

X-MULTI: VLM-based Imaging Factor Disentanglement for Factor-Aware Image Synthesis

Sonali Godavarthy^{1*}[0009-0001-5379-1139], Matthias
Neuwirth-Trapp^{2,3}[0009-0006-5320-6354], Tim-Felix Faasch³[0009-0004-7410-5232],
Maarten Bieshaar³[0000-0002-6471-6062], Michael Moeller¹[0000-0002-0492-6527],
Kristof Van Laerhoven¹[0000-0001-5296-5347], and Danda Pani
Paudel⁴[0000-0002-1739-1867]

¹ University of Siegen, Siegen, Germany

² ETH Zürich, Zürich, Switzerland

³ Bosch Research, Hildesheim, Germany

⁴ INSAIT, Sofia University “St. Kliment Ohridski”, Sofia, Bulgaria

Abstract. Imaging factor disentanglement in text-to-image generation aims to independently control image acquisition properties such as types of camera lenses, sensor types, viewpoints, and domains to enable combinatorial generalization. This should let the model synthesize novel factor combinations unobserved in the training data, such as pairing a fisheye lens with an event sensor never observed in training data. Recent work, MULTI, introduced learnable, factor-specific embeddings to disentangle imaging factors, along with the Factor Alignment Accuracy (FAA) metric to evaluate disentanglement quality. We identify and address two independent limitations. First, MULTI’s pixel-level reconstruction objective supervises the model only on observed imaging factor combinations, providing no direct training signal for novel combinations. We therefore propose X-MULTI, which uses a pretrained vision-language model (VLM) to supervise novel factor combinations synthesized during training. Second, we show the FAA metric exhibits severe cross-factor correlation leakage, misrepresenting true disentanglement quality. We therefore propose Improved-FAA (I-FAA), which employs factor-specific augmentation strategies to break these correlations and enables more rigorous evaluation. Experiments demonstrate that X-MULTI achieves improved factor alignment on novel combinations compared to MULTI. Moreover, we show that correlation leakage in FAA distorts the evaluation of true factor disentanglement and I-FAA reduces this leakage and therefore provides a more robust assessment of factor alignment.

Keywords: Text-to-Image · Diffusion Models · Factor Disentanglement · Textual Inversion · Vision-Language Model

* This work was conducted as part of a Master’s thesis in collaboration with Bosch Research.

1 Introduction

Text-to-image (T2I) diffusion models have achieved remarkable visual fidelity in synthesizing images from natural language descriptions [9, 29–31, 33]. However, precise control over low-level image acquisition properties such as types of camera lens, sensor type, viewpoint, and domain remains challenging [13]. These factors define different aspects of the image formation process, but in real datasets they often appear only in limited and correlated combinations. Therefore, a model can be considered factor-disentangled only if it can modify each factor independently and synthesize valid novel combinations that were not observed during training. This problem, termed *imaging factor disentanglement*, requires learning representations that capture the semantic and physical identity of each factor while allowing them to be modified independently [13] to generate images with novel factor combinations that are not present during training. Controlling these imaging factors is important because collecting real data for all acquisition conditions is expensive, while vision systems still need to remain robust across sensor types, lenses, viewpoints, and domains.

Recent work, MULTI [13], introduced learnable factor-specific text embeddings optimized using Textual Inversion (TI) [11]. These embeddings are inserted into the text prompt to represent imaging factors, and are optimized with the standard diffusion reconstruction loss while keeping the T2I diffusion model frozen. However, MULTI’s reconstruction-based training objective is applied only for factor combinations that are represented in the training data, so the training does not receive any signal for novel, unseen combinations. This leaves the model unable to reliably distinguish semantically distinct factors (e.g., **thermal sensor** from **rgb sensor**), resulting in factor confusion at inference time. To encourage disentanglement and thereby allow the synthesis of images for novel factor combinations, a supervisory signal is required that can provide meaningful signal during training. We therefore propose **X-MULTI** (eXtended MULTI) which introduces supervision from a zero-shot Vision-Language Model (VLM) [2, 27]. During training, images with novel factor combinations are generated and the VLM acts as external semantic classifier, independently predicting factors for these generated images and thereby providing factor-level supervisory signals.

To evaluate whether generated images contain the specified factors, authors of MULTI [13] proposed Factor Alignment Accuracy (FAA), a novel classifier-based metric measuring factor correctness. We demonstrate that this metric is contaminated by severe correlations between the factor classifiers, enabling the metric to exploit these correlations rather than independently taking each one into account. This evaluation misrepresents the true quality of imaging factor disentanglement. To address this critical limitation, we introduce **I-FAA** (Improved Factor Alignment Accuracy), a novel evaluation metric, which employs class-balanced under-sampling and diverse factor-specific augmentation strategies during training of the imaging factor classifiers to break correlations between factors and mitigate shortcut learning of dataset-specific cues. This design provides more accurate factor disentanglement evaluation by eliminating the reliance on inter-factor patterns.

