How Fragile Is Vision-Language Alignment? Mapping Concept Disruption Under Text-to-Image Personalization

Mujtaba Hasan
New Delhi, India
mujtaba.hasan@live.com

Abstract

Text-to-image diffusion models learn a mapping from natural language to visual structure, but how robust is this mapping to perturbation? We use personalization—fine-tuning a model to learn a new face, object, or style—as a controlled stress test to probe the fragility of learned vision-language alignment. We find that fine-tuning for *one* concept systematically shifts the model's ability to faithfully render *unrelated* concepts, and that this disruption follows structured, predictable patterns. To measure this fragility, we construct **Concept Entanglement Maps**: per-prompt, per-model disruption matrices that reveal which concepts are most affected and why. Using Stable Diffusion v1.5 as a controlled testbed, we evaluate 15 subjects across three personalization methods on 200 prompts and report three findings about the organization of vision-language alignment: (1) aggregate disruption is larger for vision-backbone and cross-attention perturbations than for text-embedding perturbations, despite the latter directly modifying the language representation; (2) abstract and compositional language is significantly more fragile than concrete, object-specific language; and (3) disruption does *not* follow semantic proximity—personalizing for a face does not preferentially disrupt other face-related prompts ($p = 1.0$), suggesting that alignment vulnerability is organized globally rather than purely by semantic category. These findings expose a structural vulnerability in current text-to-image personalization: the same cross-attention mechanism that enables compositional generalization also creates pathways through which local fine-tuning can propagate as global alignment shift.

1 Introduction

The ability of text-to-image diffusion models to translate natural language descriptions into coherent visual outputs depends on *vision-language alignment*—the learned correspondence between text token representations and spatial visual features. This alignment, mediated by cross-attention layers that project text embeddings into the visual feature space, enables remarkable compositional generalization: a model can render "a koala wearing a top hat on the moon" despite never observing this combination, because it has learned how each word anchors to visual structure (Rombach et al., 2022; Hertz et al., 2023).

But how robust is this learned alignment? If we perturb the model's parameters to teach it a single new concept—a specific person's face, a particular object, an artistic style—does the alignment for all other concepts remain intact? Or does local fine-tuning propagate as global alignment shift?

This question is both scientifically fundamental and practically urgent. Scientifically, the answer reveals how vision-language grounding is organized in these models: is it modular, with each concept grounded independently, or entangled, with concepts sharing representational substrate such that modifying one affects others? Practically, personalized adapters are increasingly used as general-purpose creative tools. A user who fine-tunes an adapter for a product, character, or visual style typically expects that unrelated prompts remain faithful. If a personalization silently shifts how the model grounds unrelated language—for example, abstract prompts, people prompts, or scene descriptions—then failures may surface only after deployment, where they are difficult to diagnose.

We formalize this question through the lens of **alignment stability**. Let $A(\theta, p)$ denote the alignment quality between a model with parameters θ and a text prompt p, measured as the CLIP text-image cosine similarity between p and the model's output. An ideal personalization $\theta \to \theta + \Delta_s$ for subject s would satisfy:

$$A(\theta + \Delta_s, p) = A(\theta, p) \quad \forall\, p \notin \mathcal{P}_s, \qquad (1)$$

where $\mathcal{P}_s$ is the set of prompts related to s. That

is, alignment for non-target prompts should be invariant to personalization. We show empirically that this invariance is systematically violated in aggregate, and that the violations have predictable structure.

To measure these violations, we introduce **Concept Entanglement Maps**: a structured diagnostic framework that produces a disruption matrix $\mathbf{M} \in \mathbb{R}^{P \times S}$, where entry $M_{p,s}$ quantifies the alignment shift for prompt p after personalizing for subject s. Unlike scalar evaluation metrics that assess target-concept fidelity, the entanglement map reveals the collateral structure of alignment disruption—which prompts are fragile, which methods cause broader shifts, and whether disruption follows semantic proximity.

