.77	43.96 / 42.88 / 0.70
Quantity			
Method and strength	$K=8$	$K=16$	$K=32$
HP $\alpha=+5$	1.16 / 1.14 / 1.03	1.29 / 1.27 / 1.04	1.62 / 1.55 / 0.91
$\text{DASH}_{\text{mean}}$ $\beta=-1$	1.01 / 1.01 / 1.00	1.01 / 1.01 / 1.00	1.02 / 1.02 / 1.00
$\text{DASH}_{\text{mean}}$ $\beta=-3$	1.01 / 1.01 / 1.00	1.04 / 1.04 / 1.00	1.11 / 1.11 / 1.00
$\text{DASH}_{\text{mean}}$ $\beta=-5$	1.01 / 1.01 / 1.00	1.06 / 1.06 / 1.00	1.18 / 1.15 / 0.99
$\text{DASH}_{\text{mean}}$ $\beta=-10$	1.03 / 1.03 / 1.01	1.10 / 1.09 / 0.99	1.45 / 1.31 / 0.92
$\text{DASH}_{\text{multi}}$ $\beta=-1$	1.05 / 1.05 / 1.00	1.12 / 1.12 / 1.01	1.23 / 1.23 / 1.00
$\text{DASH}_{\text{multi}}$ $\beta=-3$	1.12 / 1.11 / 1.00	1.30 / 1.28 / 1.00	1.60 / 1.57 / 0.94
$\text{DASH}_{\text{multi}}$ $\beta=-5$	1.18 / 1.17 / 1.01	1.43 / 1.39 / 0.99	1.93 / 1.80 / 0.88
$\text{DASH}_{\text{multi}}$ $\beta=-10$	1.44 / 1.30 / 1.04	1.89 / 1.55 / 0.97	5.24 / 2.09 / 0.73

D.3.3 ADDITIONAL QUALITATIVE EXAMPLES

Figures 9 and 10 present eight selected cases from a 50-image Flickr30k pool, illustrating selective edits and failure modes in color steering. All use LLaVA-NeXT-7B with $K = 32$ SOMP-selected heads. Suppression uses HP $\alpha = -1$ and DASH $\beta = 5$; amplification uses HP $\alpha = 5$ and DASH $\beta = -5$. Both DASH variants use these strengths, with $r = 10$ for the multi-direction variant. Captions are reproduced as exported; bold words match the color lexicon. The examples show how vocabulary changes relate to preservation of entities, actions, and scene relations.

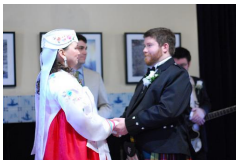
(a) Color removal with action preservation *Suppression*



Original	woman in **blue** sweater bending over to pick up her purse.
HP	a woman is standing outside a store, holding a bag.
$\text{DASH}_{\text{mean}}$	woman bending over to pick up her purse.
$\text{DASH}_{\text{multi}}$	a woman is bending over to pick up her purse.

Both DASH variants remove 'blue' while retaining the bending action and purse; HP changes the action to standing and holding a bag.

(b) Preservation when no color is mentioned *Suppression*



Original	a bride and groom shaking hands.
HP	two people holding hands in front of a wall.
$\text{DASH}_{\text{mean}}$	a bride and groom shaking hands.
$\text{DASH}_{\text{multi}}$	a bride and groom shaking hands.

Both DASH variants retain the bride–groom description and action. HP replaces these roles with 'two people'.

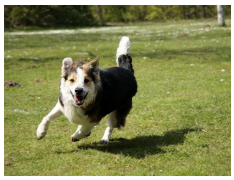
(c) Color addition with scene preservation *Amplification*



Original	a dog jumping over a hurdle.
HP	a **black** and **white** dog jumping over a **yellow** and **red** hurdle.
$\text{DASH}_{\text{mean}}$	a dog jumping over a hurdle.
$\text{DASH}_{\text{multi}}$	a **black** and **white** dog jumping over a **yellow** and **red** hurdle.

HP and the multi-direction variant add dog and hurdle colors while retaining the action; the mean variant leaves the caption unchanged.

(d) Complementary behavior across variants *Amplification*



Original	a dog running in a field.
HP	a **black** and **white** dog running in a grassy field.
$\text{DASH}_{\text{mean}}$	a **black** and **white** dog running in a field.
$\text{DASH}_{\text{multi}}$	a dog running through a grassy field.

The mean variant adds 'black and white' while preserving the original scene description. The multi-direction variant adds no color terms.

Figure 9: Selected color-steering examples with preserved scene descriptions. All methods use the strengths specified above; bold words match the color lexicon. Panels illustrate selective removal, preservation of a caption without color terms, and different amplification responses across methods.

(e) Loss of descriptive detail *Suppression*



Original	a boy in a **red** shirt is standing on a fountain.
HP	a boy is playing with a frisbee in a park.
$\text{DASH}_{\text{mean}}$	a boy in a park.
$\text{DASH}_{\text{multi}}$	a boy is standing on a fountain.

The mean variant removes the color but also drops the original action and fountain reference. The multi-direction variant preserves both; HP introduces a frisbee.

(f) Amplification can reduce color mentions *Amplification*



Original	a young girl in a **pink** hoodie and **pink** pants is running on a sidewalk.
HP	a young girl in a **pink** polka dot hoodie and **pink** pants is running on a sidewalk.
$\text{DASH}_{\text{mean}}$	a young girl running on a sidewalk.
$\text{DASH}_{\text{multi}}$	a young girl in a **pink** and **white** hoodie and **pink** pants is running on a sidewalk.

The mean variant removes both 'pink' mentions and the clothing details. HP retains the colors, while the multi-direction variant adds 'white'.

(g) Lexical ambiguity and object substitution *Amplification*



Original	a person holding an egg in a bowl.
HP	a child cracking an egg into a bowl.
$\text{DASH}_{\text{mean}}$	a person holding an **orange** in a **silver** bowl.
$\text{DASH}_{\text{multi}}$	a person holding an **orange** in a **silver** bowl.

Both DASH variants replace 'egg' with 'orange'. Here 'orange' denotes a fruit, but contributes to the color count alongside 'silver'.

(h) Color addition with a changed scene relation *Amplification*



Original	two men shaking hands under an umbrella.
HP	two men shaking hands under an **orange** building.
$\text{DASH}_{\text{mean}}$	two men shaking hands under an **orange** building.
$\text{DASH}_{\text{multi}}$	two men shaking hands under an **orange** building.

All three edits add 'orange' but replace 'under an umbrella' with 'under an orange building', altering the spatial description.

Figure 10: Selected color-steering failure cases at the same strengths. Failures include loss of descriptive detail, reduced color mentions under amplification, a fruit noun counted as a color, and altered spatial relations. Panels compare both DASH variants with HP and the unedited caption.

D.4 Additional multi-attribute captioning results

We extend the main-text composition experiments to all three pairs of captioning attributes: colors–sentiments, colors–quantity, and sentiments–quantity. For each pair, we evaluate joint suppression, joint amplification, and mixed steering. We report all available DASH strengths to characterize the range of control–preservation trade-offs, rather than selecting a single operating point.

D.4.1 Setup

We follow the LLaVA-NeXT-7B captioning protocol in Appendix D.3, using 1,000 calibration images and 1,000 evaluation images from Flickr30k, with greedy decoding limited to 50 new tokens. Each concept uses $K = 32$ SOMP-selected heads. Colors–sentiments shares six heads, giving 58 distinct intervention sites. The other two pairs each share three heads, giving 61 sites. We retain these overlaps and sum DASH's concept-specific updates within shared heads.

Intervention strengths. Combining the fixed-strength experiments and additional sweeps gives magnitudes $b \in \{3, 5, 7, 10, 15\}$ for $\text{DASH}_{\text{mean}}$ and $b \in \{1, 2, 3, 5, 7, 10\}$ for $\text{DASH}_{\text{multi}}$, with $r = 10$. For the ordered concept pair (A, B), we use $(\beta_A, \beta_B) = (b, b)$ for joint suppression, $(-b, -b)$ for joint amplification, and $(b, -b)$ for mixed steering. These grids are available for eight of the nine pair–mode combinations. For sentiments–quantity mixed steering, only the fixed strengths $(10, -10)$ for the mean variant and $(5, -5)$ for the multi-direction variant are available. Every evaluated DASH strength is included in the tables below.

The HP grids are $\alpha \in \{-1, -0.5, -0.25, 0.5\}$ for joint suppression and $\alpha \in \{1.5, 2, 3, 5, 10\}$ for joint amplification. For mixed steering, HP uses $(-1, 5)$, with both suppression priority and amplification priority on shared heads. The HP mixed-steering results therefore cover the fixed configuration, not a corresponding strength sweep.

Metrics. We report normalized mean vocabulary-match counts for all three attributes and normalized CIDEr. The untargeted attribute provides an additional preservation measure: values near one indicate retention of its baseline mention frequency. Each result is normalized by its run's unedited baseline, so identity equals one for all reported ratios. The older colors–sentiments fixed-strength runs have slightly different identity metrics from the new sweeps despite matching recorded evaluation settings; we therefore interpret small cross-run differences cautiously. The other two pairs share identical identity metrics across fixed-strength and sweep runs. All runs contain only five sentiment-vocabulary matches in the 1,000 baseline captions, so one match changes the normalized sentiment count by 0.2. Zero denotes no observed matches in this sample.

D.4.2 Results

Joint suppression. The full sweeps show that DASH can improve the trade-off between suppressing both target attributes and preserving captioning at some intervention strengths. For colors–sentiments, the mean variant at $(\beta_C, \beta_S) = (7, 7)$ reaches normalized counts of 0.095 and zero, with quantity counts of 0.997 and CIDEr of 0.970. HP at $\alpha = -0.25$ gives weaker suppression of both targets (0.265 and 0.400), lower CIDEr (0.904), and a larger deviation in quantity counts (1.033). For colors–quantity, the multi-direction variant at $(2, 2)$ gives target ratios of 0.028 and 0.235 with CIDEr 0.866, compared with 0.232 and 0.390 at CIDEr 0.844 for HP at $\alpha = -0.25$. Its untargeted sentiment ratio is 1.20 rather than HP's 1.00, corresponding to one additional match. Thus, improvements in target control and overall captioning quality do not always improve every preservation measure.

