SAGE: Semantically Disentangled Representation Learning through Latent Geometry Constraint and Large Language Model

Qiuyu Chen[1,2] **Liang Xu**[3] **Yunnan Wang**[1,2] **Mingqi Yuan**[4]
Qi Wang[5] **Jiahao Li**[6] **Zhicheng Wang**[2,7] **Hao Zheng**[2]
Mulin Chen[8] **Baao Xie**[2,9,*] **Xin Jin**[2,9] **Wenjun Zeng**[2,9]
[1]Shanghai Jiao Tong University
[2]Ningbo Key Laboratory of Spatial Intelligence and Digital Derivative,
Ningbo Institute of Digital Twin, Eastern Institute of Technology, Zhejiang, China
[3]Zhongguancun Academy, Beijing, China
[4]The University of Hong Kong
[5]Institute of Automation, Chinese Academy of Sciences
[6]Fudan University
[7]The Hong Kong Polytechnic University [8]Tsinghua University
[9]Zhejiang Key Laboratory of Industrial Intelligence and Digital Twin,
Ningbo, Zhejiang, China
[*]Corresponding author.

Abstract

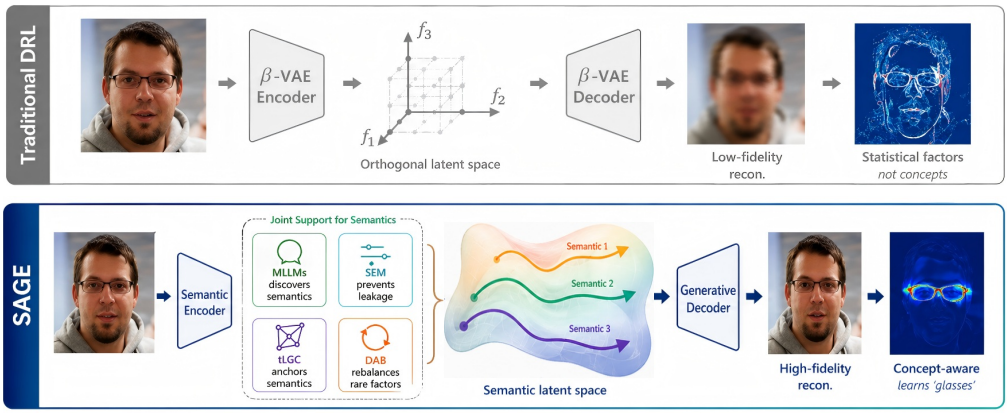
Disentangled representation learning (DRL) aims to identify and decompose the interpretable underlying factors of observations. However, most DRL methods rely on independence-oriented regularization to pursue statistical factorization, leaving learned factors without explicit semantic grounding. Such independence-oriented regularization induces an intrinsic trade-off between disentanglement and reconstruction. To address these issues, we establish a spectral learnability threshold for independence-oriented DRL, theoretically revealing a mismatch between the factors favored by the learning objective and human-perceivable semantics. We further introduce SAGE, a semantic-guided DRL framework that leverages MLLMs to automatically extract semantic signals from unlabeled data and uses a triangular latent-geometry constraint to anchor each semantic factor to a designated latent dimension, enabling factor-wise control over human-perceivable semantic variations in generated images. Extensive experiments support our theoretical analysis and demonstrate that SAGE achieves state-of-the-art performance across multiple benchmarks. In addition, our framework exhibits high flexibility and can be integrated with most advanced generative models.

1 Introduction

Disentangled Representation Learning (DRL) aims to decompose observations into semantically interpretable and independent factors, a paradigm inspired by the way biological intelligence perceives and comprehends the real world. In this form, representational attributes (*e.g.*, expressions, wrinkles, and accessories of human face) exhibit exclusive sensitivity to the changes of specific factors. Learning such disentangled representations is crucial across a wide range of domains, *e.g.*, computer vision [1–3], generative modeling [4–6], neuroscience [7–9], and interpretable artificial intelligence [10].

As shown in Figure 1, most DRL methods adopt a β-VAE-inspired paradigm [11], using independence-oriented regularization to learn statistically independent latent factors. This paradigm is mathematically well motivated and has substantially advanced disentanglement research. However, such

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



SAGE: Semantics-Aware and Geometry-Enhanced DRL framework

Figure 1: SAGE is a semantic-guided DRL framework that constructs a semantically disentangled and interpretable latent space, enabling factor-wise control over human-perceivable semantics while preserving high-fidelity reconstruction.

statistical factorization does not ensure that each learned factor is grounded in human-perceivable semantics. This limitation leaves two longstanding and fundamental problems unresolved:

- *The **conflation** of statistical factors with semantic factors*
  Current methods typically achieve disentanglement by constraining the latent distribution toward independence. From a compression perspective, this objective encourages the retained factors to capture information-rich variations. Although such factors may coincide with human-perceivable semantics, this correspondence is not guaranteed, as high-information statistical patterns may entangle multiple semantics or lack a coherent semantic interpretation.
- *The **intrinsic trade-off** between disentanglement and reconstruction fidelity*
  Strengthening independence-oriented regularization can improve disentanglement by suppressing dependencies among latent factors. However, this suppression also restricts the information capacity of the latent code, causing fine-grained reconstruction details to be discarded [12]. We refer to this phenomenon as the disentanglement–reconstruction fidelity trade-off, or DR-tradeoff.

The above observations prompt us to rethink the very notion of 'disentanglement':

Is it possible to move beyond independence-oriented DRL and instead construct a semantically disentangled latent space that mitigates the DR-tradeoff?

To this end, we first establish a spectral learnability threshold for independence-oriented DRL, theoretically explaining why such objectives may favor statistically dominant factors over human-perceivable semantics and induce the DR-tradeoff. We further propose a novel **S**emantics-**A**ware and **G**eometry-**E**nhanced DRL framework, termed **SAGE**, to construct a semantically disentangled latent space without enforcing statistical independence. Specifically, a Factor Discovery Module (FDM) uses Multimodal Large Language Models (MLLMs) as semantic proxies to discover semantic factors from unlabeled data. A Triangular Latent Geometry Constraint (tLGC) restricts latent rotational freedom and provides a stable coordinate frame, allowing semantic guidance to anchor each discovered factor to a designated latent dimension. To guarantee a bijective mapping between latent dimensions and semantic factors, we further introduce a Semantic Exclusiveness Mechanism (SEM) that penalizes redundant encoding of the same concept in non-designated dimensions. In addition, a distribution-aware rebalancing principle is proposed to address the long-tailed challenge during training. With these designs, SAGE enables factor-wise control over human-perceivable semantics while preserving high-fidelity reconstruction. Our contributions can be summarized as:

- **Theoretically,** we establish a spectral learnability threshold for independence-oriented DRL, revealing how statistical optimization objectives diverge from semantic factor learning and give rise to the DR-tradeoff.
- **Methodologically,** we propose SAGE, a semantic-guided DRL framework that constructs a semantically disentangled and interpretable latent space via a set of synergistic modules. Our model enables semantic-aware disentanglement while effectively mitigating the intrinsic DR-tradeoff.
- **Empirically,** extensive experiments demonstrate that SAGE achieves state-of-the-art disentanglement and reconstruction across multiple benchmarks. Output-space probing further isolates

the contribution of our geometric framework, while backbone studies verify its flexibility across advanced generative models.

2 Related Work

Independence-oriented DRL. DRL aims to recover underlying factors, aligning each latent axis with a distinct attribute [13]. Most DRL methods are grounded in the β-VAE [11] framework, which encourages posterior factorization and statistically independent latent factors by up-weighting the KL penalty. This paradigm has been advanced by variants such as β-TCVAE [14], DIP-VAE [15], FactorVAE [16], RF-VAE [17], and α-TCVAE [18] via regularization optimizations. However, such independence-oriented objectives use statistical factorization for disentanglement and do not ensure that the learned factors are grounded in semantics.

Semantic guidance and latent geometry constraint. To address this mismatch and the induced DR-tradeoff [19, 20], two complementary lines of work have emerged. The first introduces semantic guidance from limited labels [21, 22], paired observations [23, 24], or foundation-model signals [25, 26]. The second imposes structured latent geometries [27], such as Hyperspherical [28, 29] or Riemannian VAEs [30, 31], or factor-graph priors [32, 33]. However, conventional disentanglement metrics, such as MIG [14], DCI [34], FactorVAE Score (F-score) [16], and TAD [35], mainly evaluate whether factor information is recoverable from the latent representation, but do not verify whether decoded visual changes are consistent with the supervised semantics. SAGE addresses the problem through latent geometry constraints and introduces a new metric to validate its effectiveness.

Automatic semantic factor discovery. Concept-based vision methods seek human-interpretable concepts through concept activation, automatic concept discovery, or language-assisted concept labeling [36–41]. Recent DRL work further uses MLLMs to mine semantic attributes and construct graph priors [42]. FDM further uses MLLMs as semantic proxies to automatically induce a data-adaptive factor schema and prompt dictionary from unlabeled images. Unlike fixed attribute vocabularies or direct CLIP labeling, this enables semantic guidance to be constructed adaptively across datasets and visual domains.

3 Theoretical Analysis

This section lays the theoretical groundwork for SAGE. Following the preliminaries (Section 3.1), we analyze how traditional DRL simplifies disentanglement to rate-constrained compression, leading to semantic dropouts and a DR-tradeoff (Section 3.2). We then elucidate the risks of naive semantic supervision (Section 3.3) and conclude with the theoretical justification for our paradigm (Section 3.4).

3.1 Preliminaries

Let $\mathbf{x} \in \mathbb{R}^{D_x}$ denote an observation and $\mathbf{y} = (y_1, \dots, y_M) \in \{0,1\}^M$ denote binary semantic factor indicators. We consider a β-VAE with isotropic Gaussian prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. The latent code $\mathbf{z} \in \mathbb{R}^d$ is partitioned into semantic coordinates $\mathbf{z}_S \in \mathbb{R}^{d_s}$ and unguided coordinates $\mathbf{z}_U \in \mathbb{R}^{d_u}$, with $d_s + d_u = d$; in SAGE, this partition is realized after the triangular latent transform $\mathbf{z}^{sem} = \mathbf{L}\mathbf{z}$. For analysis, we decompose the observation into a semantic conditional mean and a residual $\mathbf{r}(\mathbf{x}) \triangleq \mathbf{x} - \mathbb{E}[\mathbf{x} \mid \mathbf{y}]$, with $\mathbf{\Sigma}_r = \mathbb{E}[\mathbf{r}\mathbf{r}^\top]$. Throughout, $\mathrm{CE}(p, y) \triangleq -y \log p - (1-y)\log(1-p)$ denotes binary cross-entropy / negative log-likelihood; for the multi-class predictors used by SEM, the same symbol denotes its categorical generalization $-\log q(\mathbf{y} \mid \cdot)$. Following Lucas et al. [43], we study local optimization dynamics with a linearized decoder $g_\theta(\mathbf{z}) \approx g_\theta(\mathbf{z}_0) + \mathbf{J}(\mathbf{z} - \mathbf{z}_0)$, consistent with the Neural Tangent Kernel regime [44].

3.2 Unsupervised DRL from a Compression Perspective

Theorem 3.1. *Assume the linearized Gaussian regime $\mathbf{x} \approx \mathbf{J}\mathbf{z} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\mathrm{noise}}\mathbf{I})$. Let $\gamma_1 \geq \cdots \geq \gamma_r$ denote the eigenvalues of the local data covariance $\mathbf{\Sigma}_x \triangleq \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ expressed in the singular basis of $\mathbf{J}$, where $r = \min(D_x, d)$ is the rank of the linearized decoder. For a β-VAE with a diagonal Gaussian variational family, the globally optimal solution activates mode k if and only if*

$$\gamma_k > \beta\sigma^2_{\mathrm{noise}}. \tag{1}$$

Otherwise, the mode collapses to the prior.

Remark 3.2 (Information-Semantic Misalignment). Theorem 3.1 suggests that independence-oriented unsupervised DRL can impose an intrinsic information lower bound for factor learnability. Stronger independence-oriented or rate-constraining regularization raises this effective bound (analogous to increasing β), thereby inducing posterior collapse and the DR-tradeoff. Crucially, as this compression is semantic-agnostic, the surviving factors are biased toward statistically dominant directions, which may not align with human-perceivable semantic concepts (see Appendix D.2 for a controlled counterexample). While the exact threshold is established in the linearized Gaussian regime, the same energy-ordering intuition provides a useful perspective for nonlinear models: local Jacobian linearization motivates the effective factor energy $\bar{E}_m = \mathrm{Var}(s_m)\,\mathbb{E}_{\mathbf{s}}\|\partial g/\partial s_m\|_2^2$ as a nonlinear counterpart to the spectral energy γ_k in Equation (1). The resulting collapse-ordering tendency is further supported by our nonlinear FactorVAE experiments in Figure 3(d). A full proof of the linear threshold and the derivation of the nonlinear effective factor energy are provided in Appendix C.1.

3.3 The Dilemma of Naive Semantic Supervision

Theorem 3.3 (Information leakage)**.** *Let $\mathbf{c}_U = h(\mathbf{P}_U\mathbf{r})$ be a finite-valued statistic of the residual projection $\mathbf{P}_U\mathbf{r}$ along reconstruction-favored directions, and suppose it is class-dependent with $I(\mathbf{y};\mathbf{c}_U) = \Delta_U > 0$. If the unguided code retains this statistic with $H(\mathbf{c}_U \mid \mathbf{z}_U) \le \epsilon_U < \Delta_U$ (Assumption C.2), then*

$$I(\mathbf{y};\mathbf{z}_U) \;\ge\; \Delta_U - \epsilon_U \;>\; 0, \tag{2}$$

i.e., semantic information leaks into the unguided subspace.

Remark 3.4. Theorem 3.3 identifies rotational symmetry as the structural origin of leakage. The model exploits the orthogonal degrees of freedom in $\mathbf{z}_U$ to align with variance-dense residuals, thereby bypassing supervision. Hence, a latent geometric constraint is necessary to break this symmetry and prevent leakage for semantically grounded DRL (Appendix C.3).

Another common supervision scheme also falls short: *full supervision* forces $\mathbf{z}$ to be a function of $\mathbf{y}$ alone, so reconstruction error is lower-bounded by the residual energy $\mathrm{Tr}(\mathbf{\Sigma}_r)$ (Proposition C.3, Appendix C.4).

3.4 Our Solution

To resolve the issues in Section 3.2 and Section 3.3, and to construct a semantically disentangled and interpretable latent space without enforcing statistical independence, this section presents the theoretical justification of the core modules of SAGE.

Theorem 3.5 (Break Rotational Symmetry in Latent Space)**.** *Let $\mathcal{T}_K^+$ denote the group of $K \times K$ lower-triangular matrices with positive diagonal entries. Then*

$$\mathcal{T}_K^+ \cap O(K) = \{\mathbf{I}_K\}. \tag{3}$$

Hence, when the semantic latent transform $\mathbf{L}$ is restricted to $\mathcal{T}_K^+$, no non-trivial orthogonal rotation $\mathbf{R} \in O(K)$ can be absorbed within the same transform family.

Remark 3.6. Theorem 3.5 provides the geometric foundation for preventing the information leakage identified in Theorem 3.3. By constraining the latent transformation to $\mathcal{T}_K^+$, we eliminate the zero-cost orthogonal degrees of freedom that the reconstruction objective exploits to align unguided dimensions with residual variance. Thus, tLGC fixes a canonical semantic coordinate frame and helps prevent the leakage mechanism identified in Theorem 3.3 (Appendix C.5).