Personalization is not our topic; it is our experimental instrument. Three personalization methods—Textual Inversion (Gal et al., 2023), DreamBooth-LoRA (Hu et al., 2022; Ruiz et al., 2023), and Custom Diffusion (Kumari et al., 2023)—modify the model at different points in the vision-language pipeline: text embeddings, UNet attention layers, and cross-attention projections respectively. By comparing the disruption patterns they produce, we can localize where alignment is most fragile within the architecture.

Our contributions are:

1. We formalize the alignment stability problem and introduce Concept Entanglement Maps as a structured diagnostic (§3).

2. We show that aggregate non-target disruption is smallest for text-space perturbations and largest for cross-attention/vision-backbone perturbations, revealing the cross-attention interface as a fragile architectural link (§4).

3. We demonstrate that abstract, compositional language is significantly more fragile than concrete language, and that disruption does not follow semantic proximity (§5).

4. We provide preliminary cross-attention analysis visualizing how personalization can shift the spatial grounding of unrelated text tokens (§6).

Our findings map to a core concern of the ALVR community: shortcomings of existing large vision and language models on downstream tasks. We show that alignment is not merely imperfect; in SD v1.5 personalization, it is structurally fragile in measurable and predictable ways.

2 Related Work

Vision-language grounding in generative models. Text-to-image diffusion models ground language in visual structure primarily through cross-attention, where text token embeddings serve as keys and values while spatial visual features serve as queries (Rombach et al., 2022). Hertz et al. (2023) showed that manipulating these attention maps enables fine-grained image editing, demonstrating that they encode spatial correspondence between words and visual regions. Tang et al. (2023) developed attribution maps that trace which image regions are influenced by which text tokens, revealing that grounding quality varies across concept types. Our work extends this line of inquiry by asking how stable this grounding is under model perturbation.

Alignment robustness and concept interference. The robustness of learned representations to perturbation is a well-studied concern in NLP (Ribeiro et al., 2020) and vision-language modeling. Fine-tuning can also produce interference or forgetting (Kirkpatrick et al., 2017; Li and Hoiem, 2016). In diffusion models, Kumari et al. (2023) note that customization can cause language drift, where common words change meaning, and recent work studies memorization and robustness issues in generative models (Wen et al., 2024). Our work differs in providing a structured, per-prompt diagnostic rather than aggregate metrics, and in explicitly testing whether disruption follows semantic organization.

Text-to-image personalization as a probe. Personalization methods span the full text-to-image pipeline. Textual Inversion (Gal et al., 2023) modifies only the text embedding, adding a learned token to CLIP's vocabulary. LoRA (Hu et al., 2022) applies low-rank perturbations to the UNet's attention projections. Custom Diffusion (Kumari et al., 2023) fine-tunes cross-attention key/value matrices, operating directly at the language-vision interface. DreamBooth (Ruiz et al., 2023) fine-tunes the model for subject-driven generation. These methods create a natural set of controlled perturbations at known architectural locations, making them useful probes for studying where alignment is most fragile. We do not evaluate these methods' personalization quality; we use them as instruments to

stress-test alignment.

Evaluation of personalization and alignment. Standard metrics—CLIP-I similarity, DINO identity preservation (Caron et al., 2021), CLIPScore (Hessel et al., 2021), LPIPS (Zhang et al., 2018), and FID (Heusel et al., 2017)—assess target concept fidelity, semantic alignment, perceptual distance, or distributional shift. No existing metric systematically measures what happens to non-target concepts after personalization. Our entanglement map fills this gap by measuring the collateral alignment cost that personalization imposes on the rest of the model's concept vocabulary.