Joint amplification. Favorable comparisons also occur in the amplification sweeps. For colors–sentiments, the mean variant at $(\beta_C, \beta_S) = (-7, -7)$ matches HP at $\alpha = 3$ in sentiment amplification (3.80), while achieving greater color amplification (1.78 versus 1.51), higher CIDEr (1.00 versus 0.97), and quantity counts closer to baseline (0.99 versus 0.89). The multi-direction variant supports stronger joint amplification as strength increases, but can incur larger preservation costs. For sentiments–quantity, changing its strengths from $(-2, -2)$ to $(-5, -5)$ increases the target ratios from 6.20/1.45 to 21.20/2.15, while CIDEr falls from 0.95 to 0.82 and untargeted color counts decrease from 0.83 to 0.65.

Mixed steering. DASH supports opposing edits across a range of strengths. For colors–quantity, the multi-direction variant at $(2, -2)$ reaches color and quantity ratios of 0.014 and 1.45 with CIDEr 0.95; at $(5, -5)$, these become 0.005 and 2.07 with CIDEr 0.85. The latter setting provides stronger control of both attributes and higher CIDEr than either fixed-strength HP priority rule. This comparison is limited to the evaluated HP configuration. For sentiments–quantity, both fixed-strength DASH variants eliminate the observed sentiment matches. The mean variant retains higher CIDEr, while the multi-direction variant achieves stronger quantity amplification.

The benefits depend on strength, concept pair, and preservation criterion. Stronger interventions do not uniformly improve even the target effects: in colors–quantity joint suppression, increasing multi-direction strength from $(3, 3)$ to $(10, 10)$ raises both target ratios (colors: 0.019 to 0.246; quantity: 0.092 to 0.167) while reducing CIDEr from 0.818 to 0.519. The full results therefore support improved control–preservation trade-offs for DASH at some tested strengths, rather than uniform superiority across the entire sweep.

Table 8: Colors–sentiments joint suppression (C ↓, S ↓) on Flickr30k. C/S/Q are normalized mean color/sentiment/quantity vocabulary-match counts; quantity is untargeted. CIDEr is normalized to the unedited model. Strength pairs follow the concept order in the title: (α_A, α_B) for HP and (β_A, β_B) for DASH. All available DASH strengths are reported.

Method	Strengths	C	S	Q	CIDEr
Identity	—	1.000	1.000	1.000	1.000
HP	$(-1, -1)$	0.725	0.600	1.131	0.685
HP	$(-0.5, -0.5)$	0.251	0.400	1.062	0.840
HP	$(-0.25, -0.25)$	0.265	0.400	1.033	0.904
HP	$(0.5, 0.5)$	0.782	0.600	1.009	0.986
$\text{DASH}_{\text{mean}}$	$(3, 3)$	0.517	0.400	1.000	0.985
$\text{DASH}_{\text{mean}}$	$(5, 5)$	0.194	0.400	1.018	0.977
$\text{DASH}_{\text{mean}}$	$(7, 7)$	0.095	0.000	0.997	0.970
$\text{DASH}_{\text{mean}}$	$(10, 10)$	0.024	0.200	1.009	0.979
$\text{DASH}_{\text{mean}}$	$(15, 15)$	0.028	0.200	1.042	0.955
$\text{DASH}_{\text{multi}}$	$(1, 1)$	0.204	0.000	0.958	0.985
$\text{DASH}_{\text{multi}}$	$(2, 2)$	0.043	0.000	0.961	0.970
$\text{DASH}_{\text{multi}}$	$(3, 3)$	0.009	0.000	0.955	0.949
$\text{DASH}_{\text{multi}}$	$(5, 5)$	0.034	0.000	0.943	0.894
$\text{DASH}_{\text{multi}}$	$(7, 7)$	0.071	0.200	0.955	0.853
$\text{DASH}_{\text{multi}}$	$(10, 10)$	0.318	0.800	0.961	0.715

Table 9: Colors–sentiments joint amplification (C ↑, S ↑) on Flickr30k. C/S/Q are normalized mean color/sentiment/quantity vocabulary-match counts; quantity is untargeted. CIDEr is normalized to the unedited model. Strength pairs follow the concept order in the title: (α_A, α_B) for HP and (β_A, β_B) for DASH. All available DASH strengths are reported.

Method	Strengths	C	S	Q	CIDEr
Identity	—	1.000	1.000	1.000	1.000
HP	$(1.5, 1.5)$	1.152	1.200	0.970	1.000
HP	$(2, 2)$	1.299	1.400	0.970	0.998
HP	$(3, 3)$	1.512	3.800	0.890	0.974
HP	$(5, 5)$	1.865	3.000	0.701	0.899
HP	$(10, 10)$	25.687	7.200	0.196	0.316
$\text{DASH}_{\text{mean}}$	$(-3, -3)$	1.374	1.400	0.988	1.012
$\text{DASH}_{\text{mean}}$	$(-5, -5)$	1.607	1.600	0.979	1.017
$\text{DASH}_{\text{mean}}$	$(-7, -7)$	1.782	3.800	0.988	1.001
$\text{DASH}_{\text{mean}}$	$(-10, -10)$	2.048	3.800	0.985	0.958
$\text{DASH}_{\text{mean}}$	$(-15, -15)$	7.090	4.600	0.943	0.923
$\text{DASH}_{\text{multi}}$	$(-1, -1)$	1.502	2.800	0.997	1.004
$\text{DASH}_{\text{multi}}$	$(-2, -2)$	1.905	5.800	0.988	0.972
$\text{DASH}_{\text{multi}}$	$(-3, -3)$	2.171	10.000	1.033	0.942
$\text{DASH}_{\text{multi}}$	$(-5, -5)$	2.688	19.000	1.063	0.860
$\text{DASH}_{\text{multi}}$	$(-7, -7)$	4.280	26.400	1.039	0.765
$\text{DASH}_{\text{multi}}$	$(-10, -10)$	11.986	38.000	0.926	0.593

Table 10: Colors–sentiments mixed steering (C ↓, S ↑) on Flickr30k. C/S/Q are normalized mean color/sentiment/quantity vocabulary-match counts; quantity is untargeted. CIDEr is normalized to the unedited model. Strength pairs follow the concept order in the title: (α_A, α_B) for HP and (β_A, β_B) for DASH. HP^{amp}/HP^{supp} prioritize amplification/suppression on shared heads. All available DASH strengths are reported.

Method	Strengths	C	S	Q	CIDEr
Identity	—	1.000	1.000	1.000	1.000
HP^{amp}	$(-1, 5)$	0.053	5.800	1.024	0.783
HP^{supp}	$(-1, 5)$	0.000	10.000	1.027	0.723
$\text{DASH}_{\text{mean}}$	$(3, -3)$	0.526	1.600	0.991	0.991
$\text{DASH}_{\text{mean}}$	$(5, -5)$	0.204	3.200	0.997	0.962
$\text{DASH}_{\text{mean}}$	$(7, -7)$	0.076	4.800	0.991	0.965
$\text{DASH}_{\text{mean}}$	$(10, -10)$	0.019	6.200	0.988	0.949
$\text{DASH}_{\text{mean}}$	$(15, -15)$	0.014	9.400	0.991	0.902
$\text{DASH}_{\text{multi}}$	$(1, -1)$	0.190	3.400	1.006	0.964
$\text{DASH}_{\text{multi}}$	$(2, -2)$	0.014	8.200	0.997	0.932
$\text{DASH}_{\text{multi}}$	$(3, -3)$	0.000	9.600	1.018	0.900
$\text{DASH}_{\text{multi}}$	$(5, -5)$	0.005	21.400	1.057	0.848
$\text{DASH}_{\text{multi}}$	$(7, -7)$	0.009	33.200	1.045	0.761
$\text{DASH}_{\text{multi}}$	$(10, -10)$	0.118	54.200	1.089	0.579

Table 11: Colors–quantity joint suppression (C ↓, Q ↓) on Flickr30k. C/S/Q are normalized mean color/sentiment/quantity vocabulary-match counts; sentiments are untargeted. CIDEr is normalized to the unedited model. Strength pairs follow the concept order in the title: (α_A, α_B) for HP and (β_A, β_B) for DASH. All available DASH strengths are reported.

Method	Strengths	C	S	Q	CIDEr
Identity	—	1.000	1.000	1.000	1.000
HP	$(-1, -1)$	0.322	3.200	0.113	0.651
HP	$(-0.5, -0.5)$	0.161	1.800	0.286	0.777
HP	$(-0.25, -0.25)$	0.232	1.000	0.390	0.844
HP	$(0.5, 0.5)$	0.744	1.200	0.807	0.970
$\text{DASH}_{\text{mean}}$	$(3, 3)$	0.526	1.000	0.905	0.961
$\text{DASH}_{\text{mean}}$	$(5, 5)$	0.204	1.200	0.857	0.955
$\text{DASH}_{\text{mean}}$	$(7, 7)$	0.095	1.000	0.827	0.944
$\text{DASH}_{\text{mean}}$	$(10, 10)$	0.028	0.800	0.747	0.939
$\text{DASH}_{\text{mean}}$	$(15, 15)$	0.005	0.600	0.545	0.904
$\text{DASH}_{\text{multi}}$	$(1, 1)$	0.242	1.400	0.646	0.953
$\text{DASH}_{\text{multi}}$	$(2, 2)$	0.028	1.200	0.235	0.866
$\text{DASH}_{\text{multi}}$	$(3, 3)$	0.019	1.000	0.092	0.818
$\text{DASH}_{\text{multi}}$	$(5, 5)$	0.033	1.000	0.030	0.713
$\text{DASH}_{\text{multi}}$	$(7, 7)$	0.052	0.400	0.018	0.659
$\text{DASH}_{\text{multi}}$	$(10, 10)$	0.246	0.600	0.167	0.519

Table 12: Colors–quantity joint amplification (C ↑, Q ↑) on Flickr30k. C/S/Q are normalized mean color/sentiment/quantity vocabulary-match counts; sentiments are untargeted. CIDEr is normalized to the unedited model. Strength pairs follow the concept order in the title: (α_A, α_B) for HP and (β_A, β_B) for DASH. All available DASH strengths are reported.

