CROSS-MODAL STEERING VECTORS EXIST ACROSS BOTH IMAGE AND TEXT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Are concepts and mechanisms unified across modalities in current AI models, or instead implemented separately for each one? We operationalize this broad question of modality unification by studying *cross-modal steering*: 1) finding vectors that successfully steer generation on tokens of one modality, which then 2) transfer zero-shot to steer generation when added to tokens of another modality. Our study goes beyond simple correlational alignment (e.g. representational similarity) because steering is fundamentally causal. We study cross-modal steering in several unified multimodal models (UMMs), since they possess shared parameters for language and image generation, focusing on these two modalities. Across diverse steered concepts such as emotion and object size, we show that steering vectors transfer zero-shot in both directions (text-to-image & image-to-text). Even for concepts that are highly unique to one modality (e.g. formal vs. colloquial language), steering vectors often induce a consistent, interpretable effect in the other modality, such as formal clothing and design. This transfer has a geometric signature too: despite a clear modality gap separating text and image representations, steering vectors with strong cross-modal transferability tend to point in similar directions for the same concept. Our findings shed light on how vision and language interact inside unified multimodal models and point towards more general, modality-agnostic approaches for controlling their behavior.

1 INTRODUCTION

Despite the complexity of modern generative models, such as LLMs and diffusion models, many high-level concepts can be controlled through surprisingly simple interventions: adding a linear direction, or *steering vector* (Subramani et al., 2022), to a model's intermediate representations. Steering vectors have been used to modulate a wide range of behaviors in both language models (Li et al., 2023; Todd et al., 2024; Turner et al., 2024; Dunefsky & Cohan, 2025; Bayat et al., 2025) and image generation systems (Bau et al., 2019; Goetschalckx et al., 2019; Spingarn et al., 2021; Park et al., 2023; Gandikota et al., 2024; Konovalova et al., 2026), providing a lightweight and causal means of controlling model outputs. Yet steering in these two modalities has largely been studied separately. Here we ask: are steering directions themselves modality-specific, or can a direction discovered in one modality also control generation in another?

Consider a steering vector extracted entirely from text that, when added to a model's hidden states, reliably causes it to write more formally. If the same model also generates images, what happens when that very same vector is instead added to its generated image tokens? Could it induce a corresponding notion of *visual formality*, such as more formal clothing or design? And vice versa: could a direction extracted purely from images steer how the model writes? We call this *cross-modal steering*, illustrated in Figure 1, and ask: **Can a steering vector constructed only on tokens from one modality generalize zero-shot when added to tokens from the other modality?**

This question connects to a broader debate about how concepts are represented across modalities. On the one hand, recent evidence from multimodal models suggests substantial internal separation between language and vision: text and image token representations remain nearly orthogonal throughout the layers (Serra et al., 2025), knowledge edited on the text side mostly fails to show

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

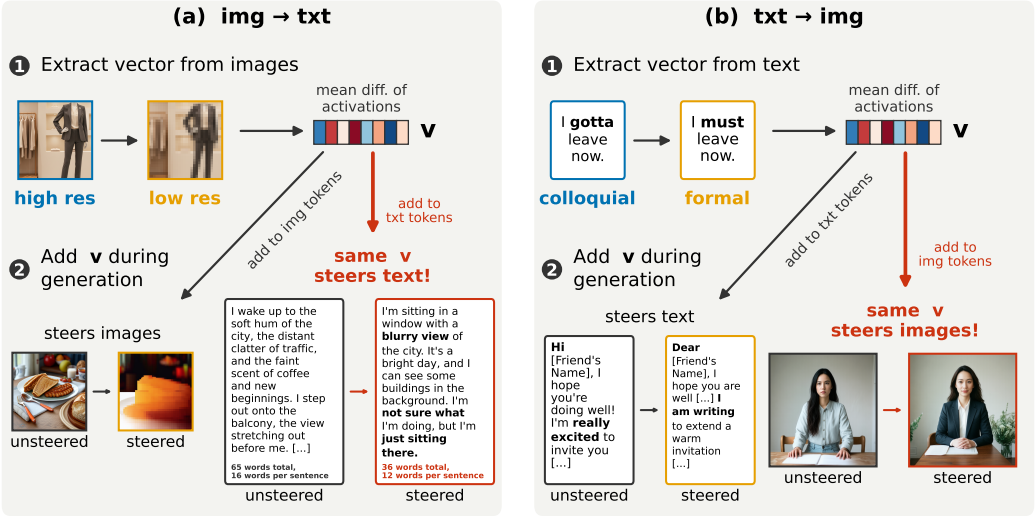


Figure 1: **Cross-modal steering vectors transfer zero-shot in both directions.** A steering vector $\mathbf{v}$ extracted entirely from one modality can steer generation both within that modality and, without any adaptation, in the other.

up in generated images (Gao et al., 2026), and circuits in multi-modal LLMs for the same task are often disjoint (Nikankin et al., 2025). On the other hand, cognitive science has long proposed a modality-agnostic "semantic hub" where concepts live regardless of whether they arrive through language or perception (Patterson et al., 2007). We investigate this through cross-modal steering: successful transfer would demonstrate that a common direction causally modulates a concept across modalities, even when their representations may occupy distinct regions in activation space.

In the mainstream paradigms of multimodal modeling (Rombach et al., 2022; Liu et al., 2023; Wang et al., 2024a), it would not be possible to answer this question since image-to-text and text-to-image generation are separately parameterized and modeled either via autoregression (Radford et al., 2019) or diffusion (Sohl-Dickstein et al., 2015). Thus, we turn to unified multimodal models (UMMs) as a natural testbed for cross-modal steering (Wang et al., 2024b; Team, 2025; Zhou et al., 2025; Chen et al., 2025; Jiao et al., 2025; Xie et al., 2025). Some frontier proprietary systems have similarly moved to generating images natively within a shared language-model architecture rather than calling a separate image model (Gemini Team, 2023; OpenAI, 2025). Yet shared parameters do not imply shared internal mechanisms, making cross-modal transfer far from guaranteed. Indeed, prior work studying cross-modal steering in UMMs reports strong asymmetry: text-derived vectors transfer well to image generation, but image-derived vectors struggle to transfer to text generation including limited transfer in autoregressive models (Wang & Li, 2026).

We find that steering vectors transfer across modalities in both directions. We tested three UMMs that autoregressively generate text and image tokens from a shared weight space: UniAR (Peng et al., 2026), Emu3.5 (Cui et al., 2025) and Liquid (Wu et al., 2026). In UniAR, the transferred vector is almost as good as the intra-modality vector for steering, recovering 84% of its steering effect on images and 74% on text. In Emu3.5 and Liquid, text-derived vectors sometimes even steer images better than image-derived ones. This illustrates the promise of studying cross-modal steering: a steering vector can be constructed in whichever modality is easier, and then used in the other. This matters most for concepts that are highly unique to one modality (see Figure 1). How such a vector would express itself in the other modality gives a glimpse of how the model bridges vision and language: steering an image generator with a vector for formal vs. colloquial *language* causes people or rooms to be depicted more formally, and steering a language model with a vector for low image *resolution* causes the model to write shorter text and express uncertainty. Finally, we show that cross-modal steering has a geometric signature: despite a clear modality gap between text and image representations, the text and image vectors of many concepts point in similar directions. The closer they are, the better they facilitate transfer (§3.4).

Together, our results show that modality-specific representations can nevertheless contain shared causal directions of control, revealing a form of internal structure that is neither fully modality-specific nor captured simply by representational alignment. Beyond offering a new lens on how vision and language interact inside unified models, cross-modal steering points toward modality-agnostic approaches for controlling multimodal generation. Code and demo will be released.

2 Cross-Modal Steering: Methodology

2.1 Preliminaries

Unified Multimodal Models. In our study, we consider unified multimodal models (UMMs) that generate both language and visual content in one autoregressive stream, from an underlying shared weight space (Wang et al., 2024b; Team, 2025; Cui et al., 2025; Wu et al., 2026; Peng et al., 2026; Zhao et al., 2026). Concretely, a UMM models a mixed sequence $x = (x_1, \ldots, x_T)$, in which every token x_t is either a text token or a discrete visual token, with a single next-token objective:

$$p_\theta(x) = \prod_{t=1}^{T} p_\theta(x_t \mid x_{<t}),$$

where the parameters θ are shared across modalities, The term *unified* is used broadly in the literature, and can even include models with mostly separate parameters and mixed objectives (Deng et al., 2025). For our study, we require models to share the weights responsible for generation. A shared autoregressive objective additionally lets us extract and apply steering vectors in the same manner on both token types. We study three such UMMs: UniAR (8B) (Peng et al., 2026), Emu3.5 (34B) (Cui et al., 2025) and Liquid (7B) (Wu et al., 2026). They differ most in how images are tokenized: UniAR's image tokens are semantic, derived from a language-aligned vision encoder, whereas those of Emu3.5 and Liquid are trained to reconstruct pixels.

Steering Vectors. Steering is any intervention on model internals that allows continuous modulation of the output generation or a behavior by adding a vector to the intermediate hidden states during generation (Subramani et al., 2019; 2022; Li et al., 2023; Turner et al., 2024):

$$h'_l = h_l + \alpha\, v_l.$$

Here h_l is a hidden representation at layer l, v_l is the steering vector, and the scalar α sets the strength and sign of the intervention. The two key design choices are a) how to find such a vector and b) how to apply it during generation. In LLMs, steering vectors are typically found from contrastive prompt pairs or from probes and sparse dictionaries trained on the residual stream, and applied to one or many layers at every step of token-by-token decoding (Li et al., 2023; Rimsky et al., 2024; Huben et al., 2024). In image generation, directions are instead often discovered in the latent space of a diffusion model (e.g., by unsupervised decomposition) or learned as lightweight adapters, and applied along the denoising trajectory (Kwon et al., 2023; Park et al., 2023; Haas et al., 2024; Gandikota et al., 2024). The contrastive recipe that is standard for LLMs is less common in vision (Haas et al., 2024; Gaintseva et al., 2026; Konovalova et al., 2026), in part because minimal image pairs are much harder to curate than minimal text pairs.

2.2 Our approach

We adopt one established approach for both language and vision: we find steering vectors via a dataset of minimally different inputs, and add them to representations at either a subset of layers or whole ranges of layers. Crucially, finding and applying the vector need not use the same modality: for cross-modal steering, we construct a vector entirely from examples in one modality and then apply that *same vector, without any adaptation or target-modality data*, to representations of the other modality during generation. This gives four steering settings, corresponding to all combinations of finding a vector in either modality and applying it to either modality, as illustrated in Figure 1. Throughout the paper, we refer to them as txt→txt, txt→img, img→txt, and img→img: txt→img and img→txt are the transfer settings we primarily study, and the intra-modality settings txt→txt and img→img serve as reference points. We color source and target modalities throughout (txt, img).

Extracting steering vectors. We extract steering vectors in a very similar way for text and images. Given a concept and its associated poles (e.g., emotion: happy/sad), we generate $N = 100$ minimal pairs (x_i^+, x_i^-) that differ only in the target pole. We pass both inputs through the model and extract one representation per input at each layer ℓ. The steering vector is then computed as the average difference between paired representations:

$$v_\ell = \frac{1}{N} \sum_{i=1}^{N} \big(h_{i,\ell}^+ - h_{i,\ell}^-\big).$$

We normalize each layer's steering vector to unit length, $\hat{v}_\ell = v_\ell / \|v_\ell\|$, so that the steering strength is set by a single coefficient that is comparable across layers and concepts. The two modalities differ only in how the pairs are built and how an input is reduced to a single representation h.

Text. We automatically generate the pairs with Gemini 2.5 Flash (Gemini Team, 2025) that produces a diverse list of subjects that we fill into a fixed contrastive template. Each pair consists of two short declarative prompts that only differ in the target attribute, which is placed as the last token, e.g., *a person is happy* and *a person is sad*. We use the hidden state of the last token as h.

Image. We first generate a set of baseline images with Gemini 2.5 Flash Image. We then edit each baseline image towards the two opposite values of the target attribute with an instruction-based image-editing model, for example *make the person happier* and *make the person sadder*, while mentioning in the prompt to keep the same subject, pose, framing, and background. For most concepts we edit with the same model. For size, depth, and spatial, where its edits are less reliable, we use Gemini 3 Pro Image and Reve[1]. For simplicity, we mean-pool the hidden states across all M image tokens, $h = \frac{1}{M} \sum_{j=1}^{M} h_j$.

Applying steering vectors. Given a steering vector, we intervene on the hidden states of all generated tokens, and for text, also on those of the subject phrase in the prompt, in the shared transformer backbone: at layer ℓ, we set $h'_\ell = h_\ell + \alpha\, \hat{v}_\ell$, where α controls the steering strength and $\hat{v}_\ell$ is the per-layer unit steering vector defined above. The intervention is identical whether the generated tokens are text or image tokens, and whether $\hat{v}_\ell$ was extracted from text or from images; in the transfer settings, the vector is simply added to tokens of the other modality. We initially explore several strategies for *where* in the model to apply this steering (single layers, contiguous ranges of layers, or all layers at once). We focus on ranges of layers in §3.2 — early, in the middle, or late in the model — since they steer more strongly than single layers and, unlike steering all layers at once, show *where* in the model the two modalities align[2]. We sweep the steering strength α over a range chosen per model and setting, because usable strengths differ greatly between text and image generation and are larger for transferred vectors. Each range extends just past the point where outputs break down (incoherent text, broken images).

Implementation details. *Concepts:* We study eight concepts: four *semantic* concepts that naturally occur in both modalities, emotion (sad/happy), age (young/old), cleanness (clean/dirty) and chaos (peaceful/chaotic), and four *visual* concepts that are spatial, geometric or appearance properties of a scene, size (small/large), depth (far/near), spatial (left/right) and color (blue/red). We choose concepts that are commonly steered or studied in language models (Turner et al., 2024; Zou et al., 2025) or in image generation (Shen et al., 2020; Gandikota et al., 2024; Baumann et al., 2025). *Minimal pairs:* The neutral baseline images are generated from short scene descriptions written by Gemini 2.5 Flash. We manually inspected all image pairs for whether the intended pole is shown while other aspects are preserved; the few failed edits (fewer than 5%) were regenerated with a revised edit instruction or a stronger model. *Layers:* We steer early, middle and late ranges of four consecutive layers (eight for Emu3.5, which is twice as deep) at one quarter, one half and three quarters of the model's depth, e.g., layers 7–10, 16–19 and 25–28 of 36 in UniAR. *Evaluation prompts:* For each concept, we generate 20 neutral scene descriptions around one subject with Gemini 2.5 Flash (e.g., *a wooden chair in a large empty room*). Image generation uses the description as is; for text generation, we iterate on various open-ended prompts, finding that *"Describe your day in the city. Your description must include {subject}."* elicits diverse open-ended descriptions where steered concept can surface in whatever the model chooses to mention. Full details are in Appendix A.

[1] `https://blog.reve.com/posts/the-reve-api/`

[2] Appendix B also shows results for single-layer and all-layer settings.

3 ARE STEERING DIRECTIONS SHARED?

3.1 MEASURING CROSS-MODAL TRANSFER

Given a concept c, a steering strength α and an evaluation prompt p, we write $\tilde{x}_{c,\alpha,p}$ for the steered output and x_p for the unsteered output at $\alpha = 0$. We measure steering success with an LLM judge that compares the two, and further quantify cross-modal transfer by comparing how well a transferred vector steers with how well the intra-modality vector does.