3 Method

3.1 The Alignment Stability Problem

Let $G(\cdot;\theta)$ be a text-to-image diffusion model and let f_T, f_I be the text and image encoders of a pretrained CLIP model (Radford et al., 2021). The alignment function for prompt p under model θ is:

$$A(\theta, p) = \mathbb{E}_{z\sim\mathcal{N}}\left[\cos\left(f_T(p), f_I(G(p, z; \theta))\right)\right], \tag{2}$$

where z is the latent noise initialization. Personalization produces a perturbation Δ_s for subject s, modifying the model to $\theta+\Delta_s$. We define the operational semantic disruption for a non-target prompt p using seed-matched generations:

$$E_{\text{sem}}(p, s) = \frac{1}{K}\sum_{k=1}^{K}\left|a^{\text{pers}}_{p,s,k} - a^{\text{base}}_{p,k}\right|, \tag{3}$$

where

$$a^{\text{base}}_{p,k} = \cos\left(f_T(p), f_I(G(p, z_k; \theta))\right), \tag{4}$$
$$a^{\text{pers}}_{p,s,k} = \cos\left(f_T(p), f_I(G(p, z_k; \theta+\Delta_s))\right). \tag{5}$$

The absolute value measures instability or shift, not necessarily signed degradation. We use the term *disruption* for this absolute semantic shift, and reserve degradation for cases where the personalized score is lower than the base score. This distinction is important because a non-target prompt can shift in either direction even when the output remains plausible.

The perturbation Δ_s occupies different subspaces of the model's parameter space depending on the method:

$$\Delta_s^{\text{TI}} \in \mathbb{R}^{d_{\text{emb}}} \quad \text{(text embedding)}, \tag{6}$$

$$\Delta_s^{\text{LoRA}} \in \{B_l A_l\}_{l=1}^{L}, \quad A_l \in \mathbb{R}^{r\times d_l}, \; B_l \in \mathbb{R}^{d_l\times r}, \tag{7}$$

$$\Delta_s^{\text{CD}} \in \{W_l^K, W_l^V\}_{l\in\mathcal{L}_x} \quad \text{(cross-attn. K/V)}. \tag{8}$$

These perturbations form a hierarchy along the vision-language pipeline: Textual Inversion perturbs the language representation before it enters the UNet; LoRA perturbs the UNet's internal processing; Custom Diffusion perturbs the cross-attention projections where text meets vision. Comparing $E_{\text{sem}}(p, s)$ across these perturbation types localizes the architectural locus of alignment fragility.

3.2 Entanglement Map Construction

Given evaluation prompts $\mathcal{P} = \{p_1, \ldots, p_P\}$ and personalized models $\mathcal{S} = \{\Delta_{s_1}, \ldots, \Delta_{s_S}\}$, we construct the entanglement map $\mathbf{M} \in \mathbb{R}^{P\times S}$. For each (p, s) pair, we generate K images from both the base and personalized models using identical seeds:

$$\hat{x}_k^{\text{base}} = G(p, z_k; \theta), \quad \hat{x}_k^{\text{pers}} = G(p, z_k; \theta+\Delta_s). \tag{9}$$

The seed-matched design controls for stochastic variation, isolating the effect of Δ_s.

Non-target prompt protocol. For each non-target evaluation prompt, the base and personalized models receive the exact same text prompt p. We do not prepend the learned personalization token, subject identifier, or style token during non-target evaluation. This distinction is essential: adding the subject token would measure triggered personalization behavior, whereas our goal is to measure collateral changes in the model's ordinary prompt-following behavior after the parameter perturbation Δ_s. The only difference between the two generations in Eq. 9 is therefore the model parameters, not the text input or random seed. Crucially, non-target evaluation uses the same prompt p for the base and personalized model; no personalization token is inserted into non-target prompts.

3.3 Disruption Metrics

We define two complementary metrics capturing semantic and perceptual dimensions of alignment disruption.

Semantic disruption. Equation 3 measures whether personalization changes how well the output matches the text prompt—a direct probe of alignment integrity.