With the rotational freedom restricted by tLGC, semantic guidance can attach discovered factors to designated coordinates, but non-designated coordinates may still re-encode the same semantics for reconstruction. SAGE therefore introduces SEM to suppress semantic leakage into non-designated dimensions, and distribution-aware balancing to preserve rare factors under long-tailed supervision. Formal statements and proofs are provided in Theorems C.4 and C.6 and Appendices C.6, C.7.

4 Methodology

4.1 Overview

Guided by Section 3, SAGE constructs a semantically disentangled and interpretable latent space through four modules. FDM adaptively discovers semantic factors and prompt dictionaries from

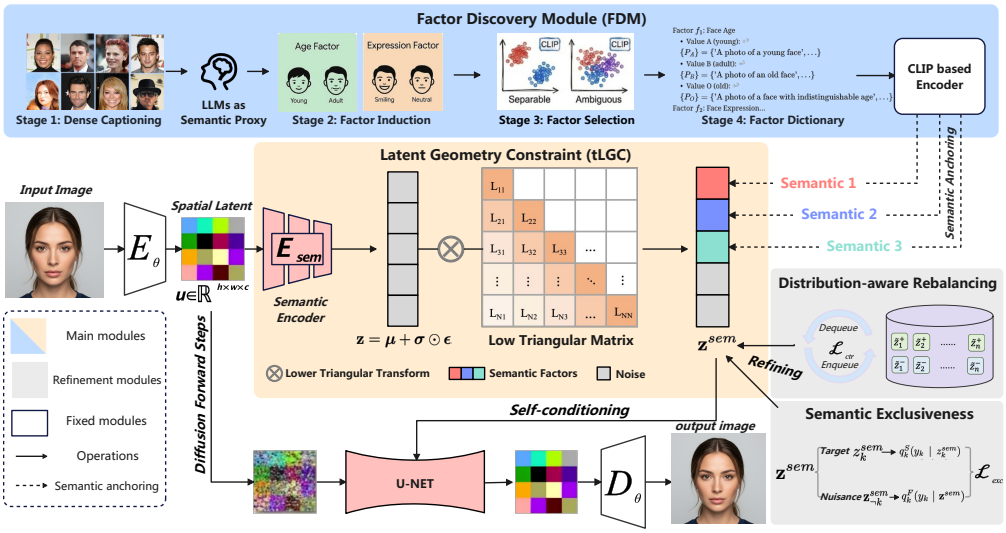


Figure 2: **Overview of the SAGE framework.** The model comprises several synergistic modules designed to construct a semantically disentangled and interpretable latent code z^{sem}. Specifically, the FDM is proposed for semantic extraction, while tLGC stabilizes the latent coordinate frame for semantic guidance. Furthermore, the semantic exclusiveness mechanism and distribution-aware balancing principle are employed to refine the latent space. While we implemented our framework using a diffusion model in this work, the architecture exhibits high flexibility and can be seamlessly integrated with other generative models (see Appendix H.1 for a VAE-based implementation).

unlabeled images; tLGC is used for stabilizing the semantic coordinate frame, allowing semantic guidance to assign discovered factors to designated latent dimensions; SEM suppresses semantic leakage into non-designated dimensions; and a distribution-aware balancing principle stabilizes rare factors. We instantiate SAGE on a latent diffusion backbone, with a VAE-based implementation provided in Appendix H.1.

4.2 Framework Backbone

As shown in Figure 2, we present the diffusion implementation version of SAGE in this work. Specifically, given an input image $\mathbf{x}$, we obtain a spatial latent $\mathbf{u}_0 = E_\theta(\mathbf{x})$ via the encoder of the latent diffusion model. Then, a lightweight semantic encoder E_{sem} maps $\mathbf{u}_0$ to a Gaussian posterior $(\boldsymbol{\mu}(\mathbf{x}), \log \boldsymbol{\Sigma}(\mathbf{x})) = E_{sem}(\mathbf{u}_0)$, and $\mathbf{z}$ is sampled by $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Let $\{\alpha_t\}_{t=1}^T$ be a fixed variance schedule and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$. We sample $t \sim \mathrm{Unif}\{1, \ldots, T\}$ and noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and formulate the forward noised latent as $\mathbf{u}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{u}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}$. A conditional U-Net ϵ_ψ is employed to predict the noise given the noisy latent $\mathbf{u}_t$, current timestep t, and the semantic condition $\mathbf{z}^{sem} = \mathbf{L}\mathbf{z}$. The model is trained using the standard noise-prediction loss:

$$\mathcal{L}_{\text{diff}} \;=\; \mathbb{E}_{t,\boldsymbol{\epsilon},\mathbf{x}}\Big[\big\|\boldsymbol{\epsilon} - \epsilon_\psi(\mathbf{u}_t, t, \mathbf{z}^{sem})\big\|_2^2\Big]. \tag{4}$$

The reconstructed image is obtained via $\hat{\mathbf{x}} = D_\varphi(\mathbf{u}_0)$.

4.3 Factor Discovery Module

Most disentanglement pipelines depend on manually curated attribute lists, which severely limit scalability and cross-domain transfer. SAGE removes this dependency via a Factor Discovery Module (FDM) that, from *unlabeled* images $\{\mathbf{x}_i\}_{i=1}^N$, *adaptively discovers* a compact set of semantic factors and *constructs* a robust prompt dictionary using an MLLM. The resulting factor set generalizes across datasets and visual domains rather than committing to any fixed attribute vocabulary. As shown at the top of Figure 2, FDM runs offline and consists of four stages: structured captioning, factor induction, factor selection, and factor dictionary construction.

Structured captioning and factor induction. We first query an MLLM for detailed descriptions of the visible attributes in each image $\mathbf{x}_i$ (e.g., color and shape), which are then organized into consistent

attribute-centric fields across images [45]. An LLM aggregates these descriptions by identifying recurring attributes, merging synonymous expressions, and normalizing their values into a canonical schema $\tilde{\mathcal{F}} = \{\tilde{f}_k\}$, where each $\tilde{f}_k$ has a canonical name and a set of value-descriptive captions $\tilde{\mathcal{V}}_k$, with one canonical caption per candidate value. These captions serve as prompts for CLIP-based image–text matching in the next stage.

Factor selection. For each candidate $\tilde{f}_k$ and value $v \in \tilde{\mathcal{V}}_k$, we estimate the empirical prevalence $\hat{\pi}_k(v) \triangleq \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[\tilde{f}_k(\mathbf{x}_i) = v]$. To discard sporadic or nearly constant attributes, FDM requires sufficient coverage ($C_k = \sum_v \hat{\pi}_k(v) \geq \tau_{\text{cov}}$) and non-degeneracy ($B_k = |\{v : \hat{\pi}_k(v) > \tau_{\min}\}| \geq 2$), ensuring that a factor is observable and that at least two of its values have meaningful support. We further assess whether these values are visually distinguishable using the CLIP margin on a held-out subset $\mathcal{H}_k$:

$$S_k = \frac{1}{|\mathcal{H}_k|} \sum_{i \in \mathcal{H}_k} \left[\max_{v \in \tilde{\mathcal{V}}_k} \rho_{i,k}(v) - \max_{v' \neq v_i^\star} \rho_{i,k}(v') \right], \qquad v_i^\star = \arg\max_{v \in \tilde{\mathcal{V}}_k} \rho_{i,k}(v), \tag{5}$$

where $\rho_{i,k}(v)$ is the CLIP image–text matching score between $\mathbf{x}_i$ and the caption v. Combining these three criteria, the retained factor set is $\mathcal{F} = \{f_k : C_k \geq \tau_{\text{cov}},\ B_k \geq 2,\ S_k \geq \tau_{\text{sep}}\}_{k=1}^{K}$.

Factor dictionary construction. The final FDM output is the factor dictionary $\mathcal{D} = \{(f_k,\ \mathcal{V}_k)\}_{k=1}^{K}$, where each retained factor f_k is paired with its value-descriptive caption set $\mathcal{V}_k$. These captions ground each value in the intended visual subject, allowing a frozen CLIP module to match every training image $\mathbf{x}_i$ against the factor descriptions. For binary guidance, two representative values are selected, while the remaining values form an "Other" category; CLIP matching then produces pseudo-labels $y_{i,k}$ and validity masks $m_{i,k}$, with "Other" assignments excluded from the downstream semantic losses. Further details are deferred to Appendix E.1.

4.4 Latent Geometry Constraint and Semantic Guidance

Triangular latent geometry. To instantiate the geometric anchoring principle in Eq. (3), SAGE applies an invertible lower-triangular transform with positive diagonal entries:

$$\mathbf{z}^{sem} = \mathbf{L}\mathbf{z}, \qquad \mathbf{L} \in \mathbb{R}^{d \times d}, \quad \mathbf{L}_{ii} = \exp(\ell_{ii}). \tag{6}$$

The first K coordinates of $\mathbf{z}^{sem}$ are used as designated semantic dimensions, with $z_{i,k}^{sem}$ corresponding to factor f_k.

Guidance and contrastive separation. Given pseudo-labels $y_{i,k}$ and masks $m_{i,k}$ from FDM, the designated coordinate $z_{i,k}^{sem}$ is supervised by a masked BCE loss:

$$\mathcal{L}_{\text{bce}}^{(k)} = \frac{1}{\sum_i m_{i,k}} \sum_i m_{i,k}\, \text{CE}\big(\sigma(z_{i,k}^{sem}), y_{i,k}\big). \tag{7}$$

We further use a margin contrastive loss on each semantic coordinate:

$$\mathcal{L}_{\text{ctr}}^{(k)} = \frac{1}{|\mathcal{I}_k^+|} \sum_{(i,j) \in \mathcal{I}_k^+} \big(z_{i,k}^{sem} - z_{j,k}^{sem}\big)^2 + \frac{1}{|\mathcal{I}_k^-|} \sum_{(i,j) \in \mathcal{I}_k^-} \Big[\max\big(0, m_{ctr} - |z_{i,k}^{sem} - z_{j,k}^{sem}|\big)\Big]^2, \tag{8}$$

where $\mathcal{I}_k^+$ and $\mathcal{I}_k^-$ are same-label and different-label pairs sampled from the memory queues.

Semantic exclusiveness. Following the CE-gain principle in Eq. (30), SEM uses a semantic-only predictor $q_k^S(y_k|z_k^{sem})$ and a full-latent predictor $q_k^F(y_k|z_k^{sem}, \mathbf{z}_{\neg k}^{sem})$, where $\mathbf{z}_{i,\neg k}^{sem}$ denotes all non-designated coordinates and $\text{sg}(\cdot)$ denotes stop-gradient. The predictors are trained on detached latents:

$$\mathcal{L}_{\text{pred}} = \frac{1}{\sum_{i,k} m_{i,k}} \sum_{i,k} m_{i,k} \Big[\text{CE}\big(q_k^S(\text{sg}(z_{i,k}^{sem})), y_{i,k}\big) + \text{CE}\big(q_k^F(\text{sg}(z_{i,k}^{sem}), \text{sg}(\mathbf{z}_{i,\neg k}^{sem})), y_{i,k}\big)\Big]. \tag{9}$$

With predictor weights frozen, the encoder minimizes the hinged CE-gain:

$$\mathcal{L}_{\text{exc}} = \frac{1}{\sum_{i,k} m_{i,k}} \sum_{i,k} m_{i,k} \Big[\text{CE}\big(q_k^S(\text{sg}(z_{i,k}^{sem})), y_{i,k}\big) - \text{CE}\big(q_k^F(\text{sg}(z_{i,k}^{sem}), \mathbf{z}_{i,\neg k}^{sem}), y_{i,k}\big)\Big]_+, \tag{10}$$

where $[x]_+ = \max(x, 0)$.

Distribution-aware balancing. For each factor, SAGE maintains FIFO queues $\mathcal{Q}_k^+$ and $\mathcal{Q}_k^-$ to estimate the positive prevalence $\hat{\pi}_k$ and sample balanced contrastive pairs. We define $w_k = 1/\sqrt{\hat{\pi}_k(1-\hat{\pi}_k)+\epsilon}$ and combine the guided objectives as $\mathcal{L}_{\text{guid}} = \frac{1}{K}\sum_{k=1}^{K} w_k\big(\mathcal{L}_{\text{bce}}^{(k)} + \mathcal{L}_{\text{ctr}}^{(k)}\big)$.

Overall objective. SAGE is trained end-to-end with

$$\mathcal{L}_{\text{total}} = \lambda_{\text{diff}}\mathcal{L}_{\text{diff}} + \lambda_{\text{guid}}\mathcal{L}_{\text{guid}} + \lambda_{\text{exc}}\mathcal{L}_{\text{exc}}. \tag{11}$$

Table 1: Disentanglement and output-space controllability. For conventional baselines, vanilla → +guid reports results before and after applying the same semantic pseudo-label supervision used by SAGE. A dash denotes that the corresponding unguided configuration is not applicable to native semantic-guided methods.

Family	Method	Shapes3D			CelebA		
		F-score ↑	DCI ↑	Probe ↑	TAD ↑	DCI ↑	Probe ↑
VAE	β-VAE	0.820 → 0.989	0.585 → 0.614	0.317 → 0.264	0.088 → 0.534	0.090 → 0.094	0.221 → 0.249
	FactorVAE	0.840 → 0.893	0.611 → 0.514	0.270 → 0.442	0.129 → 0.298	0.079 → 0.086	0.179 → 0.172
	β-TCVAE	0.873 → 0.749	0.613 → 0.717	0.412 → 0.426	0.353 → 0.123	0.039 → 0.037	0.169 → 0.147
	α-TCVAE	0.846 → 0.869	0.663 → 0.671	0.371 → 0.312	0.083 → 0.094	0.035 → 0.036	0.251 → 0.236
GAN	InfoGAN-CR	0.587 → **0.999**	0.478 → **0.999**	0.493 → 0.220	0.052 → 0.652	0.076 → 0.237	0.189 → 0.552
	DisCo	0.877 → **0.999**	0.708 → **0.999**	0.127 → 0.127	0.330 → 0.400	0.354 → 0.577	0.154 → 0.155
Diffusion	InfoDiffusion	0.925 → **0.999**	0.690 → **0.999**	0.080 → 0.076	0.299 → 0.415	0.341 → 0.481	0.205 → 0.192
	DisDiff	0.902 → 0.997	0.723 → 0.906	0.073 → 0.417	0.305 → 0.785	0.183 → 0.641	0.175 → 0.078
	EncDiff	**0.999** → **0.999**	0.969 → 0.997	0.124 → 0.246	0.638 → 0.681	0.076 → 0.268	0.192 → 0.114
	DiffAE	0.496 → **0.999**	0.322 → **0.999**	0.112 → 0.172	0.055 → 0.170	0.354 → 0.588	0.032 → 0.168
Semantic-guided	GEM	- → 0.928	- → 0.796	- → 0.635	- → 0.464	- → 0.148	- → 0.586
	ZeroDIM	- → 0.783	- → **0.999**	- → 0.763	- → 0.408	- → 0.173	- → 0.338
	TA-Dis	- → **0.999**	- → 0.920	- → 0.613	- → 0.710	- → 0.424	- → 0.381
Ours	**SAGE**	**0.999**	**0.999**	**0.807**	**0.803**	**0.655**	**0.762**

5 Experiments

In this section, SAGE is empirically validated to answer four core questions derived from our theoretical contributions:

- **Q1:** Does the spectral learnability threshold hold in linear models, and can it provide a useful principle for extending to nonlinear disentanglement models?
- **Q2:** Can SAGE achieve strong disentanglement and decoder-side controllability across diverse datasets, and do matched-supervision comparisons show that the gains come from its architecture rather than supervision alone?
- **Q3:** Can SAGE improve semantic controllability while preserving high-fidelity reconstruction, thereby mitigating the DR-tradeoff?
- **Q4:** How does each module individually contribute to the effectiveness of SAGE?

