Appendix

A. Related Work

A.1. Interpretability of Generative Models

Understanding what neural networks learn and how they represent information has been a longstanding challenge in deep learning [17]. For generative models specifically, several approaches have been developed to interpret their internal representations.

Attention and Attribution Methods. Several methods interpret diffusion models by analyzing their existing computations. Prompt-to-Prompt [9] and DAAM [26] visualize cross-attention maps to understand which text tokens influence which image regions, while TextSpan [5] decomposes CLIP's attention heads into text-interpretable directions for patch-level attribution. ConceptAttention [8] extends this to multi-modal DiTs, producing saliency maps for arbitrary concepts beyond those in the input prompt. Earlier work on CNN feature visualization [17] has also inspired similar approaches for diffusion models. However, these methods are largely observational, providing post-hoc visualization rather than enabling targeted concept detection or control.

Representation Decomposition. A complementary line of work seeks to decompose the model's latent space into interpretable components. Sparse autoencoders (SAEs) [3] have been applied to uncover monosemantic features in neural networks; for diffusion models, TIDE [11] introduces temporal-aware SAEs that reveal hierarchical features across DiT layers and timesteps, while Revelio [14] uses k-sparse autoencoders to analyze how architectures and pre-training datasets affect representation granularity. Complementary work [7] discovers interpretable directions in the bottleneck activations of diffusion models via PCA. Our approach also estimates directions in representation space, but does so from contrastive prompt pairs with explicit layer-timestep characterization, and provides a framework for both analysis and practical steering.

A.2. Activation Steering and Control

Steering in Language Models. Representation engineering [29] has shown that concepts in language models can be represented as directions in activation space. Inference-Time Intervention (ITI) [15] demonstrated that shifting activations along learned directions can elicit truthful answers, while Contrastive Activation Addition (CAA) [18] showed that steering vectors from contrastive prompt pairs can modulate LLM behavior without retraining. Concept Algebra [27] further established that concepts in score-based generative models are amenable to algebraic manipulation. More recently, Activation Transport (ACT) [23] applies optimal transport to compute maps between activation distributions, and LinEAS [24] learns steering vectors end-to-end via a distributional loss. Our work brings contrastive steering to diffusion transformers, but remains entirely training-free: we extract concept directions via mean difference or PCA without learning transport maps or optimizing losses.

Diffusion Model Control. Various control methods have been proposed for finer control of diffusion models. ControlNet [28] adds spatial conditioning through learned adapters, SDEdit [16] edits images by injecting noise and redenoising, and Concept Sliders [6] train lightweight LoRA adapters as continuous knobs for semantic attributes. CASTeer [4] explores projection-based concept removal. These methods either require additional training, are limited to specific editing scenarios, or do not analyze concept encoding and the behavior of the representation space. In contrast, our approach requires no training, provides continuous concept strength control via a single scalar, and includes analysis of concept vector properties across layers.

B. Background

B.1. Diffusion Models

Diffusion models generate images by iteratively denoising Gaussian noise. The forward diffusion process gradually adds noise to data:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{6}$$

where $\mathbf{x}_0$ is the original data, $\mathbf{x}_t$ is the noised version at timestep t, and $\bar{\alpha}_t$ controls the noise schedule.

The reverse process learns to denoise, parameterized by a neural network ϵ_θ that predicts the noise:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t, c)\right) + \sigma_t\mathbf{z} \tag{7}$$

where c is a text conditioning signal and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Classifier-Free Guidance (CFG). To improve text-image alignment, classifier-free guidance [10] interpolates between conditional and unconditional predictions:

$$\tilde{\epsilon}_\theta(\mathbf{x}_t, t, c) = \epsilon_\theta(\mathbf{x}_t, t, \emptyset) + w \cdot (\epsilon_\theta(\mathbf{x}_t, t, c) - \epsilon_\theta(\mathbf{x}_t, t, \emptyset)) \tag{8}$$

where w is the guidance scale. In practice, conditional and unconditional samples are batched together during inference.