Perceptual disruption. The perceptual distance between seed-matched outputs is:

$$E_{\text{per}}(p, s) = \frac{1}{K} \sum_{k=1}^{K} \text{LPIPS}(\hat{x}_k^{\text{base}}, \hat{x}_k^{\text{pers}}). \quad (10)$$

E_{per} captures how much the visual output changes regardless of whether it still matches the prompt. We report E_{per} for completeness and emphasize E_{sem} as the primary alignment metric.

Aggregate entanglement score. For method comparison, we aggregate over non-target prompts $\mathcal{P}_{\text{non}}$ excluding the subject's own semantic category:

$$E(s) = \frac{1}{|\mathcal{P}_{\text{non}}|} \sum_{p \in \mathcal{P}_{\text{non}}} \frac{1}{2} \left(\widehat{E_{\text{sem}}}(p, s) + \widehat{E_{\text{per}}}(p, s) \right), \quad (11)$$

where $\hat{\cdot}$ denotes min-max normalization. We use E for aggregate method and subject-category comparisons, while E_{sem} is used for semantic disruption analyses.

3.4 Noise Floor

Even without perturbation, stochastic variation in the diffusion process produces non-zero semantic and perceptual shifts between different seeds of the same model. We establish a noise floor by computing E_{sem} and E_{per} between different seed pairs of the unperturbed base model across all 200 prompts (600 measurements). The 95th percentiles set thresholds above which disruption exceeds random variation:

$$\tau_{\text{sem}} = 0.044, \quad \tau_{\text{per}} = 0.780. \quad (12)$$

The mean E_{sem} across all personalizations is 0.021, below τ_{sem}. Therefore, individual per-prompt disruptions are typically indistinguishable from stochastic variation. Our claims are aggregate claims over many prompts, seeds, and personalized models; we do not claim reliable single-prompt diagnosis unless disruption exceeds the noise floor.

3.5 Experimental Setup

Base model. Stable Diffusion v1.5 (Rombach et al., 2022). We use it as a controlled testbed because its single CLIP-L text encoder and UNet cross-attention structure make architectural attribution relatively clean. Generalization to newer architectures such as SDXL (Podell et al., 2024) and diffusion transformers (Esser et al., 2024) remains an important direction for future work.

Subjects. 15 subjects across three semantic categories: 5 faces (LFW dataset), 5 objects (DreamBooth dataset: backpack, teapot, sneaker, vase, clock), and 5 styles (WikiArt: impressionist, cubist, ukiyo-e, pop art, geometric abstraction). Each subject is trained with 3–6 reference images.

Methods. Textual Inversion (text embedding), DreamBooth-LoRA rank-4 (UNet attention), and Custom Diffusion (cross-attention K/V). This yields 45 personalized models.

Prompts. 200 prompts across 6 semantic categories: people (33), animals (33), objects (34), scenes (34), styles (33), and abstract (33). We use $K = 10$ seeds per prompt per model, yielding more than 90,000 generated images.

4 Results

4.1 Where in the Architecture Is Alignment Fragile?

Table 1 reports disruption by method and subject category. The aggregate entanglement score E follows a consistent ordering across all three subject categories:

$$E_{\text{TI}} < E_{\text{LoRA}} < E_{\text{CD}}. \quad (13)$$

This ordering is architecturally informative. Recall that TI perturbs the text embedding space, LoRA perturbs UNet attention, and CD perturbs cross-attention K/V projections. The aggregate ordering reveals that alignment is most fragile at the vision-language interface, intermediate within the vision backbone, and most robust in the text embedding space.

This result is counterintuitive. TI directly modifies the language representation by adding a new token to CLIP's vocabulary. If alignment fragility were determined by proximity to the language side, TI should cause the most disruption. Instead, the CLIP text encoder's learned representation is remarkably robust to vocabulary-level perturbation:

Table 1: Alignment disruption by method and subject category ($N = 5$ subjects each). E is the aggregate semantic-plus-perceptual entanglement score. The clean TI < LoRA < CD ordering holds for E across all subject categories; semantic-only E_{sem} is more subtle for faces.