To study imaging factor disentanglement, we conduct experiments on DFRICO [13] benchmark. First, we show improvements of our method over MULTI and other baselines, particularly for novel factor combinations. Second, we show cross-factor leakage in FAA, and I-FAA through correlation analysis.

Our main contributions are:

- **X-MULTI Approach:** Introduces zero-shot VLM-based supervision for synthetically generated unseen imaging-factor combinations, providing explicit factor-level supervision beyond the reconstruction-only training.
- **I-FAA Metric:** Introduces an improved factor-alignment evaluation metric that reduces cross-factor shortcut learning through class-balanced training and factor-specific augmentations, enabling more robust assessment of imaging factor disentanglement.

2 Related Work

Disentanglement in Generative Models: T2I models such as SDXL [30] provide strong generative backbones, while DreamBooth [32] adapts these models to specific visual concepts. Disentanglement in T2I models typically focuses on separating objects from their attributes, such as color or texture. Early approaches utilized Low-Rank Adaptation (LoRA) blocks to decouple style from content [10, 23]. Textual Inversion (TI) [11] optimizes a new token embedding for a target concept while keeping the model weights frozen, which is then inserted into prompts to generate that concept. This idea has been applied to learn embeddings for visual concepts such as color, shape, style, texture and object appearance [3, 28, 38, 39, 43]. To prevent attribute leakage, several methods steer cross-attention maps by constraining concept tokens to attend to their corresponding image regions [5, 15, 18, 36, 47] or employ explicit masks to isolate foreground objects from their backgrounds [1, 19, 22]. More recently, transformer-based architectures have leveraged the modulation space for multi-concept disentanglement [12, 49]. However, these methods mainly disentangle semantic or personalized concepts rather than imaging acquisition factors, making direct comparison with X-MULTI difficult.

Imaging-Factor Disentanglement: The image synthesis conditioned on imaging factors has received less attention. MULTI [13] addressed this by utilizing TI to optimize imaging factor embeddings. While MULTI improved over baselines for novel factor combinations, its factor alignment remained limited. We improve upon MULTI by introducing VLM-guided supervision to encourage disentanglement of imaging factors.

Discriminator-Based Guidance and VLM Supervision: Large-Language Models (LLMs) and Vision-Language Models (VLMs) have emerged as versatile zero-shot discriminators [27]. LLMs and VLMs are increasingly repurposed for complex tasks such as object detection [2], optical character recognition [20], and

more [7, 41]. Recent frameworks have integrated these models as frozen discriminators to provide guidance, ensuring that generative models follow high-level conceptual constraints [2, 4, 42]. X-MULTI leverages this paradigm by utilizing a VLM as an external classifier to enforce factor identity, improving image synthesis of novel factor combinations.

Disentanglement Evaluation Metrics: Evaluating factor disentanglement requires metrics that assess each factor in isolation. [13] introduced FAA, which evaluates imaging-factor disentanglement using factor classifiers trained on the DF-RICO benchmark [13]. While classifier-based evaluation metrics like FAA often rely on highly correlated, co-occurring dataset shortcuts, they fail to isolate individual target attributes. To address this, I-FAA introduces factor-specific augmentations that break these correlations to ensure reliable disentanglement assessment.

3 Background

Imaging Factor Disentanglement: Following MULTI [13], imaging factor disentanglement aims to decompose image formation into $\mathcal{K}$ factors such as lens, sensor, viewpoint, and domain. We consider a collection of datasets $\mathcal{D} = D^{(1)}, \dots, D^{(N)}$ with shared factor annotations, where each image is associated with a factor tuple $\mathbf{f} = (f^{(k)})_{k \in \mathcal{K}}$. Each factor k takes values from a discrete category $\mathcal{F}^{(k)}$, defining the full factor space $\mathcal{F} = \mathcal{F}^{(1)} \times \dots \times \mathcal{F}^{(\mathcal{K})}$. In practice, training data usually only covers a sparse subset $\mathcal{F}_{\text{obs}} \subset \mathcal{F}$. The challenge is to learn factor representations that produce images which reflect any target tuple $\mathbf{f}$, including combinations absent from training.

Learnable Embeddings: Each factor value $f^{(k)}$ is represented by a learnable token $t^{(k)}$, implemented as a sequence of n embedding vectors, $\mathcal{V}(t^{(k)}) = \{v_1^{(k)}, \dots, v_n^{(k)}\} \in \mathbb{R}^{n \times d}$. Following MULTI [13], these tokens are inserted into a structured prompt together with an image caption and optimized using the standard diffusion loss,

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{z_t, x, \epsilon, t} \left[\|\epsilon - \mathcal{U}(z_t, t, \mathbf{c})\|_2^2 \right] \quad (1)$$

where ϵ is the sampled noise, z_t is the noisy latent at timestep t , $\mathcal{U}$ is the denoising network, and $\mathbf{c}$ is the text-conditioning embedding from the structured prompt. The diffusion backbone remains frozen.