Method	Strengths	C	S	Q	CIDEr
Identity	—	1.000	1.000	1.000	1.000
HP	$(1.5, 1.5)$	1.218	0.800	1.119	1.034
HP	$(2, 2)$	1.284	1.400	1.223	1.037
HP	$(3, 3)$	1.521	2.000	1.366	1.013
HP	$(5, 5)$	2.474	2.800	1.631	0.971
HP	$(10, 10)$	28.654	11.800	10.161	0.185
$\text{DASH}_{\text{mean}}$	$(-3, -3)$	1.370	0.600	1.104	1.005
$\text{DASH}_{\text{mean}}$	$(-5, -5)$	1.474	0.600	1.179	1.000
$\text{DASH}_{\text{mean}}$	$(-7, -7)$	1.597	0.600	1.310	0.967
$\text{DASH}_{\text{mean}}$	$(-10, -10)$	2.185	0.400	1.521	0.914
$\text{DASH}_{\text{mean}}$	$(-15, -15)$	5.493	0.400	2.887	0.760
$\text{DASH}_{\text{multi}}$	$(-1, -1)$	1.531	0.800	1.277	1.034
$\text{DASH}_{\text{multi}}$	$(-2, -2)$	1.872	0.800	1.449	1.000
$\text{DASH}_{\text{multi}}$	$(-3, -3)$	2.137	1.000	1.619	0.973
$\text{DASH}_{\text{multi}}$	$(-5, -5)$	2.569	1.000	2.036	0.898
$\text{DASH}_{\text{multi}}$	$(-7, -7)$	4.517	1.400	2.479	0.801
$\text{DASH}_{\text{multi}}$	$(-10, -10)$	16.512	1.000	8.884	0.644

Table 13: Colors–quantity mixed steering (C ↓, Q ↑) on Flickr30k. C/S/Q are normalized mean color/sentiment/quantity vocabulary-match counts; sentiments are untargeted. CIDEr is normalized to the unedited model. Strength pairs follow the concept order in the title: (α_A, α_B) for HP and (β_A, β_B) for DASH. HP^{amp}/HP^{supp} prioritize amplification/suppression on shared heads. All available DASH strengths are reported.

Method	Strengths	C	S	Q	CIDEr
Identity	—	1.000	1.000	1.000	1.000
HP^{amp}	$(-1, 5)$	0.237	6.400	1.821	0.697
HP^{supp}	$(-1, 5)$	0.047	4.800	1.878	0.697
$\text{DASH}_{\text{mean}}$	$(3, -3)$	0.474	1.000	1.122	0.983
$\text{DASH}_{\text{mean}}$	$(5, -5)$	0.175	1.000	1.182	0.961
$\text{DASH}_{\text{mean}}$	$(7, -7)$	0.090	0.600	1.274	0.941
$\text{DASH}_{\text{mean}}$	$(10, -10)$	0.019	0.200	1.658	0.880
$\text{DASH}_{\text{mean}}$	$(15, -15)$	0.033	0.200	2.940	0.767
$\text{DASH}_{\text{multi}}$	$(1, -1)$	0.209	0.800	1.283	0.990
$\text{DASH}_{\text{multi}}$	$(2, -2)$	0.014	0.600	1.446	0.954
$\text{DASH}_{\text{multi}}$	$(3, -3)$	0.009	0.800	1.625	0.918
$\text{DASH}_{\text{multi}}$	$(5, -5)$	0.005	0.400	2.065	0.850
$\text{DASH}_{\text{multi}}$	$(7, -7)$	0.009	0.400	2.527	0.774
$\text{DASH}_{\text{multi}}$	$(10, -10)$	0.052	0.400	6.804	0.601

Table 14: Sentiments–quantity joint suppression (S ↓, Q ↓) on Flickr30k. C/S/Q are normalized mean color/sentiment/quantity vocabulary-match counts; colors are untargeted. CIDEr is normalized to the unedited model. Strength pairs follow the concept order in the title: (α_A, α_B) for HP and (β_A, β_B) for DASH. All available DASH strengths are reported.

Method	Strengths	C	S	Q	CIDEr
Identity	—	1.000	1.000	1.000	1.000
HP	$(-1, -1)$	2.246	0.000	0.074	0.747
HP	$(-0.5, -0.5)$	1.877	0.000	0.188	0.847
HP	$(-0.25, -0.25)$	1.664	0.000	0.333	0.894
HP	$(0.5, 0.5)$	1.175	0.000	0.807	0.995
$\text{DASH}_{\text{mean}}$	$(3, 3)$	1.057	0.200	0.911	0.994
$\text{DASH}_{\text{mean}}$	$(5, 5)$	1.062	0.000	0.866	0.990
$\text{DASH}_{\text{mean}}$	$(7, 7)$	1.085	0.000	0.812	0.977
$\text{DASH}_{\text{mean}}$	$(10, 10)$	1.095	0.000	0.750	0.973
$\text{DASH}_{\text{mean}}$	$(15, 15)$	1.024	0.000	0.613	0.930
$\text{DASH}_{\text{multi}}$	$(1, 1)$	1.171	0.000	0.613	0.986
$\text{DASH}_{\text{multi}}$	$(2, 2)$	1.370	0.000	0.190	0.917
$\text{DASH}_{\text{multi}}$	$(3, 3)$	1.536	0.000	0.077	0.842
$\text{DASH}_{\text{multi}}$	$(5, 5)$	1.730	0.200	0.027	0.742
$\text{DASH}_{\text{multi}}$	$(7, 7)$	1.877	0.200	0.027	0.658
$\text{DASH}_{\text{multi}}$	$(10, 10)$	1.474	0.600	0.482	0.492

Table 15: Sentiments–quantity joint amplification (S ↑, Q ↑) on Flickr30k. C/S/Q are normalized mean color/sentiment/quantity vocabulary-match counts; colors are untargeted. CIDEr is normalized to the unedited model. Strength pairs follow the concept order in the title: (α_A, α_B) for HP and (β_A, β_B) for DASH. All available DASH strengths are reported.

Method	Strengths	C	S	Q	CIDEr
Identity	—	1.000	1.000	1.000	1.000
HP	$(1.5, 1.5)$	0.853	1.600	1.119	1.002
HP	$(2, 2)$	0.773	2.800	1.220	0.987
HP	$(3, 3)$	0.664	4.200	1.399	0.946
HP	$(5, 5)$	0.744	7.800	1.726	0.842
HP	$(10, 10)$	0.043	9.000	6.640	0.040
$\text{DASH}_{\text{mean}}$	$(-3, -3)$	0.905	1.400	1.110	0.991
$\text{DASH}_{\text{mean}}$	$(-5, -5)$	0.872	2.200	1.202	0.983
$\text{DASH}_{\text{mean}}$	$(-7, -7)$	0.886	3.800	1.301	0.948
$\text{DASH}_{\text{mean}}$	$(-10, -10)$	0.867	5.800	1.631	0.898
$\text{DASH}_{\text{mean}}$	$(-15, -15)$	0.697	9.600	3.149	0.731
$\text{DASH}_{\text{multi}}$	$(-1, -1)$	0.858	2.600	1.280	0.988
$\text{DASH}_{\text{multi}}$	$(-2, -2)$	0.834	6.200	1.449	0.951
$\text{DASH}_{\text{multi}}$	$(-3, -3)$	0.725	9.200	1.655	0.914
$\text{DASH}_{\text{multi}}$	$(-5, -5)$	0.649	21.200	2.149	0.821
$\text{DASH}_{\text{multi}}$	$(-7, -7)$	0.559	33.200	2.917	0.715
$\text{DASH}_{\text{multi}}$	$(-10, -10)$	0.422	71.200	9.292	0.476

Table 16: Sentiments–quantity mixed steering (S ↓, Q ↑) on Flickr30k. C/S/Q are normalized mean color/sentiment/quantity vocabulary-match counts; colors are untargeted. CIDEr is normalized to the unedited model. Strength pairs follow the concept order in the title: (α_A, α_B) for HP and (β_A, β_B) for DASH. HP^{amp}/HP^{supp} prioritize amplification/suppression on shared heads. Only fixed-strength results are available for this configuration.

Method	Strengths	C	S	Q	CIDEr
Identity	—	1.000	1.000	1.000	1.000
HP^{amp}	$(-1, 5)$	1.071	0.600	1.679	0.875
HP^{supp}	$(-1, 5)$	1.052	0.200	1.655	0.883
$\text{DASH}_{\text{mean}}$	$(10, -10)$	0.915	0.000	1.485	0.922
$\text{DASH}_{\text{multi}}$	$(5, -5)$	0.858	0.000	1.964	0.869

D.5 FACTUAL QUESTION ANSWERING: TRIVIAQA

We evaluate country-answer suppression on TriviaQA while measuring preservation of performance on other questions and of general language modelling. Following Basile et al. (2025), we use Mistral-7B-Instruct-v0.2 and a concept dictionary constructed from country names. This setting tests whether direction-aware editing can selectively reduce a factual answering capability without broadly degrading the model.

D.5.1 SETUP

Dataset and splits. We use TriviaQA (Joshi et al., 2017) dataset. Calibration questions are drawn from `train` set and evaluation questions from `validation` set, using independently generated permutations with seed 42. We verify that the calibration and evaluation question IDs are disjoint. SOMP head selection uses 1,000 calibration questions. Fixed-strength results use 5,000 evaluation questions, while intervention-strength sweeps use 1,000.

Prompting and generation. Head selection uses raw, unwrapped calibration questions, following Head Pursuit's encoding protocol. Evaluation questions use the prompt:

> Answer the following question in 1–3 words only. Do not provide any additional explanation for your answer. Question: {question} Answer:

We generate answers with greedy decoding, `max_new_tokens=200`, and a batch size of 8. Each prediction is evaluated against the question's gold-answer aliases.

Metrics. We distinguish suppression of country-answer performance from preservation of other QA performance and language modelling:

- **Normalized country-answer F_1, $F_1^{\text{country}}/F_1^{\text{country},0}$.** A question belongs to the country-answer subset if any gold alias equals a country name or contains one as a word-boundary match, ignoring case. For each question, we compute SQuAD-style token F_1 and take the maximum over aliases. We average over the country-answer subset and divide by the unedited-model score on that subset. Lower ratios indicate stronger suppression of correct country answers.
- **Normalized non-country F_1, $F_1^{\text{other}}/F_1^{\text{other},0}$.** We compute the same score over the remaining questions and normalize by the corresponding unedited baseline. Values near 1 indicate preservation of QA performance on this subset.
- **Country mention count and presence.** We additionally record country-name matches in lowercased predictions using word-boundary regular expressions. We normalize both the total number of matches and the fraction of answers containing at least one match by their unedited baselines. These lexical measures distinguish changes in country mentions from changes in answer correctness.
- **Relative WikiText perplexity change, $\Delta\text{ppl}_{\text{wt}}$.** We use the WikiText-2 evaluation and relative percentage-change metric defined in Section D.1.1, measured on 1,000 test sentences.

D.5.2 RESULTS

Table 17 reports fixed-strength DASH results, and Figure 11 compares DASH with Head Pursuit through intervention-strength sweeps at $K = 32$. Unlike the visual tasks, country-answer suppression does not show a consistent control–preservation advantage for DASH. The mean direction largely preserves the model but has little effect on the target capability, while stronger multi-direction edits substantially degrade both other QA performance and language modelling.