Concept rate. We use Gemini 3.7 Flash (Google DeepMind, 2026) as the judge. It sees x_p and $\tilde{x}_{c,\alpha,p}$ in random order, without knowing which one is steered, and is asked which one shows more of a pole of the concept, for example *happy* for emotion, with the asked pole chosen at random between the concept's two poles. The correct answer is the steered output when the judge is asked for the pole that α steers toward, and the baseline otherwise. We score each pair $(x_p, \tilde{x}_{c,\alpha,p})$ as

$$J(x_p, \tilde{x}_{c,\alpha,p}) = \begin{cases} 1, & \text{if the judge picks the correct output,} \\ 0, & \text{if it picks the other output or answers } \texttt{Tie}. \end{cases}$$

The concept rate is the mean score over the 20 prompts of a concept,

$$\mathrm{CR}_{c,\alpha} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} J(x_p, \tilde{x}_{c,\alpha,p}), \qquad |\mathcal{P}| = 20.$$

To summarize a concept-rate curve across steering strengths α with one number, we take its normalized area under the curve, $\mathrm{AUC}_c \in [0, 1]$. An AUC of 1 means the judge identified the concept in every output at every strength, and 0 that it never did. [3]

Cross-modal recovery. To judge *transfer*, we measure how much of the intra-modality steering performance (i.e., AUC_c) a transferred vector recovers. We consider the intra-modality vector as a reference point since we expect a vector to steer the modality it was found in best. Thus recovery is defined for image generation as $\mathrm{Rec}_{\text{img}}$ and for text generation as $\mathrm{Rec}_{\text{txt}}$:

$$\mathrm{Rec}_{\text{img}} = \frac{\mathrm{AUC}_c(\text{txt} \to \text{img})}{\mathrm{AUC}_c(\text{img} \to \text{img})}, \qquad \mathrm{Rec}_{\text{txt}} = \frac{\mathrm{AUC}_c(\text{img} \to \text{txt})}{\mathrm{AUC}_c(\text{txt} \to \text{txt})}.$$

A recovery of 1 means the transferred vector steers as well as the intra-modality one, 0 means it does not steer at all, and above 1 means it steers better. Since each setting has its own strength range (§2.2), recovery compares each vector over the strength ranges it tolerates, not at equal α.

Further details. The judge agrees with human raters at Cohen's $\kappa = 0.62$ (Cohen, 1960). On top of concept rate, we separately report a metric for output quality, namely the fluency of steered text and the visual quality of steered images, adopted from (Wu et al., 2025b). Details in Appendix A.

3.2 RESULTS

Across our three models and eight concepts, we report Cross-modal Recovery (both $\mathrm{Rec}_{\text{img}}$ and $\mathrm{Rec}_{\text{txt}}$) in Figure 3 and table 1. The concept-rate curves underlying these numbers are shown in Figure 2 for UniAR and in the Appendix for Emu3.5 and Liquid (Figures 11 and 12).

Steering vectors transfer in both directions. In all three models, vectors found in one modality also steer generation in the other for most concepts (Figure 3 and table 1). For every concept except spatial, there is a layer range at which text-derived vectors recover more than 80% of the intra-modality effect, in all three models. UniAR exhibits the strongest cross-modal steering effects: at middle layers, transferred vectors recover on average 84% of the intra-modality effect on images and 74% on text, and its concept-rate curves in Figure 2 follow the intra-modality curves closely for most concepts.[4] txt→img is strongest at middle layers in all three models; transfer into text peaks at early layers in UniAR and at late layers in Emu3.5 and Liquid, where their mean img→txt recovery reaches 0.87 and 0.91. Examples for txt→img and img→txt are shown in Figure 4.

[3] AUC measures are commonly used as summary statistics in machine learning (e.g. for benchmark evaluation (Subramani et al., 2025b), classification (AUROC; Marcum, 1960), and calibration (El-Yaniv & Wiener, 2010; Subramani et al., 2025a; 2026), and in the sciences (Wagner & Ayres, 1977; Hanley & McNeil, 1982).

[4] We speculate this is due to the difference in image encoders (see §2.2).

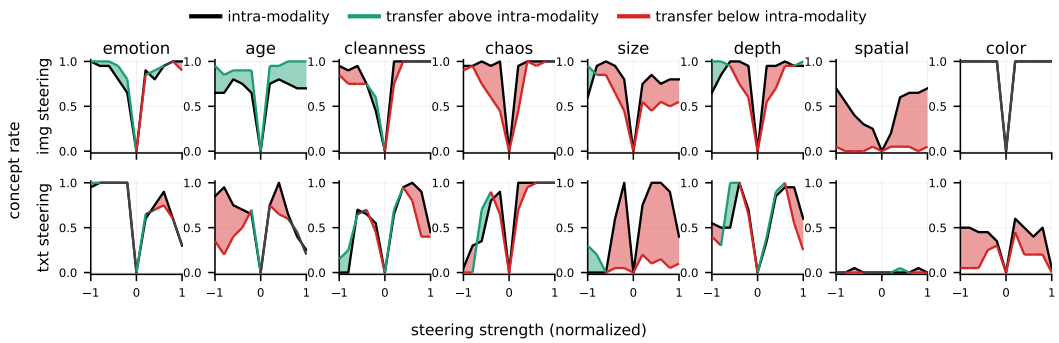


Figure 2: **Concept rate on UniAR (Peng et al., 2026) at middle layer range.** Concept rate is the fraction of steered outputs in which an LLM judge finds the concept shifted toward the steered pole (§3.1), plotted against normalized steering strength. Black curves show intra-modality steering (top: img→img, bottom: txt→txt). The colored areas indicate the gap between the two, either when transfer performs better (green) or worse (red). Recovery in Figure 3 and table 1 quantifies this comparison. For most concepts, the transfer curve comes close to the intra-modality curve.

txt→img is easier than img→txt. Text-derived vectors steer images for nearly every concept, and in Emu3.5 and Liquid even often better than intra-modality steering (img→img): their txt→img recovery averages 1.10 and 1.20, i.e., above 1. We hypothesize this is mostly due to weaker image-derived vectors in these two models rather than particularly strong transfer: img→img reaches an AUC of only 0.44 and 0.43 at middle layers, compared to 0.83 in UniAR. Practically, for these two models, text vectors are the better vectors for steering images, despite image-derived vectors being "native" to the space. In the other direction, img→txt is only widely successful in UniAR, while more mixed in Emu3.5 and Liquid. UniAR recovers on average up to 0.96 of the intra-modality effect, whereas Emu3.5 and Liquid come close only at late layers (0.87 and 0.91).

Emotion transfers best, size and spatial worst. Emotion transfers most evenly, with a mean recovery of 0.95 in both directions, and age transfers into images better than other concepts (1.02–1.87 across models and layer ranges). Visual concepts tend to work better in the txt→img setting than in img→txt. Size shows this most clearly with overall mean recovery of 1.12 for txt→img but only 0.32 for img→txt. The only concept that fails across most settings is spatial (left-right), with moderate success only for img→img in UniAR (see curve in Figure 2) and minor effects for txt→txt in Emu3.5 and Liquid.

Output quality. We check whether there is a strong trade-off between high concept rates and output quality. With an AxBench-like judge (Wu et al., 2025b), we find that output quality noticeably drops only beyond the strengths where the concept rate plateaus (and, for text, then drops): at 0.4 of the strength range, where the text concept rate peaks, text fluency is still at about three quarters of the level of unsteered outputs, and it collapses beyond that point (Appendix A.7). On the image generation side the trade-off is smaller: image coherence is still at about 90% of its unsteered level at 0.6 of the range, where the concept rate has already reached over 90% of its maximum.

3.3 Qualitative analysis: modality-specific concepts

So far we have studied concepts that occur in both language and vision, such as emotion or age, and some that are more central to vision, such as left-right. We next study concepts that are unique to one modality, but harder to define in the other one. For example, what would an image look like if steered along a uniquely linguistic concept? We qualitatively inspect and discuss two examples for each modality in UniAR, as illustrated in Figure 1. Detailed sample outputs are in Appendix D.

Language. We construct sets of minimal pairs in language that should elicit two purely linguistic steering axes: Formal vs. colloquial language ("I need to go" vs. "I gotta go") and language identity ("I love apples" vs."J'adore des pommes"). We verify that these work as intended on language, e.g.,

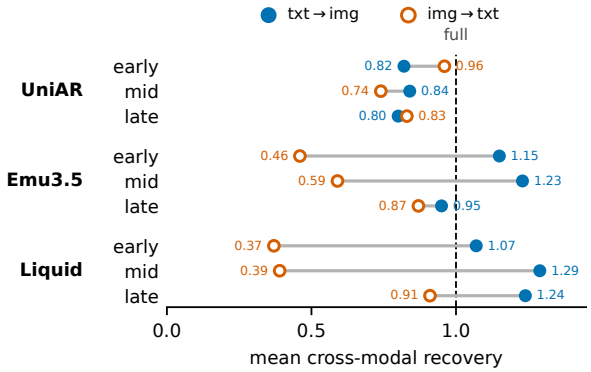
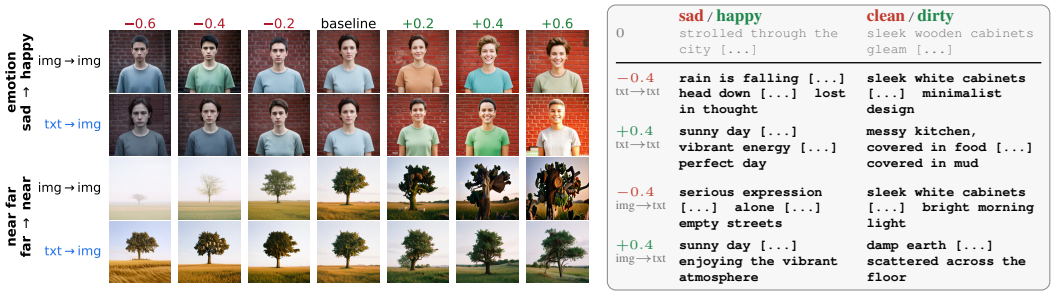


Figure 3: **Cross-modal recovery** (averaged over concepts, showing steered models and layer ranges): ratio of transferred vector's AUC to intra-modality AUC of concept-rate curve (see §3.1). 1.0 (dashed) means the transferred vector steers as well as the intra-modality one.

Concept	txt→img	img→txt
emotion	0.95	0.95
age	1.45	0.60
cleanness	0.97	0.69
chaos	1.06	0.70
size	1.12	0.32
depth	0.93	0.82
spatial	0.08	0.01
color	0.97	0.88
mean	1.04	0.68

Table 1: **Cross-modal recovery per concept**, averaged over models and layer ranges. Gray marks weak transfer (recovery < 0.5).



(a) Image generation

(b) Text generation

Figure 4: **Cross-modal and intra-modal steering in UniAR** (middle layer range). We show examples for emotion, distance and cleanness steering. In both modalities the transferred vector produces a similar shift as the intra-modality vector.

producing more and more colloquial e-mails, or switching languages. Despite our minimal pairs only defining the *linguistic* aspect of formality and not, e.g., suits vs. sportswear, we find that this steering direction transfers to visual properties such as clothing, black-and-white vs. colorful, and luxurious minimalist style. The English–French direction is also purely linguistic, translating the same semantic content into a different form. Nonetheless, the steered images exhibit an increasingly French style that is historical and ornate rather than modern (mansard roofs, Art Nouveau posters, a vintage car), and, surprisingly, people express more feminine-presenting visual characteristics. Writing in the image turns into French-looking but non-existent words.

Vision. It is less trivial to find exclusively visual concepts since any concept, almost by definition, has words associated to it or might not allow reasonable minimal pair construction. However, we identify two successful case studies. First, applying a steering vector found in images for the axis of low-to-high resolution makes generated text less (or more) detailed: fewer generated tokens, shorter sentences, mentions of "blurry" and even expressions of confusion[5]; see Figure 1. We observe similar effects when applying a "motion blur" vector to text, with descriptions showing either calmer words on one end of the spectrum or faster movement-related ones on the end. Finally, we successfully construct a steering vector for camera angle, i.e., whether the camera is positioned above the subject or near the floor; it steers images cleanly, but reaches text only in minor ways, shifting the next-token probability distribution without ever changing the predicted token.

[5]We confirm that overall shorter responses is not an artifact of steering: applying other steering vectors lead too much less reduction in length, if any at all.

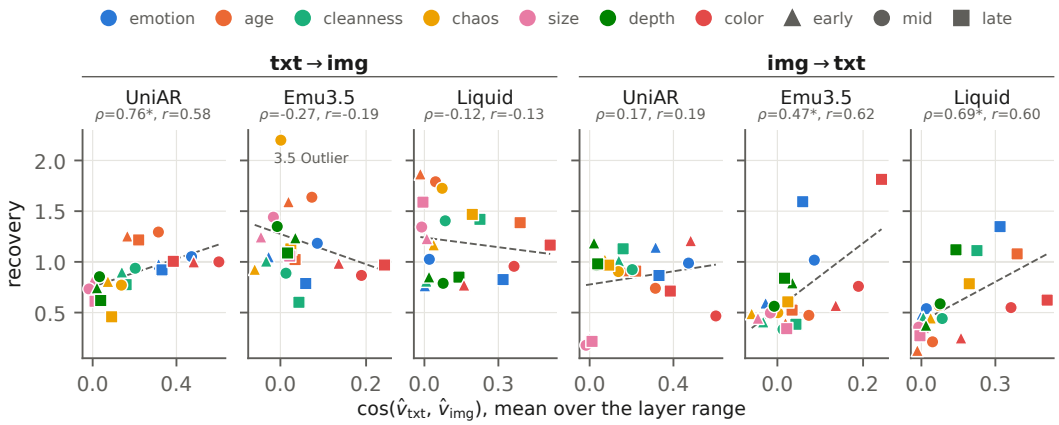


Figure 5: **Cross-modal cosine similarity vs. cross-modal steering success.** Each point is one (concept, layer range): x is cosine between text- and image-derived steering vectors; y is our recovery metric (§3.1) for the given transfer direction (left: txt→img, right: img→txt). Spatial is excluded because it does not steer in most settings; ρ: Spearman, r: Pearson; * marks $p < 0.05$.

3.4 Geometry of steering vectors

Do text-derived and image-derived steering vectors point in similar directions? And if that is the case, would their cosine similarity be predictive how well cross-modal steering works?

Finding: Cosine similarities correlate with cross-modal steering (Figure 5). For UniAR and Liquid, we find that cosine similarities for most concepts are significantly above zero[6], up to 0.60 (UniAR) and 0.51 (Liquid). On the other hand, Emu3.5's steering vectors are close to zero cosine similarity for all concepts except color (maximum 0.24). Concretely, averaging across all (model, concept, layer range) points, we find a moderate correlation between cosine similarity and our main metric of steering transfer (*recovery*, averaged over both transfer directions): Spearman $\rho = 0.38$ ($p < 0.01$), Pearson $r = 0.22$ ($n = 63$, the 21 points per model of Figure 5). However, this correlation varies significantly across studied models and between img→txt vs. txt→img. For example, in UniAR the correlation is strongest for txt→img transfer ($\rho = 0.76$). In all three models the correlating direction is the one whose *intra-modality* steering is strongest (img→img in UniAR, txt→txt in Liquid and Emu3.5), i.e. where the denominator of the recovery ratio is reliable.

Steering vectors show some alignment despite modality gap. For example, the happy–sad vectors found in text and in images point in somewhat similar directions, with a cosine similarity of 0.47 in UniAR's mid layers. Broadly, there could be two explanations for this: 1) The text and image activations for happy inputs are close, text and image activations for sad inputs are close, therefore the two difference vectors are close. 2) Text and image occupy separate subspaces but the two subspace share a similar structure, i.e., relations between the activations in each space are similar. **Setup:** We test both explanations in UniAR's representation space: each concept pole in each modality (e.g., happy images) is summarized by the mean activation over its minimal-pair examples. Activations are strongly anisotropic: two random activations of the same modality have a cosine similarity of about 0.75 at the middle layers, and two from opposite modalities still 0.65. Therefore, we rescale them to unit length and subtract the overall mean before computing cosines (Appendix C); the steering vectors themselves are unaffected by this centering.

We find evidence for the second explanation: Image and text occupy separate narrow spaces inside the model, while the relative directions between activations within each region exhibits some alignment. If the first explanation were right, a happy image and a happy sentence would sit close together in the model. However, in UniAR's middle layers, the average happy image is more than ten times closer to the average *sad* image than to the average happy sentence, with similar or stronger trend

[6] 30 of the 42 (concept, layer range) points of UniAR and Liquid in Figure 5 lie more than two standard deviations (0.03) above random baseline.

for every concept. On the other hand, the steering directions extracted in text and in images have a cosine similarity between 0.14 and 0.60 for five of the eight concepts, far above a random-direction baseline of ± 0.03, while size, near-far and spatial-lr are around zero. Emu3.5 and Liquid show the same separation by modality. Our findings illustrate that steering vectors can provide a separate avenue for quantifying whether a unified model is *actually* unified. Aligned directions despite separate subspaces is also related to modality-gap studies in contrastive embedding spaces (Liang et al., 2022; Zhang et al., 2024) and the vector-offset regularities in word embeddings, i.e., the famous king-queen example (Mikolov et al., 2013b; Park et al., 2024) and the Platonic Representation Hypothesis (Huh et al., 2024; Gröger et al., 2026).