5.1 Experimental Setup

Datasets and baselines. We evaluate SAGE on five typical datasets: Shapes3D [16], CelebA [46], FFHQ [47], AFHQ-v2 [48], and CompCars [49]. We benchmark SAGE against FactorVAE [16], β-VAE [11], β-TCVAE [14], DIP-VAE [15], RF-VAE [17], α-TCVAE [18], InfoGAN-CR [50], GANspace [51], DisCo [52], Diff-AE [53], InfoDiffusion [54], DisDiff [55], CL-Dis [56], EncDiff [57], GEM [42], ZeroDIM [58], and TA-Dis [59]. Details of the dataset and implementation settings are provided in Appendix G.

5.2 Theoretical Validation

To answer Q1, Figure 3 validates Theorem 3.1 from linear exactness to nonlinear extension. Panels (a,b) use a one-dimensional linear β-VAE to show that the normalized ratio $\gamma/(\beta\sigma^2)$ governs both decoder activation w^2 and posterior variance ν^2, where w is the decoder weight from latent to observation and ν^2 is the variance of $q(z|x)$. Panel (c) extends this mechanism to an 8D staircase model: as β increases, only high-information modes remain active, yielding stronger statistical

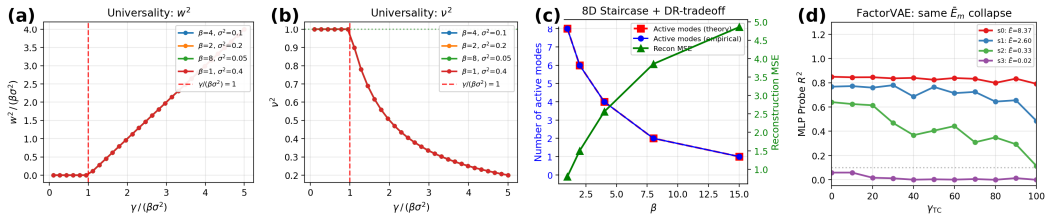


Figure 3: **Empirical validation of the spectral threshold.** (a) Decoder activation w^2 collapses under the normalized threshold $\gamma/(\beta\sigma^2)$. (b) Posterior variance ν^2 follows the same threshold behavior. (c) 8D staircase: increasing β prunes low-energy modes and raises reconstruction MSE, revealing the DR-tradeoff. (d) Nonlinear FactorVAE: factor collapse follows the effective-energy ordering $\bar{E}_m$.

disentanglement but higher reconstruction MSE, which exposes the source of the DR-tradeoff. Panel (d) evaluates the nonlinear case using the effective factor energy $\bar{E}_m$ derived from Theorem 3.1; under increasing disentanglement strength, factor collapse follows the predicted energy ordering. These results show that the theorem is exact in the linear regime and remains useful for understanding factor learnability in nonlinear DRL models. Full details are provided in Appendix D.1.

5.3 Qualitative and Quantitative Results

To answer Q2, we evaluate whether semantic guidance is expressed as controllable visual changes in decoded images, rather than merely improving latent–label alignment. To this end, we introduce **Probe**, an output-space metric defined as $\text{Probe} = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_i\left[|A_k(\hat{\mathbf{x}}^+_{i,k}) - A_k(\hat{\mathbf{x}}^-_{i,k})|\right]$, where A_k is an attribute classifier and $\hat{\mathbf{x}}^+_{i,k}, \hat{\mathbf{x}}^-_{i,k}$ are decoded samples obtained by traversing z^{sem}_k in opposite directions. Unlike DCI or F-score, which evaluate whether semantic information can be read from the latent code, Probe directly tests whether the decoder produces visual changes consistent with the injected semantics.

The qualitative results first reveal why output-space evaluation is necessary. As shown in Figure 4, SAGE produces smooth target-attribute changes while preserving irrelevant factors, and its sensitivity maps localize to semantically meaningful regions. More importantly, Figure 5 shows that competitive latent-side alignment does not necessarily imply decoder-side controllability: baseline models can exhibit diffuse responses or fail to express the target concept in decoded images. This visual gap motivates Probe, which directly measures whether latent traversal induces semantic visual changes in the output space.

Table 2: Quantitative comparison of reconstruction quality.

Method	CelebA (FID ↓)	CompCars (FID ↓)
FactorVAE	92.1 ± 3.1	112.5 ± 3.9
β-VAE	99.8 ± 2.4	107.4 ± 3.1
β-TCVAE	79.7 ± 2.6	87.2 ± 2.3
InfoGAN-CR	58.2 ± 1.5	63.7 ± 5.3
GAN Edit v1	24.5 ± 0.6	35.3 ± 4.1
Diff-AE	12.9 ± 2.1	11.4 ± 1.2
InfoDiffusion	29.6 ± 1.3	27.4 ± 2.6
EncDiff	14.8 ± 2.3	13.2 ± 1.4
TA-Dis	29.8 ± 0.2	34.5 ± 2.7
SAGE (Ours)	**9.3** ± 0.6	**10.3** ± 1.0

Table 1 evaluates conventional baselines with matched semantic pseudo-label supervision (vanilla → +guid) alongside native semantic-guided methods. The +guid results vary across architectures and metrics: for example, CelebA TAD for β-TCVAE decreases from 0.353 to 0.123, and Shapes3D Probe for InfoGAN-CR decreases from 0.493 to 0.220. These observations show that direct semantic supervision alone does not consistently improve latent disentanglement or decoder-side controllability.

We additionally compare against the native semantic-guided methods GEM [42], ZeroDIM [58], and TA-Dis [59] under the same dataset splits and metric implementations. SAGE matches or exceeds their latent-side disentanglement scores and achieves the highest output-space Probe on both Shapes3D (0.807) and CelebA (0.762). Together, these comparisons support the value of SAGE's full design beyond directly injecting pseudo-label supervision.

We further examine scalability on ImageNet-100, where SAGE achieves the lowest FID (13.404) among the compared methods, along with favorable peak-memory and iteration-time results. Detailed ImageNet-100 comparisons and a matched-setting training-efficiency study on CelebA are provided in Appendix H.4.



Figure 4: **Semantic disentanglement results.** Left: latent traversals — SAGE smoothly manipulates the target attribute while preserving others. Right: difference-based sensitivity maps (red/white = high sensitivity) localize each factor to semantically meaningful regions across diverse datasets.

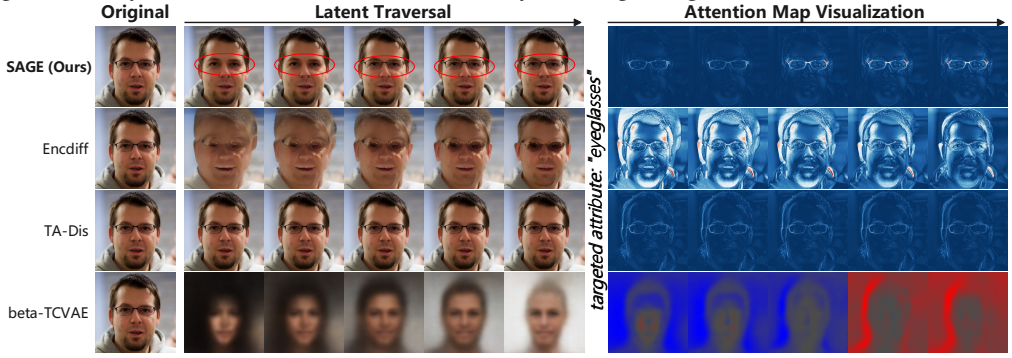


Figure 5: Qualitative comparison of semantic awareness and reconstruction quality. SAGE captures precise spatial semantics, while baselines exhibit diffuse responses or fail to express the target concept.

5.4 DR-Tradeoff Mitigation

To answer Q3, we combine the controllability evidence from Table 1 with the reconstruction results in Table 2. SAGE achieves strong disentanglement and the highest decoder-side Probe, while also obtaining the lowest FID on both CelebA and CompCars. These results show that SAGE improves disentanglement performance without sacrificing reconstruction fidelity, thereby effectively mitigating the DR-tradeoff.

5.5 Ablation Study

To answer Q4, we ablate the core components of SAGE in Table 3. Removing tLGC causes the most severe drop in Probe, from 0.762 to 0.423, while TAD remains moderate. This indicates

Table 3: Ablation study results.

CONFIGURATION	TAD ↑	PROBE ↑
SAGE (FULL)	**0.803** $\pm$ **0.011**	**0.762** $\pm$ **0.014**
(A) W/O TLGC	0.732 $\pm$ 0.015	0.423 $\pm$ 0.022
(B) W/O EXCLUSIVENESS	0.647 $\pm$ 0.014	0.738 $\pm$ 0.016
(C) W/O BALANCING	0.751 $\pm$ 0.012	0.714 $\pm$ 0.013

that semantic guidance alone fails to yield reliable decoder-side controllability without geometric anchoring. Removing Exclusiveness leads to the largest TAD degradation, from 0.803 to 0.647, while Probe remains relatively high. This suggests that tLGC still enables semantic traversal, but without SEM the factor–axis assignment becomes less exclusive, causing redundant semantic encoding and weaker disentanglement. Removing distribution-aware balancing moderately decreases both TAD and Probe, consistent with its role in stabilizing long-tailed semantic factors rather than defining the latent coordinate structure itself. Overall, these results confirm that tLGC is crucial for decoder-side controllability, SEM enforces cleaner semantic factorization, and balancing improves robustness to rare factors. Finally, to ensure that our method does not rely on specific proprietary models, a comprehensive robustness analysis of LLM selection is provided in Appendix H.2, confirming that SAGE maintains high effectiveness even when powered by open-source alternatives.

6 Discussion

Beyond proposing a specific model, this work aspires to provoke a fundamental shift in perspective within the community: *Is it time to rethink disentanglement beyond compression and statistical independence, and instead pursue a semantically disentangled latent space?* Although SAGE still has limitations, which we discuss in Appendix A, it provides a concrete step toward the latter view. Specifically, FDM uses MLLMs as semantic proxies to discover semantic factors from unlabeled data. In this sense, SAGE moves beyond enforcing global statistical independence among latent variables and instead pursues factor-wise independence on a semantic manifold. We hope this perspective encourages future DRL research to look past purely statistical proxies and develop more semantic-aware, interpretable, and reconstruction-preserving disentangled representation learning frameworks.

7 Conclusion

To address the oversimplification problem and DR-tradeoff prevalent in current DRL, we introduce SAGE, a novel DRL paradigm that constructs a semantically disentangled latent space via specialized modules. Comprehensive analyses demonstrate that SAGE achieves state-of-the-art disentanglement and reconstruction performance while capturing semantics overlooked by existing baselines. Moreover, we introduce an output-space probing metric to verify whether latent traversals induce semantic visual changes in decoded images, further confirming SAGE's superior decoder-side semantic controllability.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant Nos. 62672260 and 62302246; the Zhejiang Provincial Natural Science Foundation of China under Grant Nos. LQN26F030006 and LQ23F010008; the Ningbo Natural Science Foundation of China under Grant Nos. 2025J037 and NB002024010; and research grants from Ningbo under Grant Nos. 2023D-001-01, 2023Z237, 2024Z284, 2024Z289, 2023CX050011, 2025Z038, 2025Z028, and 2025Z059. We also acknowledge the support provided by the High Performance Computing Center at Eastern Institute of Technology and the Ningbo Institute of Digital Twin.

References

[1] Baao Xie, Zequn Zhang, Huanting Guo, Qiuyu Chen, Hu Zhu, Bohan Li, Wenjun Zeng, and Xin Jin. Navinerf++: Towards interpretable 3d reconstruction via unsupervised disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[2] Yunnan Wang, Fan Lu, Kecheng Zheng, Ziyuan Huang, Ziqiang Li, Wenjun Zeng, and Xin Jin. Vision-centric activation and coordination for multimodal large language models. *arXiv preprint arXiv:2510.14349*, 2025.

[3] Qi Wang, Zhipeng Zhang, Baao Xie, Xin Jin, Yunbo Wang, Shiyu Wang, Liaomo Zheng, Xiaokang Yang, and Wenjun Zeng. Disentangled world models: Learning to transfer semantic knowledge from distracting videos for reinforcement learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2599–2608, 2025. doi: 10.1109/ICCV51701.2025.00250.

[4] Yunnan Wang, Ziqiang Li, Wenyao Zhang, Zequn Zhang, Baao Xie, Xihui Liu, Wenjun Zeng, and Xin Jin. Scene graph disentanglement and composition for generalizable complex image generation. *Advances in Neural Information Processing Systems*, 37:98478–98504, 2024.

[5] Qiuyu Chen, Xin Jin, Yue Song, Xihui Liu, Shuai Yang, Tao Yang, Ziqiang Li, Jianguo Huang, Yuntao Wei, Ba'ao Xie, Nicu Sebe, Wenjun (Kevin) Zeng, Jooyeol Yun, Davide Abati, Mohamed Omran, Jaegul Choo, Amir Habibian, Auke Wiggers, Masato Kobayashi, Ning Ding, Toru Tamaki, Marzieh Gheisari, Auguste Genovesio, Yuheng Chen, Dingkun Liu, Xinyao Yang, Xinping Xu, Baicheng Chen, Dongrui Wu, Junhao Geng, Lexiang Lv, Jianxin Lin, Hanzhe Liang, Jie Zhou, Xuanxin Chen, Jinbao Wang, Can Gao, Zhangyi Wang, Zongze Li, Bihan Wen, Yixin Gao, Xiaohan Pan, Xin Li, Zhibo Chen, Baorui Peng, Zhongming Chen, and Haoran Jin. The 1st international workshop on disentangled representation learning for controllable generation (DRL4Real): Methods and results, 2025. URL `https://arxiv.org/abs/2509.10463`.

[6] Xiangchen Yin, Jiahui Yuan, Zhangchi Hu, Wenzhang Sun, Jie Chen, Xiaozhen Qiao, Hao Li, and Xiaoyan Sun. DeCo-VAE: Learning compact latents for video reconstruction via decoupled representation, 2025. URL `https://arxiv.org/abs/2511.14530`.

[7] Dingkun Liu, Siyang Li, Ziwei Wang, Wei Li, and Dongrui Wu. SDDA: Spatial distillation based distribution alignment for cross-headset EEG classification. *IEEE Transactions on Biomedical Engineering*, 73(7):2400–2411, 2026. doi: 10.1109/TBME.2025.3631604.

[8] Dingkun Liu, Yuheng Chen, Zhu Chen, Zhenyao Cui, Yaozhi Wen, Jiayu An, Jingwei Luo, and Dongrui Wu. EEG-FM-Compass: Progress, benchmarking, and future directions for EEG foundation models. *National Science Review*, page nwag466, 2026. doi: 10.1093/nsr/nwag466.

[9] Dingkun Liu, Zhu Chen, Jingwei Luo, Shijie Lian, Yuheng Chen, Shaojie Hou, Xiaolian Zhu, and Dongrui Wu. MIRepNet: A pipeline and pre-trained model for EEG-based motor imagery classification. *Knowledge-Based Systems*, 343:115966, 2026. doi: 10.1016/j.knosys.2026.115966.

[10] Xinqi Zhu, Chang Xu, and Dacheng Tao. Where and what? examining interpretable disentangled representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5861–5870, 2021.

[11] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

[12] Jinming Liu, Ruoyu Feng, Yunpeng Qi, Qiuyu Chen, Zhibo Chen, Wenjun Zeng, and Xin Jin. Rate-distortion-cognition controllable versatile neural image compression. In *Computer Vision – ECCV 2024*, pages 329–348. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72992-8_19.