B.2. Diffusion Transformers and the Residual Stream

Diffusion transformers (DiTs) replace the UNet backbone with a stack of L transformer blocks. Each block ℓ applies self-attention, cross-attention (conditioned on text embeddings), and an MLP sub-layer, each followed by a residual connection. Denoising timestep t as $\mathbf{h}_{\ell,t} \in \mathbb{R}^d$, a single block computes:

$$\mathbf{h}_{\ell+1,t} = \mathbf{h}_{\ell,t} + g_\ell(\mathbf{h}_{\ell,t},\ c,\ t) \tag{9}$$

where g_ℓ encapsulates the attention and MLP sub-layers. The sequence $\mathbf{h}_{1,t}, \mathbf{h}_{2,t}, \ldots, \mathbf{h}_{L,t}$ forms the *residual stream*: the primary information pathway through the network that each block reads from and writes to.

We intercepted the cross-attention output at each block i.e., the additive contribution produced by the cross-attention sub-layer before it is added to the residual stream, to extract and steer concept directions. The resulting activation tensor at each layer is mean-pooled over spatial (patch-token) positions to yield a single vector $\mathbf{h}_{\ell,t} \in \mathbb{R}^d$.

B.3. Linear Representation of Concepts

The linear representation hypothesis [19] posits that high-level semantic concepts are encoded as directions in a model's activation space: for a concept c, there exists a direction $\mathbf{v}_c \in \mathbb{R}^d$ such that $\text{presence}(c, \mathbf{h}) \approx \mathbf{v}_c^\top \mathbf{h}$, where $\mathbf{h}$ is an intermediate activation. A practical consequence is that adding $\alpha \mathbf{v}_c$ to activations should modulate the presence of c in the output. In language models, contrastive methods have successfully extracted and steered concepts such as truthfulness and sentiment [15, 18], with theoretical support showing that linear representations emerge naturally under the next-token objective [13]. The shared architectural backbone of stacked attention and MLP layers operating on a residual stream (Section B.2) suggests that similar structure may arise in diffusion transformers.

Extending this hypothesis to DiTs introduces challenges absent in the language setting: activations are distributed over *spatial* token grids that must be aggregated to obtain a single concept direction, representations evolve across a *temporal* denoising trajectory so the same concept may manifest differently at different noise levels, and text conditioning enters via architecture-dependent mechanisms (cross-attention in DiT, joint attention in MMDiT). Recent analysis by [12] reveals a coarse-to-fine semantic evolution across layers and timesteps, motivating the layer-wise and timestep-specific approach we develop in Section 2.

C. Notation

Table 1 summarizes the notation used throughout this paper.

Notational conventions. Throughout this paper, we use subscripts to index layer ℓ and timestep t, reserving superscripts for semantic labels (e.g., $+/-$ for positive/negative prompts). When discussing a specific concept c, we write $\mathbf{v}_{\ell,t}$ with the concept implicit from context. In settings comparing multiple concepts, we use $\mathbf{v}_{c,\ell,t}$ to disambiguate. Our general formulation allows schedules $\alpha_{\ell,t}$, but all experiments in this paper use the constant case $\alpha_{\ell,t} \equiv \alpha$.

D. Extended Method Details

The main text presents our extraction and steering framework using the mean difference estimator for brevity. Here we expand on the problem formulation and contrastive data collection, describe the PCA-based direction estimation strategy referenced in Section 2.1.1, provide the full motivation for our joint normalization scheme, and detail how steering interacts with classifier-free guidance and available steering schedules.

Table 1. Summary of notation.

Symbol	Description
ℓ	Layer index, $\ell \in \{1, \ldots, L\}$
t	Diffusion timestep, $t \in \{0, \ldots, T-1\}$
i	Prompt pair index, $i \in \{1, \ldots, N\}$
d	Activation dimension
N	Number of contrastive prompt pairs
$\mathbf{h}^{\pm}_{i,\ell,t}$	Activation vector for pair i at layer ℓ, timestep t
$\boldsymbol{\delta}_{i,\ell,t}$	Difference vector: $\mathbf{h}^{+}_{i,\ell,t} - \mathbf{h}^{-}_{i,\ell,t}$
$\bar{\boldsymbol{\delta}}_{\ell,t}$	Mean difference vector across all pairs
$\mathbf{v}_{\ell,t}$	Concept vector at layer ℓ, timestep t
$\mathbf{u}_{\ell,t}$	Unit-norm PCA direction at layer ℓ, timestep t
$\alpha_{\ell,t}$	Steering strength at layer ℓ, timestep t
$\mathbf{H}^{\pm}_{\ell,t}$	Stacked activation matrices for positive/negative prompts
$\mathbf{X}_{\ell,t}$	Concatenated activation matrix: $[\mathbf{H}^{+}_{\ell,t}; \mathbf{H}^{-}_{\ell,t}]$