4 Related Work

Unification and Alignment of Modalities. Approaches and tasks for connecting vision and language differ widely, from multimodal language models (Lu et al., 2019; Tsimpoukelli et al., 2021; Alayrac et al., 2022), text-conditioned image generation (Rombach et al., 2022), multimodal retrieval (Radford et al., 2021), to the many variants of recent unified multimodal models (Team, 2025; Wang et al., 2024b; Deng et al., 2025). Across all, it is widely studied how vision and language representations interact and align (Merullo et al., 2023; Neo et al., 2025; Wu et al., 2025a), ranging from modality gaps studies in CLIP models (Liang et al., 2022) to circuit analyses and vocabulary space projections in multimodal LLMs (Nikankin et al., 2025; Krojer et al., 2026; Feucht et al., 2026). Only few works have studied internals of unified models, via attention patterns and embedding geometry Serra et al. (2025), transfer of knowledge edits Gao et al. (2026) and steering vectors Wang & Li (2026). They often report a lack of unification (Serra et al., 2025; Gao et al., 2026), in contrast to our more positive findings. Most similar to us is Wang & Li (2026) who report cross-modal steering vectors in only one direction (txt→img). They primarily study a different type of UMM, BAGEL (Deng et al., 2025) with separate language and image generation weights and objectives, while relying on minimal text inputs for both text and image vectors. Finally, they study closed-form QA text outputs and not our setting of free-form text.

Visual steering. Most related to our transfer question, Gan et al. (2026) and Deng et al. (2026) each study one direction of transfer (text to image tokens and image to text, respectively) to alter the visual *understanding* of MLLMs, which only take images as conditioning. Other continuous attribute control in image generation (Baumann et al., 2025; Ekin & Gandelsman, 2026; Dalva & Yanardag, 2024; Kamenetsky et al., 2025; Wang et al., 2026) likewise targets diffusion models rather than the shared space of a unified model. Steering the visual-token stream of an *autoregressive* image generator, as we do, is rarely explored (Chrabaszcz et al., 2026; Wang & Li, 2026).

Textual steering. Identifying and applying linear and non-linear "concept directions" started with operating on embeddings themselves (Mikolov et al., 2013a; Bolukbasi et al., 2016). Now, we operate in the latent spaces of language models for both exact and attribute-based steering (Subramani et al., 2019; 2022; You et al., 2026). For these causal steering operations, data often comes from contrastive prompt pairs (Turner et al., 2024; Rimsky et al., 2024; Zou et al., 2025), in-context task demonstrations (Todd et al., 2024), or sparse-autoencoder features (Huben et al., 2024), all grounded in the linear representation hypothesis (Park et al., 2024).

5 Conclusion

In this paper we have shown cross-modal steering vectors in both directions (txt→img and img→txt). We also characterized and provided intuition how such cross-modal steering can manifest in the other modality. As unified multimodal models become more capable and mature as a field, we see potential for methods that leverage this unified nature for model control. If we identify a promising (steering) intervention to make models safer (Arditi et al., 2024), it should automatically transfer to visual capabilities; likewise, if we want to regularly update knowledge in the model, such knowledge edits should transfer to visual capabilities (Gao et al., 2026). We encourage future work to expand our findings to other families of unified models and novel concepts. Scientifically, we see unified models as a fruitful testbed for developing modality-agnostic interpretability tools and theories of multimodal interactions, as exemplified by our study of cross-modal steering.

ETHICS STATEMENT

Our work studies internal steering directions of open-weight unified multimodal models on synthetic minimal pairs; we foresee no direct ethical concerns beyond the plethora of general risks of controllable generation, including generation of harmful content targeting and/or exploiting vulnerable populations.

REPRODUCIBILITY STATEMENT

Appendix A contains everything needed to reproduce our results: the construction of the minimal pairs with all edit instructions, the evaluation prompts, decoding settings per model, layer and strength selection, the judge prompts and metrics, and the human validation. Appendix C details the geometry analyses. We will release the code, minimal pairs, steering vectors, and all generated outputs with their judge scores.

AI USE STATEMENT

We used generative AI tools throughout this project, always in a back-and-forth with the authors, and not simply adopting what they suggest.

Research and experiments. We used coding agents to implement the methods and to run most experiments reported in the paper, including cleaning and reformatting the generated data. Some methodological decisions, the hypotheses we tested, and how prior work has approached the problem were discussed with AI tools. We never blindly adopted their suggestions, it was a long back-and-forth in which we often disagreed with the AI assistant and often stuck to our own scientific intuition. AI assistants also helped interpret results and describe qualitative patterns in steered outputs; every such interpretation was checked against the underlying numbers and generations by the authors. We did not review agent-written code line by line. Instead, we discussed the code and its outputs with the agent, re-ran experiments, and sanity-checked results (e.g., for strange or implausible values) before adopting them into the paper.

Data and evaluation. We used generative AI models to automatically generate the minimal pairs for concept directions, including the French sentences of the language-identity pairs, and the evaluation prompts (Appendices A.1 and A.2). The authors inspected the generated pairs and measured how minimal they are before using them. An LLM (Gemini 3.7 Flash) served as the judge for all reported metrics; its agreement with human raters is reported in Appendix A.8.

Writing, figures and literature. The final paper is the full responsibility of, and was primarily driven by, us authors. We sometimes adopted isolated sentence-level suggestions from AI assistants, especially for sentences reporting technical details or for rephrasing a suboptimal phrase, and used them much more sparingly for writing than for coding and experiments. AI tools helped create the figures, and helped identify and summarize related work; when an AI tool suggested citations, we checked them against its original source.

We did not use AI for developing theoretical models or for formulating or proving mathematical claims (not applicable here). All AI-assisted content was reviewed by the authors, who take responsibility for the final content of this work.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 9

Andy Arditi, Oscar Balcells 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 (NeurIPS)*, 2024. URL https://arxiv.org/abs/2406.11717. 9

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. In *ICLR (Poster)*, 2019. URL https://openreview.net/forum?id=Hyg_X2C5FX. 1

Stefan Andreas Baumann, Felix Krause, Michael Neumayr, Nick Stracke, Melvin Sevi, Vincent Tao Hu, and Björn Ommer. Continuous, subject-specific attribute control in T2I models by identifying semantic directions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13231–13241, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Baumann_Continuous_Subject-Specific_Attribute_Control_in_T2I_Models_by_Identifying_Semantic_CVPR_2025_paper.html. 4, 9, 19

Reza Bayat, Ali Rahimi-Kalahroudi, Mohammad Pezeshki, Sarath Chandar, and Pascal Vincent. Steering large language model activations in sparse spaces. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=VGw1viYliK. 1

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016. 9

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=xYGNO86OWDH. 38

Agneet Chatterjee, Gabriela Ben Melech Stan, Estelle Aflalo, Sayak Paul, Dhruba Ghosh, Tejas Gokhale, Ludwig Schmidt, Hannaneh Hajishirzi, Vasudev Lal, Chitta Baral, and Yezhou Yang. Getting it right: Improving spatial consistency in text-to-image models. In *Computer Vision – ECCV 2024*, pp. 204–222. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72670-5_12. 19

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling, 2025. URL https://arxiv.org/abs/2501.17811. 2

Maciej Chrabaszcz, Aleksander Szymczyk, Jan Dubiński, Tomasz Trzciński, Franziska Boenisch, and Adam Dziedzic. Conditioned activation transport for t2i safety steering, 2026. URL https://arxiv.org/abs/2603.03163. 9

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, April 1960. ISSN 1552-3888. doi: 10.1177/001316446002000104. URL http://dx.doi.org/10.1177/001316446002000104. 5

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, Yueze Wang, Chengyuan Wang, Fan Zhang, Yingli Zhao, Ting Pan, Xianduo Li, Zecheng Hao, Wenxuan Ma, Zhuo Chen, Yulong Ao, Tiejun Huang, Zhongyuan Wang, and Xinlong Wang. Emu3.5: Native multimodal models are world learners, 2025. URL https://arxiv.org/abs/2510.26583. 2, 3

Yusuf Dalva and Pinar Yanardag. Noiseclr: A contrastive learning approach for unsupervised discovery of interpretable directions in diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24209–24218. IEEE, 2024. URL https://doi.org/10.1109/CVPR52733.2024.02285. 9

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining, 2025. URL https://arxiv.org/abs/2505.14683. 3, 9

Zehao Deng, Tianjie Ju, Zheng Wu, Liangbo He, Jun Lan, Huijia Zhu, Weiqiang Wang, and Zhuosheng Zhang. Causal probing for internal visual representations in multimodal large language models, 2026. URL https://arxiv.org/abs/2605.05593. 9

Jacob Dunefsky and Arman Cohan. One-shot optimized steering vectors mediate safety-relevant behaviors in LLMs. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=teW4nIZ1gy. 1

Yigit Ekin and Yossi Gandelsman. The unreasonable effectiveness of text embedding interpolation for continuous image steering, 2026. URL https://arxiv.org/abs/2603.17998. 9

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL http://jmlr.org/papers/v11/el-yaniv10a.html. 5

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 55–65, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1006. URL https://aclanthology.org/D19-1006/. 26, 38

Sheridan Feucht, Benno Krojer, Sarah Wang, Henry Abrahamsen, Byron C. Wallace, and David Bau. Using ocr heads to verbalize image semantics, 2026. URL https://arxiv.org/abs/2609.18823. 9

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 28

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, 2026. URL https://arxiv.org/abs/2503.09630. 3, 26

Woody Haosheng Gan, Deqing Fu, Julian Asilis, Ollie Liu, Vatsal Sharan, Robin Jia, and Willie Neiswanger. Textual steering vectors can improve visual understanding in multimodal large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 40056–40087. Association for Computational Linguistics, 2026. URL https://aclanthology.org/2026.acl-long.1861/. 9

Rohit Gandikota, Joanna Materzynska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: LoRA adaptors for precise control in diffusion models. In *Computer Vision – ECCV 2024*, volume 15098 of *Lecture Notes in Computer Science*, pp. 172–188. Springer, 2024. URL https://doi.org/10.1007/978-3-031-73661-2_10. 1, 3, 4, 19

Xin Gao, Cheng Yang, Chufan Shi, and Taylor Berg-Kirkpatrick. Do text edits generalize to visual generation? benchmarking cross-modal knowledge editing in umms. *arXiv preprint arXiv:2606.00477*, 2026. 2, 9

Gemini Team. Gemini: A family of highly capable multimodal models, 2023. URL https://arxiv.org/abs/2312.11805. 2

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261. 4

Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5743–5752. IEEE, 2019. 1

Google DeepMind. Gemini 3.7 flash model card. https://deepmind.google/models/model-cards/gemini-3-7-flash/, August 2026. Accessed 2026-09-21. 5

Fabian Gröger, Shuo Wen, and Maria Brbic. Revisiting the platonic representation hypothesis: An aristotelian view. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=uz0gAAYydl. 9

René Haas, Inbar Huberman-Spiegelglas, Rotem Mulayoff, Stella Graßhof, Sami S. Brandt, and Tomer Michaeli. Discovering interpretable directions in the semantic latent space of diffusion models. In *18th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–9. IEEE, 2024. URL https://doi.org/10.1109/FG59268.2024.10581912. 3

James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982. 5

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations (ICLR)*. OpenReview.net, 2024. URL https://openreview.net/forum?id=F76bwRSLeK. 3, 9

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 2024. URL https://proceedings.mlr.press/v235/huh24a.html. 9

Yang Jiao, Haibo Qiu, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3600–3610. Computer Vision Foundation / IEEE, 2025. URL https://openaccess.thecvf.com/content/CVPR2025W/MMFM/html/Jiao_UniToken_Harmonizing_Multimodal_Understanding_and_Generation_through_Unified_Visual_Encoding_CVPRW_2025_paper.html. 2

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9161–9175, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.568. URL https://aclanthology.org/2023.emnlp-main.568/. 19

Ronen Kamenetsky, Sara Dorfman, Daniel Garibi, Roni Paiss, Or Patashnik, and Daniel Cohen-Or. Saedit: Token-level control for continuous image editing via sparse autoencoder, 2025. URL https://arxiv.org/abs/2510.05081. 9

Nina Konovalova, Andrey Kuznetsov, and Aibek Alanov. Shift: Steering hidden intermediates in flow transformers, 2026. URL https://arxiv.org/abs/2604.09213. 1, 3

Benno Krojer, Shravan Nayak, Oscar Mañas, Vaibhav Adlakha, Desmond Elliott, Siva Reddy, and Marius Mosbach. Latentlens: Revealing highly interpretable visual tokens in llms, 2026. URL https://arxiv.org/abs/2602.00462. 9

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *The Eleventh International Conference on Learning Representations (ICLR)*. OpenReview.net, 2023. URL https://openreview.net/forum?id=pd1P2eUBVfq. 3

Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html. 1, 3

Michael Li and Nishant Subramani. Model internal sleuthing: Finding lexical identity and inflectional features in modern language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15825–15864, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.720. URL https://aclanthology.org/2026.acl-long.720/. 26

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17612–17625. Curran Associates, Inc., 2022. doi: 10.52202/068431-1280. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/702f4db7543a7432431df588d57bc7c9-Paper-Conference.pdf. 9, 38

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36*, NeurIPS 2023, pp. 34892–34916. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-1516. URL http://dx.doi.org/10.52202/075280-1516. 2

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 9

J. Marcum. A statistical theory of target detection by pulsed radar. *IRE Transactions on Information Theory*, 6(2):59–267, 1960. doi: 10.1109/TIT.1960.1057560. 5

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=8tYRqb05pVn. 9

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013a. 9

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff (eds.), *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751, Atlanta, Georgia, June 2013b. Association for Computational Linguistics. URL https://aclanthology.org/N13-1090/. 9

Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HkuGJ3kCb. 38

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 57172–57189, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/900fb3439e4968df79a6f2bfedec49cd-Paper-Conference.pdf. 9

Yaniv Nikankin, Dana Arad, Yossi Gandelsman, and Yonatan Belinkov. Same task, different circuits: Disentangling modality-specific mechanisms in VLMs. In *Advances in Neural Information Processing Systems*, 2025. 2, 9

OpenAI. Addendum to GPT-4o system card: Native image generation. https://cdn.openai.com/11998be9-5319-4302-bfbf-1167e093f1fb/Native_Image_Generation_System_Card.pdf, March 2025. 2

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Forty-first International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 2024. URL https://proceedings.mlr.press/v235/park24c.html. 9

Yong-Hyun Park, Mingi Kwon, Junghyo Jo, and Youngjung Uh. Unsupervised discovery of semantic latent directions in diffusion models, 2023. URL https://arxiv.org/abs/2302.12469. 1, 3

Karalyn Patterson, Peter J. Nestor, and Timothy T. Rogers. Where do you know what you know? the representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, 8 (12):976–987, December 2007. ISSN 1471-0048. doi: 10.1038/nrn2277. URL http://dx.doi.org/10.1038/nrn2277. 2

Wujian Peng, Lingchen Meng, Yuxuan Cai, Xianwei Zhuang, Yuhuan Yang, Rongyao Fang, Chenfei Wu, Junyang Lin, Zuxuan Wu, and Shuai Bai. Unified multimodal autoregressive modeling with shared context-visual tokenizer is key to unification. *arXiv preprint arXiv:2606.18249*, 2026. 2, 3, 6

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 2

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 9

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 15504–15522. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.acl-long.828. 3, 9

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020. doi: 10.1162/tacl_a_00349. URL https://aclanthology.org/2020.tacl-1.54/. 26

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022. 2, 9

Alessandro Serra, Francesco Ortu, Emanuele Panizon, Lucrezia Valeriani, Lorenzo Basile, Alessio Ansuini, Diego Doimo, and Alberto Cazzaniga. The narrow gate: Localized image-text communication in native multimodal models. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/6c420d6d2f19acbd44c219d244d892ee-Abstract-Conference.html. 1, 9

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of GANs for semantic face editing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9240–9249. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00926. 4, 19

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=WGXb7UdvTX. 26

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html. 2

Nurit Spingarn, Ron Banner, and Tomer Michaeli. {GAN} "steerability" without optimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=zDy_nQCXiIj. 1