Method	Category	$E_{\text{sem}} \downarrow$	$E \downarrow$	Std
TI	Face	0.021	0.220	0.024
	Object	0.021	0.235	0.018
	Style	0.019	0.267	0.034
LoRA	Face	0.020	0.222	0.016
	Object	0.021	0.235	0.015
	Style	0.026	0.313	0.007
CD	Face	0.019	0.244	0.012
	Object	0.027	0.276	0.019
	Style	0.030	0.354	0.011

adding a new token does not significantly destabilize existing tokens' grounding. The cross-attention projections W_K, W_V—where text features are projected into the visual feature space—are the fragile link. Perturbing these projections creates broad collateral effects because every text token's spatial grounding passes through these shared projection matrices.

Figure 1 complements Table 1 by separating semantic shift from perceptual drift. Most models remain below the semantic noise threshold on average, confirming that per-prompt semantic effects are subtle. At the same time, the distribution is not random: style and cross-attention-heavy personalizations occupy the high-drift region more often than face/text-space personalizations. This is why the paper treats entanglement as an aggregate diagnostic rather than a single-output failure detector.

4.2 What Kinds of Language Are Fragile?

The entanglement map (Figure 2) reveals that alignment fragility is not uniform across language types. Table 2 shows the most and least disrupted prompts. A clear dichotomy emerges: abstract, compositional language is fragile; concrete, specific language is robust. Prompts requiring compositional reasoning ("time frozen in mid-motion," "a landscape made of food") or stylistic interpretation ("a pop art comic panel") are highly disrupted. Prompts describing concrete visual scenes with strong prototypical structure ("a cup of coffee with latte art," "a crystal chandelier") are barely affected.

We hypothesize that this asymmetry reflects two distinct modes of vision-language grounding. Concrete prompts activate well-established prototypical grounding—strong, localized attention patterns learned from many training examples that are resistant to perturbation. Abstract prompts depend on compositional grounding—diffuse, context-dependent attention patterns assembled at inference time and therefore more sensitive to changes in the model's internal representations.

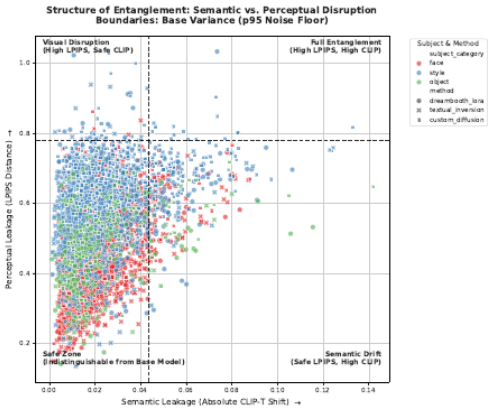

Figure 1: **Semantic vs. perceptual disruption.** The p95 base-model noise thresholds partition personalized models into safe, semantic-drift, visual-drift, and full-entanglement regions. The figure shows that the aggregate signal is structured even when many individual semantic shifts are below the noise floor.

Table 2: Alignment fragility varies systematically across prompt types. Prompts requiring compositional or abstract grounding are most fragile; concrete, visually specific prompts are most robust.

Most fragile prompts	$\bar{E}_{\text{sem}}$
a set of watercolor paints	0.053
a portrait of a young woman smiling	0.052
time frozen in mid-motion	0.050
a landscape made of food	0.048
a pop art comic panel	0.046
Most robust prompts	$\bar{E}_{\text{sem}}$
a chef cooking in a kitchen	0.009
a surfer riding a wave	0.009
a koala in a eucalyptus tree	0.009
a crystal chandelier	0.009
a cup of coffee with latte art	0.008

4.3 Does Disruption Follow Semantic Proximity?

The category-level analysis reveals an additional dimension of alignment structure. Style personalizations cause the most disruption (mean $E = 0.312$), followed by objects ($E = 0.248$) and faces ($E = 0.228$). This ordering, $E_{\text{face}} < E_{\text{object}} <$