D.1. Problem Formulation

Given a diffusion model ϵ_θ and a semantic concept c (e.g., "happiness", "watercolor style"), our method addresses three interconnected objectives. First, we seek to **extract** concept vectors $\mathbf{v}_{\ell,t}$ that capture how the model represents c at each layer ℓ and timestep t along the residual stream (Section B.2). Second, we aim to **characterize** how these representations evolve across both network depth and the denoising trajectory. Finally, we enable **steering** of the generation process by manipulating activations according to $\mathbf{h}'_{\ell,t} = f(\mathbf{h}_{\ell,t}, \mathbf{v}_{\ell,t}, \alpha)$, where α controls the steering strength.

These objectives are complicated by several properties of diffusion transformers: activations live on high-dimensional spatial token grids that must be aggregated to obtain a single concept direction, the same concept may manifest differently at different noise levels along the denoising trajectory, and a small number of contrastive pairs must suffice for practical deployment.

D.2. Contrastive Data Collection

The main text introduces the contrastive dataset $\mathcal{P}_c = \{(p_i^+, p_i^-)\}_{i=1}^N$ in compact form. Here we provide the full formalization. We posit that a concept c can be isolated by analyzing the difference in activations between prompts that differ only by the presence of c. To robustly estimate this

direction and marginalize over uncorrelated semantic contexts, each pair consists of a positive prompt p_i^+ exhibiting the concept and a negative prompt p_i^- lacking it (or exhibiting its antonym), while sharing a common context K_i. Formally, $p_i^+ = T(K_i, c)$ and $p_i^- = T(K_i, \varnothing)$, where T is a template function that inserts the concept into a context-specific sentence frame. For example, if c is "happiness," a pair might be ("A happy person in a park", "A sad person in a park"), where K_i = "person in a park." By averaging over N diverse contexts K_i, the computed vector captures the direction of c rather than the specifics of any single scene.

D.3. PCA-Based Direction Estimation

An alternative to the mean difference estimator relaxes the isotropic noise assumption by identifying the direction of maximum variance in the data. For each layer ℓ and timestep t, we concatenate all positive and negative activations into a single matrix $\mathbf{X}_{\ell,t} = [\mathbf{H}^+_{\ell,t}; \mathbf{H}^-_{\ell,t}] \in \mathbb{R}^{2N \times d}$ and compute the first principal component:

$$\mathbf{u}_{\ell,t} = \arg\max_{\|\mathbf{u}\|=1} \mathbf{u}^\top \bar{\mathbf{X}}^\top_{\ell,t} \bar{\mathbf{X}}_{\ell,t} \mathbf{u} \tag{10}$$

where $\bar{\mathbf{X}}_{\ell,t}$ denotes the mean-centered data. This unit-norm direction captures the dominant mode of variation between concept-present and concept-absent samples.

Since PCA components are defined up to sign, we align $\mathbf{u}_{\ell,t}$ with the mean difference direction for consistent steering. We then rescale by the mean difference magnitude to preserve relative concept strength across layers:

$$\tilde{\mathbf{v}}_{\text{PCA},\ell,t} = \|\bar{\boldsymbol{\delta}}_{\ell,t}\|_2 \cdot \mathbf{u}_{\ell,t} \tag{11}$$

where $\bar{\boldsymbol{\delta}}_{\ell,t} = \mathbb{E}[\mathbf{h}^+_{\ell,t}] - \mathbb{E}[\mathbf{h}^-_{\ell,t}]$. Finally, we apply joint normalization across layers as in the mean difference method:

$$\mathbf{v}_{\text{PCA},\ell,t} = \frac{\tilde{\mathbf{v}}_{\text{PCA},\ell,t}}{S_t}, \quad S_t = \|\tilde{\boldsymbol{V}}_t\|_2 \tag{12}$$

where $\tilde{\boldsymbol{V}}_t = [\tilde{\mathbf{v}}_{\text{PCA},1,t}; \ldots; \tilde{\mathbf{v}}_{\text{PCA},L,t}]$. This allows a single steering coefficient α to control intervention strength across all layers while preserving their relative importance. Steering results using this estimator are reported in Section E.3.