[13] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[14] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

[15] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017.

[16] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International conference on machine learning*, pages 2649–2658. PMLR, 2018.

[17] Minyoung Kim, Yuting Wang, Pritish Sahu, and Vladimir Pavlovic. Relevance factor vae: Learning and identifying disentangled factors. *arXiv preprint arXiv:1902.01568*, 2019.

[18] Cristian Meo, Louis Mahon, Anirudh Goyal, and Justin Dauwels. α-tcvae: On the relationship between disentanglement and diversity. *arXiv preprint arXiv:2411.00588*, 2024.

[19] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018.

[20] Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational autoencoders pursue pca directions (by accident). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12406–12415, 2019.

[21] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International conference on machine learning*, pages 6348–6359. PMLR, 2020.

[22] Diederik P Kingma, Danilo J Rezende, Shakir Mohamed, and Max Welling. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27, 2014.

[23] Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. *arXiv preprint arXiv:1910.09772*, 2019.

[24] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[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*, pages 10684–10695, 2022.

[27] Tao Sun, Huanting Guo, Baao Xie, Zhou Fang, Qiuyu Chen, Hao Zheng, and Wenjun Zeng. GeoNaviDiff: Transferring disentangled semantic controls to diffusion features via local riemannian navigation, 2026. URL `https://arxiv.org/abs/2610.26091`.

[28] Tim R Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M Tomczak. Hyperspherical variational auto-encoders. *arXiv preprint arXiv:1804.00891*, 2018.

[29] Jiacheng Xu and Greg Durrett. Spherical latent spaces for stable variational autoencoders. *arXiv preprint arXiv:1808.10805*, 2018.

[30] Hang Shao, Abhishek Kumar, and P Thomas Fletcher. The riemannian geometry of deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 315–323, 2018.

[31] Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models. *arXiv preprint arXiv:1710.11379*, 2017.

[32] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9593–9602, 2021.

[33] Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, 23(241):1–55, 2022.

[34] Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *International conference on learning representations*, 2018.

[35] Eric Yeats, Frank Liu, David Womble, and Hai Li. Nashae: Disentangling representations through adversarial covariance minimization. In *European Conference on Computer Vision*, pages 36–51. Springer, 2022.

[36] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[37] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

[38] Tuomas Oikarinen and Tsui-Wei Weng. Clip-dissect: Automatic description of neuron representations in deep vision networks. *arXiv preprint arXiv:2204.10965*, 2022.

[39] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19187–19197, 2023.

[40] Zhaoda Ye, Xiangteng He, and Yuxin Peng. Unsupervised cross-media hashing learning via knowledge graph. *Chinese Journal of Electronics*, 31(6):1081–1091, 2022. doi: 10.1049/cje.2021.00.455.

[41] Jiahao Li, Qingwang Zhang, Qiuyu Chen, Guozhan Qiu, Yunzhong Lou, and Xiangdong Zhou. Towards high-fidelity CAD generation via LLM-driven program generation and text-based B-Rep primitive grounding. In Tong Zhang, Miroslav Dudik, Martin Jaggi, Alekh Agarwal, Sharon Li, Dale Schuurmans, Jerry Zhu, Felix Berkenkamp, Hanze Dong, and Alberto Bietti, editors, *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 72315–72330. PMLR, 2026. URL `https://proceedings.mlr.press/v306/li26jd.html`.

[42] Baao Xie, Qiuyu Chen, Yunnan Wang, Zequn Zhang, Xin Jin, and Wenjun Zeng. Graph-based unsupervised disentangled representation learning via multimodal large language models. *Advances in Neural Information Processing Systems*, 37:103101–103130, 2024.

[43] James Lucas, George Tucker, Roger B Grosse, and Mohammad Norouzi. Don't blame the elbo! a linear vae perspective on posterior collapse. *Advances in Neural Information Processing Systems*, 32, 2019.

[44] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[45] Yuxin Peng, Zishuo Wang, Geng Li, Xiangtian Zheng, Sibo Yin, and Hulingxiao He. A survey on fine-grained multimodal large language models. *Chinese Journal of Electronics*, 35(2): 771–803, 2026. doi: 10.23919/cje.2025.00.336.

[46] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[47] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[48] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020.

[49] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3973–3981, 2015.

[50] Zinan Lin, Kiran Thekumparampil, Giulia Fanti, and Sewoong Oh. Infogan-cr and modelcentrality: Self-supervised model training and selection for disentangling gans. In *international conference on machine learning*, pages 6127–6139. PMLR, 2020.

[51] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in neural information processing systems*, 33:9841–9850, 2020.

[52] Xuanchi Ren, Tao Yang, Yuwang Wang, and Wenjun Zeng. Learning disentangled representation by exploiting pretrained generative models: A contrastive learning view. *arXiv preprint arXiv:2102.10543*, 2021.

[53] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10619–10629, 2022.

[54] Yingheng Wang, Yair Schiff, Aaron Gokaslan, Weishen Pan, Fei Wang, Christopher De Sa, and Volodymyr Kuleshov. Infodiffusion: Representation learning using information maximizing diffusion models. In *International conference on machine learning*, pages 36336–36354. PMLR, 2023.

[55] Tao Yang, Yuwang Wang, Yan Lv, and Nanning Zheng. Disdiff: Unsupervised disentanglement of diffusion probabilistic models. *arXiv preprint arXiv:2301.13721*, 2023.

[56] Xin Jin, Bohan Li, Baao Xie, Wenyao Zhang, Jinming Liu, Ziqiang Li, Tao Yang, and Wenjun Zeng. Closed-loop unsupervised representation disentanglement with β-vae distillation and diffusion probabilistic feedback. In *European Conference on Computer Vision*, pages 270–289. Springer, 2024.

[57] Tao Yang, Cuiling Lan, Yan Lu, and Nanning Zheng. Diffusion model with cross attention as an inductive bias for disentanglement. *Advances in Neural Information Processing Systems*, 37: 82465–82492, 2024.

[58] Aviv Gabbay, Niv Cohen, and Yedid Hoshen. An image is worth more than a thousand words: Towards disentanglement in the wild. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 9216–9228. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/4cef5b5e6ff1b3445db4c013f1d452e0-Paper.pdf`.

[59] Junhao Geng, Lexiang Lv, and Jianxin Lin. Textual semantics matters: Unsupervised representation disentanglement in realistic scenarios with language inductive bias. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 361–370, 2025.

A Limitations and Future Work

SAGE offers a semantic-guided alternative to independence-oriented DRL, but it is still an initial step. Since FDM uses MLLMs and vision-language models as semantic proxies, the discovered factors may inherit biases, ambiguities, or coverage limitations from these foundation models, even though this design removes the need for manually curated attribute lists. A promising direction is to make factor discovery more reliable through uncertainty-aware selection, agreement among multiple semantic proxies, or lightweight human verification. In addition, our current latent geometry provides a simple and effective way to anchor semantic factors, but more expressive geometries, adaptive semantic coordinate systems, and integration with larger generative backbones remain valuable directions for future work. We view SAGE as a step toward replacing purely statistical disentanglement objectives with semantic-factor-centered representation learning.

B Broader Impact

SAGE aims to improve interpretable and controllable representation learning, which may benefit transparent generative modeling and human-aligned visual editing. Like other controllable generation techniques, it may also be misused for biased or deceptive image manipulation, and therefore should be deployed with appropriate safeguards and responsible-use considerations.

C Proofs and Additional Discussions

C.1 Proof of Theorem 3.1

Proof. Under $\mathbf{x} = \mathbf{J}\mathbf{z} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_{\text{noise}}^2 \mathbf{I})$, isotropic prior, and diagonal Gaussian variational family, an optimum exists where encoder and decoder are diagonal in the singular basis of $\mathbf{J}$; hence the β-ELBO decomposes across $r = \min(D_x, d)$ modes (modes beyond rank r of $\mathbf{J}$ contribute no reconstruction signal and collapse trivially to the prior; we therefore focus on a single active mode below).

Let $x \in \mathbb{R}$ denote one mode with variance $\mathbb{E}[x^2] = \gamma$. Consider a 1D linear decoder $p(x \mid z) = \mathcal{N}(wz, \sigma_{\text{noise}}^2)$ and Gaussian encoder $q(z \mid x) = \mathcal{N}(ax, \nu^2)$. The expected negative β-ELBO (up to additive constants independent of (a, ν, w)) is

$$\mathcal{L}(a, \nu, w) = \frac{1}{2\sigma_{\text{noise}}^2} \mathbb{E}[(x - wz)^2] + \frac{\beta}{2} \mathbb{E}[(ax)^2 + \nu^2 - \log \nu^2 - 1] \tag{12}$$

Since $z \sim \mathcal{N}(ax, \nu^2)$,

$$\mathbb{E}[(x - wz)^2 \mid x] = (1 - wa)^2 x^2 + w^2 \nu^2 \quad \Rightarrow \quad \mathbb{E}[(x - wz)^2] = \gamma(1 - wa)^2 + w^2 \nu^2 \tag{13}$$

Also $\mathbb{E}[(ax)^2] = a^2 \gamma$, so

$$\mathcal{L}(a, \nu, w) = \frac{\gamma(1 - wa)^2 + w^2 \nu^2}{2\sigma_{\text{noise}}^2} + \frac{\beta}{2}\big(a^2 \gamma + \nu^2 - \log \nu^2 - 1\big) \tag{14}$$

Optimize (a, ν) given w. Setting derivatives to zero yields

$$a^\star(w) = \frac{w}{w^2 + \beta \sigma_{\text{noise}}^2}, \qquad \nu^{2\star}(w) = \frac{\beta \sigma_{\text{noise}}^2}{w^2 + \beta \sigma_{\text{noise}}^2} \tag{15}$$

Substitute $u = w^2 \geq 0$ to obtain an objective $\mathcal{L}(u)$:

$$\mathcal{L}(u) = \frac{\beta}{2} \log\Big(1 + \frac{u}{\beta \sigma_{\text{noise}}^2}\Big) + \frac{\beta \gamma}{2(\beta \sigma_{\text{noise}}^2 + u)} + \text{const} \tag{16}$$

Optimize u.

$$\frac{d\mathcal{L}}{du} = \frac{\beta}{2(\beta \sigma_{\text{noise}}^2 + u)} - \frac{\beta \gamma}{2(\beta \sigma_{\text{noise}}^2 + u)^2} \tag{17}$$

Setting $d\mathcal{L}/du = 0$ gives $\beta \sigma_{\text{noise}}^2 + u = \gamma$, i.e., $u^\star = \gamma - \beta \sigma_{\text{noise}}^2$. Since $u \geq 0$, the global minimizer is $u^\star = \max(0, \gamma - \beta \sigma_{\text{noise}}^2)$. If $\gamma \leq \beta \sigma_{\text{noise}}^2$, then $u^\star = 0$ implies $w^\star = 0$, $a^\star = 0$, and $\nu^{2\star} = 1$ (exact collapse). If $\gamma > \beta \sigma_{\text{noise}}^2$, then $u^\star > 0$ and

$$\nu^{2\star} = \frac{\beta \sigma_{\text{noise}}^2}{\beta \sigma_{\text{noise}}^2 + u^\star} = \frac{\beta \sigma_{\text{noise}}^2}{\gamma} < 1 \tag{18}$$

This proves the thresholding statement. □

Numerical verification of the closed-form solution. The derivation above produces the closed-form $w^{2*} = \max(0, \gamma - \beta\sigma^2_{\text{noise}})$ analytically, but it relies on a specific factorization of the ELBO. As a numerical sanity check that the analytical solution is indeed the global optimum of the original 1D ELBO (rather than an artifact of the factorization), we directly minimize the analytical 1D objective in Eq. 12 using *L-BFGS* (Limited-memory BFGS, a quasi-Newton optimizer that approximates the Hessian from gradient history and makes no prior assumption on the form of the optimum). With $\beta = 4$, $\sigma^2_{\text{noise}} = 0.1$ (theoretical threshold 0.4), we sweep $\gamma \in [0.05, 2.0]$ over 23 values and compare the L-BFGS optimum against the closed-form prediction.

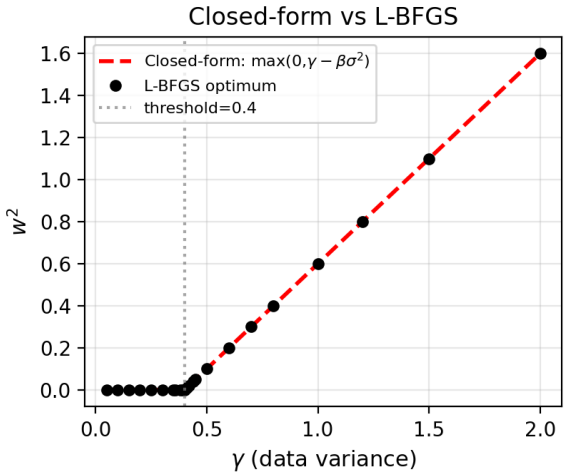


Figure 6: **Closed-form vs. numerical optimum.** The L-BFGS optimum (black dots) overlays the closed-form curve $w^{2*} = \max(0, \gamma - \beta\sigma^2)$ (red dashed) with maximum deviation $< 10^{-5}$, providing an independent numerical confirmation that the analytical thresholding solution above is the global minimizer of the ELBO.

The two curves overlay at machine-precision agreement (max deviation $< 10^{-5}$): the analytical threshold is indeed reached by direct numerical optimization, with no a-priori assumption about the form of the solution. This independent verification establishes the closed-form as the foundation upon which the multi-dimensional and nonlinear extensions in Figure 3 build.

From spectral modes to nonlinear factor energy. Theorem 3.1 characterizes factor learnability in a linearized Gaussian regime, where each learnable mode is governed by the spectral energy γ_k of the local data covariance in the singular basis of $\mathbf{J}$. For a nonlinear generator $\mathbf{x} = g(\mathbf{s}) + \boldsymbol{\epsilon}$, a local linearization gives

$$g(\mathbf{s} + \Delta\mathbf{s}) \approx g(\mathbf{s}) + \sum_m \frac{\partial g(\mathbf{s})}{\partial s_m} \Delta s_m.$$

Therefore, the output-space covariance contribution of factor s_m is controlled by the effective signal energy

$$\bar{E}_m \triangleq \operatorname{Var}(s_m)\, \mathbb{E}_{\mathbf{s}} \left[\left\| \frac{\partial g(\mathbf{s})}{\partial s_m} \right\|_2^2 \right].$$

In the linear orthogonal case $\mathbf{x} = \mathbf{A}\mathbf{s} + \boldsymbol{\epsilon}$, this reduces to $\operatorname{Var}(s_m)\|\mathbf{a}_m\|_2^2$, which coincides with the eigenvalue-like energy γ_k when factor directions align with the singular basis of $\mathbf{J}$. In nonlinear models, the exact threshold may be softened by variations in the Jacobian across the data manifold, while $\bar{E}_m$ provides a useful energy-ordering perspective: lower-energy factors tend to collapse earlier, whereas higher-energy factors tend to survive stronger regularization, as observed in our nonlinear FactorVAE experiments in Figure 3(d). This observation suggests that the energy-dependent factor-retention mechanism characterized by Theorem 3.1 remains informative beyond the linear Gaussian regime and motivates the geometry-anchored design used by SAGE downstream.