Nishant Subramani, Samuel Bowman, and Kyunghyun Cho. Can unconditional language models recover arbitrary sentences? *Advances in Neural Information Processing Systems*, 32, 2019. 3, 9

Nishant Subramani, Nivedita Suresh, and Matthew Peters. Extracting latent steering vectors from pretrained language models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 566–581, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.48. URL https://aclanthology.org/2022.findings-acl.48/. 1, 3, 9, 26

Nishant Subramani, Jason Eisner, Justin Svegliato, Benjamin Van Durme, Yu Su, and Sam Thomson. MICE for CATs: Model-internal confidence estimation for calibrating agents with tools. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 12362–12375, Albuquerque, New Mexico, April 2025a. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.615. URL https://aclanthology.org/2025.naacl-long.615/. 5

Nishant Subramani, Alfredo Gomez, and Mona T. Diab. SimBA: Simplifying benchmark analysis using performance matrices alone. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13220–13233, Suzhou, China, November 2025b. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.711. URL https://aclanthology.org/2025.findings-emnlp.711/. 5

Nishant Subramani, Palash Goyal, Yiwen Song, Mani Malek, Yuan Xue, Tomas Pfister, and Hamid Palangi. The ACUTE protocol: Operationalizing language model activations for better calibration, utility, and trust. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=67lpPXbYf6. 5, 26

Hao Sun, Huailiang Peng, Qiong Dai, Xu Bai, and Yanan Cao. Layernavigator: Finding promising intervention layers for efficient activation steering in large language models. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/9280bfb61f159c768dfdeb729a41ccd0-Abstract-Conference.html. 26

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2025. URL https://arxiv.org/abs/2405.09818. 2, 3, 9

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL https://aclanthology.org/P19-1452/. 26

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations (ICLR)*. OpenReview.net, 2024. URL https://openreview.net/forum?id=AwyxtyMwaG. 1, 9

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 9

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248. 1, 3, 4, 9, 19

John G Wagner and James W Ayres. Bioavailability assessment: methods to estimate total area (auc 0-∞) and total amount excreted (ae$_\infty$) and importance of blood and urine sampling scheme with application to digoxin. *Journal of pharmacokinetics and Biopharmaceutics*, 5(5):533–557, 1977. 5

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024a. URL https://arxiv.org/abs/2409.12191. 2

Qingsong Wang, Mikhail Belkin, and Yusu Wang. General and efficient steering of diffusion models, 2026. URL https://arxiv.org/abs/2602.11395. 9

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need, 2024b. URL https://arxiv.org/abs/2409.18869. 2, 3, 9

Yu Wang and Sharon Li. Do unified multimodal models think in one space? a lens through cross-branch steering, 2026. URL https://arxiv.org/abs/2607.26411. 2, 9

Junfeng Wu, Yi Jiang, Chuofan Ma, Yuliang Liu, Hengshuang Zhao, Zehuan Yuan, Song Bai, and Xiang Bai. Liquid: Language models are scalable and unified multi-modal generators. *International Journal of Computer Vision*, 134(1), January 2026. ISSN 1573-1405. doi: 10.1007/s11263-025-02639-5. URL http://dx.doi.org/10.1007/s11263-025-02639-5. 2, 3

Zhaofeng Wu, Xinyan Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 53705–53723, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/84fb08ae30ab266953d1ff72eb0611b1-Paper-Conference.pdf. 9

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? even simple baselines outperform sparse autoencoders. *arXiv preprint arXiv:2501.17148*, 2025b. 5, 6, 29

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *The Thirteenth International Conference on Learning Representations (ICLR)*. OpenReview.net, 2025. URL https://openreview.net/forum?id=o6Ynz6OIQ6. 2

Zejia You, Chunyuan Deng, and Hanjie Chen. Spherical steering: Geometry-aware activation rotation for language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=9PPeJsVLdB. 9

Yuhui Zhang, Elaine Sui, and Serena Yeung. Connect, collapse, corrupt: Learning cross-modal tasks with uni-modal data. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 51711–51736, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/e26f31de8b13ec569bf507e6ae2cd952-Paper-Conference.pdf. 9

Shanshan Zhao, Xinjie Zhang, Jintao Guo, Jiakui Hu, Lunhao Duan, Minghao Fu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. Unified multimodal understanding and generation models: Advances, challenges, and opportunities, 2026. URL https://arxiv.org/abs/2505.02567. 3

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *The Thirteenth International Conference on Learning Representations (ICLR)*. OpenReview.net, 2025. URL https://openreview.net/forum?id=SI2hI0frk6. 2

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, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency, 2025. URL https://arxiv.org/abs/2310.01405. 4, 9, 19, 26

Appendix

A Experimental Details **19**

B Detailed results **32**

C Details for: geometry of steering vectors **38**

D Concepts unique to one modality **40**

E Averaging text and image steering vectors **46**

A Experimental Details

Concepts. Our concepts are chosen from those commonly studied in either the language or vision community. Emotion (sad/happy) is commonly steered in language models (Turner et al., 2024; Zou et al., 2025), and age (young/old) and expression are canonical targets of latent-direction editing in image generation (Shen et al., 2020; Gandikota et al., 2024; Baumann et al., 2025). Cleanness and chaos are semantic concepts that describe the state of a whole scene. In text a single word expresses them, while in an image they are spread over many objects, which makes them an interesting test of whether one direction can capture both. The visual concepts describe basic properties of objects in a scene, such as their size, depth, spatial position or color. We expect spatial (left/right) to be the hardest concept to steer, since spatial relations are a well-documented failure mode of image generation models and VLMs (Chatterjee et al., 2024; Kamath et al., 2023).

A.1 Minimal Pairs

Each steering vector is the mean difference in hidden states over a concept's minimal pairs, which are either all text or all images. We describe how both kinds of pair are constructed below.

Text pairs. We use Gemini 2.5 Flash to generate $N = 100$ distinct subjects that fit the concept and fill each into the template "a {*subject*} is {*pole*}", once with each pole. For people in the color concept, the template is "a {*subject*} is dressed in {*pole*}". Each text sample is thus an unpunctuated short declarative sentence ending in the pole itself, so the two sentences of a text minimal pair differ only there, for instance *a chef is happy* and *a chef is sad*. Table 2 gives more examples.

Table 2: Example text minimal pairs. Each subject gives one pair that differs only in the pole word.

Concept	Example subjects	− pole	+ pole
emotion	teacher, chef, coach	*a teacher is sad*	*a teacher is happy*
age	engineer, cook, merchant	*an engineer is young*	*an engineer is old*
cleanness	oven, fork, glove	*an oven is clean*	*an oven is dirty*
chaos	street, hospital, party	*a street is peaceful*	*a street is chaotic*
size	flower, pen, statue	*a flower is small*	*a flower is large*
depth	flower, pen, statue	*a flower is far away*	*a flower is nearby*
spatial	flower, pen, statue	*a flower is on the left*	*a flower is on the right*
color	hoodie, bus, chair	*a hoodie is blue*	*a hoodie is red*

Image pairs. An image pair consists of two images of the same scene that differ only in the pole. We start from 100 short image descriptions per concept, written by Gemini 2.5 Flash. Each names a subject and, if needed, what it is doing and where. We generate a baseline image from each description and edit it toward each pole. For most concepts, both steps use Gemini 2.5 Flash Image. However, depth, size and spatial edits are harder and Gemini 2.5 Flash Image often gets them wrong, so for these we use Gemini 3 Pro Image and Reve. Figure 6 shows one image pair per concept with its description. The edit instructions are listed below for each concept and pole.

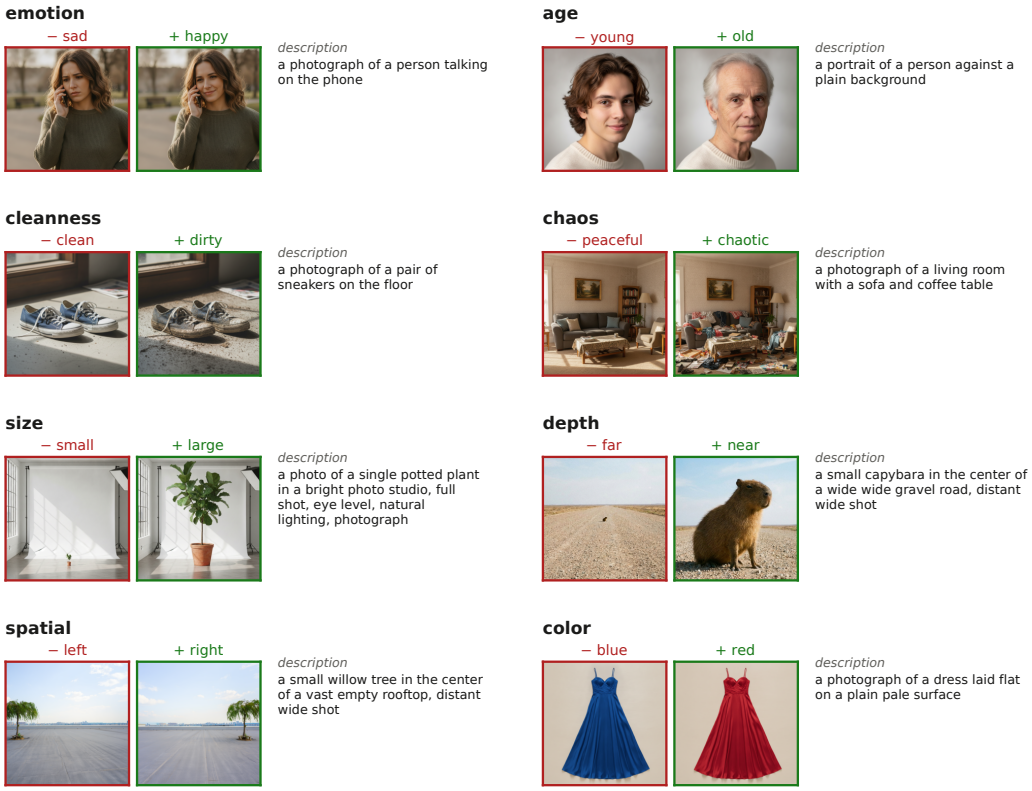


Figure 6: One image pair per concept, with the description used to generate its baseline image.

Edit instruction: emotion

− **(sad):** Subtly modify this image to make the person look sad. Everything else, such as identity, pose, hair, clothing, background and lighting, should be preserved exactly as before. Please only change the facial expression.

\+ **(happy):** Subtly modify this image to make the person look happy. Everything else, such as identity, pose, hair, clothing, background and lighting, should be preserved exactly as before. Please only change the facial expression.

Edit instruction: age

− **(young):** Modify this image to make the person look much younger, a young child or young teenager, with smooth wrinkle-free skin, a fuller rounder face and bright clear eyes. Everything else, such as the person, pose, framing, clothing and background, should be preserved exactly as before. Please only change the apparent age.

\+ **(old):** Modify this image to make the person look much older, elderly, in their 70s or 80s, with deep wrinkles, grey or white thinning hair, sagging skin and age spots. Everything else, such as the person, pose, framing, clothing and background, should be preserved exactly as before. Please only change the apparent age.

Edit instruction: cleanness

− **(clean):** Make everything in this image look spotless, freshly cleaned, polished and pristine, with no dirt, dust or stains anywhere. Everything else, such as the objects, layout, framing, camera angle and composition, should be preserved exactly as before. Please only change how clean or dirty the surfaces look.

\+ **(dirty):** Make everything in this image look dirty, grimy, dusty and stained, covered in mud, dirt, grease and grime. Everything else, such as the objects, layout, framing, camera angle and composition, should be preserved exactly as before. Please only change how clean or dirty the surfaces look.

Edit instruction: chaos

− **(peaceful):** Modify this scene to look peaceful, calm and serene: orderly and tidy, open and uncluttered, still and quiet. Everything else, such as the scene, layout, viewpoint and subjects, should be preserved exactly as before. Please only change how tidy or chaotic the scene is.

\+ **(chaotic):** Modify this scene to look chaotic, turbulent and disordered: cluttered and messy, objects scattered, knocked over and in disarray, crowded and frantic. Everything else, such as the scene, layout, viewpoint and subjects, should be preserved exactly as before. Please only change how tidy or chaotic the scene is.

Edit instruction: size

− **(small):** Make the {*subject*} much smaller, tiny, only a few centimetres, taking up just a small part of the scene. Keep exactly one {*subject*}, do not add or duplicate it. Everything else, such as the spot, the room, the camera and the size and position of everything around it, should be preserved exactly as before. Please only change the size of the {*subject*}.

+ **(large):** Make the {*subject*} a giant {*subject*}, several times larger than normal, towering and reaching high up near the top of the frame. Keep exactly one {*subject*}, do not add or duplicate it. Everything else, such as the spot, the room, the camera and the size and position of everything around it, should be preserved exactly as before. Please only change the size of the {*subject*}.

Edit instruction: depth

− **(far):** Make the single main subject much smaller and more distant, so it occupies only a tiny part of the frame. Keep exactly one subject visible, do not add, remove, clone or duplicate it. Everything else, such as the subject, background scene, camera framing, lighting and style, should be preserved exactly as before. Please only change how close the subject is.

+ **(near):** Make the single main subject much larger so it fills most of the frame, like a close-up. Keep exactly one subject visible, do not add, remove, clone or duplicate it. Everything else, such as the subject, background scene, camera framing, lighting and style, should be preserved exactly as before. Please only change how close the subject is.

Edit instruction: spatial

− **(left):** Move the single main subject to the far left side of the frame. Keep exactly one subject visible, do not add, remove, clone or duplicate it. Everything else, such as the subject, background scene, camera framing, lighting and style, should be preserved exactly as before. Please only change where the subject is in the frame, left or right.

+ **(right):** Move the single main subject to the far right side of the frame. Keep exactly one subject visible, do not add, remove, clone or duplicate it. Everything else, such as the subject, background scene, camera framing, lighting and style, should be preserved exactly as before. Please only change where the subject is in the frame, left or right.

Edit instruction: color

```
− (blue): Modify this image so the subject is clearly blue, its
main surfaces repainted, re-dyed or remade in a saturated blue.
Everything else, such as the subject's shape, material, placement and
scale, the background, framing, camera and lighting, should be
preserved exactly as before. The rest of the scene keeps its
original colours and brightness. Please only change the colour of
the subject.

+ (red): Modify this image so the subject is clearly red, its main
surfaces repainted, re-dyed or remade in a saturated red. Everything
else, such as the subject's shape, material, placement and scale, the
background, framing, camera and lighting, should be preserved exactly
as before. The rest of the scene keeps its original colours and
brightness. Please only change the colour of the subject.
```

A.2 Evaluation Prompts

Every concept gets its own set of 20 evaluation prompts, with subjects and scenes that make sense for it. We reuse them across all layer configurations and steering strengths.

Each evaluation prompt is a description of one to three complete sentences that names a subject and describes the scene around it. For image generation, we use the full description ({*image description*}) to give the model more to condition on and to keep more of the scene fixed. For text generation, we keep only its main subject ({*subject*}) and place it in an open instruction, leaving the model free to describe the scene in as much visual detail as an image would show (Appendix A.3). Table 3 shows a few examples. Subjects are not limited to people, but include other subjects where the concept can show, such as toys and figurines with a face for emotion, or a bike for age.

Table 3: Example evaluation prompts. Image generation uses the full description, and text generation uses only its subject, shown in bold.

Concept	Image description	Text subject
emotion	A head-and-shoulders portrait of **a person with short cropped hair**, wearing a plain crew-neck shirt and holding a relaxed, neutral expression. They stand in front of a weathered red brick wall in soft natural daylight. It is shot as a sharp, realistic photograph with an even, unremarkable mood.	*a person with short cropped hair*
cleanness	A photograph of **a modern kitchen with wooden cabinets**, stone countertops, and a stainless steel sink. Even natural daylight from a window falls across the counters, backsplash, and floor. It is a plain, realistic interior photograph with no people.	*a modern kitchen with wooden cabinets*
size	A single **traffic cone** in the center of a wide empty asphalt parking lot with faded painted markings, under flat even daylight.	*a traffic cone*
color	A photograph of **a car parked** alone on a wide empty asphalt lot, the car centred and seen from the side, under soft even daylight and a pale overcast sky. A plain, realistic photograph with no people.	*a parked car*

We generate the evaluation prompts with Gemini 2.5 Flash, using the prompt below. It asks for scenes that are neutral with respect to the concept, so that the unsteered outputs are not biased toward either pole.