D.4. Joint Normalization

As described in Section 2.1.1, we apply a joint normalization scheme rather than normalizing each layer independently. The motivation is twofold: (1) it is computationally efficient, requiring only a single global norm computation, and (2) it ensures a single steering coefficient α can modulate all layers simultaneously while each layer's contribution is weighted by its natural concept magnitude. Concretely, we stack all layer-wise mean differences into a single vector $\boldsymbol{\Delta}_t = [\bar{\boldsymbol{\delta}}_{1,t}; \ldots; \bar{\boldsymbol{\delta}}_{L,t}] \in \mathbb{R}^{Ld}$ and compute the global scale factor $S_t = \|\boldsymbol{\Delta}_t\|_2$. The normalized concept vectors are then $\mathbf{v}_{\text{MD},\ell,t} = \bar{\boldsymbol{\delta}}_{\ell,t}/S_t$. This preserves relative magnitudes: layers where the concept manifests more strongly contribute proportionally more to steering, matching our empirical observation that concept encoding varies substantially across layers (Figure 3 (c)).

Algorithm 1 Concept Vector Extraction

Require: Diffusion model ϵ_θ, concept c, paired prompts $\mathcal{P}_c = \{(p_i^+, p_i^-)\}_{i=1}^N$
Ensure: Concept vectors $\{\mathbf{v}_{\ell,t}\}$ for all layers ℓ and timesteps t

```
Initialize activation extractor on the cross-attention output at each block
for each pair (p_i^+, p_i^-) ∈ P_c do
    Run forward pass with p_i^+, collect {h^+_{i,ℓ,t}}
    Run forward pass with p_i^-, collect {h^-_{i,ℓ,t}}
end for
for each layer ℓ and timestep t do
    Aggregate: H^+_{ℓ,t} = [h^+_{1,ℓ,t}, ..., h^+_{N,ℓ,t}]
    Aggregate: H^-_{ℓ,t} = [h^-_{1,ℓ,t}, ..., h^-_{N,ℓ,t}]
    Compute: v_{ℓ,t} = EstimateDirection(H^+_{ℓ,t}, H^-_{ℓ,t})
end for
return {v_{ℓ,t}}
```

D.5. Classifier-Free Guidance Compatibility

Since CFG processes conditional and unconditional predictions in a single batched forward pass (Section B), steering must be applied selectively. We intervene only on the conditional branch while leaving the unconditional baseline untouched:

$$\mathbf{h}'_{\text{batch}} = \begin{bmatrix} \mathbf{h}_{\text{uncond}} \\ \mathbf{h}_{\text{cond}} + \alpha_{\ell,t} \cdot \mathbf{v}_{\ell,t} \end{bmatrix} \tag{13}$$

This ensures that steering modulates the semantic content introduced by the text prompt, while preserving the unconditional reference that CFG uses to amplify prompt adherence.

EstimateDirection is instantiated as either the mean difference estimator (Equation (3)) with joint normalization, or the PCA-based estimator (Section D.3).

E. Additional Experimental Results

E.1. Hyperparameters

E.2. Convergence Measurement

To assess how quickly concept vectors stabilize as more contrastive pairs are added, we measure convergence via a running update procedure. We initialize the concept vector estimate using the mean difference computed from the first 5 prompt pairs, then incrementally incorporate additional

Table 2. Hyperparameter settings for all experiments.

Parameter	Value
Sampling steps	20
CFG scale	4.5
Scheduler	DDPM
Number of prompt pairs (N)	35
Default α	2.0
Extraction point	Cross-attention output
Image resolution	1024×1024
Batch size	1
Random seed	42

pairs one at a time, updating the running mean. At each step k, we compute the cosine similarity between the current estimate and the final estimate obtained after all N pairs have been included. The convergence curves in Figure 3 (a) plot this similarity as a function of k, while Figure 3 (b) reports the number of pairs required to reach 99% similarity for each concept.

E.3. PCA-Based Steering Results

In the main text, we report steering results using the mean difference estimator. Here we present results using the PCA-based estimator (described in Section D.3), which identifies the direction of maximum variance across concept-present and concept-absent activations. Figure 5 shows a steering strength sweep using PCA-derived concept vectors. The qualitative behavior is consistent with the mean difference results: concept presence increases smoothly with α, and the method provides graded control across all evaluated concepts. This confirms that both estimation strategies yield viable concept directions for steering. We further observe that combining the first two principal components (PC1 and PC2) as the steering direction produces qualitatively better results than using PC1 alone, hinting that concepts may be encoded not as single directions but as low-dimensional subspaces in the residual stream. A more thorough investigation of multi-component concept subspaces is an interesting direction for future work.