C.2 Auxiliary Lemma for Residual Leakage

Lemma C.1. *Let the residual be* $\mathbf{r} = \mathbf{x} - \mathbb{E}[\mathbf{x}|\mathbf{y}]$. *Assume the class-conditional distribution* $p(\mathbf{r} \mid \mathbf{y} = y)$ *has covariance* $\mathbf{\Sigma}_y$. *Let* $\mathbf{\Sigma}_r = \mathbb{E}[\mathbf{r}\mathbf{r}^\top]$ *be the marginal residual covariance. If there exists a class* y *such that* $\mathbf{\Sigma}_y \neq \mathbf{\Sigma}_r$ *, then* $I(\mathbf{y}; \mathbf{r}) > 0$.

Proof. If $\mathbf{\Sigma}_y \neq \mathbf{\Sigma}_r$, then $p(\mathbf{r} \mid \mathbf{y} = y)$ differs from the marginal $p(\mathbf{r})$ in its second moment. Since independence requires $p(\mathbf{r} \mid \mathbf{y}) = p(\mathbf{r})$ almost everywhere, the mismatch confirms dependence. Thus, $I(\mathbf{y}; \mathbf{r}) = D_{\mathrm{KL}}(p(\mathbf{r}, \mathbf{y}) \| p(\mathbf{r})p(\mathbf{y})) > 0$. □

C.3 Proof of Theorem 3.3

Assumption C.2 (Residual retention by unguided coordinates)**.** Let $\mathbf{c}_U = h(\mathbf{P}_U \mathbf{r})$ be a finite-valued statistic of the residual projection $\mathbf{P}_U \mathbf{r}$ along reconstruction-favored directions. We say that the unguided code $\mathbf{z}_U$ retains this residual statistic up to error ϵ_U if

$$H(\mathbf{c}_U \mid \mathbf{z}_U) \leq \epsilon_U. \tag{19}$$

Proof. By the chain rule of mutual information, we decompose $I(\mathbf{y}; \mathbf{c}_U, \mathbf{z}_U)$ in two ways:

$$I(\mathbf{y}; \mathbf{c}_U, \mathbf{z}_U) = I(\mathbf{y}; \mathbf{z}_U) + I(\mathbf{y}; \mathbf{c}_U \mid \mathbf{z}_U) \tag{20}$$

$$= I(\mathbf{y}; \mathbf{c}_U) + I(\mathbf{y}; \mathbf{z}_U \mid \mathbf{c}_U). \tag{21}$$

Subtracting and using $I(\mathbf{y}; \mathbf{z}_U \mid \mathbf{c}_U) \geq 0$,

$$I(\mathbf{y}; \mathbf{z}_U) \;=\; I(\mathbf{y}; \mathbf{c}_U) + I(\mathbf{y}; \mathbf{z}_U \mid \mathbf{c}_U) - I(\mathbf{y}; \mathbf{c}_U \mid \mathbf{z}_U) \;\geq\; I(\mathbf{y}; \mathbf{c}_U) - I(\mathbf{y}; \mathbf{c}_U \mid \mathbf{z}_U). \tag{22}$$

Since $\mathbf{c}_U$ is finite-valued, $I(\mathbf{y}; \mathbf{c}_U \mid \mathbf{z}_U) \leq H(\mathbf{c}_U \mid \mathbf{z}_U) \leq \epsilon_U$ by Assumption C.2. Combined with $I(\mathbf{y}; \mathbf{c}_U) = \Delta_U$,

$$I(\mathbf{y}; \mathbf{z}_U) \;\geq\; \Delta_U - \epsilon_U, \tag{23}$$

which is strictly positive because $\epsilon_U < \Delta_U$. Hence semantic information leaks into the unguided subspace. □

C.4 Proof of Proposition C.3

Proposition C.3 (Fidelity Collapse in Full Supervision)**.** *If the latent code is constrained to be a function of* $\mathbf{y}$ *alone, i.e.,* $\mathbf{x} \to \mathbf{y} \to \mathbf{z}$ *forms a Markov chain, then for squared-error reconstruction the minimum achievable distortion is lower-bounded by the residual energy:*

$$\inf_{\hat{\mathbf{x}}(\mathbf{z})} \mathbb{E}\|\mathbf{x} - \hat{\mathbf{x}}(\mathbf{z})\|^2 \;\geq\; \mathbb{E}\|\mathbf{x} - \mathbb{E}[\mathbf{x} \mid \mathbf{y}]\|^2 \;=\; \mathrm{Tr}(\mathbf{\Sigma}_r). \tag{24}$$

Proof. Assume $\mathbf{x} \to \mathbf{y} \to \mathbf{z}$ is a Markov chain. For squared-error distortion, the Bayes-optimal estimator is the conditional mean, hence for any measurable decoder $\hat{\mathbf{x}}(\mathbf{z})$,

$$\mathbb{E}\|\mathbf{x} - \hat{\mathbf{x}}(\mathbf{z})\|^2 \;\geq\; \mathbb{E}\|\mathbf{x} - \mathbb{E}[\mathbf{x} \mid \mathbf{z}]\|^2. \tag{25}$$

We now compare the optimal MSE attainable from $\mathbf{z}$ with that attainable from $\mathbf{y}$. Since $\mathbf{z}$ is generated solely from $\mathbf{y}$ (Markov chain), the conditional independence $\mathbf{x} \perp \mathbf{z} \mid \mathbf{y}$ implies, by the tower property of conditional expectation,

$$\mathbb{E}[\mathbf{x} \mid \mathbf{z}] \;=\; \mathbb{E}\big[\mathbb{E}[\mathbf{x} \mid \mathbf{y}, \mathbf{z}] \,\big|\, \mathbf{z}\big] \;=\; \mathbb{E}\big[\mathbb{E}[\mathbf{x} \mid \mathbf{y}] \,\big|\, \mathbf{z}\big]\,. \tag{26}$$

In other words, the optimal Bayes estimator from $\mathbf{z}$ is a coarsening (conditional expectation) of the optimal estimator from $\mathbf{y}$. Because conditional expectation is the orthogonal projection in L^2, projecting $\mathbb{E}[\mathbf{x} \mid \mathbf{y}]$ onto the smaller σ-algebra generated by $\mathbf{z}$ cannot reduce its L^2-distance to $\mathbf{x}$:

$$\mathbb{E}\|\mathbf{x} - \mathbb{E}[\mathbf{x} \mid \mathbf{z}]\|^2 \;\geq\; \mathbb{E}\|\mathbf{x} - \mathbb{E}[\mathbf{x} \mid \mathbf{y}]\|^2. \tag{27}$$

Combining the two inequalities,

$$\mathbb{E}\|\mathbf{x} - \hat{\mathbf{x}}(\mathbf{z})\|^2 \;\geq\; \mathbb{E}\|\mathbf{x} - \mathbb{E}[\mathbf{x} \mid \mathbf{y}]\|^2 = \mathbb{E}\|\mathbf{r}(\mathbf{x})\|^2 = \mathrm{Tr}(\mathbf{\Sigma}_r), \tag{28}$$

where the last two equalities use $\mathbf{r}(\mathbf{x}) = \mathbf{x} - \mathbb{E}[\mathbf{x} \mid \mathbf{y}]$ and $\mathbf{\Sigma}_r = \mathbb{E}[\mathbf{r}\mathbf{r}^\top]$. □

C.5 Proof of Theorem 3.5

Proof. Let $\mathbf{A} \in \mathcal{T}_K^+ \cap O(K)$. Since $\mathbf{A} \in O(K)$, $\mathbf{A}^{-1} = \mathbf{A}^\top$. Since $\mathbf{A}$ is lower triangular with positive diagonal, $\mathbf{A}^{-1}$ is also lower triangular, hence $\mathbf{A}^\top$ is lower triangular. Therefore, $\mathbf{A}$ is both lower and upper triangular, hence diagonal. Orthogonality implies diagonal entries are ± 1, while $\mathcal{T}_K^+$ requires positive diagonals. Thus $\mathbf{A} = \mathbf{I}_K$. □

C.6 Proof of Proposition C.4

Proposition C.4 (CE-gain equals conditional leakage at Bayes risk)**.** *Let $\mathbf{s} = \mathbf{z}^{sem}$ and let $s_k = z_k^{sem}$ be the designated coordinate for factor y_k. For any input variable $\mathcal{U}$, define the Bayes cross-entropy risk*

$$\mathcal{C}_k(\mathcal{U}) \triangleq \inf_q \mathbb{E}\big[\mathrm{CE}\big(q(\mathcal{U}), y_k\big)\big] = H(y_k \mid \mathcal{U}). \tag{29}$$

Then

$$\mathcal{C}_k(s_k) - \mathcal{C}_k(s_k, \mathbf{s}_{\setminus k}) = I(y_k; \mathbf{s}_{\setminus k} \mid s_k). \tag{30}$$

If semantic guidance makes s_k predictive,

$$\mathcal{C}_k(s_k) \le \epsilon_k, \tag{31}$$

and SEM drives the CE-gain below δ_k,

$$\mathcal{C}_k(s_k) - \mathcal{C}_k(s_k, \mathbf{s}_{\setminus k}) \le \delta_k, \tag{32}$$

then

$$H(y_k \mid s_k) \le \epsilon_k, \qquad I(y_k; \mathbf{s}_{\setminus k} \mid s_k) \le \delta_k. \tag{33}$$

Hence, s_k is an ϵ_k-sufficient coordinate for predicting y_k, and the remaining coordinates contain at most δ_k additional information about y_k beyond s_k.

Remark C.5. Together with Theorem 3.5 (which fixes the coordinate frame), Proposition C.4 provides a clean information-theoretic target for SEM: minimizing the empirical CE-gain pushes the conditional leakage $I(y_k; \mathbf{s}_{\setminus k} \mid s_k)$ toward zero as the predictors approach Bayes risk, without invoking an unsigned approximation gap.

Proof. For any predictor q and target y_k, cross-entropy admits the identity

$$\mathbb{E}\big[\mathrm{CE}\big(q(\mathcal{U}), y_k\big)\big] = H(y_k \mid \mathcal{U}) + \mathbb{E}_{\mathcal{U}}\big[D_{\mathrm{KL}}\big(p(y_k \mid \mathcal{U}) \,\|\, q(\mathcal{U})\big)\big], \tag{34}$$

where the second term is non-negative and vanishes iff $q(\mathcal{U}) = p(y_k \mid \mathcal{U})$ almost surely. Taking the infimum over q therefore gives $\mathcal{C}_k(\mathcal{U}) = H(y_k \mid \mathcal{U})$, establishing the first equality.

Substituting $\mathcal{U} \in \{s_k,\ (s_k, \mathbf{s}_{\setminus k})\}$ and applying the chain rule of conditional entropy,

$$\mathcal{C}_k(s_k) - \mathcal{C}_k(s_k, \mathbf{s}_{\setminus k}) = H(y_k \mid s_k) - H(y_k \mid s_k, \mathbf{s}_{\setminus k}) = I(y_k; \mathbf{s}_{\setminus k} \mid s_k), \tag{35}$$

which is Equation (30). The two consequences then follow immediately: $\mathcal{C}_k(s_k) \le \epsilon_k$ is exactly $H(y_k \mid s_k) \le \epsilon_k$, and the CE-gain bound δ_k transfers to $I(y_k; \mathbf{s}_{\setminus k} \mid s_k) \le \delta_k$ via the equality. □

C.7 Proof of Proposition C.6

Proposition C.6 (Re-balance the Long-tailed Distribution)**.** *For a mini-batch of size B and an attribute with prevalence π, the expected Fisher information of the guided gradient is upper-bounded by the label variance:*

$$\mathcal{I}(\nabla_{\mathbf{z}} \mathcal{L}_{sup}) \ \lesssim\ B \cdot \pi(1-\pi). \tag{36}$$

Remark C.7. Theorem C.6 shows that rare attributes ($\pi \to 0$) inject a vanishing supervisory signal into the gradient, so their effective signal variance falls below the activation threshold of Theorem 3.1 ($\gamma < \beta \sigma_{\mathrm{noise}}^2$) and they are pruned as stochastic noise. Distribution-aware balancing amplifies these factors so that they remain learnable.

Proof. Consider a discriminative classifier that takes a latent $\mathbf{z}$ and produces a logit $z_\theta(\mathbf{z})$ with parameters θ; the predicted probability is $p_\theta = \sigma(z_\theta)$, where $\sigma(\cdot)$ is the sigmoid function. The Bernoulli log-likelihood for label $y \in \{0, 1\}$ is

$$\log P(y \mid \theta) \ =\ y \log p_\theta + (1-y) \log(1 - p_\theta). \tag{37}$$

Step 1 — Score function. Using $\partial p_\theta / \partial z_\theta = p_\theta(1-p_\theta)$:

$$\frac{\partial \log P(y \mid \theta)}{\partial z_\theta} \;=\; \frac{y}{p_\theta}\, p_\theta(1-p_\theta) - \frac{1-y}{1-p_\theta}\, p_\theta(1-p_\theta) \;=\; y - p_\theta. \tag{38}$$

By the chain rule, the score with respect to any parameter θ_j is $s_j(y;\theta) = (y-p_\theta)\,\partial_{\theta_j} z_\theta$.

Step 2 — Per-example Fisher information. Conditional on $\mathbf{z}$, let $\pi_{\text{true}}(\mathbf{z}) = \Pr(y=1 \mid \mathbf{z})$ and $y \sim \text{Bern}(\pi_{\text{true}}(\mathbf{z}))$. A bias–variance decomposition gives

$$\begin{aligned}\mathbb{E}\big[(y-p_\theta)^2 \mid \mathbf{z}\big] &= \text{Var}(y \mid \mathbf{z}) + \big(\mathbb{E}[y \mid \mathbf{z}] - p_\theta\big)^2 \\ &= \pi_{\text{true}}(\mathbf{z})\,(1-\pi_{\text{true}}(\mathbf{z})) + (\pi_{\text{true}}(\mathbf{z}) - p_\theta(\mathbf{z}))^2,\end{aligned} \tag{39}$$

so

$$\mathcal{I}(\theta \mid \mathbf{z}) = \mathbb{E}_y\big[s_j(y;\theta)\, s_k(y;\theta)\big] = \mathbb{E}\big[(y-p_\theta)^2 \mid \mathbf{z}\big]\,(\partial z_\theta)(\partial z_\theta)^\top. \tag{40}$$

Step 3 — Expected Fisher information. For a well-calibrated predictor, $p_\theta(\mathbf{z}) \to \pi_{\text{true}}(\mathbf{z})$ on the training distribution; the residual bias term $(\pi_{\text{true}} - p_\theta)^2$ vanishes and the per-example Fisher reduces to the Bernoulli variance times the Jacobian outer product. Taking expectation over a mini-batch of size B and applying Jensen's inequality to the marginal prevalence $\pi = \mathbb{E}_{\mathbf{z}}[\pi_{\text{true}}(\mathbf{z})]$:

$$\mathbb{E}\big[\mathcal{I}_B(\theta)\big] \;\propto\; B \cdot \mathbb{E}_{\mathbf{z}}\big[\pi_{\text{true}}(\mathbf{z})(1-\pi_{\text{true}}(\mathbf{z}))\big] \;\le\; B \cdot \pi(1-\pi). \tag{41}$$

The upper bound is tight when π_{true} is (locally) constant, which is the standard working assumption under which the proposition is stated.

Conclusion. For rare attributes ($\pi \to 0$ or $\pi \to 1$), $\pi(1-\pi) \to 0$, so the expected Fisher information per mini-batch collapses to zero even when the latent $\mathbf{z}$ is informative about y. The resulting score gradient has vanishingly small magnitude, empirically manifesting as "the rare attribute is never learned." The distribution-aware rebalancing mechanism inflates the effective sampling weights of rare classes so that the $B \cdot \pi(1-\pi)$ term does not vanish, exactly as stated in Proposition C.6. □

D Empirical Validation of the Spectral Threshold

Theorem 3.1 gives an exact activation threshold in a linearized Gaussian regime. This appendix presents three complementary checks. First, we verify the closed-form threshold in linear β-VAE toy models. Second, we examine whether the induced energy-ordering tendency is observed in nonlinear disentanglement models. Third, we test whether observation-space factor energy can conflict with task-relevant semantic importance in a controlled Shapes3D experiment. For brevity, we write σ^2 for σ^2_{noise} throughout this appendix.