$\text{DASH}_{\text{mean}}$ remains close to the unedited model across the reported strengths. Even at $\beta = 10$, normalized country-answer F_1 is 0.91 at $K = 16$ and 0.93 at $K = 32$, while non-country F_1 remains near baseline and WikiText perplexity decreases slightly. Thus, its strong preservation is accompanied by limited suppression. $\text{DASH}_{\text{multi}}$ reaches a wider suppression range, but with increasing preservation costs. At $K = 32$ and $\beta = 5$, normalized country-answer F_1 falls to 0.48, while non-country F_1 falls to 0.63 and perplexity increases by 32.7%. At $\beta = 10$, country-answer

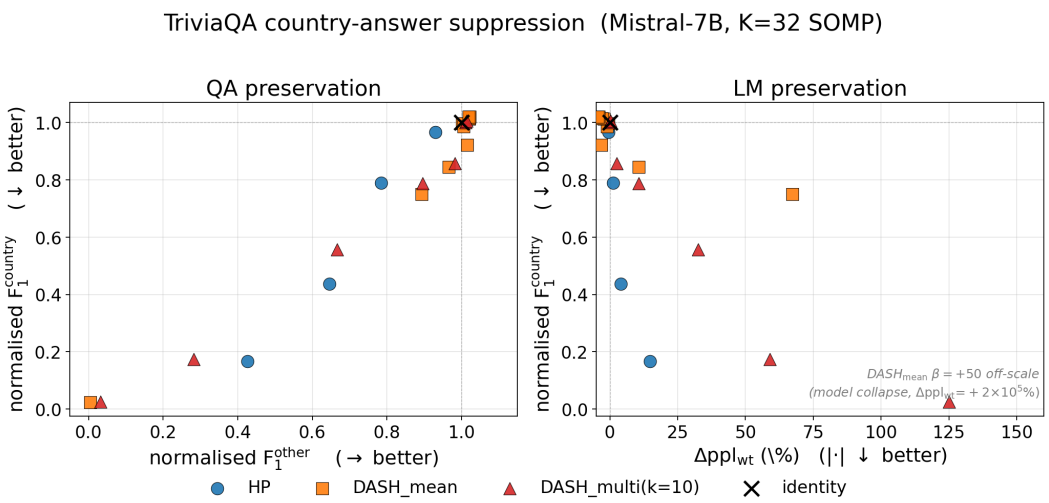


Figure 11: Country-answer suppression on TriviaQA at $K = 32$, sweeping α for Head Pursuit and β for DASH ($\text{DASH}_{\text{multi}}$: $r = 10$). Both panels report normalized country-answer F_1 against preservation: non-country F_1 on the left and relative WikiText perplexity change on the right. Lower country-answer F_1, non-country F_1 near 1, and perplexity change near zero indicate the desired trade-off. The sweep uses 1,000 evaluation questions; crosses mark the unedited model.

Table 17: DASH country-answer suppression on Mistral-7B-Instruct-v0.2 with SOMP, evaluated on 5,000 TriviaQA questions. Entries report normalized country-answer F_1 / normalized non-country F_1 / relative WikiText perplexity change (%), measured on 1,000 WikiText-2 sentences. Unedited country-answer F_1 is 0.55, 0.54, and 0.54 for $K = 8, 16, 32$, respectively; non-country F_1 is 0.46, 0.44, and 0.44. $\text{DASH}_{\text{multi}}$ uses $r = 10$.

Method and strength	K=8	K=16	K=32
$\text{DASH}_{\text{mean}}$ β=+3	1.01 / 1.00 / -2.1	1.00 / 1.01 / -2.1	1.01 / 1.01 / -3.0
$\text{DASH}_{\text{mean}}$ β=+5	1.02 / 1.02 / -3.3	0.98 / 1.01 / -3.1	1.00 / 1.01 / -4.1
$\text{DASH}_{\text{mean}}$ β=+10	1.01 / 1.00 / -4.6	0.91 / 0.98 / -4.2	0.93 / 1.00 / -3.0
$\text{DASH}_{\text{multi}}$ β=+3	0.87 / 0.94 / +7.2	0.75 / 0.90 / +8.8	0.73 / 0.85 / +10.7
$\text{DASH}_{\text{multi}}$ β=+5	0.75 / 0.85 / +26.3	0.56 / 0.77 / +29.1	0.48 / 0.63 / +32.7
$\text{DASH}_{\text{multi}}$ β=+10	0.41 / 0.48 / +102.1	0.19 / 0.35 / +105.8	0.02 / 0.03 / +125.2

F_1 is nearly eliminated, but non-country F_1 also falls to 0.03 and perplexity increases by 125.2%. This extreme operating point reflects broad degradation rather than selective suppression.

The sweeps show why both preservation measures are needed. At moderate suppression, multi-direction editing can retain more non-country QA performance than whole-head rescaling at a similar target effect, but incurs a larger perplexity cost. At stronger suppression, this QA advantage also disappears. Head Pursuit itself has substantial collateral effects: at $K = 32$ and $\alpha = -1$, normalized country-answer and non-country F_1 fall to 0.19 and 0.41, respectively, with a 14.8% perplexity increase. Nevertheless, its observed language-modelling trade-off is more favorable than that of strong multi-direction edits. These results identify a limitation of the current country dictionary and editing directions: substantial country-answer suppression remains coupled to broader capability loss, and the evaluated directional interventions do not consistently improve that trade-off.

E ABLATIONS AND DIAGNOSTIC ANALYSES

We provide full protocols and results for ablations of DASH-specific components and complementary diagnostic analyses. We examine head and direction specificity, subspace rank, calibration size, and the effect of extending interventions beyond a sparse set of heads. We also consider alternative direction constructions and analyze the activation energy and geometry of the resulting concept subspaces.

Our experiments cover toxicity suppression on RTP with Mistral-7B-Instruct-v0.2 and sentiment steering on Flickr30k with LLaVA-NeXT-7B. Task-specific splits and metrics follow Sections D.1.1 and D.3.1; sample sizes and protocol changes are specified in each subsection.

Unless otherwise specified, we use $K = 32$ SOMP-selected heads and the concept dictionaries from the corresponding task experiments. $\text{DASH}_{\text{mean}}$ edits one mean direction per head, while $\text{DASH}_{\text{multi}}$ edits a rank-r subspace, with $r = 10$ except in the rank sweep.

E.1 HEAD AND DIRECTION SPECIFICITY

Here, we test head selection, direction construction, and head–direction correspondence.

Setup. We use the RTP protocol in Section D.1.1, with 5,000 calibration and 1,000 evaluation prompts, and evaluate both DASH variants at $\beta \in \{1, 3, 5, 10\}$. We compare the following four controls:

1. **Random heads.** Replace the selected heads with $K = 32$ heads sampled uniformly without replacement from all $32 \times 32 = 1{,}024$ heads. Construct concept directions in the sampled heads using the original dictionary and calibration inputs.
2. **Permuted directions.** Keep the selected heads and the collection of directions or subspaces, but randomly reassign them across heads using a derangement. Every head receives another head's direction or subspace. This tests the head–direction assignment while retaining the original collection of concept representations.
3. **Random directions.** Keep the selected heads but replace the mean direction with a normalized Gaussian vector, or the multi-direction basis with the orthonormal basis of a Gaussian matrix of matching rank. This tests whether an arbitrary subspace at the selected intervention sites provides similar control.
4. **Random-vocabulary SVD.** Keep the selected heads and project a size-matched dictionary of uniformly sampled vocabulary token atoms into each head through $(W^O_{\ell,h})^\top$. Stack the projected atoms as columns and apply SVD: use its leading left singular vector for the rank-one control and its first $r = 10$ left singular vectors for the multi-direction control. This control changes both the dictionary and the direction construction relative to the DASH defaults. Section E.5 additionally compares concept and random vocabularies under the same rank-ten SVD construction.

Each randomized control uses five seeds (42–46), varying only the randomized component while holding calibration and evaluation prompts fixed. We report means and sample standard deviations across these construction seeds. The unmodified DASH reference uses the same inputs and calibration size. The DASH reference is deterministic under this fixed setup and is shown without seed error bars.

Metrics. We use lexical toxicity τ_{lex}, classifier toxicity τ_{rob}, and relative percentage change in perplexity on 1,000 WikiText-2 sentences, as defined in Section D.1.1. Tables report raw toxicity rates; the budget curves normalize them by the unedited baseline.

Comparisons across strengths. To avoid confounding construction quality with intervention magnitude, we report results using the lowest observed target metric achievable within a preservation budget. For construction m, construction seed s, and budget B, the component curve is

$$g_{m,s}(B) = \min_{\substack{\beta \in \{1,3,5,10\} \\ |\Delta \text{ppl}_{\text{wt},m,s}(\beta)| \le B}} \frac{\tau_{m,s}(\beta)}{\tau_{m,s}(0)}.$$

We compute this minimum separately for each seed before averaging.

Results. Tables 18 and 19 report results at all tested β values. At $\beta = 10$, the mean variant reduces lexical toxicity to 0.043, while the four randomized-control means range from 0.093 to 0.105. At $\beta = 5$, the multi-direction variant reaches 0.042, compared with control means between 0.086 and 0.098. However, preservation costs differ, motivating the budget comparisons in Figure 12.

Figure 12 shows that within a 5% perplexity budget, the mean and multi-direction variants reach normalized lexical rates of approximately 0.43 and 0.56, respectively, while all randomized-control averages remain above 0.85. At a fixed head budget, this supports the importance of which heads are chosen, how their directions are constructed, and how those directions are assigned to heads.

The β sweeps also show why suppression should be interpreted alongside preservation. Random-head multi-direction edits can lower toxicity at $\beta = 10$, but incur an average perplexity increase of about 73%.