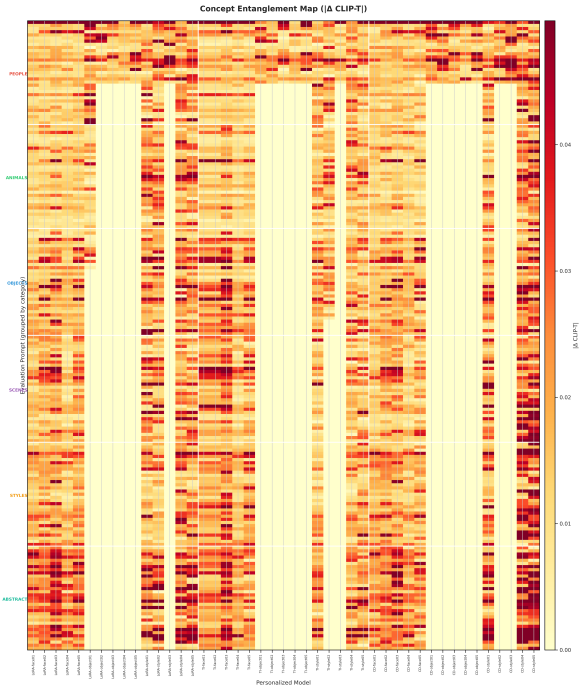

Figure 2: **Concept Entanglement Map**. Rows are prompts, columns are personalized models, and warmer colors indicate larger E_{sem}. The map reveals broad style disruption and high fragility for abstract prompts.

E_{style}, reflects the spatial scope of each concept type: faces and objects are local, while styles are global, affecting texture, palette, and composition everywhere. Global perturbations propagate more broadly through the cross-attention mechanism, disrupting alignment for a wider range of prompts.

4.4 FID Validation

To confirm that entanglement scores reflect meaningful visual change, we compute per-category FID between base and personalized model outputs. The FID ordering (style-CD $>$ style-LoRA $>$ face-LoRA $>$ face-TI) is fully consistent with our entanglement ordering. Style-CD produces the highest FID (abstract: 218–239; people: 226–275), while face-TI produces the lowest (animals: 44–49; scenes: 63–109), providing independent validation that the entanglement map captures real distributional change in output space.

5 Is Alignment Organized by Semantic Similarity?

The structure of the entanglement map allows us to test hypotheses about how vision-language alignment is organized.

5.1 H1: Semantic Proximity Predicts Fragility

Hypothesis. If alignment is organized by semantic similarity, face personalizations should disproportionately disrupt people-category prompts. We operationalize semantic proximity using the people prompt category; this is an intentionally coarse test because not every people prompt is face-specific.

Result: rejected ($p = 1.0$, Mann-Whitney U). Face personalizations produce lower disruption on people prompts ($\bar{E}_{\text{sem}} = 0.016$) than on non-people prompts ($\bar{E}_{\text{sem}} = 0.020$). The direction of the effect is opposite to the prediction.

This finding challenges a natural assumption about how vision-language models organize concept representations. Semantic similarity in language space does not predict alignment vulnerability in cross-attention space under this operationalization. Instead, personalization causes global perturbations to the vision backbone that affect prompts according to their grounding mode (compositional vs. prototypical) rather than their semantic relationship to the personalized subject.

5.2 H2: Styles Cause Broader Disruption

Hypothesis. Style personalizations cause more uniform cross-category disruption than face or object personalizations, because style is a global image property.

Result: confirmed ($p = 7.77 \times 10^{-15}$, Mann-Whitney U). Style personalizations produce significantly higher non-target disruption (mean $\bar{E}_{\text{sem}} = 0.024$) than face personalizations ($\bar{E}_{\text{sem}} = 0.020$). The effect is consistent across all six prompt categories.