D.1 Implementation Details for Spectral-Threshold Validation

This section details how the four panels in Figure 3 are computed. The goal is to validate the exact spectral threshold in the linear regime and to test its energy-ordering implication in nonlinear disentanglement models.

Panels (a) and (b): universality of the threshold. To test whether the threshold depends on any specific parameter choice, we evaluate four one-dimensional configurations:

$$(\beta, \sigma^2_{\text{noise}}) \in \{(4, 0.1), (2, 0.2), (8, 0.05), (1, 0.4)\}.$$

All four configurations share the same product $\beta\sigma^2_{\text{noise}} = 0.4$. For each setting we directly use the closed-form solution (no training is required) and plot the normalized variables

$$x = \frac{\gamma}{\beta\sigma^2_{\text{noise}}}, \qquad y_w = \frac{w^{2*}}{\beta\sigma^2_{\text{noise}}}, \qquad y_\nu = \nu^{2*}.$$

Panel (a) shows the normalized decoder weight y_w collapsing onto a single curve with a kink at $x = 1$; panel (b) shows the posterior variance ν^{2*} holding flat at 1 for $x \le 1$ and dropping below 1 for $x > 1$, again on a single curve across all four settings. Both panels demonstrate that the learnability boundary is controlled by the dimensionless ratio $\gamma/(\beta\sigma^2_{\text{noise}})$ rather than by γ, β, or σ^2_{noise} alone, confirming that Theorem 3.1 captures a parameter-invariant relationship rather than a coincidence at a single (β, σ^2).

Panel (c): multi-dimensional staircase and DR-tradeoff. We construct an 8-dimensional diagonal linear β-VAE with independent observation modes whose variances are

$$\gamma = (1.6, 1.1, 0.75, 0.55, 0.38, 0.28, 0.18, 0.12),$$

and fix $\sigma^2_{\text{noise}} = 0.1$. We sweep $\beta \in \{1, 2, 4, 8, 15\}$, corresponding to thresholds $\beta\sigma^2_{\text{noise}} \in \{0.1, 0.2, 0.4, 0.8, 1.5\}$. For each (β, seed) pair we generate 50000 samples and fit the diagonal model with L-BFGS on the analytic objective; results are averaged over 5 seeds. A mode is counted as active when $w_k^2 > 0.01$. The theoretical active-mode count is obtained directly from the condition $\gamma_k > \beta\sigma^2_{\text{noise}}$, while the empirical count is obtained after optimization. We also compute the analytical reconstruction error

$$\text{MSE}_{\text{rec}} = \sum_k \left[\gamma_k(1 - w_k a_k)^2 + w_k^2 \nu_k^2\right].$$

The panel verifies that the threshold predicts multi-mode pruning and shows that stronger regularization increases reconstruction error, thereby visualizing the DR-tradeoff.

Panel (d): nonlinear FactorVAE collapse ordering. To test the nonlinear extension, we use a controlled four-factor ReLU generator

$$\mathbf{x} = \mathbf{W}_2 \,\text{ReLU}(\mathbf{W}_1 \mathbf{s} + \mathbf{b}_1) + \boldsymbol{\epsilon},$$

where $\mathbf{W}_1 \in \mathbb{R}^{32\times 4}$ and $\mathbf{W}_2 \in \mathbb{R}^{20\times 32}$ are fixed structured weights (each factor has 8 dedicated hidden units plus 10% cross-coupling). The factor variances are set to

$$\text{Var}(\mathbf{s}) = (1.0, 0.6, 0.2, 0.05).$$

For each factor s_m, we compute the effective nonlinear factor energy via the generator Jacobian:

$$\bar{E}_m = \text{Var}(s_m)\, \mathbb{E}_{\mathbf{s}}\left[\left\| \frac{\partial g(\mathbf{s})}{\partial s_m} \right\|_2^2\right],$$

yielding $\bar{E}_0 = 8.37$, $\bar{E}_1 = 2.60$, $\bar{E}_2 = 0.33$, and $\bar{E}_3 = 0.02$. We train an MLP FactorVAE (encoder/decoder hidden width 64, Gaussian likelihood with fixed σ^2, latent dim $K = 4$) while sweeping the total-correlation regularization weight

$$\gamma_{\text{TC}} \in \{0, 1, 5, 10, 20, 50, 100, 200, 500\},$$

2 seeds per setting (shaded $\pm$std band in the plot). For each trained model, we fit an MLP probe (sklearn `MLPRegressor`) from the learned latent code $\mathbf{z}$ to each ground-truth factor s_m and report the probe R^2. Lower R^2 indicates that the factor information has collapsed from the latent representation. The collapse order follows the effective energy ranking: the lowest-energy factor s_3 stays near $R^2{=}0$ even at $\gamma_{\text{TC}}{=}0$; s_2 collapses around $\gamma_{\text{TC}}{\sim}50$; s_1 collapses near $\gamma_{\text{TC}}{\sim}500$; the highest-energy factor s_0 remains recoverable across the entire sweep. This empirically realizes the nonlinear energy-ordering implication of Theorem 3.1: $\bar{E}_m$ replaces the linear spectral energy γ_k as the practical control variable for factor retention in independence-oriented nonlinear DRL.

Summary. The linear toy validates the exact spectral threshold, while the nonlinear toy exhibits a corresponding energy-ordering tendency. Together, these results support the use of effective factor energy as a practical guide for analyzing nonlinear models, without asserting an exact nonlinear threshold.

D.2 Controlled Semantic–Nuisance Misalignment on Shapes3D

Experimental setup. To examine whether output-space energy ordering can conflict with semantic priority, we conduct a controlled semantic–nuisance study on Shapes3D. We use object shape as a human-interpretable ground-truth factor and the target of a balanced four-class shape-recognition task. We introduce an independently sampled global pixel-intensity nuisance $u \sim \mathcal{U}[-1, 1]$, independent of all native generative factors, and apply the same offset su to every RGB pixel:

$$x_{\text{obs}} = \text{clip}(x + su,\, 0,\, 1). \tag{42}$$

Here s controls the nuisance magnitude, and clipping is performed elementwise. This nuisance changes low-level pixel intensities without changing the object-shape label or contributing to the shape-recognition task.

Table 4: Controlled energy–retention study on Shapes3D at $\beta = 16$. The ratio compares the output-space variation energy of the pixel-intensity nuisance with that of object shape; nuisance R^2 measures recoverability from held-out latent representations.

Condition	$E_{\text{nuisance}}/E_{\text{shape}}$	Nuisance R^2 ($\beta = 16$)
Low	0.243	-0.017
Matched	0.962	0.073
High	1.916	0.617

Controlled measurement of factor energy. By varying only the nuisance magnitude s, while keeping its distribution, independence, and task role unchanged, we construct three conditions with different ratios of nuisance output energy to shape-factor output energy. For each condition, we train the same convolutional β-VAE with three random seeds. We quantify nuisance retention by fitting a two-layer MLP on held-out latent representations to predict u, reporting its coefficient of determination R^2; $R^2 > 0$ indicates that nuisance information is recoverable. At a matched compression strength of $\beta = 16$, we obtain the results in Table 4.

A linear Ridge probe gives the same ordering. Thus, increasing only the output-space energy of the same nuisance changes it from unrecoverable to strongly retained under an otherwise identical compression setting, providing a controlled energy–retention relationship.

A direct semantic–nuisance counterexample. Using matched image pairs, we measure the average squared RGB change caused by varying each factor while holding all other native factors fixed. The resulting output-space variation energies are

$$E_{\text{shape}} = 0.018137, \qquad E_{\text{nuisance}} = 0.089535, \qquad \frac{E_{\text{nuisance}}}{E_{\text{shape}}} = 4.936. \tag{43}$$

The task-irrelevant pixel-intensity nuisance therefore has substantially higher observation-space energy than the human-interpretable shape factor. We train the same convolutional β-VAE over $\beta \in \{16, 24, 32, 48, 64\}$. At $\beta = 48$, the two factors show sharply different recoverability from sampled latent codes z:

Factor	Recoverability from sampled z ($\beta = 48$)
Object shape	Accuracy $= 0.257 \pm 0.007$
Pixel-intensity nuisance	$R^2 = 0.659 \pm 0.004$

Object-shape classification accuracy approaches the chance level of 0.25 for the balanced four-class task. In contrast, the nuisance remains strongly recoverable in every seed, with all bootstrap confidence-interval lower bounds exceeding 0.64. Thus, under the same compression strength, the lower-energy, task-relevant shape factor becomes nearly unrecoverable while the higher-energy, task-irrelevant nuisance remains encoded.

Interpretation and scope. This experiment provides a direct empirical counterpart to Theorem 3.1. The theorem establishes an energy-dependent activation criterion in its stated linear Gaussian setting, while the controlled nonlinear counterexample demonstrates that output-space energy ordering can conflict with task-specific semantic priority. The evidence supports a possible semantic-agnostic failure mode of compression-driven representation learning, without asserting a universal ordering of semantic factor retention for arbitrary nonlinear models. The matched-pair output-energy measurements are finite-change quantities; they are related to, but not identical to, the local Jacobian-based effective factor energy $\bar{E}_m$ discussed in Appendix C.1.

E Factor Discovery Module: Implementation Details

This appendix provides supplementary details for the four FDM stages described in Section 4.3. The formal definitions of the schema $\tilde{\mathcal{F}}$, prevalence $\hat{\pi}_k(v)$, prompt set $\mathcal{P}_k(v)$, and dictionary $\mathcal{D}$ are introduced in the main text; here we focus on design rationale and examples that justify each step.

Step 1 (Structured captioning): why structured fields. Compared to free-form captions, structured fields enforce consistent keys across images, which reduces synonym noise (*e.g.*, mapping "golden" and "blonde" into a shared field `hair_color`) and prevents brittle factor definitions during downstream aggregation.

Step 2 (Factor induction): canonicalization. The MLLM/LLM is prompted to (i) propose candidate factors with clear semantic scope (one concept per factor), (ii) normalize field names into canonical identifiers (*e.g.*, `hair_color`, `fur_length`, `car_body_type`), and (iii) merge semantically equivalent values into a small set of discrete states. This is crucial for avoiding spurious multi-meaning factors.

Step 3 (Factor selection): retention criteria. A candidate factor $\tilde{f}_k$ is retained if it satisfies all of the following:

- **Coverage:** $\sum_{v \in \tilde{\mathcal{V}}_k} \hat{\pi}_k(v)$ is high (frequently observable, not sporadic).
- **Non-degeneracy:** at least two values have non-trivial support (to avoid near-constants).
- **Visual separability:** the factor admits reliable CLIP separation on a held-out subset.

Factors that are too rare, unstable, or visually ambiguous are discarded.

Step 4 (Dictionary construction): sentence-level prompts and ternary templates. Using isolated words (*e.g.*, "blonde", "SUV") often yields unstable CLIP scores because the text embedding can be dominated by spurious context or background correlations; sentence-level prompts $p_{k,j}(v)$ instead bind the attribute to the intended subject. To support robust validity masking, a ternary template is adopted for each factor: two visually dominant values are designated as A and B, while remaining values are pooled into O ("Other"), giving prototype prompt sets $\{\mathcal{P}_k^{\mathrm{A}}, \mathcal{P}_k^{\mathrm{B}}, \mathcal{P}_k^{\mathrm{O}}\}$ used by the CLIP scoring module (Appendix E.1).

Example on CelebA. For FDM on CelebA, we randomly sample 80% of its $202{,}599$ images ($162{,}079$) and use GPT-4o for dense captioning and factor induction. Each image is described with the open-ended prompt "Describe this image in as much detail as possible." The resulting descriptions are consolidated into canonical attributes and values: for example, "golden hair" and "blonde hair" are merged into "blonde" under `hair_color`, alongside values such as "black", "brown", and "gray". This yields 43 candidate factors, of which 21 are retained after coverage, non-degeneracy, and CLIP-separability filtering and used to construct semantic guidance. The retained values are paired with subject-grounded descriptions, such as "The person has blonde hair", for subsequent CLIP matching.

E.1 CLIP-based Semantic Signal Construction

This section details the standard CLIP-based instantiation step used to convert the FDM prompt dictionary into pseudo-labels and validity masks. The main methodological novelty lies in the FDM construction of $\mathcal{D}$, while this module provides a lightweight image–text scoring interface for training.

Given the prompt dictionary $\mathcal{D}$ from FDM, semantic signals are derived utilizing a frozen CLIP encoder [25]. For each factor f_k, the corresponding prompt set $\mathcal{P}_k$ is partitioned into three subsets $\mathcal{P}_k^{\mathrm{A}}, \mathcal{P}_k^{\mathrm{B}}, \mathcal{P}_k^{\mathrm{O}}$. Here, A and B correspond to two salient, different factor values, while O is an explicit reject class describing instances that are neither A nor B. Let $\phi_I(\mathbf{x}) \in \mathbb{R}^{d_{\mathrm{clip}}}$ and $\phi_T(p) \in \mathbb{R}^{d_{\mathrm{clip}}}$ denote the ℓ_2-normalized image and text embeddings produced by CLIP. For each factor k and class $c \in \{\mathrm{A}, \mathrm{B}, \mathrm{O}\}$, prompts are aggregated into a single prototype embedding:

$$\bar{\phi}_k^c \;=\; \mathrm{normalize}\left(\frac{1}{|\mathcal{P}_k^c|} \sum_{p \in \mathcal{P}_k^c} \phi_T(p)\right). \tag{44}$$

Given an image $\mathbf{x}_i$ and factor f_k, the cosine similarities to these three prototypes are computed:

$$s_{i,k}^c \;=\; \left\langle \phi_I(\mathbf{x}_i),\, \bar{\phi}_k^c \right\rangle, \qquad c \in \{\mathrm{A}, \mathrm{B}, \mathrm{O}\}. \tag{45}$$

We then predict $\hat{c}_{i,k} = \arg\max_c s_{i,k}^c$ and convert it into a binary target and a validity mask:

$$m_{i,k} \;=\; \mathbb{I}[\hat{c}_{i,k} \in \{\mathrm{A}, \mathrm{B}\}], \quad y_{i,k} \;=\; \mathbb{I}[\hat{c}_{i,k} = \mathrm{A}]. \tag{46}$$

Table 5: Per-attribute Probe accuracy of SAGE-tLGC (7k iters) on CelebA. Mean Probe accuracy of the supervision-only variant is reported in the caption for reference.

Attribute	Accuracy ↑
Eyeglasses	0.93
Bangs	0.78
Young	0.71
Bald	0.62
Heavy_Makeup	0.85
Male	0.95
Mean	**0.81**

Reference: SAGE-no-tLGC (158k iters) mean Probe accuracy $= 0.42$.

Crucially, only the pairs with $m_{i,k} = 1$ contribute to downstream guidance, effectively utilizing $\hat{c}_{i,k} = \mathrm{O}$ as an explicit rejection to filter ambiguous samples. The resulting $(y_{i,k}, m_{i,k})$ pairs are used in the masked guidance loss and contrastive objective in Section 4.3.

F Output-Space Probing for tLGC Verification

This appendix contains the full protocol and results of the output-space probing study referenced in Section 5.3. The goal is to isolate the contribution of the triangular latent geometry constraint (tLGC) from the BCE supervision signal alone, using an end-to-end probe that measures *decoder-side* attribute controllability rather than encoder–label alignment.