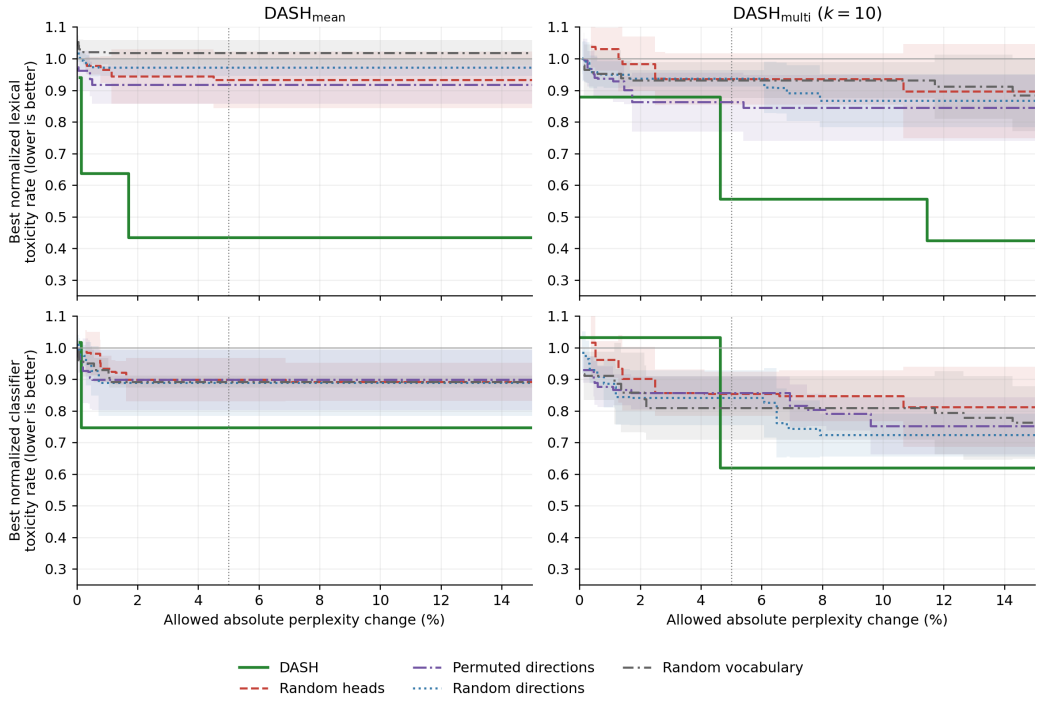


Figure 12: RTP component controls under a perplexity budget. The top row measures normalized lexical toxicity and the bottom row normalized classifier toxicity. Each curve selects among $\beta \in \{1, 3, 5, 10\}$ separately per seed. Bands show sample standard deviations over five construction seeds. The horizontal line marks the unedited ratio of one, and the vertical line a 5% budget.

Table 18: RTP component controls for DASH$_{\text{mean}}$. Entries are classifier and lexical toxicity rates and relative WikiText perplexity change (%). Randomized controls report mean ± sample standard deviation across five construction seeds.

Construction	β	τ_{rob}	τ_{lex}	$\Delta\text{ppl}_{\text{wt}}$
DASH	1	0.064	0.093	−0.1
	3	0.058	0.070	−0.2
	5	0.047	0.063	−0.1
	10	0.051	0.043	+1.7
Random heads	1	0.065 ± 0.003	0.098 ± 0.003	−0.1 ± 0.1
	3	0.062 ± 0.002	0.100 ± 0.009	−0.1 ± 0.4
	5	0.059 ± 0.004	0.100 ± 0.013	+0.1 ± 0.7
	10	0.058 ± 0.005	0.102 ± 0.008	+2.7 ± 3.1
Permuted directions	1	0.063 ± 0.002	0.098 ± 0.005	+0.0 ± 0.0
	3	0.063 ± 0.001	0.099 ± 0.004	+0.0 ± 0.0
	5	0.061 ± 0.003	0.096 ± 0.007	+0.0 ± 0.1
	10	0.057 ± 0.006	0.093 ± 0.007	+0.3 ± 0.1
Random directions	1	0.064 ± 0.002	0.101 ± 0.002	+0.0 ± 0.0
	3	0.064 ± 0.003	0.102 ± 0.002	+0.0 ± 0.0
	5	0.062 ± 0.004	0.098 ± 0.003	+0.1 ± 0.1
	10	0.057 ± 0.007	0.097 ± 0.003	+0.5 ± 0.2
Random vocabulary	1	0.065 ± 0.003	0.101 ± 0.003	+0.0 ± 0.0
	3	0.065 ± 0.004	0.101 ± 0.004	+0.0 ± 0.1
	5	0.065 ± 0.005	0.102 ± 0.005	+0.2 ± 0.1
	10	0.058 ± 0.001	0.105 ± 0.007	+0.9 ± 0.2

Table 19: RTP component controls for $\mathrm{DASH}_{\mathrm{multi}}$ ($r = 10$). Entries are classifier and lexical toxicity rates and relative WikiText perplexity change (%). Randomized controls report mean $\pm$ sample standard deviation across five construction seeds.

Construction	β	τ_{rob}	τ_{lex}	$\Delta\mathrm{ppl}_{\mathrm{wt}}$
DASH	1	0.065	0.087	+0.0
	3	0.039	0.055	+4.6
	5	0.042	0.042	+11.5
	10	0.080	0.035	+64.8
Random heads	1	0.061 ± 0.004	0.102 ± 0.007	+0.0 ± 0.3
	3	0.058 ± 0.011	0.094 ± 0.009	+1.3 ± 1.0
	5	0.052 ± 0.008	0.092 ± 0.018	+5.9 ± 3.0
	10	0.030 ± 0.013	0.047 ± 0.021	+72.7 ± 31.8
Permuted directions	1	0.060 ± 0.005	0.098 ± 0.007	+0.0 ± 0.1
	3	0.058 ± 0.005	0.095 ± 0.004	+0.3 ± 0.3
	5	0.055 ± 0.003	0.086 ± 0.010	+1.3 ± 0.5
	10	0.047 ± 0.006	0.088 ± 0.012	+7.5 ± 1.5
Random directions	1	0.063 ± 0.003	0.098 ± 0.006	+0.0 ± 0.1
	3	0.057 ± 0.002	0.095 ± 0.004	+0.3 ± 0.2
	5	0.053 ± 0.005	0.098 ± 0.006	+1.2 ± 0.3
	10	0.046 ± 0.004	0.089 ± 0.013	+6.8 ± 0.7
Random vocabulary	1	0.062 ± 0.002	0.100 ± 0.006	−0.1 ± 0.1
	3	0.061 ± 0.006	0.094 ± 0.008	+0.3 ± 0.2
	5	0.053 ± 0.008	0.094 ± 0.005	+1.7 ± 0.5
	10	0.044 ± 0.008	0.097 ± 0.022	+14.2 ± 2.2

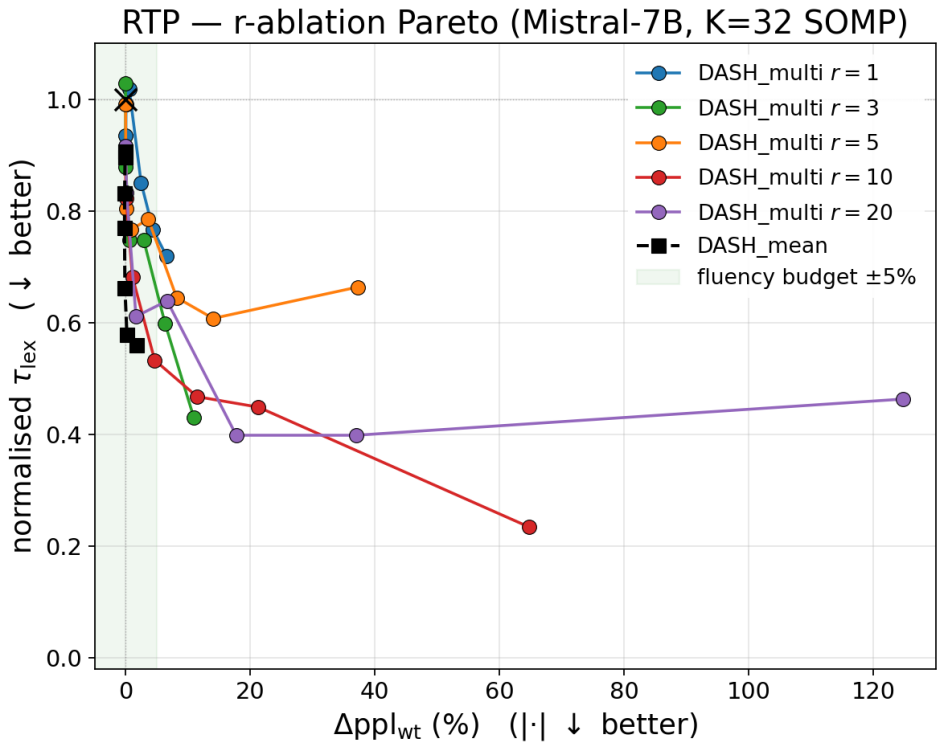


Figure 13: RTP rank-sweep operating points on 500 prompts, using the same selected heads and a generation limit of 200 new tokens for all ranks. Each series is normalized by its recorded unedited baseline. Points correspond to tested strengths, with lines connecting successive strengths. Ranks 1 and 3 include $\beta \in \{0.5, 1, 2, 3, 5, 7\}$; ranks 5, 10, and 20 additionally include $\beta = 10$.

E.2 Subspace rank

Here, we vary the editing rank of $\text{DASH}_{\text{multi}}$.

Setup. On RTP, we vary $r \in \{1, 3, 5, 10, 20\}$ and sweep $\beta \in \{0.5, 1, 2, 3, 5, 7\}$, with an additional $\beta = 10$ for ranks 5, 10, and 20. At rank one, the multi-direction construction uses the first SOMP-selected atom, whereas $\text{DASH}_{\text{mean}}$ uses the full-dictionary mean.

All ranks use the same 32 selected heads, 5,000 calibration prompts, 500 evaluation prompts, and a generation limit of 200 new tokens. We use the toxicity and WikiText-2 metrics defined in Section E.1, normalizing toxicity rates by each run's recorded unedited baseline.

Results. Table 20 reports lexical toxicity and the available classifier scores, and Fig. 13 compares the lexical toxicity–preservation trade-offs. These results show that increasing rank or strength does not produce a uniform improvement. For example, at rank 10, increasing β from 3 to 10 lowers the normalized lexical rate from approximately 0.53 to 0.23, but raises the classifier ratio from 0.64 to 1.18 and the perplexity increase from 4.6% to 64.8%.

We use $r = 10$ throughout our experiments as a shared default that supports strong suppression. This choice avoids task-specific rank tuning, but is not guaranteed to be optimal for every task or metric.