Figure 3 visualizes this category-level structure. The most important pattern is not a single outlier cell, but the row-wise trend: abstract prompts are high across subject categories, while people prompts are not maximally affected by face personalization. This supports the interpretation that fragility is governed more by grounding mode and spatial scope than by simple semantic proximity in the text labels.

5.3 H3: Objects Are Intermediate

Object personalizations ($E = 0.248$) fall between faces ($E = 0.228$) and styles ($E = 0.312$), consistent with objects being spatially localized like faces but visually distinctive from the base model's priors, requiring stronger weight updates like styles.

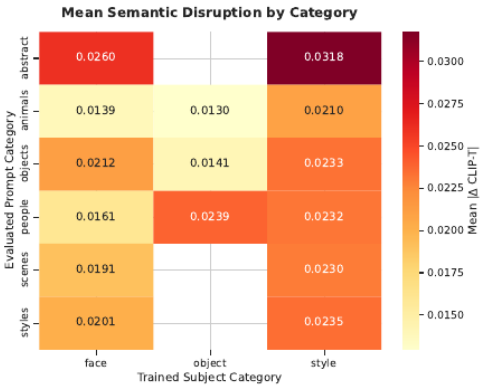

Figure 3: **Category × category disruption.** Mean semantic disruption for each trained subject category and evaluated prompt category. Abstract prompts are consistently fragile; the face→people cell is not the largest in the face column, rejecting H1 under our coarse people-vs.-non-people operationalization.

Table 3: Cross-attention shift for high- vs. low-disruption prompts (one face, one style subject). Values are $|\Delta\mathcal{A}| \times 10^4$. The analysis is preliminary and not used as standalone evidence for the main claims.

Prompt	Face	Style
High-disruption		
time frozen in mid-motion	4.65	3.67
portrait of young woman	3.67	4.09
watercolor paints	2.94	3.05
Low-disruption		
koala in eucalyptus	3.35	2.93
cup of coffee, latte art	3.50	3.25
crystal chandelier	2.31	2.81

6 Visualizing Fragility at the Attention Level

The alignment function $A(\theta, p)$ is ultimately mediated by cross-attention maps $\mathcal{A} \in \mathbb{R}^{H\times W\times T}$, where $\mathcal{A}[i, j, t]$ indicates how much spatial position (i, j) attends to text token t. If personalization disrupts alignment, it should manifest as shifts in these maps even for unrelated prompts. We test this by extracting cross-attention maps from both the base and personalized models during denoising for identical prompt-seed pairs and computing the mean absolute shift $|\Delta\mathcal{A}|$ across spatial positions and tokens.

Table 3 shows results for 6 prompts and 2 subjects. The magnitudes are small ($\sim 10^{-4}$), consistent with the subtle nature of per-prompt disruption. The ordering is suggestive but imperfect: the abstract prompt "time frozen in mid-motion" shows the largest attention shift for the face subject, while "crystal chandelier" shows the smallest; however, several low-disruption prompts have shifts comparable to high-disruption prompts. We therefore present this section as preliminary mechanistic evidence rather than definitive causal explanation. A comprehensive analysis across layers, timesteps, prompts, and subjects is needed to establish the attention mechanism conclusively.

7 Discussion

7.1 Three Properties of Vision-Language Alignment

Our results reveal three structural properties of vision-language alignment in SD v1.5 personalization.

Property 1: Text-space alignment is robust; vision-space alignment is fragile. The CLIP text encoder provides a stable grounding substrate. Adding a new token through Textual Inversion does not significantly destabilize existing tokens' representations, likely because CLIP's large-scale image-text pretraining creates a well-regularized embedding space (Radford et al., 2021). In contrast, the UNet's cross-attention projections—shared matrices W_K, W_V through which every text token must pass to reach visual features—are fragile. Modifying these projections for one concept disrupts the grounding pathway for many concepts.

Property 2: Compositional grounding is fragile; prototypical grounding is robust. Abstract prompts requiring novel visual arrangements are more fragile than concrete prompts that activate prototypical visual patterns. This suggests that diffusion models rely on at least two grounding regimes: a robust, lookup-like mechanism for common visual prototypes, and a fragile, inference-time compositional mechanism for novel combinations.