Output-space attribute probing protocol. For each supervised latent dimension k with attribute y_k, we (i) DDIM-invert each test image to its noise latent, (ii) sweep z_k^{sem} over $[-2, +2]$ in 9 equal steps, (iii) run DDIM sampling and VAE decoding, and (iv) feed the decoded images into an independently trained CelebA attribute classifier (ResNet-18 backbone, 91.7% mean validation accuracy at 256×256). We record $|\Delta P_k| = |\, P(y_k{=}1 \mid z_k{=}{+}2) - P(y_k{=}1 \mid z_k{=}{-}2)\,|$ per sample and report the per-attribute mean as the *Probe accuracy*.

Isolating tLGC from supervision. We probe two variants of our framework on the same six CelebA attributes (Eyeglasses, Bangs, Young, Bald, Heavy_Makeup, Male) using identical pseudo-label supervision: (a) SAGE-tLGC with the triangular transform, after $7{,}000$ iterations; and (b) SAGE-NO-tLGC without the transform, trained to $158{,}000$ iterations ($22\times$ more compute). Table 5 reports per-attribute Probe accuracy across the two variants. The tLGC model wins on every attribute despite $22\times$ less training. The mean accuracy reaches 0.81 and substantially exceeds the supervision-only variant (mean 0.42).

Implication. With the identical BCE supervision signal, adding the tLGC module produces a statistically significant improvement in end-to-end decoder controllability that $22\times$ more training on the supervision-only model cannot achieve. Per-dim latent std also reflects this: tLGC spans $[0.0002, 0.99]$ while supervision-only spans $[0.0003, 0.44]$, indicating tLGC resists posterior collapse on unguided axes. The Bald attribute is severely long-tailed ($\sim 2\%$ prevalence) and is addressed by the Distribution-Aware Balancing module, whose effect is reported separately in the ablation study (Section 5.5).

G Implementation Details

In this section, we provide comprehensive implementation details of our SAGE framework, including network architectures, training hyperparameters, and dataset configurations. The main training experiments were conducted on four NVIDIA A100 (40 GB) GPUs with PyTorch 2.0, CUDA 11.8, and Python 3.10; end-to-end training takes roughly two days per dataset. The training-efficiency experiments used a single RTX 4090, as detailed in Appendix H.4. SAGE is implemented on a Latent Diffusion backbone. `GPT-4o` is used for factor discovery and CLIP (ViT-L/14) for pseudo-labeling. The default loss weights are $\lambda_{\mathrm{diff}} = 2.0,\ \lambda_{\mathrm{guid}} = 5.0,\ \lambda_{\mathrm{exc}} = 2.0$, selected by the sensitivity sweeps reported in Figure 7.

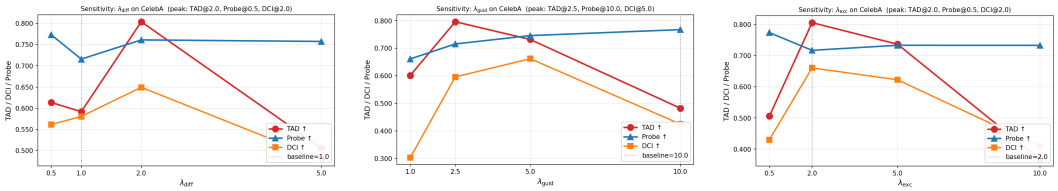


Figure 7: **Hyperparameter sensitivity on CelebA.** Single-factor sweeps over λ_{diff} (left), λ_{guid} (middle), and λ_{exc} (right). Each curve reports TAD, DCI, Probe, and their average; the recommended setting (peak of the average) is λ_{diff}=2.0, λ_{guid}=5.0, λ_{exc}=2.0.

Hyperparameter sensitivity. We perform single-factor sweeps on CelebA over λ_{diff}, λ_{guid}, and λ_{exc}, holding the remaining weights at their default values. As shown in Figure 7, all three coefficients exhibit a clear unimodal trend on the average of TAD, DCI, and Probe: λ_{diff} peaks at 2.0 (TAD/DCI both maximize while Probe stays nearly flat), λ_{guid} peaks at 5.0 (with 2.5 within 0.01 of optimal), and λ_{exc} peaks at 2.0. Both λ_{guid} and λ_{exc} recover the values used in our default configuration, while raising λ_{diff} from the previously reported 1.0 to 2.0 further improves the latent-side metrics without trading off output-space Probe. Performance degrades smoothly when any coefficient is moved $\pm$one tick away from the optimum, indicating that SAGE is robust to moderate perturbations in the loss weights.

Architecture. The encoder follows a residual convolutional design as follows:

- **Input projection**: Conv2d$(C_{\text{in}}, C_{\text{base}}, 3 \times 3)$ with padding 1
- **Downsampling blocks**: Sequential residual blocks with channel multipliers $[1, 2, 4, 4]$, each containing:
  - GroupNorm(32 groups) $\to$ SiLU $\to$ Conv2d$(C, C, 3 \times 3)$
  - GroupNorm(32 groups) $\to$ SiLU $\to$ Conv2d$(C, C, 3 \times 3)$
  - Residual connection with optional 1×1 convolution for channel matching
  - AvgPool2d(2×2) for spatial downsampling
- **Attention block**: Self-attention at the bottleneck resolution
- **Output heads**: Parallel linear projections:

$$\mathbf{a} = \text{fc}_a(\text{flatten}(\mathbf{h})) \in \mathbb{R}^D \tag{47}$$

$$\boldsymbol{\mu} = \text{fc}_\mu(\text{flatten}(\mathbf{h})) \in \mathbb{R}^D \tag{48}$$

$$\log \boldsymbol{\sigma}^2 = \text{fc}_{\text{var}}(\text{flatten}(\mathbf{h})) \in \mathbb{R}^D \tag{49}$$

The semantic latent is sampled via reparameterization:

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{50}$$

The diffusion backbone is a conditional U-Net that generates images conditioned on semantic latents. We employ two conditioning mechanisms: Adaptive Group Normalization (AdaGN) and Cross-Attention.

Adaptive Group Normalization (AdaGN). Given feature maps $\mathbf{h} \in \mathbb{R}^{C \times H \times W}$ and condition $\mathbf{c} \in \mathbb{R}^D$, AdaGN modulates normalized features:

$$\boldsymbol{\gamma}, \boldsymbol{\beta} = \text{split}(\text{Linear}(\mathbf{c}), \text{dim} = C) \tag{51}$$

$$\text{AdaGN}(\mathbf{h}, \mathbf{c}) = \text{GroupNorm}(\mathbf{h}) \cdot (1 + \boldsymbol{\gamma}) + \boldsymbol{\beta} \tag{52}$$

Cross-Attention Block. Cross-attention allows spatial features to attend to semantic conditions:

$$\mathbf{Q} = \mathbf{W}_Q \cdot \text{reshape}(\mathbf{c}) \tag{53}$$

$$\mathbf{K} = \mathbf{W}_K \cdot \text{reshape}(\mathbf{h}) \tag{54}$$

$$\mathbf{V} = \mathbf{W}_V \cdot \text{reshape}(\mathbf{h}) \tag{55}$$

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{56}$$

U-Net Architecture. The U-Net follows an encoder-decoder structure with skip connections:

- **Encoder**: 4 resolution levels with channel multipliers $[1, 2, 4, 4]$
- **Bottleneck**: ResBlock $\rightarrow$ Attention $\rightarrow$ ResBlock
- **Decoder**: Symmetric to encoder with skip connections
- **Time embedding**: Sinusoidal positional encoding projected to hidden dimension
- **Condition injection**: AdaGN in all ResBlocks, CrossAttention at selected resolutions

The base channel count is 128, resulting in the following channel progression: $128 \rightarrow 256 \rightarrow 512 \rightarrow 512$.

G.1 Dataset Configurations

Datasets. We evaluate our method on five diverse datasets covering synthetic objects, human faces, animal faces, and fine-grained vehicles:

Shapes3D [16] is a synthetic dataset consisting of 480,000 images with resolution 64×64. It is generated from 6 ground truth latent factors (floor hue, wall hue, object hue, scale, shape, and orientation), making it an ideal benchmark for disentangled representation learning.

CelebA-HQ [46] is a high-quality version of the CelebA dataset. It contains 30,000 facial images at a resolution of 1024×1024. This dataset is widely used for evaluating high-fidelity unconditional image generation and attribute editing.

FFHQ (Flickr-Faces-HQ) [47] consists of 70,000 high-quality PNG images of human faces at 1024×1024 resolution. Compared to CelebA, FFHQ offers considerably more variation in terms of age, ethnicity, and image background.

AFHQ-v2 (Animal Faces-HQ) [48] contains 15,000 high-quality images at 512×512 resolution, categorised into three domains: cat, dog, and wildlife. We use the v2 release, which improves quality by removing lower-quality samples.

CompCars [49] is a large-scale fine-grained vehicle dataset. It comprises over 136,000 web-nature images and 50,000 surveillance-nature images covering 1,716 car models. We utilize the web-nature subset for our experiments due to its rich viewpoint and attribute annotations.

H Additional Qualitative Results and Robustness Analysis

To provide a more comprehensive assessment of SAGE's capabilities and robustness, we present additional experimental results in this section. Specifically, we demonstrate: (1) The **generalizability** of our proposed framework by applying it to a VAE backbone (β-TCVAE); (2) The **robustness** of our Factor Discovery Module (FDM) across different Large Language Models (LLMs); (3) The **visual necessity** of each component through qualitative ablation studies; and (4) The **training efficiency and scalability** of SAGE on CelebA and ImageNet-100.

H.1 Generalizability to VAE Frameworks

While SAGE is implemented on a Latent Diffusion backbone in the main paper, our core theoretical contributions are model-agnostic. To demonstrate this versatile applicability, we integrated these modules into the training of β-TCVAE, a leading VAE-based baseline.

Figure 8 illustrates the latent traversals of the improved β-TCVAE. Integrating SAGE yields exceptional consistency between the learned factors and the semantic guidance, ensuring that specific axes strictly correspond to the target concepts. Notably, these geometric constraints benefit the entire latent space: even the unguided dimensions (visualized in the bottom two rows) exhibit significantly sharper factor isolation compared to the baseline. **In stark contrast**, directly imposing identical semantic constraints on vanilla β-TCVAE and β-VAE triggers optimization instability, leading to **severe entanglement or posterior collapse**, where the models fail to capture the desired attributes entirely. This evidence suggests that SAGE serves as a robust, universal stabilizer that enables controllable disentanglement across diverse generative architectures.

Table 6 reports the corresponding quantitative results on CelebA. While supervision-only guidance decreases TAD, DCI, and Probe relative to vanilla β-TCVAE, the full SAGE integration improves all three metrics.

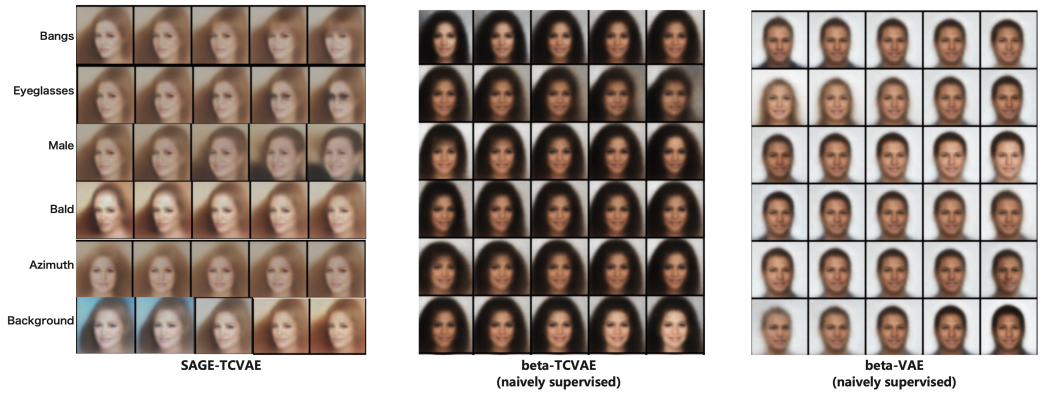


Figure 8: **Generalization to VAE Architectures.** Latent traversals of β-TCVAE trained with our proposed SAGE framework on CelebA. The results demonstrate that our geometric constraint effectively enforces semantic alignment even within a VAE latent space, validating the architectural universality of our method.

Table 6: **Quantitative Results on a VAE Backbone.** CelebA comparison of vanilla β-TCVAE, semantic supervision only, and full SAGE integration. Higher is better for all metrics.

Method	**TAD** $\uparrow$	**DCI** $\uparrow$	**Probe** $\uparrow$
β-TCVAE	0.353	0.039	0.169
β-TCVAE + supervision only	0.123	0.037	0.147
β-TCVAE + full SAGE	**0.611**	**0.493**	**0.661**

H.2 Robustness Analysis of LLM Selection

Our Factor Discovery Module leverages LLMs as semantic proxies. To ensure that SAGE is not reliant on a specific model, we conduct a multi-dimensional robustness evaluation comparing `GPT-4o`, `GPT-4v`, and `GLM-4`. The evaluation hierarchy consists of: (1) Fine-grained recognition accuracy, (2) Dictionary construction consistency, and (3) Downstream disentanglement performance.

1. Fine-grained Recognition Accuracy. We first evaluate the capability of different LLMs to accurately identify attributes against ground-truth labels on CelebA. As illustrated in Figure 9, `GPT-4o` demonstrates superior stability with an average accuracy of 88.4% and a negligible zero-scoring rate. In contrast, `GLM-4` exhibits significant variance on specific long-tailed attributes, while `GPT-4v` maintains intermediate performance (78.9%).

2. Dictionary Construction Consistency (IoU). Beyond fine-grained recognition accuracy, a critical requirement for FDM is the stability of the generated semantic basis across different trials. To quantify this, we explicitly calculate the *Inter-run Consistency* using the Intersection over Union (IoU) of the extracted factor sets across five independent runs:

$$\text{Consistency} = \frac{|\bigcap_{i=1}^{N} \mathcal{F}_i|}{|\bigcup_{i=1}^{N} \mathcal{F}_i|} \tag{57}$$

where $\mathcal{F}_i$ denotes the factor dictionary obtained from the i-th run.

As reported in Table 7, the quantitative results empirically confirm that our pipeline yields a highly stable semantic consensus. `GPT-4o` exhibits near-deterministic consistency. Crucially, even less capable models like `GLM-4` maintain a high consistency score of 0.83. This demonstrates that while weaker models may fluctuate in recognizing specific image instances (as shown in Figure 9), they reliably converge on the same set of dominant semantic concepts, ensuring the robustness of the factor dictionary construction.

3. Impact on Disentanglement. Finally, we evaluate the downstream performance on CelebA. We utilized the factor dictionaries constructed by `GPT-4o`, `GPT-4v`, and `GLM-4`, randomly selecting 5 sets (each containing 8 semantic factors) from each model to serve as semantic guidance for training. As

Table 7: **Dictionary Construction Consistency.** We report the Intersection over Union (IoU) of the factor dictionaries generated across 5 independent runs. The high IoU scores indicate that FDM converges to a stable semantic basis regardless of the underlying LLM.