Table 20: RTP rank sweeps on 500 evaluation prompts with a shared generation limit of 200 new tokens. Ratios use each run's unedited baseline. Dashes mark unavailable results: the corrected rank-1 and rank-3 sweeps lack $\beta = 10$ and RoBERTa classifier scores.

r	β	$\tau_{\mathrm{rob}}/\tau_{\mathrm{rob},0}$	$\tau_{\mathrm{lex}}/\tau_{\mathrm{lex},0}$	$\Delta\mathrm{ppl}_{\mathrm{wt}}$ (%)
1	0.5	—	0.93	+0.0
	1	—	0.99	+0.0
	2	—	1.02	+0.5
	3	—	0.85	+2.5
	5	—	0.77	+4.4
	7	—	0.72	+6.5
	10	—	—	—
3	0.5	—	1.03	+0.0
	1	—	0.88	+0.0
	2	—	0.75	+0.6
	3	—	0.75	+2.9
	5	—	0.60	+6.3
	7	—	0.43	+10.9
	10	—	—	—
5	0.5	0.91	0.99	+0.0
	1	0.64	0.80	+0.0
	2	0.58	0.77	+0.8
	3	0.55	0.79	+3.6
	5	0.67	0.64	+8.2
	7	0.67	0.61	+14.0
	10	1.42	0.66	+37.2
10	0.5	0.85	0.90	+0.0
	1	0.64	0.82	+0.0
	2	0.73	0.68	+1.1
	3	0.64	0.53	+4.6
	5	0.64	0.47	+11.4
	7	0.67	0.45	+21.3
	10	1.18	0.23	+64.8
20	0.5	0.94	0.92	+0.0
	1	0.69	0.83	+0.1
	2	0.50	0.61	+1.6
	3	0.50	0.64	+6.6
	5	0.69	0.40	+17.7
	7	0.56	0.40	+37.0
	10	1.50	0.46	+124.8

Table 21: RTP calibration-size sweep for $\text{DASH}_{\text{mean}}$, with $K = 32$ and 1,000 evaluation prompts. Toxicity rates are unnormalized, perplexity changes are percentages.

n_{calib}	β	τ_{rob}	τ_{lex}	$\Delta\text{ppl}_{\text{wt}}$
50	1	0.063	0.096	-0.1
	3	0.057	0.072	-0.2
	5	0.047	0.070	-0.1
	10	0.047	0.049	+1.8
200	1	0.064	0.097	-0.1
	3	0.060	0.070	-0.2
	5	0.052	0.066	-0.2
	10	0.045	0.050	+1.7
1,000	1	0.064	0.095	-0.1
	3	0.060	0.071	-0.3
	5	0.050	0.068	-0.2
	10	0.045	0.047	+1.7
5,000	1	0.064	0.093	-0.1
	3	0.058	0.070	-0.2
	5	0.047	0.063	-0.1
	10	0.051	0.043	+1.7

E.3 Calibration size

Here, we vary the calibration-set size used for head scoring and subspace construction.

Setup. We vary the number of RTP calibration prompts over $n_{\text{calib}} \in \{50, 200, 1000, 5000\}$, keeping $K = 32$, the dictionary, and the 1,000 evaluation prompts fixed. For each calibration size, we recompute head scores and SOMP atom selections. Calibration therefore affects both the selected heads and the multi-direction subspaces of DASH_{multi}, while the dictionary-derived mean direction of DASH_{mean} remains fixed for any given head. Both variants are evaluated at $\beta \in \{1, 3, 5, 10\}$. We use one fixed calibration sample per size and the generation protocol from Section E.1. We report raw toxicity rates and WikiText-2 perplexity changes using the metrics and preservation sample size defined in Section E.1.

Results. Tables 21 and 22 show useful suppression even with 50 calibration prompts. At $\beta = 10$, the mean variant's lexical rate is 0.049 with 50 prompts, 0.047 with 1,000, and 0.043 with 5,000, and perplexity increases remain between 1.7% and 1.8%. For the multi-direction variant at $\beta = 5$, the lexical rate is 0.051 with 50 prompts and 0.042 with either 1,000 or 5,000 prompts, with perplexity increases of 10.2%, 10.8%, and 11.5%, respectively. These observations suggest modest calibration requirements in this setting.

Table 22: RTP calibration-size sweep for $\text{DASH}_{\text{multi}}$ ($r = 10$), with $K = 32$ and 1,000 evaluation prompts. Toxicity rates are unnormalized, perplexity changes are percentages.

n_{calib}	β	τ_{rob}	τ_{lex}	$\Delta\text{ppl}_{\text{wt}}$
50	1	0.061	0.086	+0.0
	3	0.056	0.052	+3.9
	5	0.043	0.051	+10.2
	10	0.067	0.062	+58.5
200	1	0.062	0.091	+0.0
	3	0.049	0.063	+4.5
	5	0.037	0.044	+11.4
	10	0.080	0.071	+62.4
1,000	1	0.065	0.083	-0.1
	3	0.043	0.047	+4.4
	5	0.040	0.042	+10.8
	10	0.064	0.079	+59.4
5,000	1	0.065	0.087	+0.0
	3	0.039	0.055	+4.6
	5	0.042	0.042	+11.5
	10	0.080	0.035	+64.8

Table 23: Mean captured activation-energy fractions for RTP, averaged over 32 selected heads using 200 inputs per distribution. The final column is the ratio of aggregate target and preservation energy fractions, computed before rounding.

Construction	RTP	WikiText	RTP / WikiText
$\text{DASH}_{\text{mean}}$	0.024	0.019	1.25
Dictionary SVD, rank 1	0.027	0.024	1.12
Gaussian, rank 1	0.009	0.008	1.17
Random-vocabulary SVD, rank 1	0.012	0.012	1.02
$\text{DASH}_{\text{multi}}$ ($r = 10$)	0.122	0.116	1.05
Dictionary SVD, rank 10	0.119	0.109	1.09
Gaussian, rank 10	0.079	0.078	1.01
Random-vocabulary SVD, rank 10	0.108	0.103	1.05

E.4 Activation-energy and geometric diagnostics

Here, we examine how DASH's direction constructions differ in the activation energy they capture and the geometry of their concept subspaces. We measure mean–subspace alignment, dictionary spectral concentration, and sensitivity to dictionary subsampling and random-vocabulary replacement.

E.4.1 Captured activation energy

Setup. For an orthonormal basis Q_h and unedited head activation z_h, we measure the captured energy fraction

$$e_h(z_h; Q_h) = \frac{\|Q_h^\top z_h\|_2^2}{\|z_h\|_2^2}.$$

We report aggregate fractions over the recorded input sets and the $K = 32$ selected heads. We use 200 RTP prompts and 200 WikiText sentences. The target-to-preservation ratio compares the aggregate RTP and WikiText fractions.

Results. Results in Table 23 show that the mean direction captures energy fractions of 0.024 on RTP and 0.019 on WikiText, yielding a target-to-preservation ratio of 1.25. The multi-direction subspace captures more energy in both distributions, but its ratio is only 1.05.

Thus, the mean direction captures proportionally more activation energy on RTP than on WikiText, while the multi-direction subspace captures similar fractions across the two distributions. This is consistent with the mean variant providing a more selective but narrower intervention, while the multi-direction variant can affect a broader range of representations, potentially enabling stronger suppression at greater preservation cost. The ratios describe differences between the RTP and WikiText activation distributions.

E.4.2 Mean–subspace alignment and dictionary geometry

Here, we analyze the geometric relationship between DASH's mean direction and multi-direction subspace, the concentration of concept dictionary energy across directions, and the sensitivity of these representations to dictionary subsampling and random-vocabulary replacement.

Setup. We analyze 32 selected heads for RTP toxicity and Flickr30k sentiments. Each head has dimension 128, and the multi-direction basis has rank $r = 10$. For normalized mean direction v_h and basis Q_h, define

$$o_h = \|Q_h^\top v_h\|_2^2, \qquad \theta_h = \arccos\sqrt{o_h}.$$

The overlap measures how much of the mean direction lies in the multi-direction subspace.

We also inspect the singular values $\{\sigma_i\}$ of the projected dictionary. We report the energy participation ratio, the top-ten energy fraction, and the rank needed for 90% of energy:

$$r_{\text{eff}} = \frac{(\sum_i \sigma_i^2)^2}{\sum_i \sigma_i^4}, \qquad e_{10} = \frac{\sum_{i=1}^{10} \sigma_i^2}{\sum_i \sigma_i^2}, \qquad r_{90} = \min\left\{ r : \frac{\sum_{i=1}^{r} \sigma_i^2}{\sum_i \sigma_i^2} \geq 0.9 \right\}.$$

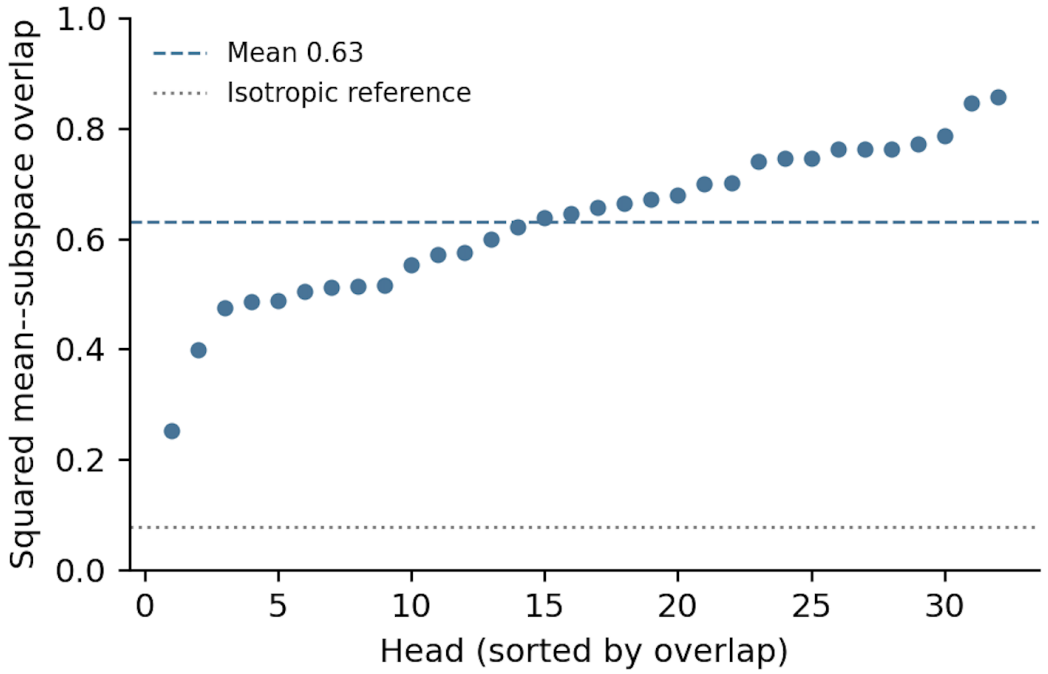


Figure 14: Per-head squared overlap between the mean concept direction and the rank-ten multi-direction subspace. RTP heads are sorted by overlap. The dashed line shows the mean and the dotted line the isotropic reference $10/128$.

Table 24: Concept geometry, averaged over 32 heads per task. M counts token atoms, o is squared mean–subspace overlap, θ its principal angle, r_{eff} the energy participation ratio, and e_{10} the dictionary energy captured by its top ten singular directions.