Evaluation prompt generation

```
Write 20 short descriptions of photographs.  Each description is one
to three sentences about one main subject, saying what it is doing
and where, when.  Be sure to keep all subjects distinct.

The descriptions will be used to test the concept {concept}, whose
two poles are {positive pole} and {negative pole}.  Every description
must be neutral on this concept:  do not describe the subject or the
scene as either pole, and avoid any word that leans toward one of
them.

Respond as JSON only, giving the subject phrase of each description
as its own field:
{
  "concept":  "{concept}",
  "prompts":  [
    {
      "subject":  "<subject>",
      "description":  "<one to three sentences>"
    },
    ...
  ]
}
```

A.3 Generation and Decoding

Text generation. All judged text outputs are generated with the same instruction, shown below in each model's own chat format. We iterated on this instruction, since description-style prompts made the model focus on a single aspect of the scene, or refuse or output JSON, often leaving little room for steering. The final instruction lets the model freely mention many concepts and visual associations, so that a steered concept can surface in whatever the model chooses to mention, while the shared subject keeps the two modalities comparable. Only the subject phrase, taken from Appendix A.2, changes from one prompt to the next. The instruction states that no image is provided, since a unified model may otherwise expect one. We steer the subject phrase and every generated token.

Text prompt: UniAR

<|im_start|>system
You are a helpful assistant.<|im_end|>

<|im_start|>user
No image is provided; write from imagination. Answer in two or three plain sentences, no title, no list. Describe your day in the city. Your description must include {subject}.**<|im_end|>**

<|im_start|>assistant
[generated tokens]

steered tokens **bold** chat template

Text prompt: Emu3.5

<|extra_203|>**You are a helpful assistant.**
USER: No image is provided; write from imagination. Answer in two or three plain sentences, no title, no list. Describe your day in the city. Your description must include {*subject*}.
ASSISTANT: <|extra_100|>[generated tokens]

steered tokens **bold** chat template

Text prompt: Liquid

<start_of_turn>user
No image is provided; write from imagination. Answer in two or three plain sentences, no title, no list. Describe your day in the city. Your description must include {*subject*}.**<end_of_turn>**
<start_of_turn>model
[generated tokens]

steered tokens **bold** chat template

Image generation. Each full image description (Appendix A.2) is given to the model in its own image-generation format, shown below. The steered tokens are only the generated visual tokens, not the prompt.

Image prompt: UniAR

<|im_start|>**user**
{*image description*}**<image_gen> generate image 16 16<|im_end|>**
<|im_start|>**assistant**
<|vision_start|>[generated visual tokens]

steered tokens **bold** chat template

Image prompt: Emu3.5

<|extra_203|>**You are a helpful assistant for t2i task.**
USER: {*image description*}
ASSISTANT: <|extra_100|>[generated visual tokens]

steered tokens **bold** chat template

Image prompt: Liquid

{*image description*} **Generate an image based on this description.**
<boi>[generated visual tokens]

steered tokens **bold** chat template

The decoding settings of every sweep are listed in Table 4.

Table 4: The strength grid is set per model and setting (Appendix A.4). We keep each model's default decoding settings other than the number of decoder steps and the image resolution for practical reasons, and fix the seed at 42.

Model	Setting	α grid	Decoding
UniAR	txt→txt, img→txt	$\pm$100, step 20	greedy, 96 tokens
	img→img	$\pm$10, step 2	16×16 tokens, guidance 1.5, temperature 1.0, SD3 decoder with 16 steps and guidance 1.5, 544×544 image
	txt→img	$\pm$15, step 3	
Emu3.5	txt→txt, img→txt	$\pm$25, step 5	greedy, 96 tokens
	img→img	$\pm$1.5, step 0.3	32 × 32 tokens, guidance 2.0, top-k 10,240, top-p 1.0, temperature 1.0, VQ decoder, 512 × 512 image
	txt→img	$\pm$5, step 1	
Liquid	txt→txt, img→txt	$\pm$25, step 5	greedy, 96 tokens
	img→img	$\pm$0.5, step 0.1	32 × 32 tokens, guidance 7.0, top-k 4,096, top-p 0.96, temperature 0.99, VQ decoder, 512 × 512 image
	txt→img	$\pm$1.5, step 0.3	

A.4 Layer and Steering Strength

Layer configurations. Traditionally, steering operates on a single layer, at one or all timesteps depending on the task (Subramani et al., 2022), but more recent work has found success steering multiple or even all layers (Zou et al., 2025; Gaintseva et al., 2026; Sun et al., 2025). Different layers encode different linguistic properties and later layers tend to encode more complex features in neural network models (Ethayarajh, 2019; Tenney et al., 2019; Rogers et al., 2020; Skean et al., 2025; Li & Subramani, 2026). We therefore compare early, middle and late layers without choosing one over the others.[7] We steer seven layer configurations per model: three single layers, three contiguous layer ranges and all layers, listed in Table 5. Each layer range spans four consecutive layers around its matching single layer, or eight in Emu3.5 since it has about twice as many layers (64, against 36 and 28), so that a layer range covers a similar share of the depth in every model. A transferred vector is injected at the same layers it was extracted from.

Table 5: Layer configurations for each model (0-indexed). Single layers are evenly spaced in depth, each contiguous layer range spans the layers around its matching single layer, and the seventh configuration steers all layers.

		Single layer			Contiguous layer range		
Model	Depth	early	middle	late	early	middle	late
UniAR	36	9	18	27	7–10	16–19	25–28
Emu3.5	64	16	32	48	12–19	28–35	44–51
Liquid	28	7	14	21	5–8	12–15	19–22

Steering strength α. Each model and setting has its own grid of steering strengths α, with five evenly spaced values on each side of the unsteered $\alpha = 0$ (Table 4). To find the range, we first sweep a fine grid of α on a few evaluation prompts at the middle layer range. We then pick the range so that, across all concepts, the one or two largest $|\alpha|$ on each side produce degenerate samples, such as incoherent text or broken images. This grid is used for all seven layer configurations.

A.5 Judge Setup and Prompts

Judge. The judge is Gemini 3.7 Flash with a low thinking level and one sample per comparison, and a JSON response schema restricts its verdict to `A`, `B` or `Tie`. It never sees any steering hyper-

[7] In interpretability, splitting layers into early, middle, and late is common practice (Subramani et al., 2026).

parameter, such as α, the layer configuration or the steering vector. We validate it against human reviewers in Appendix A.8.

Prompt. The prompts are shown below, where {*pole*} is the asked pole, {*anchor*} the opposite pole and {*answer*} the rater's answer. We prompt the judge few-shot with ten example pairs for each modality. They are held-out pairs on which the reviewers agreed, and their answers are spread evenly across the steered output, the baseline and a tie. The question names one of the two poles at random, so the judge cannot score well just by recognizing steering artifacts.

Concept rate judge: image pair

```
You are a careful annotator comparing two model outputs.

Here are examples of how this rater judges such pairs.  The image
shows numbered rows; in each row the left panel is A and the right
panel is B:
Row 1:  Which represents more of {pole} (or less of {anchor})?
Answer:  {answer}
...
Row 10:  Which represents more of {pole} (or less of {anchor})?
Answer:  {answer}
[example grid image, rows 1A|1B to 10A|10B]

Now judge this new pair the same way:
A: [image A]
B: [image B]

Which represents more of {pole} (or less of {anchor})?
Choose A or B only if one clearly and obviously shows more of the
attribute at a glance.  If the difference is small, faint, or you
have to look hard, answer Tie.
Judge only the attribute - ignore style, grammar, and quality.  A
candidate that looks odd, broken, or repetitive can still clearly
show the attribute (e.g.  it repeats the trait), and is not
automatically a Tie; decide only by whether the attribute itself is
clearly stronger in one.
Answer with one word (A, B, or Tie).
```

Concept rate judge: text pair

```
You are a careful annotator comparing two model outputs.

Here are examples of how this rater judges such pairs:
[10 examples, in the format below]

Now judge this pair the same way:

Candidate A:
[text A]

Candidate B:
[text B]

Which represents more of {pole} (or less of {anchor})?
Choose A or B only if one clearly and obviously shows more of the
attribute at a glance.  If the difference is small, faint, or you
have to look hard, answer Tie.
Judge only the attribute - ignore style, grammar, and quality.  A
candidate that looks odd, broken, or repetitive can still clearly
show the attribute (e.g.  it repeats the trait), and is not
automatically a Tie; decide only by whether the attribute itself is
clearly stronger in one.
Answer with one word (A, B, or Tie).
```

Text example format

```
Example:
Candidate A:
[example text A]

Candidate B:
[example text B]

Which text represents more of {pole} (or less of {anchor})?
Choose A or B, or Tie if it's not obvious.
Answer with one word (A, B, or Tie).
Answer:  {answer}
```

A.6 DREAMSIM POLE SCORE (IMAGES)

As a complement to the concept rate metric for images, we compare each generated image with two references, a positive-pole and a negative-pole edit of the same scene. With d the cosine distance in DreamSim, a learned perceptual image similarity (Fu et al., 2023), the pole score of an image I with references R^+, R^- is

$$\text{pole}(I) \;=\; 100 \cdot \frac{d(I,\, R^-)}{d(I,\, R^+) \,+\, d(I,\, R^-)},$$

so 100 means I matches the positive-pole reference, 0 the negative-pole reference, and 50 is equidistant. The references for a prompt are the model's own unsteered image, minimally edited toward each pole with Gemini 3 Pro Image, following the same method as the minimal pairs.

Figure 7 plots the pole score against normalized strength for the two image-output settings. It rises with α for the concepts that steer and stays flat at 50 for spatial. The same rise appears when a text-derived vector steers image generation.

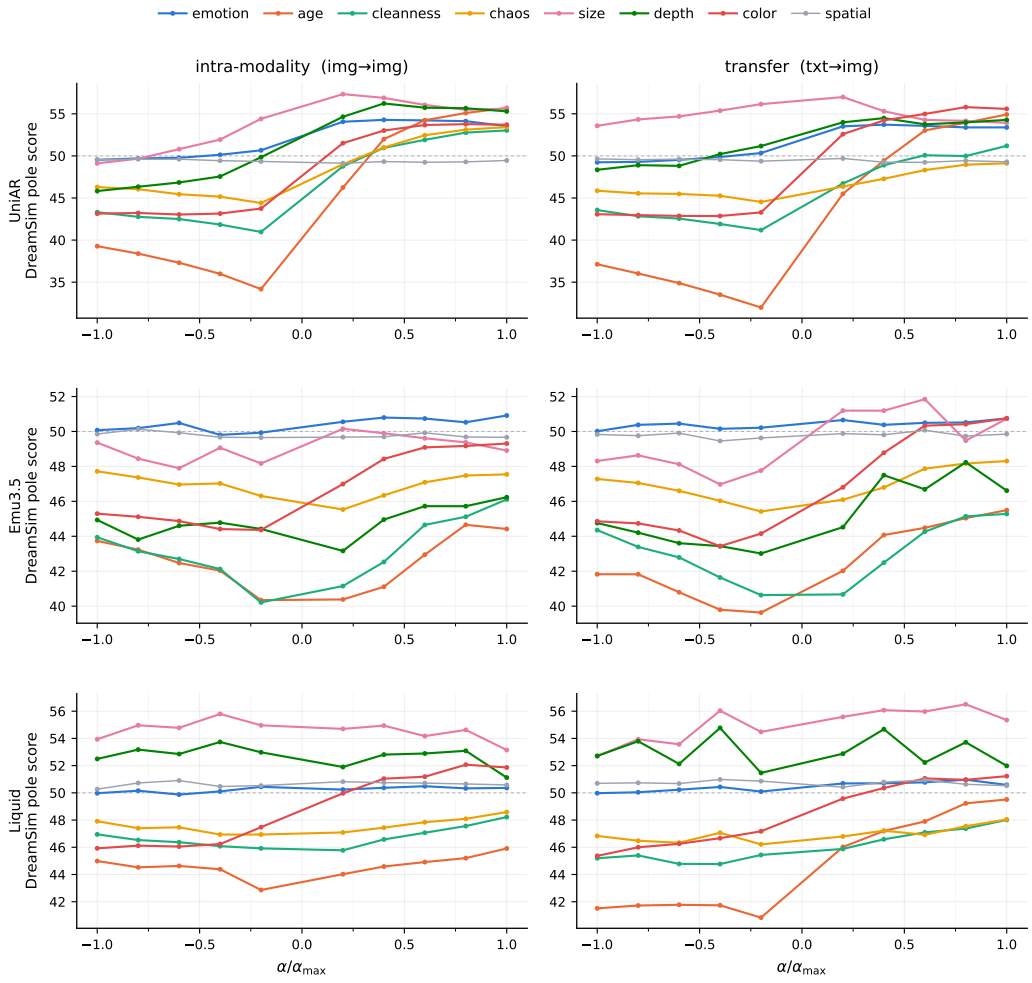


Figure 7: DreamSim pole score for image-derived (left) and text-derived (right) vectors.

UniAR moves the furthest, with age strongest and every concept except spatial rising, apart from size under the text-derived vector. In Emu3.5 and Liquid the shifts are only a few points: color moves most with both vectors, and age moves mainly with the text-derived vector. In Liquid, size and depth sit above 50 at every strength without moving, so their offset is not steering.

A.7 Fluency and Coherence Judges

Text. We use an AxBench-style pointwise fluency judge (Wu et al., 2025b). Gemini 3.7 Flash scores each output 0, 1 or 2 for fluency and sees only the output, never the prompt, the strength or the layer. We score all 22,400 steered text outputs per model plus one unsteered baseline per concept and prompt. The prompt is shown below.

Text fluency

```
You are evaluating the linguistic fluency of a language model output.

Model output:
{text}

Rate the fluency and coherence of the output.

Use the following scale:

0 = The output is substantially incoherent, ungrammatical, or
    otherwise difficult to understand.
1 = The output is understandable but contains noticeable issues with
    grammar, coherence, or naturalness.
2 = The output is fluent, coherent, grammatical, and natural.

Focus only on linguistic quality. Do not judge whether the output
    follows the original instruction or expresses the target concept.

Return JSON only:
{"score": 0 | 1 | 2, "reasoning": "<one concise sentence>"}
```

Images. For images we build the counterpart of the fluency judge: the same model sees only the image, with no prompt, concept or strength, and returns one of three verdicts, *intact*, *partly broken* or *broken*, which we map to 2/1/0. We use verdict words rather than a numeric scale, which returned 1 for nearly every image. We score all 3,200 steered images per model at the middle layer range and the 160 unsteered baselines. The prompt is shown below.

Image coherence

```
You are checking an image generated by a model for visible breakage.

Answer with one of three verdicts:

"intact": the image is coherent and well-formed. This is the
    expected verdict for a typical generated image; the usual look
    of AI-generated images (smoothness, softness, a synthetic feel)
    is not breakage.
"partly_broken": the content is recognizable, but there are clear
    artifacts, distortions, or incoherent elements, such as
    malformed hands or faces, merged or duplicated objects, garbled
    text, or corrupted regions.
"broken": severe artifacts or garbled textures dominate, or the
    content cannot be recognized.

Only visible breakage matters. Do not judge whether the image
    matches any description or expresses any concept, and do not
    judge style, aesthetics, or realism.

Return JSON only:
{"reasoning": "<one concise sentence naming any breakage you see, or
    stating that there is none>", "verdict": "intact" |
    "partly_broken" | "broken"}
```

Results. Figure 8 plots both quality scores next to the concept rate against normalized strength, each relative to the model's unsteered outputs. Text fluency holds up to about 0.4 of the range, where the concept rate peaks, and collapses beyond it. Image coherence declines gently over the

whole range while the concept rate keeps rising. Transferred and intra-modality vectors degrade quality similarly, except that image-derived vectors cost text fluency more than text-derived ones in Emu3.5 and Liquid at middle strengths.