E.4. PixArt-Σ Steering Results

We also implement our technique and repeat the steering experiments on PixArt-Σ [2]. Figure 6 shows a steering strength sweep across selected concepts. The qualitative behavior is consistent with PixArt-α: concept presence increases smoothly with α, confirming that the method generalizes across DiT variants.

E.5. Additional Steering Samples

Figure 7 presents a broader set of steered generations across multiple concepts and contexts, complementing the examples shown in Figure 1. It reveals several informative patterns across concept types. For emotion-based concepts such as happiness and surprise, steering produces a smooth gradient of expression intensity as α increases, whereas the direct-prompting reference (rightmost column) tends to yield a single high-intensity expression with less nuance. This illustrates a key advantage of activation-space steering: it offers continuous, graded control that is difficult to achieve through prompt engineering alone.

For more complex, multi-faceted concepts such as Plushie Cartoon, the steered outputs diverge visually from the direct-prompting reference. This is expected: such concepts can manifest in multiple visual forms (e.g., texture, shape, color palette), and the concept vector captures an aggregate direction that may emphasize different aspects than the text encoder. For concepts such as Young Age, the unsteered baseline already closely aligns with the target direction, leaving limited headroom for steering. Although concept expression is still visible at higher α values, the small perceptual gap between baseline and steered outputs can result in modest CLIP score improvements despite meaningful visual changes. These observations highlight the importance of careful prompt design and high-quality contrastive pairs for effective concept vector extraction.

E.6. Additional Concept Composition Results

We provide additional examples of linear concept composition with further concept pairs and subjects.

F. Discussion and Limitations

Our results show that simple linear interventions in DiT activations can provide effective control over several semantic concepts and styles. In many cases, steering matches or exceeds direct prompting on CLIP similarity, and additive composition remains qualitatively coherent across distinct concept categories.

Effectiveness of steering. Steering is particularly strong for style and certain semantic attributes, where it often outperforms direct prompting in CLIP similarity. One plausible reason is that direct prompting depends on the text encoder and prompt phrasing to express the desired visual change, whereas steering modifies internal representations more directly. At the same time, CLIP similarity is only a proxy for concept strength and does not fully capture perceptual quality or faithfulness, so broader evaluation would be useful.

Concept subspaces. Our PCA analysis (Section E.3) indicates that some concepts may be better represented as low-dimensional subspaces rather than single directions. In

Figure 5. **PCA-based steering.** Steering strength sweep using concept vectors computed via the PCA estimator. The qualitative trends match those obtained with mean difference (Figure 2), confirming that both estimation strategies produce effective concept directions.

Figure 6. **PixArt-Σ steering results.** Steering strength sweep on PixArt-Σ, demonstrating that our method generalizes across DiT variants.

particular, combining PC1 and PC2 often produces better results than PC1 alone. This points to structure beyond a single steering vector, especially for more complex concepts.

Cross-subject generalization. Concept vectors can transfer beyond the subject category used during extraction. For example, a surprise vector extracted from human prompts can still produce meaningful effects when applied to a cat (Figure 8). This suggests that some concept directions are not tied to a single subject class, although the limits of this transfer remain unclear.