Task / concept	M	o	θ (deg.)	r_{eff}	e_{10}
RTP / toxicity	94	0.631	37.1	18.12	0.572
Flickr30k / sentiments	200	0.395	51.5	34.54	0.447

Note that we compute the spectrum of the full projected dictionary, rather than the SOMP-selected subset. Consequently, e_{10} describes the energy captured by the leading ten singular directions, rather than the SOMP-selected DASH subspace.

Finally, we recompute the mean direction from five random half-dictionary subsets and measure its cosine with the full-dictionary direction, holding the selected heads fixed. We also compare against five size-matched random-vocabulary dictionaries. For two rank-ten subspaces, we report the Grassmann geodesic distance $d_{\text{Gr}} = (\sum_{i=1}^{10} \theta_i^2)^{1/2}$, using their principal angles in radians.

Results. Results are presented in Table 24 and Fig. 14. On RTP, mean–subspace overlap averages 0.631, suggesting that the mean and multi-direction constructions can edit substantially different components. The average exceeds the isotropic reference $r/d_h = 10/128 \approx 0.078$, while substantial mean-direction components remain outside the multi-direction subspace.

On RTP, dictionary energy is distributed across many directions: the mean energy participation ratio is 18.12, while the top ten singular directions capture 57.2% of projected dictionary energy. These differences motivate examining rank sensitivity.

Table 25 reports RTP dictionary perturbations. The mean cosine is 0.909 for half-dictionary mean directions and 0.486 for random-vocabulary mean directions. The mean rank-ten random-vocabulary subspace distance is 4.113, while the mathematical maximum is $\sqrt{10}\pi/2 \approx 4.97$, which is not a random-subspace expectation.

Table 25: RTP dictionary perturbations and spectral concentration. Cosines and random-vocabulary distances are averaged over 32 heads and five dictionary draws. r_{90} is the number of singular directions required to capture 90% of uncentered dictionary energy, averaged over heads.

Half-dictionary cosine	Random-vocab. cosine	d_{Gr}	r_{90}
0.909	0.486	4.113	40.22

Table 26: Alternative direction constructions on RTP. Each row reports the operating point with the lowest normalized lexical toxicity among tested strengths satisfying $|\Delta\mathrm{ppl}_{\mathrm{wt}}| \leq 5\%$. The classifier rate is reported at that same operating point. Both toxicity rates are normalized to the corresponding unedited-model baseline.

Construction	β	$\tau_{\mathrm{lex}}/\tau_{\mathrm{lex},0}$	$\tau_{\mathrm{rob}}/\tau_{\mathrm{rob},0}$	$\Delta\mathrm{ppl}_{\mathrm{wt}}$ (%)
Rank one				
Full-dictionary mean	7	0.546	0.677	+0.18
Per-head SOMP mean	7	0.574	0.733	+4.41
Dictionary PC1	5	0.731	0.933	+4.18
Rank ten				
SOMP subspace	3	0.565	0.774	+4.64
Dictionary SVD	3	0.617	0.485	+4.57
Random-vocabulary SVD	3	0.822	0.697	+0.80

E.5 Alternative direction constructions

Here, we examine whether DASH's behavior depends on how concept directions are constructed, holding the selected heads and intervention operator fixed.

Setup. We evaluate toxicity suppression on RTP with Mistral-7B-Instruct-v0.2, using $K = 32$ SOMP-selected heads, $5{,}000$ calibration prompts, and 500 evaluation prompts, with greedy generation limited to 200 new tokens. For each construction, we sweep $\beta \in \{0.5, 1, 2, 3, 5, 7, 10\}$. At rank one, we compare: (i) the full-dictionary mean used by $\mathrm{DASH}_{\mathrm{mean}}$; (ii) a *per-head SOMP mean*, formed by averaging the ten dictionary atoms selected for each head before pulling the mean into head space; (iii) *dictionary PC1*, obtained by pulling the leading singular direction of the residual-space concept dictionary into each head. Each resulting vector is normalized.

At rank ten, we compare: (i) the SOMP-selected subspace used by $\mathrm{DASH}_{\mathrm{multi}}$; (ii) *dictionary SVD*, using the ten leading singular directions of the full concept dictionary after projection into each head; (iii) *random-vocabulary SVD*, applying the same construction to a size-matched dictionary of randomly sampled vocabulary tokens. The SVD constructions are uncentered. All constructions use the same operator, $z' = (I - \beta QQ^{\top})z$, with an orthonormal basis Q.

The runs use the same nominal evaluation configuration and selected heads, but retain small differences in their unedited outputs: lexical toxicity rates range from 0.214 to 0.216, and classifier rates from 0.060 to 0.066. We normalize each run by its own baseline and interpret small differences cautiously.

Rank-one constructions. The full-dictionary mean provides the strongest lexical suppression among the tested rank-one constructions within a 5% absolute relative perplexity budget (Tab. 26). It reaches normalized lexical toxicity of 0.546 with only a 0.18% perplexity increase, compared with 0.574 at $+4.41\%$ for the per-head SOMP mean and 0.731 at $+4.18\%$ for dictionary PC1. It also achieves the lowest classifier toxicity within this budget: 0.613 at $\beta = 5$, with a 0.14% perplexity decrease.

The alternatives support stronger lexical suppression at larger preservation costs. At $\beta = 10$, the per-head SOMP mean reaches a lexical ratio of 0.380 with a 9.43% perplexity increase, but its classifier ratio remains 0.867. Dictionary PC1 reaches lexical and classifier ratios of 0.435 and 0.633, respectively, with a 20.19% perplexity increase. These results support the full-dictionary mean as a preservation-oriented default. The dictionary-subsampling diagnostics provide complementary evidence of representation stability: mean directions from half-dictionary subsets have a mean cosine similarity of approximately 0.91 with the full-dictionary directions (Tab. 25).

Rank-ten constructions. Tab. 26 shows that SOMP and dictionary SVD yield competitive tradeoffs, with their relative performance depending on the toxicity metric. At $\beta = 3$, their perplexity increases are nearly identical (4.64% and 4.57%, respectively). SOMP produces lower lexical toxicity (0.565 versus 0.617), whereas dictionary SVD produces lower classifier toxicity (0.485 versus 0.774). At $\beta = 5$, their lexical ratios are nearly equal (0.472 and 0.467), while dictionary SVD again has the lower classifier ratio (0.545 versus 0.677).

Random-vocabulary SVD gives weaker lexical suppression within the 5% budget, but is not uniformly inferior across metrics: at $\beta = 2$, it reaches a classifier ratio of 0.606 with essentially unchanged perplexity. At strong interventions, lexical and classifier outcomes can also diverge for concept-derived subspaces. At $\beta = 10$, both SOMP and dictionary SVD increase perplexity by approximately 65%, and their classifier toxicity exceeds baseline despite reduced lexical toxicity. These results reinforce the need to evaluate both toxicity measures alongside language-model preservation.

We retain the full rank-ten subspace in $\mathrm{DASH}_{\mathrm{multi}}$ based on SOMP, as it provides this basis directly from the calibration-informed atom ordering already used in the pipeline. However, our ablation results show that dictionary SVD is a viable alternative.

E.6 INTERVENTION SCOPE: SELECTED HEADS, SELECTED LAYERS, AND ALL LAYERS

Here, we examine how DASH behaves when interventions extend beyond a sparse set of concept-relevant heads. We distinguish three scopes: (i) *selected heads*, using the top $K = 32$ SOMP heads; (ii) *selected layers*, editing every head in a layer containing at least one selected head; and (iii) *all layers*, editing all $1{,}024$ heads. The selected-layer scope contains 448 heads across 14 layers on RTP and 384 heads across 12 layers for captioning. DASH retains its head-specific direction construction at every scope; $\text{DASH}_{\text{multi}}$ uses $r = 10$.

Setup. We evaluate toxicity suppression on RTP with Mistral-7B-Instruct-v0.2 and sentiment steering on Flickr30k with LLaVA-NeXT-7B. Calibration uses $5{,}000$ prompts for RTP and $1{,}000$ images for captioning; evaluation uses 500 examples per task. Greedy generation is limited to 200 new tokens for RTP and 50 for captioning. For RTP, we sweep $\beta \in \{0.5, 1, 2, 3, 5, 7, 10\}$. For captioning, we evaluate both signs of these strengths, with positive values suppressing sentiment mentions and negative values amplifying them. We report every tested DASH strength at each scope, without selecting an operating point for a particular preservation budget.

Toxicity metrics are normalized lexical and RoBERTa toxic-generation rates, together with the relative percentage change in WikiText-2 perplexity, as defined in Section D.1.1. The RTP runs retain small differences in unedited outputs (lexical rates of 0.214–0.216 and classifier rates of 0.062–0.066); each is normalized by its own baseline. The captioning runs share identical recorded identity metrics. We report normalized sentiment counts, untargeted color and quantity counts, and CIDEr. Their unedited baseline contains three sentiment matches across 500 captions, so a single match changes the normalized sentiment count by one third. Zero denotes no observed matches in this sample.

Toxicity suppression. Tables 27 and 28 show that broad directional editing can provide useful suppression at weak strengths. For example, all-layer $\text{DASH}_{\text{mean}}$ at $\beta = 2$ reaches normalized lexical toxicity of 0.402 with a 0.28% perplexity decrease, whereas selected-head editing at $\beta = 7$ reaches 0.546 with a 0.18% increase. All-layer $\text{DASH}_{\text{multi}}$ at $\beta = 1$ also attains stronger lexical suppression and lower perplexity cost than selected-head editing at $\beta = 3$: 0.523 versus 0.565, with perplexity increases of 1.77% versus 4.64%. These comparisons show that broad directional edits can provide favorable toxicity–preservation trade-offs at weak strengths.

However, broad edits can become highly disruptive as strength increases. For all-layer mean-direction editing, increasing β from 2 to 3 raises the perplexity change from -0.28% to $+15.27\%$. For all-layer multi-direction editing, increasing β from 1 to 2 raises the cost from 1.77% to 223.02%. Thus, intervention scope changes the strengths at which useful control is obtained; a fixed-strength comparison alone does not establish which scope offers the better trade-off.

For reference, whole-head rescaling at the canonical $\alpha = -1$ increases perplexity by 7.80% on the 32 selected heads, approximately $80{,}247\%$ on the selected layers, and $189{,}282\%$ on all layers. The corresponding normalized lexical toxicity rates are 0.704, 0.009, and 0.000. The extreme suppression under expanded rescaling therefore accompanies severe language-modelling degradation. This comparison concerns the canonical HP strength and does not exclude useful expanded rescaling at other strengths.