Property 3: Alignment is globally, not purely semantically, organized. The rejection of H1 indicates that disruption does not follow semantic proximity in the coarse people-vs.-non-people sense. Personalizing for a face does not preferentially degrade grounding for people prompts. Instead, concepts appear to share representational substrate in ways that do not simply mirror their semantic relationships in language.

7.2 The Entanglement Map as a Diagnostic

The entanglement map is a practical diagnostic tool. Before deploying a personalized model, practitioners can generate the map against a prompt suite relevant to their application to identify which prompts are most affected and whether disruption exceeds acceptable thresholds. The map also enables method selection: if an application requires robust style-related prompts, the map shows that text-space perturbations preserve these prompts better than cross-attention perturbations.

A practical audit workflow has three steps: choose a deployment-relevant prompt suite, compute a seed-matched entanglement map against the base model, and inspect both aggregate summaries and the highest-shift prompts. The aggregate view identifies whether a personalization method is broadly risky; the prompt-level view identifies concrete prompts that should be manually reviewed before release.

7.3 Scope and Practical Significance

The per-prompt semantic effects we measure are subtle and usually below the noise floor. The practical significance of any individual prompt-level shift should therefore not be overstated. The contribution of this work is aggregate and diagnostic: small shifts, when structured consistently across thousands of prompt-model comparisons, reveal how alignment is organized and where it is vulnerable. Whether these shifts are perceptible to end users in each application depends on deployment context and should be validated with task-specific human evaluation.

8 Conclusion

We have shown that vision-language alignment in text-to-image diffusion models is fragile in structured, predictable ways. Fine-tuning for a single concept shifts alignment for unrelated concepts, and the pattern of disruption reveals how alignment is organized: robustly in the text embedding space, fragilely at the cross-attention interface; robustly for concrete visual prototypes, fragilely for abstract compositional concepts; globally rather than purely along semantic boundaries. These findings expose a structural vulnerability in personalized generative systems and provide both a diagnostic tool (the Concept Entanglement Map) and concrete guidance (prefer less invasive perturbations, monitor abstract prompts, and evaluate non-target behavior) for practitioners working with personalized text-to-image models.

Limitations

Our study uses Stable Diffusion v1.5 exclusively; generalization to newer architectures such as SDXL, Stable Diffusion 3, Flux, or other DiT-based models is untested. We evaluate 200 prompts, which may not cover all semantic categories relevant to specific applications. The noise floor analysis reveals that individual per-prompt semantic disruptions are typically within stochastic variation; our statistical claims are based on aggregate patterns across many prompts and models. The cross-attention analysis is preliminary (6 prompts, 2 subjects), and attention shift magnitudes are small. Finally, the compositional-vs.-prototypical grounding hypothesis is observational; establishing causality would require controlled interventions on model internals.

Ethics Statement

This work studies personalized text-to-image models, including face personalization, as a diagnostic probe. Personalized generative models can be misused for impersonation, non-consensual identity generation, or style imitation. Our analysis is intended to improve auditing of such models by measuring non-target alignment shifts, not to improve deceptive generation. Any release of trained adapters or generated examples should respect dataset licenses, subject consent, and privacy constraints.

References

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, pages 12606–12633.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2023. An image is worth one word: Personalizing

text-to-image generation using textual inversion. In *International Conference on Learning Representations*.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528. Association for Computational Linguistics.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, pages 6626–6637.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1882–1891.

Zhizhong Li and Derek Hoiem. 2016. Learning without forgetting. In *European Conference on Computer Vision*, pages 614–629.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*.

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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912. Association for Computational Linguistics.

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

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. 2023. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5644–5659. Association for Computational Linguistics.

Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. 2024. Detecting, explaining, and mitigating memorization in diffusion models. In *International Conference on Learning Representations*.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595.