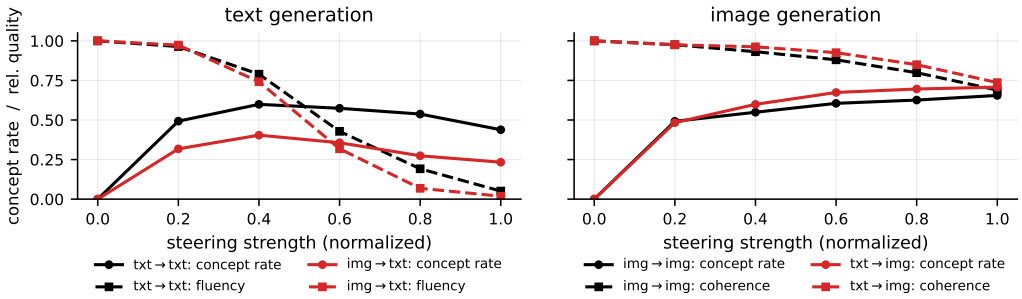


Figure 8: Concept rate (solid) and output quality (dashed) against normalized strength at the middle layer range, averaged over models, concepts, prompts and both signs of α. Quality is text fluency (left) and image coherence (right), relative to the unsteered outputs.

A.8 Human Validation

To check that the automated concept-rate judge tracks human judgment, we collect human annotations on 320 held-out pairs, 10 per concept and setting across eight concepts, none of which is part of the evaluated sweeps. Each pair receives three independent reviews. Reviewers face the same forced-choice task as the judge: the steered output and its $\alpha = 0$ baseline shown as A and B in *randomized order*, with the asked pole also *randomized* (≈50% the steered pole, ≈50% the opposite), and are blind to α, the layer, and which output is steered. Figure 9 shows the annotation interface.

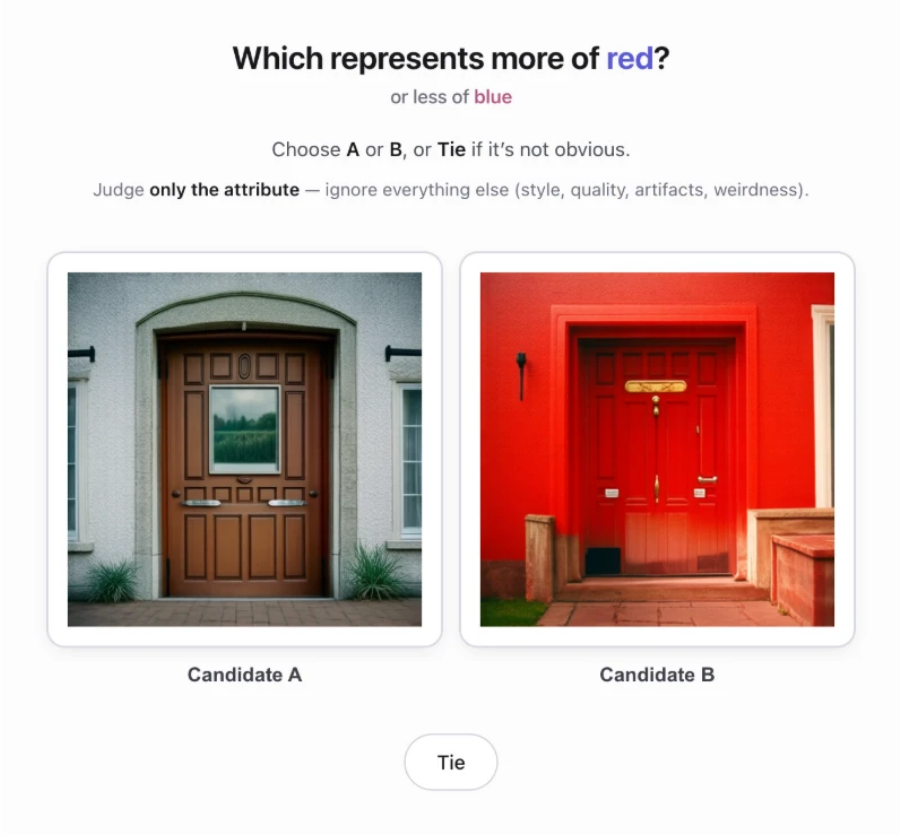


Figure 9: The annotation interface shown to human reviewers. It names the asked pole and its opposite. Two model outputs, the unsteered baseline and the steered output, are shown as Candidate A and B in random order. The human reviewer answers A, B or Tie.

Since the A/B order and the asked pole are random, we recode each answer as agreeing with the steering direction, going against it, or a tie, and combine the three reviews of a pair by majority vote, breaking three-way splits at random. Cohen's κ between the judge and this majority label is 0.62 over the 320 pairs.

B Detailed results

This section shows the concept-rate curves of every concept and setting for all three models. Figures 10 to 12 compare intra-modality and transferred vectors at the middle layer range, in the layout of Figure 2, with UniAR repeated for comparison. Figures 13 to 15 show all seven layer configurations. Additionally, Table 6 gives the cross-modal recovery of every concept, model and layer range.

Table 6: **Cross-modal recovery**: How much of intra-modality steering is *recovered* by the transferred vector, defined as ratio of transfer AUC to the intra-modality AUC of concept-rate curve (see Figure 2 and §3.1): 1.0 means the transferred vector steers as well as intra-modality, 0 no transfer; values above 1.0 mean transferred vector steers better than intra-modality. Cells are blank where intra-modality steering itself barely works (intra-modality AUC < 0.2), so the ratio is not meaningful. Gray marks weak transfer (recovery < 0.5); bold marks the mean of the best layer range, per model and setting.

Model	Layers	Direction	emotion	age	cleanness	chaos	size	depth	spatial	color	mean
UniAR	early	txt→img	0.98	1.26	0.90	0.81	0.81	0.74	0.09	1.00	0.82
		img→txt	1.14	0.91	1.01	1.02	0.25	1.19	–	1.21	**0.96**
	middle	txt→img	1.05	1.29	0.94	0.77	0.73	0.85	0.07	1.00	**0.84**
		img→txt	0.99	0.74	0.92	0.90	0.18	0.96	–	0.47	0.74
	late	txt→img	0.92	1.22	0.78	0.46	0.61	0.62	–	1.01	0.80
		img→txt	0.87	0.91	1.13	0.97	0.22	0.98	–	0.71	0.83
Emu3.5	early	txt→img	1.05	1.59	1.01	0.93	1.25	1.24	–	0.99	1.15
		img→txt	0.59	0.40	0.41	0.49	0.44	0.79	0.00	0.57	0.46
	middle	txt→img	1.18	1.64	0.89	–	1.44	1.35	–	0.87	**1.23**
		img→txt	1.02	0.47	0.33	0.50	0.50	0.56	–	0.76	0.59
	late	txt→img	0.79	1.02	0.60	1.13	1.06	1.09	–	0.97	0.95
		img→txt	1.59	0.52	0.38	0.61	0.34	0.84	–	1.81	**0.87**
Liquid	early	txt→img	0.76	1.87	0.81	1.17	1.23	0.85	–	0.77	1.07
		img→txt	0.48	0.13	0.45	0.45	0.35	0.38	–	–	0.37
	middle	txt→img	1.03	1.79	1.40	1.73	1.34	0.79	–	0.96	**1.29**
		img→txt	0.54	0.21	0.44	0.57	0.36	0.59	0.02	–	0.39
	late	txt→img	0.83	1.39	1.42	1.47	1.59	0.85	–	1.17	1.24
		img→txt	1.35	1.08	1.11	0.78	0.27	1.12	–	0.62	**0.91**
mean	all	txt→img	0.95	1.45	0.97	1.06	1.12	0.93	0.08	0.97	1.04
		img→txt	0.95	0.60	0.69	0.70	0.32	0.82	0.01	0.88	0.68

Every sweep follows the same recipe, described in Appendix A. For each model, we generate one output for every combination of concept, setting, layer configuration, strength and prompt, and judge it against its unsteered baseline with the concept-rate judge.

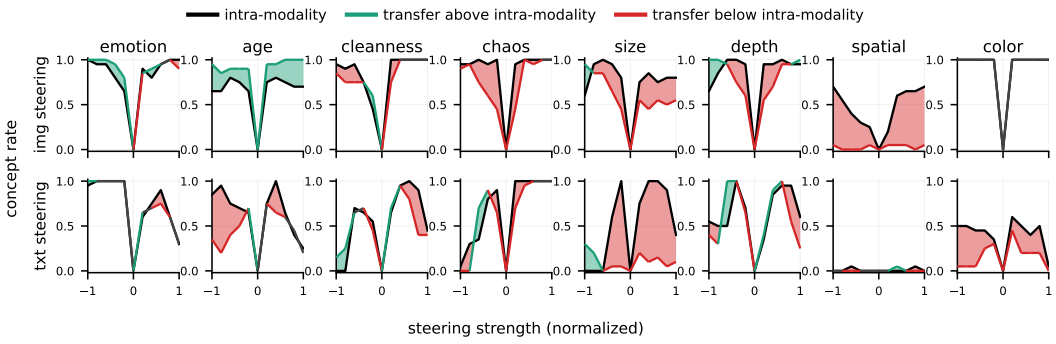


Figure 10: UniAR at the middle layer range, same as Figure 2, repeated for comparison. The black curve is the intra-modality vector and the colored curve the transferred one. The shaded area between them is their difference in AUC, green where the transferred vector steers more and red where it steers less.

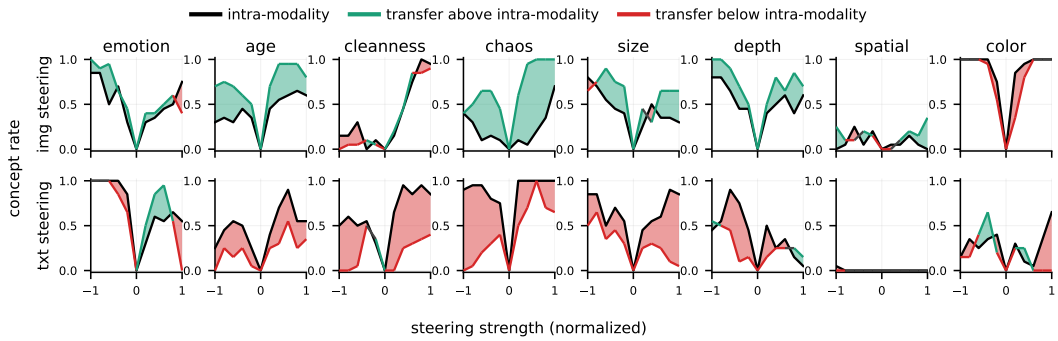


Figure 11: Emu3.5 at the middle layer range, same layout as Figure 2.

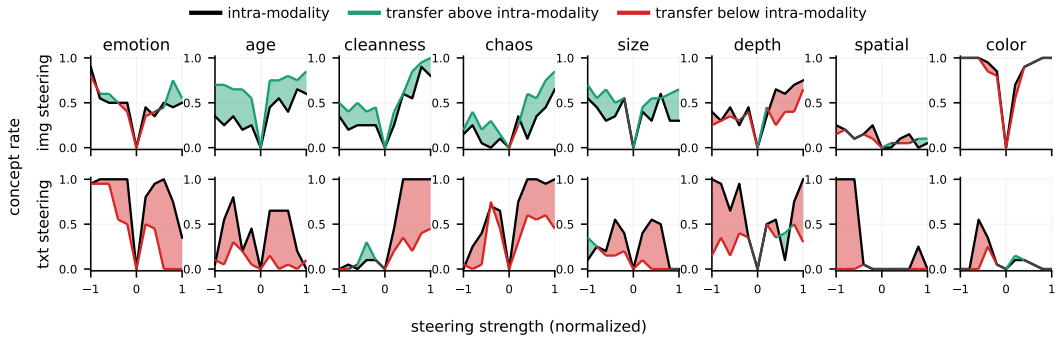


Figure 12: Liquid at the middle layer range, same layout as Figure 2.

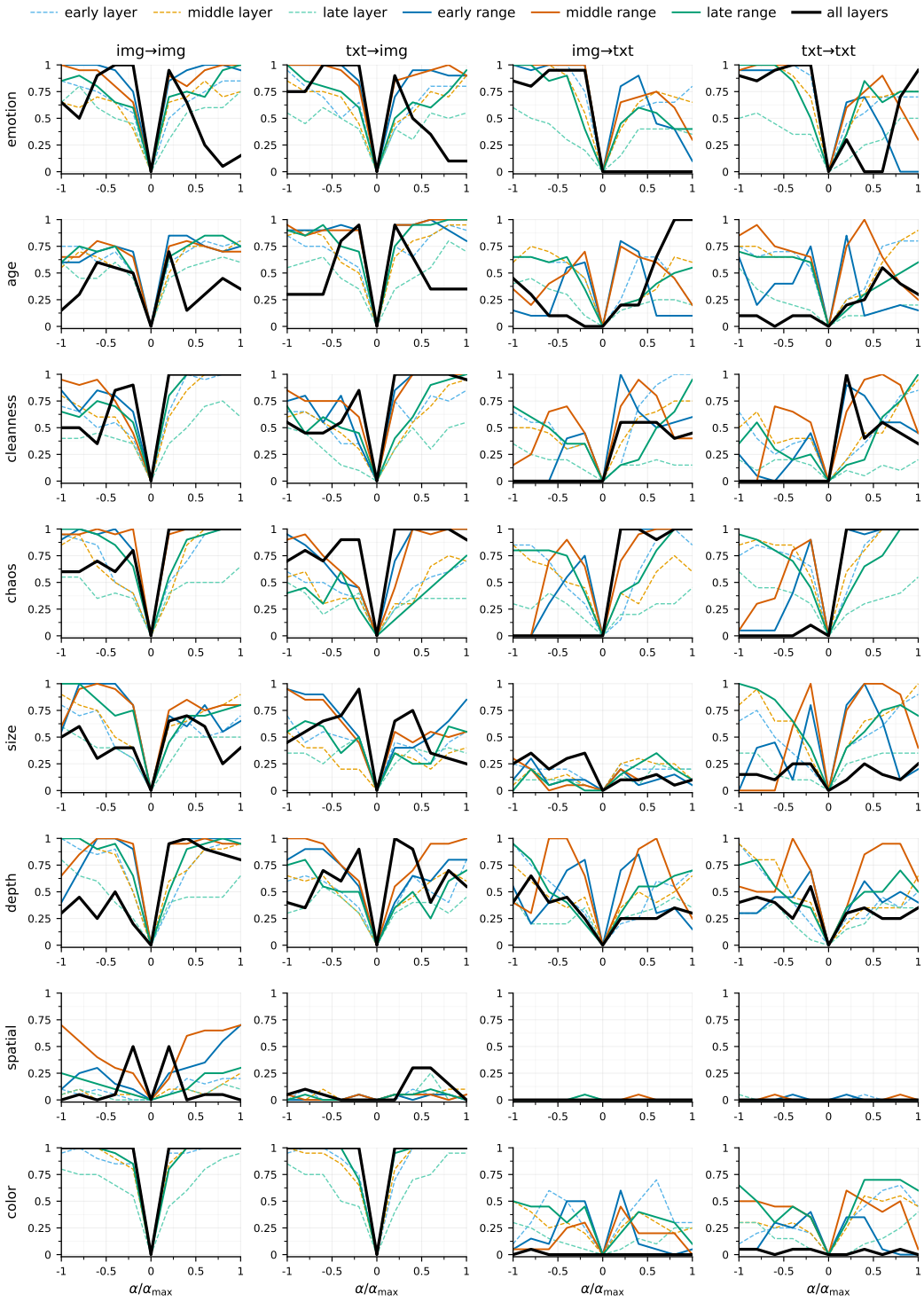


Figure 13: UniAR, every concept (rows) and setting (columns), one line per layer configuration. Single layers are light and dashed, layer ranges solid, one hue per depth (blue early, orange middle, green late), and all layers is the thick black line. Every curve starts at rate 0 at $\alpha = 0$.