Captioning and untargeted attributes. Tables 29 and 30 report the full amplification sweeps. Selected-head mean-direction editing supports amplification with little change in untargeted attributes: at $\beta = -10$, the sentiment ratio is 3.67, color and quantity ratios are 1.00 and 1.03, and CIDEr is 0.996. Selected-layer mean-direction editing at $\beta = -3$ gives a similar sentiment ratio of 3.00 and CIDEr of 1.009, but increases color mentions to 1.26. All-layer editing at $\beta = -3$ reaches a larger sentiment ratio of 8.67 with CIDEr 0.991, while increasing color mentions to 1.25. Broad mean-direction editing can therefore remain competitive on aggregate captioning quality while changing untargeted attributes.

For multi-direction editing, sparse selection provides favorable amplification–CIDEr trade-offs at some tested strengths. Selected-head editing at $\beta = -3$ reaches a sentiment ratio of 7.33 with CIDEr 0.963, compared with 8.00 and 0.946 for selected-layer editing at $\beta = -2$, and 6.00 and 0.826 for all-layer editing at $\beta = -1$. At stronger amplification, the preservation costs grow: $\beta = -10$ yields normalized CIDEr of 0.725, 0.015, and 0.003 for selected heads, selected layers, and all layers,

respectively. Sparse editing also changes untargeted attributes: at this strength, its color ratio falls to 0.685. Head selection therefore moderates some collateral effects without eliminating them.

The corresponding suppression sweeps are reported in Tables 31 and 32. Because only three sentiment matches occur at baseline, suppression quickly reaches the measurement floor; amplification provides a larger range of observed target effects for comparing scopes. Together, these results show that the benefit of sparse selection depends on the task, intervention strength, and preservation criterion. Useful broad edits exist, while sparse edits can provide better preservation of captioning quality or untargeted attributes at comparable levels of control.

Table 27: RTP intervention-scope ablation for $\mathrm{DASH}_{\mathrm{mean}}$. All tested strengths are reported. Toxicity rates are normalized to the unedited baseline within each run. The final column is relative WikiText-2 perplexity change in percent. Head counts are given in parentheses.

β	$\tau_{\mathrm{lex}}/\tau_{\mathrm{lex},0}$	$\tau_{\mathrm{rob}}/\tau_{\mathrm{rob},0}$	$\Delta\mathrm{ppl}_{\mathrm{wt}}$ (%)
Selected heads (32)			
0.5	0.907	0.968	-0.06
1	0.843	0.968	-0.11
2	0.778	0.871	-0.18
3	0.778	0.710	-0.21
5	0.648	0.613	-0.14
7	0.546	0.677	$+0.18$
10	0.565	0.645	$+1.71$
Selected layers (448)			
0.5	0.925	0.818	-0.78
1	0.888	0.818	-1.43
2	0.617	0.455	-2.21
3	0.495	0.515	-2.07
5	0.383	0.455	$+44.76$
7	0.308	0.364	$+4.29 \times 10^{4}$
10	0.019	0.000	$+4.20 \times 10^{10}$
All layers (1,024)			
0.5	0.692	0.667	-0.66
1	0.570	0.424	-1.00
2	0.402	0.364	-0.28
3	0.196	0.303	$+15.27$
5	0.000	0.333	$+1.88 \times 10^{7}$
7	0.000	0.000	$+1.24 \times 10^{9}$
10	0.000	0.000	$+1.11 \times 10^{7}$

Table 28: RTP intervention-scope ablation for $\mathrm{DASH}_{\mathrm{multi}}$. All tested strengths are reported. Toxicity rates are normalized to the unedited baseline within each run. The final column is relative WikiText-2 perplexity change in percent. Head counts are given in parentheses.

β	$\tau_{\mathrm{lex}}/\tau_{\mathrm{lex},0}$	$\tau_{\mathrm{rob}}/\tau_{\mathrm{rob},0}$	$\Delta\mathrm{ppl}_{\mathrm{wt}}$ (%)
Selected heads (32)			
0.5	0.935	0.839	-0.04
1	0.843	0.742	$+0.03$
2	0.713	0.774	$+1.07$
3	0.565	0.774	$+4.64$
5	0.472	0.677	$+11.46$
7	0.435	0.710	$+21.24$
10	0.241	1.226	$+64.77$
Selected layers (448)			
0.5	0.897	0.879	-2.00
1	0.813	0.576	-2.41
2	0.514	0.424	$+8.33$
3	0.271	0.273	$+68.74$
5	0.000	0.061	$+2.27 \times 10^{6}$
7	0.000	0.273	$+8.30 \times 10^{7}$
10	0.028	0.879	$+1.97 \times 10^{7}$
All layers (1,024)			
0.5	0.879	0.515	-1.46
1	0.523	0.364	$+1.77$
2	0.019	0.121	$+223.02$
3	0.000	0.000	$+1.15 \times 10^{6}$
5	0.000	0.000	$+2.66 \times 10^{6}$
7	0.000	2.818	$+1.82 \times 10^{5}$
10	0.000	15.152	$+1.05 \times 10^{6}$

Table 29: Captioning intervention-scope ablation for $\mathrm{DASH}_{\mathrm{mean}}$: sentiment amplification. All tested amplification strengths are reported. S/C/Q denote normalized mean sentiment/color/quantity vocabulary-match counts; colors and quantity are untargeted. CIDEr is normalized to the unedited model. Head counts are given in parentheses.

β	S	C	Q	CIDEr
Selected heads (32)				
-0.5	1.000	0.991	1.000	1.007
-1	1.000	0.972	1.000	1.005
-2	1.000	0.981	0.994	1.002
-3	1.000	0.981	0.994	1.006
-5	2.000	0.991	0.975	1.014
-7	3.333	0.972	1.000	1.012
-10	3.667	1.000	1.031	0.996
Selected layers (384)				
-0.5	1.000	0.981	0.994	1.002
-1	0.667	1.093	1.000	1.023
-2	1.333	1.157	1.013	1.021
-3	3.000	1.259	1.038	1.009
-5	5.000	1.361	1.050	0.986
-7	7.333	1.556	1.057	0.964
-10	20.667	1.509	1.101	0.866
All layers (1,024)				
-0.5	1.333	1.037	1.025	0.997
-1	2.000	1.102	1.019	1.022
-2	7.333	1.148	0.994	1.007
-3	8.667	1.250	0.969	0.991
-5	28.667	1.870	1.145	0.787
-7	88.333	1.417	1.472	0.457
-10	186.333	1.398	0.484	0.269

Table 30: Captioning intervention-scope ablation for $\mathrm{DASH}_{\mathrm{multi}}$: sentiment amplification. All tested amplification strengths are reported. S/C/Q denote normalized mean sentiment/color/quantity vocabulary-match counts; colors and quantity are untargeted. CIDEr is normalized to the unedited model. Head counts are given in parentheses.

β	S	C	Q	CIDEr
Selected heads (32)				
-0.5	1.667	0.926	1.019	0.987
-1	2.333	0.852	1.025	1.001
-2	5.000	0.796	1.044	0.981
-3	7.333	0.843	1.050	0.963
-5	14.667	0.769	1.075	0.917
-7	22.333	0.713	1.075	0.837
-10	35.000	0.685	1.082	0.725
Selected layers (384)				
-0.5	1.333	1.019	1.013	1.011
-1	2.667	1.102	1.088	0.984
-2	8.000	1.130	1.182	0.946
-3	12.000	1.306	1.226	0.878
-5	24.667	1.389	1.283	0.703
-7	54.667	1.139	1.220	0.421
-10	152.333	0.037	0.226	0.015
All layers (1,024)				
-0.5	2.000	0.889	1.013	0.960
-1	6.000	0.852	0.987	0.826
-2	10.000	0.750	0.818	0.621
-3	9.000	1.389	0.145	0.355
-5	0.333	0.000	0.019	0.001
-7	5.000	0.000	0.000	0.003
-10	186.333	0.000	0.000	0.003

Table 31: Captioning intervention-scope ablation for $\mathrm{DASH}_{\mathrm{mean}}$: sentiment suppression. All tested suppression strengths are reported. S/C/Q denote normalized mean sentiment/color/quantity vocabulary-match counts; colors and quantity are untargeted. CIDEr is normalized to the unedited model. Head counts are given in parentheses.

β	S	C	Q	CIDEr
Selected heads (32)				
0.5	0.667	1.009	1.000	1.001
1	1.000	1.019	1.006	1.001
2	0.333	1.019	0.994	1.000
3	0.667	1.056	1.000	0.991
5	0.333	1.046	1.000	0.992
7	0.333	1.019	1.006	0.986
10	0.000	1.056	1.019	0.993
Selected layers (384)				
0.5	1.333	1.019	1.019	1.000
1	0.333	0.963	1.019	1.001
2	0.333	0.954	1.038	1.020
3	0.333	1.074	1.050	1.000
5	0.667	1.074	0.862	0.706
7	6.333	0.111	0.321	0.088
10	0.000	0.000	0.000	0.000
All layers (1,024)				
0.5	0.333	1.000	1.038	0.997
1	0.000	0.944	1.025	1.006
2	0.000	1.102	1.063	1.034
3	1.667	1.306	1.132	0.566
5	0.000	0.167	0.145	0.065
7	0.000	0.000	0.000	0.000
10	0.333	0.000	0.000	0.000

Table 32: Captioning intervention-scope ablation for $\mathrm{DASH}_{\mathrm{multi}}$: sentiment suppression. All tested suppression strengths are reported. S/C/Q denote normalized mean sentiment/color/quantity vocabulary-match counts; colors and quantity are untargeted. CIDEr is normalized to the unedited model. Head counts are given in parentheses.

β	S	C	Q	CIDEr
Selected heads (32)				
0.5	0.000	1.019	1.000	1.006
1	0.000	1.056	1.013	1.008
2	0.000	1.083	0.981	1.000
3	0.000	1.046	0.969	1.009
5	0.000	1.176	0.956	0.983
7	0.000	1.278	0.969	0.968
10	0.333	1.065	0.956	0.862
Selected layers (384)				
0.5	0.000	0.917	0.987	0.995
1	0.000	0.833	0.943	0.986
2	0.000	0.935	0.906	0.942
3	0.000	0.880	0.811	0.900
5	0.000	0.593	0.396	0.366
7	0.333	0.028	0.031	0.014
10	0.000	0.000	0.000	0.000
All layers (1,024)				
0.5	0.333	0.972	1.000	1.037
1	0.333	1.083	0.918	1.025
2	0.000	0.139	4.509	0.055
3	0.000	0.574	0.126	0.003
5	0.000	0.000	0.000	0.000
7	0.000	0.000	0.000	0.000
10	0.000	0.000	0.000	0.000