METRIC	GPT-4O	GPT-4V	GLM-4
FACTOR CONSISTENCY (IOU)	0.91	0.88	0.83

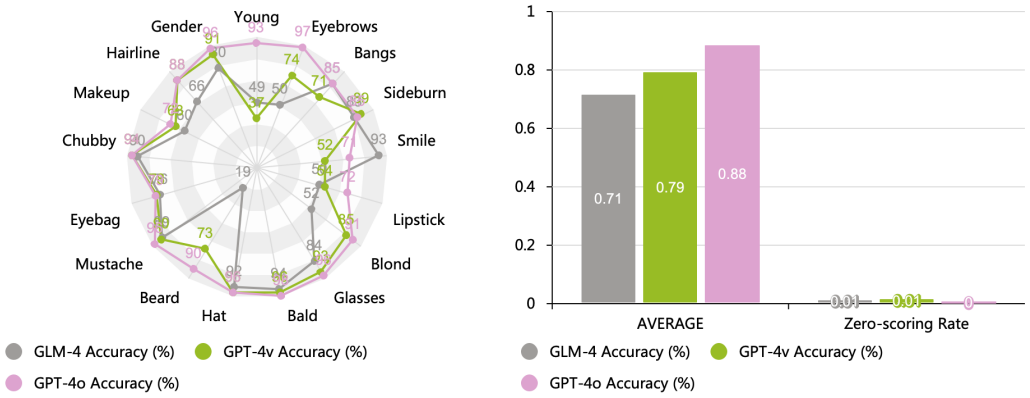


Figure 9: **Fine-grained Attribute Recognition Accuracy across LLMs.**

shown in Table 8, the resulting FID and TAD scores are statistically indistinguishable across different LLM sources. This uniformity suggests that once a reasonable semantic basis is established, the final performance is upper-bounded by the alignment precision of the CLIP encoder rather than the reasoning capability of the LLM. Consequently, SAGE maintains high effectiveness even with less capable open-source models.

Table 8: **Impact of LLM Selection on Training Performance.** We report the FID and TAD scores on CelebA using semantic prompts generated by different MLLMs.

LLM SOURCE	FID ↓	TAD ↑
GPT-4O	9.3 ± 0.6	0.803 ± 0.011
GPT-4V	9.1 ± 1.3	0.781 ± 0.014
GLM-4	9.3 ± 0.8	0.797 ± 0.020

H.3 Visual Ablation Studies

Complementing the quantitative ablation analysis in Section 5.5, we provide qualitative visualizations to intuitively reveal the specific failure modes associated with each incomplete configuration. Figure 11 presents a side-by-side comparison of latent traversals under different ablation settings.

1. Impact of Triangular Latent Geometry Constraint (tLGC). To verify the role of geometric anchoring, we attempt to manipulate the *Eye Color* attribute on the AFHQ dataset.

- **Observation:** As shown in the top row of Figure 11, applying semantic guidance for the *Eye Color* without tLGC leads to a clear semantic misalignment. Instead of modulating the iris color as intended (circled in red), the traversal induces irrelevant variations in the *Fur Color* and texture. The target attribute remains largely invariant, while the semantic signal is incorrectly mapped to global features.
- **Analysis:** This failure empirically confirms the issue of rotational symmetry. Without the triangular geometric constraint to strictly anchor the axis, the semantic signal cannot lock onto the specific, subtle direction of "eye color" and instead drifts towards dominant, high-variance directions like fur texture.

2. Impact of Semantic Exclusiveness Mechanism (SEM). We investigate the necessity of SEM by applying semantic guidance for the *Eyeglasses* attribute on CelebA.

- **Observation:** In the absence of Semantic Exclusiveness (middle row), the model suffers from severe semantic coupling. While the target attribute is generated, it is inextricably entangled with the *Beard* attribute. As the traversal adds values of the axis, it simultaneously imposes a beard on the subject (circled in red).
- **Analysis:** This indicates a leakage of semantic information into the nuisance subspace. Without the orthogonality enforced by SEM, the model relies on spurious correlations in the dataset, failing to achieve independent control.

3. Impact of Distribution-aware Balancing. To validate our balancing strategy, we first analyze the attribute distribution of CelebA. As shown in Figure 10, the dataset exhibits a long-tail distribution, where attributes like *Wearing Hat* are highly imbalanced ($< 5\%$). We select this rare attribute for testing.

- **Observation:** Without the distribution-aware balancing module (bottom row), the model exhibits feature collapse for this rare concept. The traversal fails to generate the hat distinctively or produces generic artifacts.
- **Analysis:** This validates our theoretical insight on 3.1. Rare attributes possess low variance and are spectrally pruned as noise during training. By re-weighting these components, the Full Architecture successfully preserves the weak signal, enabling precise control over the *Wearing Hat* attribute even under extreme data imbalance.

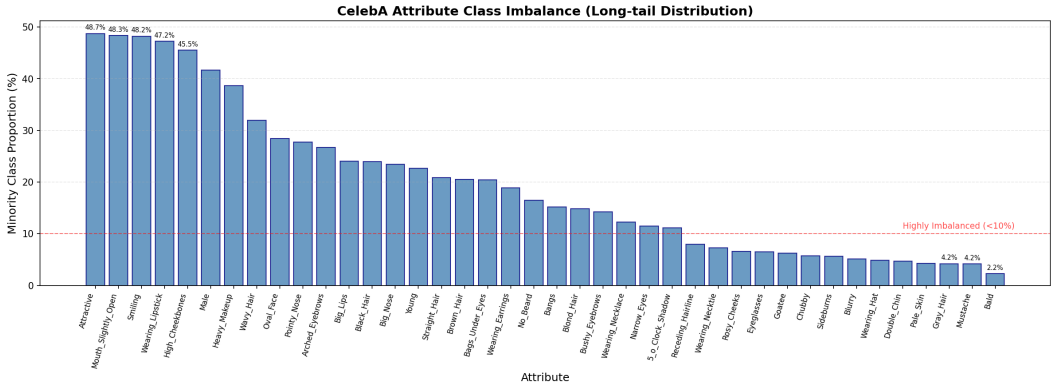


Figure 10: **Long-tail Attribute Distribution on CelebA.** The bar chart illustrates the minority class proportion for all 40 attributes. The red dashed line marks the threshold for highly imbalanced attributes ($< 10\%$).

H.4 Scalability and Computational Efficiency

We compare the training efficiency of SAGE with four representative diffusion-based DRL methods on CelebA using a single NVIDIA RTX 4090 and batch size 4. Iteration time (TT) is averaged over 50 optimization steps following 10 warm-up steps. Parameter counts exclude frozen autoencoders, whereas GFLOPs, peak GPU memory, and TT include online image encoding. Table 9 summarizes the matched-setting measurements. SAGE has the lowest measured peak memory (5,691.417 MB) and iteration time (0.0792 s/iter) among the compared methods.

We further evaluate SAGE on ImageNet-100 to examine its practical scalability to a more visually diverse dataset. As shown in Table 10, SAGE attains the lowest FID (13.404), peak memory (5,781.6 MB), and iteration time (0.0641 s/iter) among the evaluated methods. These results assess generation quality and computational feasibility rather than semantic disentanglement on ImageNet-100.

The computational measurements in both tables cover online training; offline MLLM/CLIP-based factor discovery and semantic-signal preparation costs are not included in the per-iteration timings.

I Additional Visualization Results on Diverse Datasets

To further demonstrate the generalization capability and robustness of SAGE, we provide comprehensive qualitative results across varying domains. Specifically, we select representative datasets from

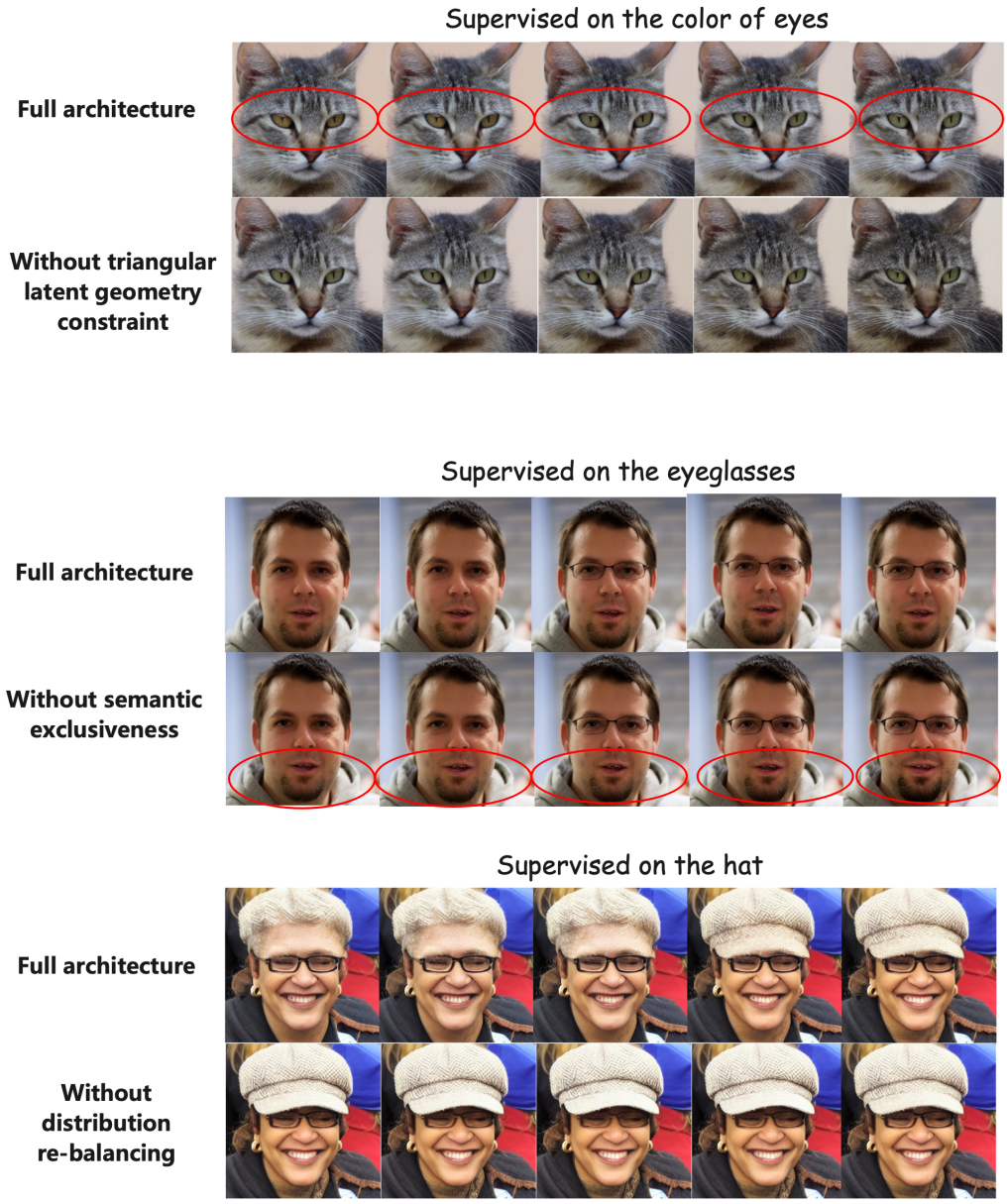


Figure 11: **Visual Ablation Study.** *Top (w/o tLGC):* The model fails to anchor the *Eye Color* attribute, resulting in irrelevant variations. *Middle (w/o Exclusiveness):* Applying semantic guidance for *Eyeglasses* leads to entanglement with *Beard*, showing semantic leakage. *Bottom (w/o Balancing):* The model fails to learn the rare *Wearing Hat* attribute, whereas the full model generates it correctly.

three distinct categories: **Shapes3D** (synthetic structures), **CelebA** (human faces), and **AFHQ-v2** (fine-grained animal images). As shown in Figure 12, SAGE effectively discovers and disentangles meaningful semantic factors in all scenarios.

1. Synthetic Benchmarks (Shapes3D). The top row of Figure 12 illustrates the latent traversals on Shapes3D. SAGE precisely captures the ground-truth generative factors, such as *Floor Hue*, *Wall Hue*, and *Object Shape*. The clean separation between geometry and texture confirms the model's geometric understanding in simplified environments.

2. Human Facial Domains (CelebA). The middle row presents results on the CelebA dataset. SAGE achieves fine-grained control over specific facial attributes like *Eyeglasses*, *Bangs*, and *Makeup*.

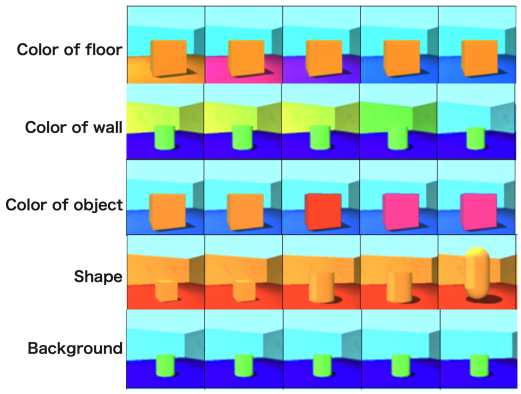
Table 9: Training efficiency on CelebA. Parameter counts exclude frozen autoencoders; GFLOPs, peak memory, and TT include online image encoding.

Method	**Params (M)**	**GFLOPs**	**Peak Mem. (MB)** $\downarrow$	**TT (s/iter)** $\downarrow$
DisDiff	54.203	369.675	11,076.911	0.9981
EncDiff	38.844	82.958	9,480.214	0.3389
Diff-AE	55.273	344.012	5,953.333	0.0876
InfoDiffusion	40.052	514.949	16,276.157	0.2654
SAGE	61.518	282.787	**5,691.417**	**0.0792**

Table 10: ImageNet-100 scalability comparison. FID measures generation quality; memory and TT measure online training efficiency.

Method	**FID** $\downarrow$	**Peak Mem. (MB)** $\downarrow$	**TT (s/iter)** $\downarrow$
DisDiff	19.773	11,615.0	0.9981
EncDiff	17.908	9,940.7	0.3389
Diff-AE	15.068	6,242.5	0.0876
InfoDiffusion	18.243	17,066.8	0.2654
SAGE	**13.404**	**5,781.6**	**0.0641**

3. Fine-grained Domains (AFHQ-v2). The bottom row showcases the performance on the challenging AFHQ-v2 dataset. SAGE clearly separates domain-specific attributes, such as *Texture* and *Color* characteristics from the global pose.



(a) **Shapes3D:** Disentanglement of primitive factors including Floor Color, Wall Color, Object Color, and Shape.



(b) **CelebA:** High-fidelity manipulation of attributes such as Eyeglasses, Bangs, and Makeup while preserving identity.



(c) **AFHQ-v2:** Disentangling fine-grained traits such as Fur Color, Eye Color, Fur Length, and Texture.

Figure 12: **Qualitative Results on Diverse Domains.** We provide visualization results across synthetic (Shapes3D), human facial (CelebA), and fine-grained animal (AFHQ-v2) datasets. SAGE consistently learns semantic factor from simple geometry to complex datasets and achieves precise disentanglement across all domains.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction are supported by Sections 3–5.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A dedicated discussion of limitations and future directions is provided in Appendix A.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Assumptions are stated with each result; proofs are provided in Appendix C.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Implementation details are provided in Appendix G.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The code is currently being prepared for public release. All datasets used are publicly available, and implementation details are provided in Appendix G.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental settings are provided in Appendix G.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Tables 2 and 3 report mean $\pm$ standard deviation across runs; Table 1 reports point estimates.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Main training used four A100 (40 GB) GPUs; efficiency experiments used one RTX 4090. See Appendices G and H.4.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Positive and negative societal impacts are discussed in Appendix B.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No pre-trained models or scraped datasets are released.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Datasets and models are cited and used under applicable licenses and terms of use, including the provider's terms for GPT-4o.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new datasets or pre-trained models are released in this submission.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing or human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve human subjects research.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLM usage is described in Section 4.3 and Appendix H.2.