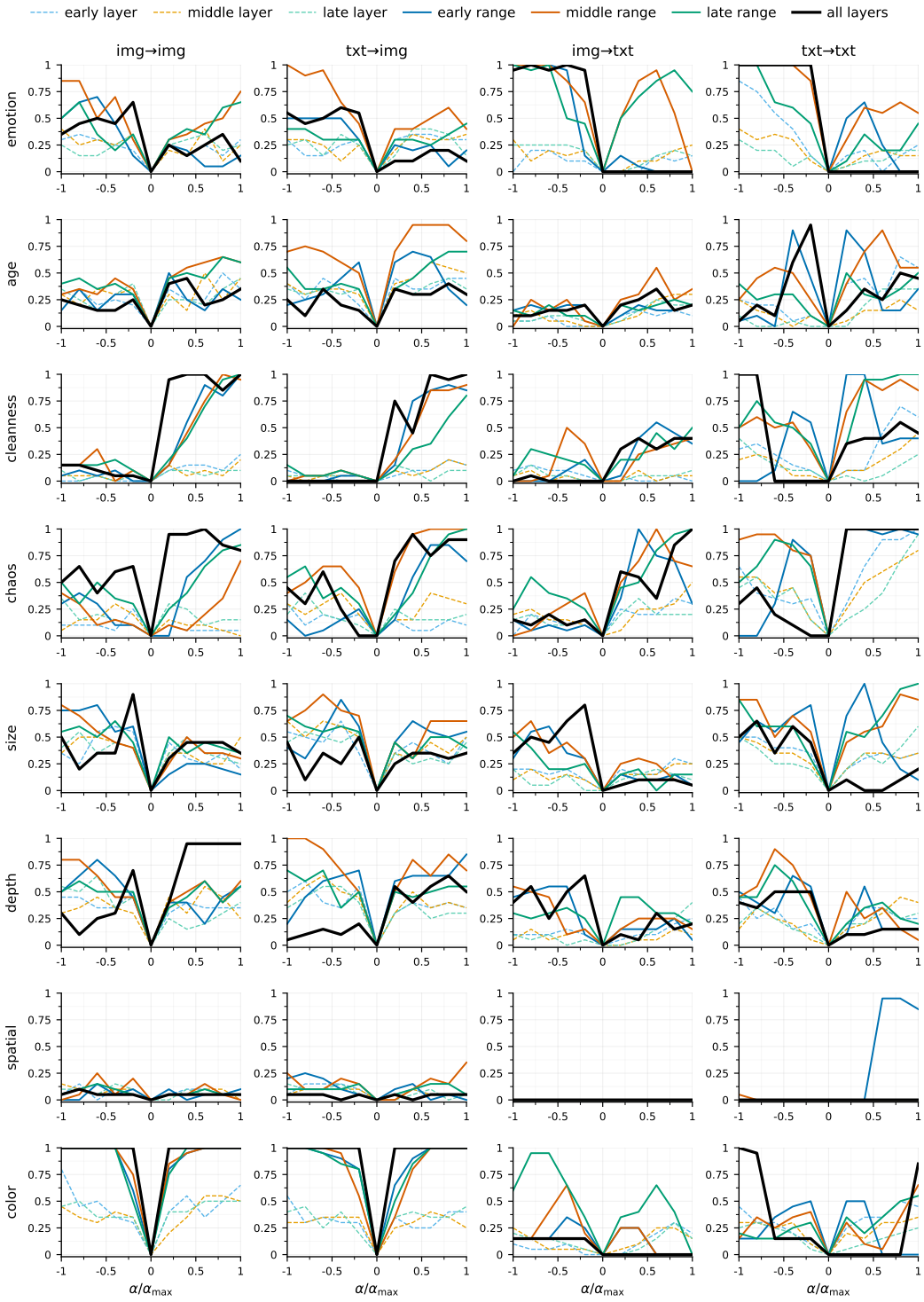


Figure 14: Same as Figure 13 for Emu3.5.

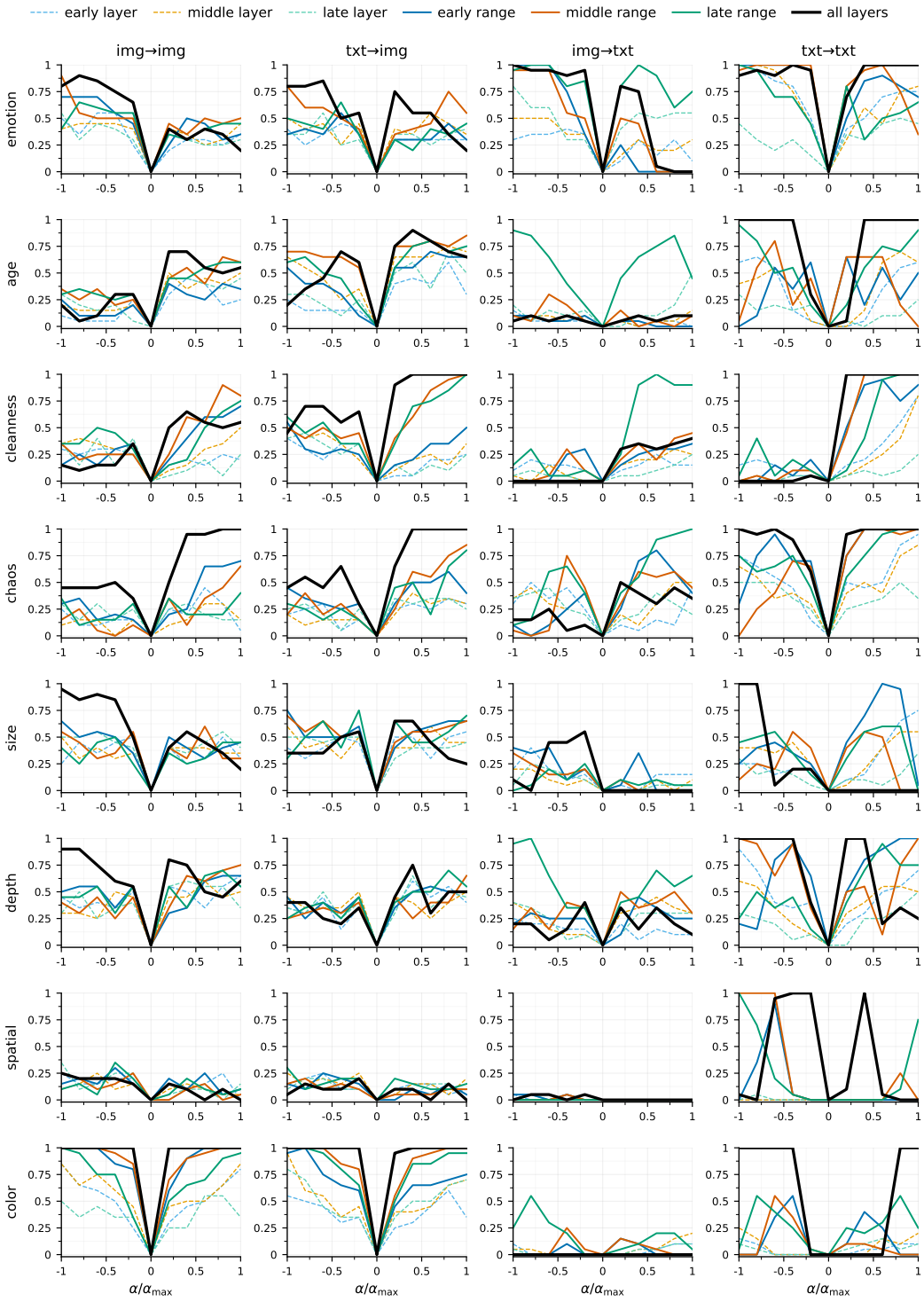


Figure 15: Same as Figure 13 for Liquid.

C Details for: geometry of steering vectors

Here we provide details behind §3.4 with full per-model results. Table 7 shows the correlation between cosine similarity and steering success for every model and transfer direction (either with recovery score or AUC). Table 8 gives the anisotropy numbers for all three models, and Table 9 the per-concept comparison of "same pole" (happy sentences, happy images) versus "same direction" (image steering vector for happy, text steering vector for happy). Figure 16 shows the per-layer cosine between the text and image steering vectors of each concept.

Cosine between steering vectors. We compare the same per-layer unit vectors that are injected during steering. A layer range is summarized by the mean of its per-layer cosines. As a baseline, the cosine between two random directions is zero on average, with a standard deviation of $1/\sqrt{d}$, which is 0.016 for UniAR, 0.014 for Emu3.5 and 0.018 for Liquid. We call a cosine above the baseline if it exceeds two standard deviations, $2/\sqrt{d}$.

Activations and centering. For the modality-gap analysis, each sentence and each image is represented by one vector per layer, pooled the same way the steering vectors are built (last token for sentences, mean over image tokens for images), and a concept pole by the mean over its 100 examples. Before comparing across modalities we subtract the midpoint between the image mean and the text mean, so that neither modality's sample size affects the center. We rescale every activation to unit length before centering, because text activations are longer than image activations in Emu3.5 and Liquid, about $19\times$ and $3\times$ per token, and would otherwise dominate the comparison. In UniAR the two have nearly the same length, so this step changes little. The anisotropy numbers use single-token activations rather than pooled ones. All steering vectors are stable across random halves of their minimal pairs (split-half cosine at least 0.91), except the emotion and spatial image vectors of Emu3.5 and Liquid (0.45 to 0.84). UniAR's activations at various layers exhibit a strong narrow cone effect (*anisotropy*) (Ethayarajh, 2019; Liang et al., 2022): two random text or image activations at the middle layers (16–19 of 36) have an average cosine similarity of 0.75 and 0.74 respectively, while even two random activations from opposite modalities still have 0.65 (Table 8). Rescaling to unit length and mean-centering is a common method to counteract anisotropy (Mu & Viswanath, 2018; Cai et al., 2021). Note that such centering leaves the steering vectors themselves unchanged: each is a difference of two means within one modality, $\hat{v} = \text{mean}(h^+) - \text{mean}(h^-)$, so any constant offset cancels.

Table 7: Correlation between the cosine similarity of the text and image steering vectors and cross-modal steering, per model and transfer direction. Each cell uses 21 points (7 concepts $\times$ 3 layer ranges; spatial excluded because it does not steer in most settings). Left: against recovery, as in Figure 5. Right: against the concept-rate AUC of the transferred vector alone, i.e. the numerator of recovery.

		vs. recovery			vs. AUC of transferred vector	
Model	Direction	ρ	r	p	ρ	p
UniAR	txt→img	$+0.76$	$+0.58$	< 0.001	$+0.80$	< 0.001
UniAR	img→txt	$+0.17$	$+0.19$	0.46	$+0.17$	0.46
Emu3.5	txt→img	-0.27	-0.19	0.23	$+0.42$	0.06
Emu3.5	img→txt	$+0.47$	$+0.62$	0.03	$+0.04$	0.88
Liquid	txt→img	-0.12	-0.13	0.60	$+0.32$	0.16
Liquid	img→txt	$+0.69$	$+0.60$	0.001	$+0.35$	0.12

Table 8: Anisotropy at the middle layers: mean cosine similarity between random pairs of single-token activations, raw and after centering (unit length, midpoint subtracted). Samples: 64 tokens from each of 5 images per concept, and every token but the first of all minimal-pair sentences.

	raw			unit length, centered		
Model	text–text	image–image	image–text	text–text	image–image	image–text
UniAR	0.75	0.74	0.65	0.17	0.16	−0.17
Emu3.5	0.36	0.45	0.17	0.15	0.18	−0.16
Liquid	0.51	0.50	0.42	0.09	0.07	−0.08

Table 9: Same concept pole vs. same direction at the middle layers, per concept and model. *pole cos*: cosine, after centering, between the positive-pole centroid computed from images and from sentences. *ratio*: distance from the positive-image centroid to the positive-text centroid, divided by its distance to the negative-image centroid (independent of centering). *dir. cos*: cosine between the text and image steering vectors, with z = cosine divided by the baseline standard deviation.

	UniAR			Emu3.5			Liquid		
Concept	pole cos	ratio	dir. cos (z)	pole cos	ratio	dir. cos (z)	pole cos	ratio	dir. cos (z)
emotion	−0.34	53×	0.47 (30)	−0.62	203×	0.09 (6)	−0.33	210×	0.02 (1)
age	−0.49	25×	0.31 (20)	−0.65	33×	0.07 (5)	−0.41	25×	0.05 (3)
cleanness	−0.46	10×	0.20 (13)	−0.65	12×	0.01 (1)	−0.36	9×	0.08 (5)
chaos	−0.42	12×	0.14 (9)	−0.61	23×	0.00 (0)	−0.39	17×	0.07 (4)
size	−0.59	15×	−0.02 (−1)	−0.76	25×	−0.02 (−1)	−0.59	11×	−0.01 (−1)
depth	−0.52	7×	0.03 (2)	−0.64	12×	−0.01 (−1)	−0.52	6×	0.08 (4)
spatial	−0.44	40×	−0.02 (−2)	−0.54	57×	0.01 (1)	−0.34	48×	0.03 (2)
color	−0.31	17×	0.60 (39)	−0.63	15×	0.19 (14)	−0.31	9×	0.36 (20)

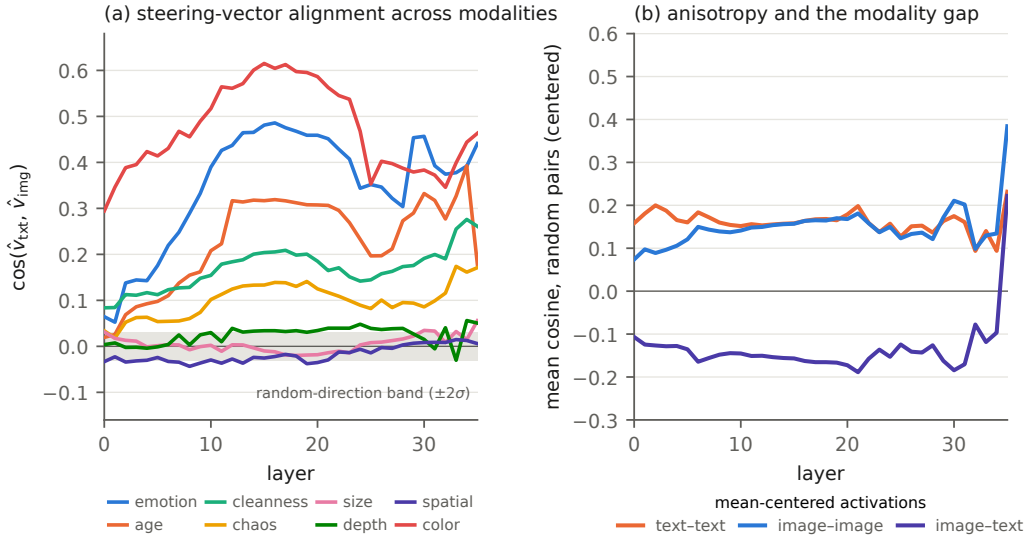


Figure 16: **(a)** Per-layer cosine similarity between each concept's textual and visual steering vectors in UniAR; the gray band is $\pm 2\sigma$ for random directions in d=4096 (σ=0.016). **(b)** Mean cosine similarity between random pairs of single-token activations across layers of UniAR, within text, within image, and across modalities, after centering (midpoint of the two modality means subtracted). The cross-modal cosine is negative after centering: the modality gap of §3.4.

D CONCEPTS UNIQUE TO ONE MODALITY

This appendix provides details on §3.3: all modality-unique concepts we could conceive, and and examples for the concepts that were successful.

D.1 ALL ATTEMPTED CONCEPTS

A candidate has to be native to one modality and still expressible as two poles of one dimension. Text pairs are 100 sentences from one template with the attribute at the final token; language identity instead uses 100 parallel translations, mean-pooled, since a language is a property of a whole sentence. Image pairs are built programmatically where the property is an image operation (a directional blur for motion blur, a reduction to 48×48 pixels for resolution) and otherwise as two edits of one generated image (camera height). We check that a direction steers its own modality before applying it, unchanged, to the other, and we report each concept on prompts where the property can appear at all: steering formality while the model describes a bowl of fruit changes little, while on a message to a friend the same vector moves contractions from 49 to 0 over twenty prompts. Where free generation shows nothing, we let the model complete a sentence whose next word must pick a pole, e.g. *"The shot was taken from ___"*, for which it says *below* 1%, 15% and 89% of the time at $\alpha = -48, 0, +48$, while eight random directions of the same size move the choice by -0.17 ± 0.59 nats against $+4.50$ here. Table 10 lists the concepts we built pairs for. Two failed at that stage: for render vs. photograph the editor either changed nothing or simply desaturated the image, and for first- vs. third-person viewpoint it could not turn an outside photograph into a participant's view (one usable pair in twelve). Visual concepts that represent a mechanical operation or a small camera move can be used as minimal pairs. But concepts that require the editing model to reimagine the whole scene do not.

Table 10: The concepts we built minimal pairs for, and what came of each.