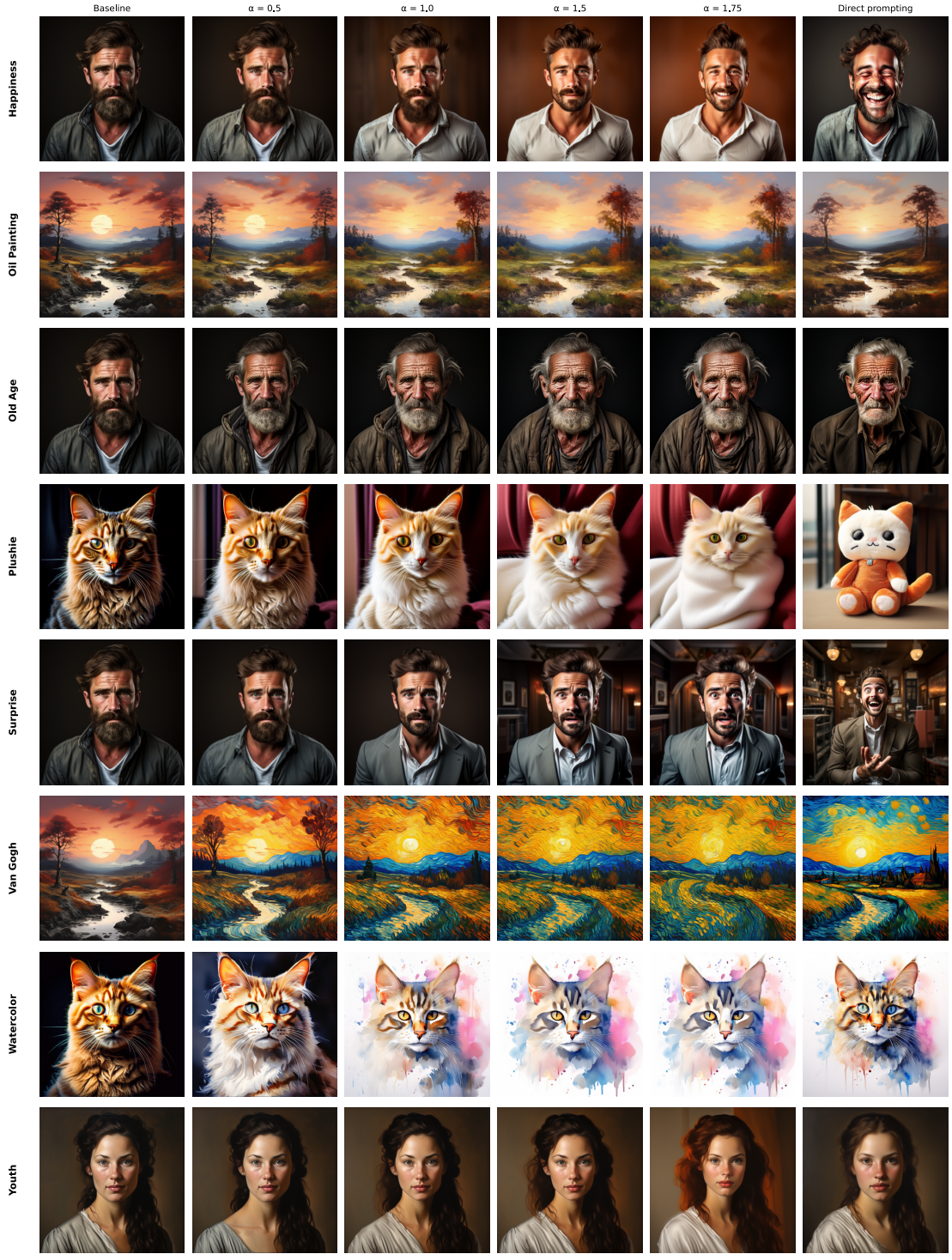

Figure 7. **Additional steering samples.** Steered generations across multiple concepts and contexts at varying steering strengths. Each row corresponds to a different concept-context combination, with α increasing from left to right.

F.1. Limitations

Our method depends on the quality of the contrastive prompts: if positive and negative prompts differ in unin-