Concept	Poles	Pairs	Outcome
Language-native			
formality	colloquial / formal	text template	works in both modalities
language identity	English / French	parallel translations	works in both modalities
language identity	English / Chinese	parallel translations	works in text; in images mostly adds lettering
Vision-native			
motion blur	sharp / blurred	directional blur of one image	works in both modalities
resolution	high / low	48×48 pixelation of one image	works in both modalities
camera height	above / below	two edits of one baseline	steers images; text only in forced choice
Tried and dropped			
render vs. photo	rendered / photo	editor	pairs failed
viewpoint	first / third person	editor	pairs failed

D.2 EXAMPLES FOR THE CONCEPTS THAT WORKED

In every figure the middle column is the unsteered baseline and the columns are steering strength relative to the largest value of that sweep, as in Figure 4; each row is one prompt, listed in the caption.

Formality. Asked to *"write an email to a friend inviting them to dinner"*, the model opens with *"Hey [Friend's Name], I'm so excited…"* at $\alpha = -16$ and with *"Dear [Friend's Name], …I am writing to extend a warm invitation…"* at $\alpha = +16$; the formal pole drops contractions and adds letter conventions. At high strength it also changes content: an apology for a noisy party comes to claim "the noise was not my fault".



Figure 17: **Formality, text → image.** Rows: a person at a desk, a basket of laundry, one person pouring water for another, a pair of shoes, a menu. None of the prompts mentions clothing or register. Toward formal: suits, documents, neatly stacked laundry, dress shoes, a plain typeset menu. Toward colloquial the scenes become warmer and more cluttered, and beyond -0.6 they start to look like illustrations; beyond $+0.6$ they wash out.

Language identity (English–French). Since an English prompt is already answered in English, we prompt in German: *"Beschreibe eine kleine Bäckerei an einer Straßenecke"* is answered in German unsteered, in English at $\alpha = -16$, and in French from $\alpha = +16$ (*"Le soleil se lève lentement derrière les toits…"*). In images the objects themselves take a French form, and any writing turns into French-looking but non-existent words. The direction supplies cultural convention only where the prompt leaves it open: scenes explicitly placed in East Asia stay East Asian at every strength.

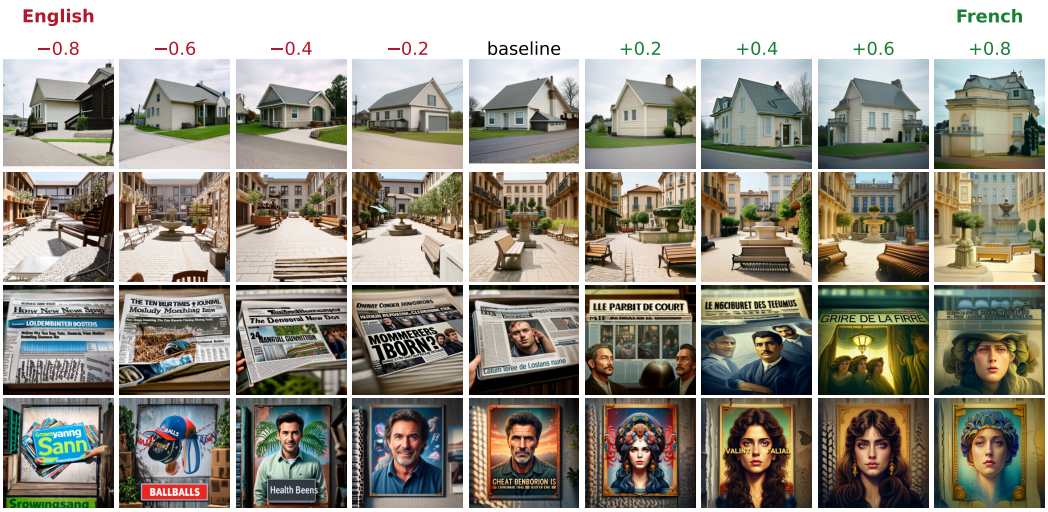


Figure 18: **English vs. French, text → image.** Rows: a house, a town square, the front page of a newspaper, a poster. Toward French: a mansard roof, a Beaux-Arts square, mastheads and posters in French-looking but invented words. Toward English: vinyl siding and a garage, plain paving, English headlines and slogans.

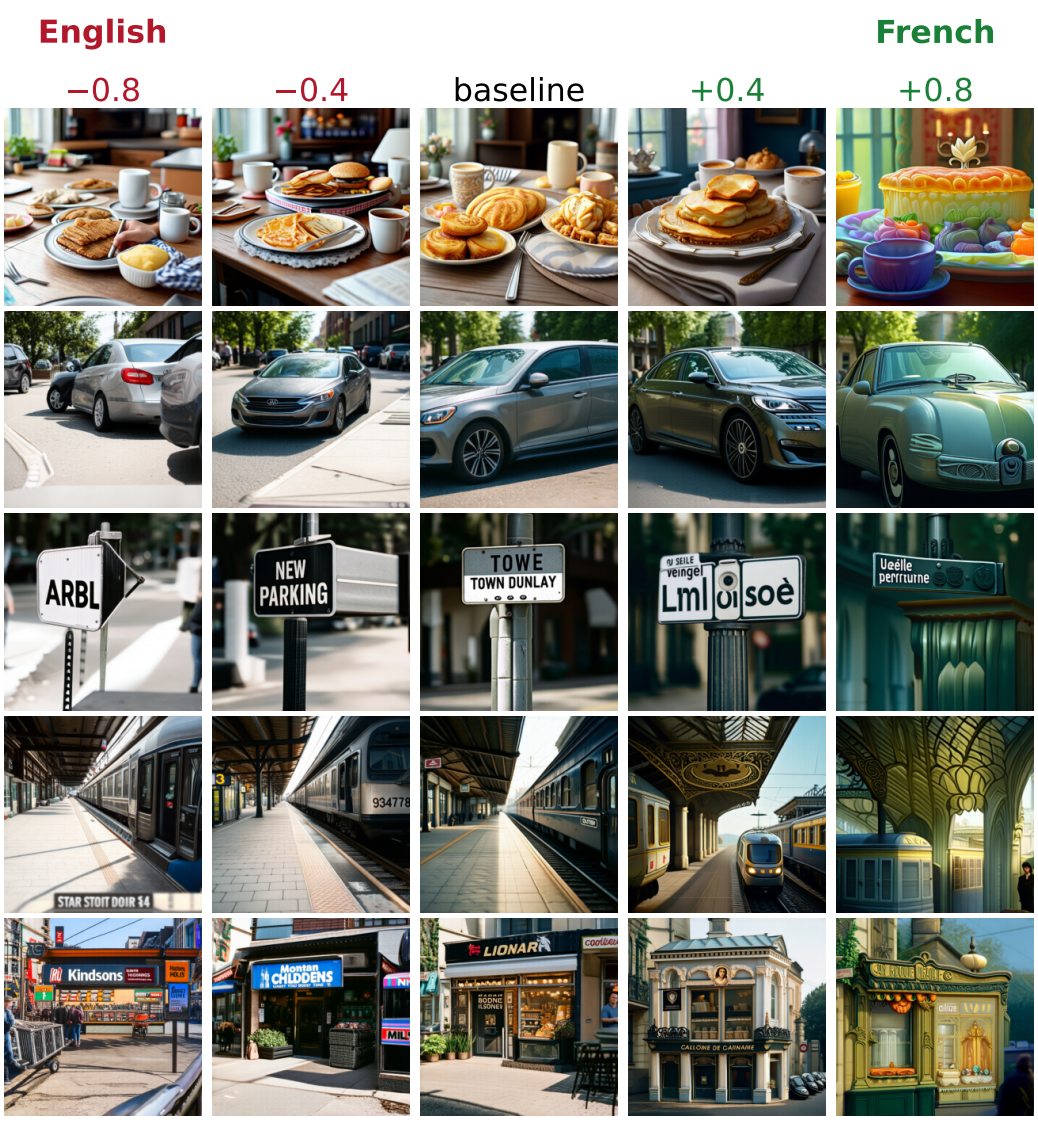


Figure 19: **The same direction on everyday objects.** Rows: a breakfast, a parked car, a street sign, a station platform, a small shop. At $+0.4$ the objects are French while the image is still photographic (pastries, a vintage car, an enamel street sign, an iron station canopy, a Haussmann shopfront); at $+0.8$ the images turn golden and painterly.

Camera height. The direction steers images cleanly in both directions but leaves free text almost untouched. We probed it with two prompts, reading the next-token distribution at layers 25–28. Completing *"The shot was taken from __"*, steering toward a low camera raises $P(\textit{below})$ from 3.7% to 12.8% and lowers $P(\textit{above})$ from 21.1% to 4.7% ($\alpha = +32$), but the most likely token remains a function word at every strength. Asked directly, *"Was the shot taken from above or below? Answer:"*, the model answers *above* unsteered (45% against 12%) and still answers *above* at every strength, with $P(\textit{below})$ peaking at 16%; swapping the order of the two words in the question moves the answer more than the steering does.



Figure 20: **Camera height, image → image.** Rows: a statue, an apartment tower, a staircase, a person. Toward one pole the camera rises until it looks straight down on the subject; toward the other it drops to near ground level. The subject stays the same.

Motion blur. In free text, steering toward sharp makes descriptions calmer (*serene*, *calm*: 3 occurrences unsteered, 11 at $\alpha = -32$) and steering toward blurred adds movement and speed words, and finally *blur* itself: asked for the view from a car, the model writes "a blur of motion and colors as the vehicle moves at high speed". In forced choice it moves "the runner looks ___" from 6% to 99% *blurry*.



Figure 21: **Motion blur, image → image.** Rows: a cyclist, a dog, a train, a sports car. Toward blurred, the background first streaks as in a panning shot while the subject stays readable, then the whole frame blurs. Toward sharp, the scene fills with extra detail.

Resolution. Steering toward low resolution makes descriptions short and vague: *"I'm sitting in a window with a blurry view of the city. . . . I'm not sure what I'm doing"* (36 words) against 65 words of specific detail unsteered; the words *blurry* and *pixelated* appear only toward low resolution. The

shortening is specific to this direction: other directions of the same size, including random ones, leave answers at 57–73 words.

must include	unsteered	toward low resolution
a street with shops	*"I strolled down the bustling street, where the scent of fresh bread and coffee filled the air. I passed by a cozy café, its inviting aroma drawing me in. [. . .]"*	*"I'm walking down a busy street in the city. I see a shop with a bright red sign."*
a view from a window	*"I wake up to the soft hum of the city, the distant clatter of traffic, and the faint scent of coffee and new beginnings. [. . .]"*	*"I'm sitting in a window with a **blurry** view of the city. It's a bright day, and I can see some buildings in the background. I'm **not sure what I'm doing**, but I'm just sitting there."*
a bridge over the river	*"I wake up to the soft hum of the city, the distant sound of traffic and the gentle lapping of water. As I step out onto the bridge, [. . .]"*	*"I was walking through the city, and I saw a bridge over the river. I was walking on the bridge, and I was looking at the city."*
a hotel lobby	*"I wake up in the hotel lobby, the soft hum of the alarm clock filling the air. The lobby is quiet, with a few guests lingering in the background. [. . .]"*	*"I'm in a hotel lobby, and I'm trying to find something. I see a **pixelated** object that looks like a **pixelated** character. I **can't tell what it is**, but I'm trying to understand it."*
a construction site	*"I wake up to the soft hum of the city, the distant clatter of traffic, and the faint scent of coffee. Today, I'm heading to the construction site, where the early morning light paints the steel and concrete in hues of gold and silver. [. . .]"*	*"I'm working on a project at a construction site. I'm trying to fix something, but I **can't see what I'm fixing**."*
a small alley	*"I strolled through the city, the air thick with the scent of coffee and the distant hum of traffic. As I walked down a narrow alley, the walls on either side were lined with small shops and cafes. [. . .]"*	*"I was walking through a small alley when I saw a person who **looked like a ghost**."*

Table 11: Six examples of applying the high–low resolution steering vector to text generation, for the prompt *"Describe your day in the city. Your description must include [X]"* ($\alpha = +32$).

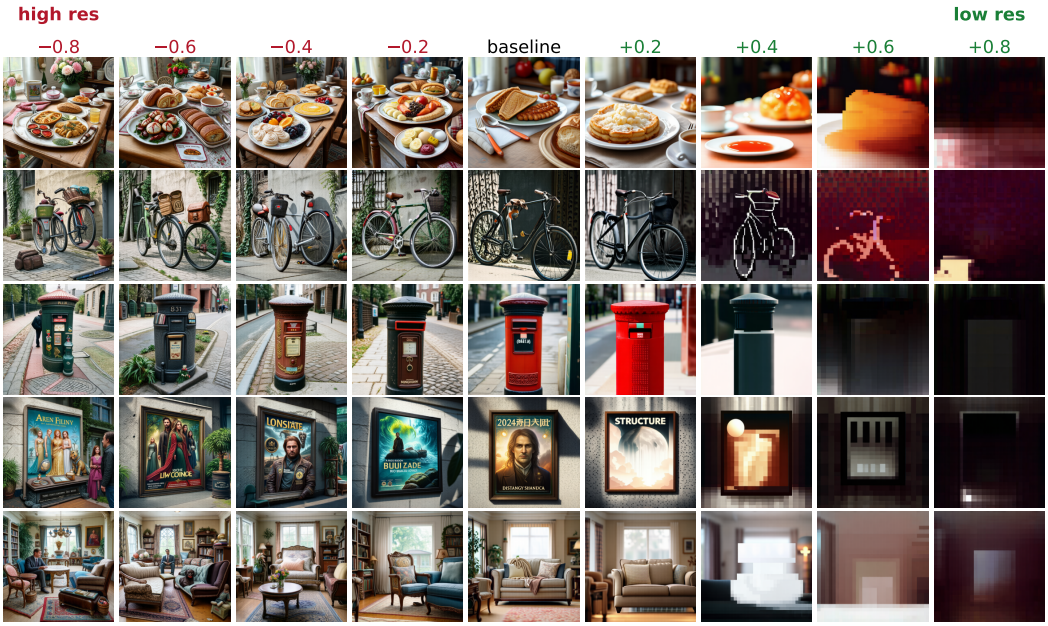


Figure 22: **Resolution, image → image.** Rows: a breakfast, a bicycle, a postbox, a poster, a living room. Toward high resolution scenes gain detail and clutter; toward low resolution they first become simpler and emptier, then turn into pixel blocks, and collapse at $+0.8$.

E AVERAGING TEXT AND IMAGE STEERING VECTORS

Here, we report a negative finding: averaging the steering vectors from the two modalities (which is equivalent to applying both at the same time) does not seem to yield any straightforward *quantitative* improvements. We did not qualitatively inspect whether averaging has any interesting effects.

For each concept we form the per-layer average $\hat{v}(\lambda) = \text{normalize}\big(\lambda\,\hat{v}_{\text{txt}} + (1-\lambda)\,\hat{v}_{\text{img}}\big)$ with $\lambda \in \{0.25, 0.5, 0.75\}$, so that $\lambda = 1$ and $\lambda = 0$ recover the single-modality vectors, and additionally a baseline that averages $\hat{v}_{\text{txt}}$ with a random direction of matched norm. We steer text generation with each vector at the early, middle, and late layer ranges over the full strength grid, and score every output with the concept-rate judge against the unsteered baseline, summarized as the per-pole maximum and the range-normalized AUC.

No average improves over the better single-modality vector: at the per-pole maximum the best average exactly ties the text vector, and differences elsewhere are within judge noise (Table 12); the same holds at the other layer ranges and on the image target, where rates saturate for most vectors.

Table 12: Averaging the two modalities' steering vectors, text target, middle layer range: concept rate summarized as per-pole maximum and range-normalized AUC over $\alpha \in [-60, 60]$, averaged over the six transfer concepts. No average improves over the better single-modality vector beyond judge noise.

Steering vector	Per-pole max	Norm. AUC
$\hat{v}_{\text{txt}}$ ($\lambda = 1$)	0.76	0.56
$\hat{v}_{\text{img}}$ ($\lambda = 0$)	0.70	0.47
Average, $\lambda = 0.75$	0.76	0.58
Average, $\lambda = 0.50$	0.73	0.55
Average, $\lambda = 0.25$	0.71	0.49
Random-direction control	0.72	0.50