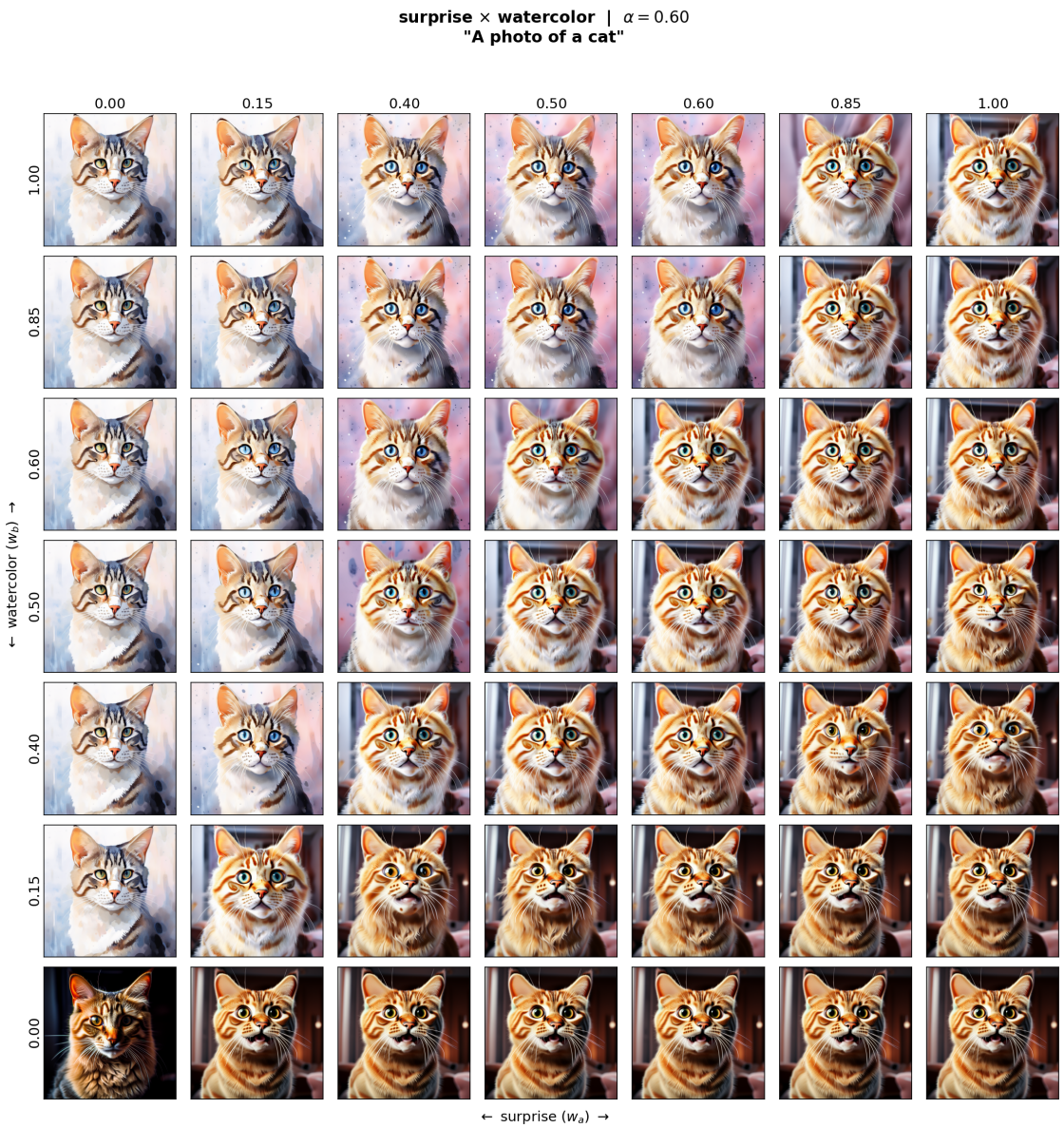

Figure 8. **Additional concept composition.** Steering with a weighted combination $w_a \mathbf{v}^{(\text{surprise})} + w_b \mathbf{v}^{(\text{watercolor})}$ on a cat subject. The horizontal axis varies the surprise weight w_a and the vertical axis varies the watercolor weight w_b, each from 0 to 1. Both concepts compose smoothly across a different subject category. Notably, the surprise concept vector was extracted exclusively from human-subject prompt pairs, yet it transfers to a non-human subject (cat), suggesting that the learned direction captures a general visual notion of surprise rather than a human-specific one.

tended ways, the extracted vector may capture spurious attributes rather than the target concept. It is also most effective for global attributes such as style, expression, or broad semantic properties; fine-grained or spatially localized concepts are harder to control with a single global intervention and may require spatially structured steering mechanisms. Finally, positive addition was consistently more reliable than negative steering or subtraction-based composition. Suppressing a concept with $-\alpha \mathbf{v}_{\ell,t}$ often underperformed, likely because the text conditioning pathway can continue to reinforce the concept, and while additive composition ($\mathbf{v}_A + \mathbf{v}_B$) usually remains coherent, subtraction-based composition ($\mathbf{v}_A - \mathbf{v}_B$) was much less stable. Together, these observations suggest that the representation is only approximately linear in practice and supports addition more robustly than inverse operations. We leave these questions for future work.