Factor Alignment Accuracy (FAA): To evaluate whether the generated images faithfully represent the intended factors, MULTI introduced Factor Alignment Accuracy (FAA) [13], where for each factor category, a classifier is trained. During evaluation, a generated image is passed through these classifiers, and each predicted factor is compared against the corresponding factor label used to condition the generation. FAA measures the fraction of predictions that match the target factor labels, thereby estimating whether the generated image visually contains the requested imaging factors.

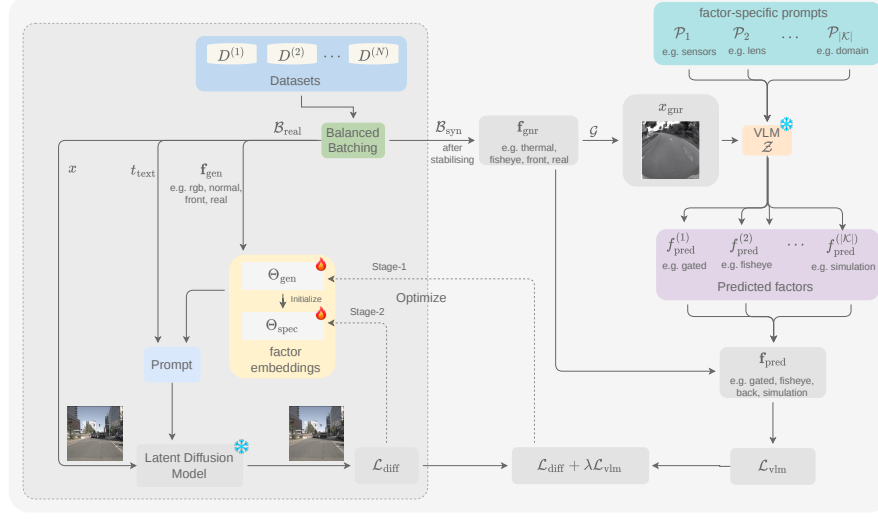


Fig. 1: X-MULTI Architecture. Overview of our framework, which augments MULTI (components enclosed within the dark grey dashed box) with zero-shot VLM supervision to optimize factor embeddings via joint diffusion and VLM-based factor alignment loss.

4 Methodology

4.1 X-MULTI: VLM-based Factor Disentanglement

To strengthen factor disentanglement and synthesize images with novel factor combinations, we introduce X-MULTI, which augments MULTI with zero-shot VLM-based supervision. MULTI learns factor embeddings in two stages: (1), it learns general embeddings Θ_{gen} for broad imaging factors such as lens, sensor, viewpoint, and domain (e.g. fisheye lens); (2), it adapts these embeddings to the specific factor values of a target dataset Θ_{spec} (e.g., fisheye lens of Woodscape dataset). X-MULTI retains this two-stage structure, and adds VLM-based supervision only at Stage-1. Fig. 1 provides an overview of the proposed pipeline.

VLM-based Supervision: Let x_{gnr} be an image generated with the target factor tuple $\mathbf{f}_{\text{gnr}} = (f_{\text{gnr}}^{(k)})_{k \in \mathcal{K}}$. We employ a frozen zero-shot VLM $\mathcal{Z}$ and factor-specific prompts $\mathcal{P} = \{\mathcal{P}_1, \dots, \mathcal{P}_{|\mathcal{K}|}\}$. For each factor category k , the VLM predicts $f_{\text{pred}}^{(k)} = \mathcal{Z}(x_{\text{gnr}}, \mathcal{P}_k)$. The VLM-based factor alignment loss is defined as

$$\mathcal{L}_{\text{vlm}} = \sum_{k \in \mathcal{K}} \ell_{\text{CE}} \left(f_{\text{pred}}^{(k)}, f_{\text{gnr}}^{(k)} \right), \quad (2)$$

where ℓ_{CE} denotes cross-entropy loss. Factor-wise VLM reliability is determined by evaluating $\mathcal{Z}$ on ground-truth images and comparing its predictions against true labels using the factor-specific prompts.

Training Objective: Following MULTI [13], we optimize the general factor embeddings Θ_{gen} with the diffusion objective. After an initial diffusion-only warm-up, in each training step, a real batch $\mathcal{B}_{\text{real}}$ is sampled from the training datasets. From the factor values present in $\mathcal{B}_{\text{real}}$, we additionally construct a synthetic factor tuple $\mathbf{f}_{\text{gnr}}$ by resampling factor values across categories. The current diffusion model $\mathcal{G}$ then produces an image $x_{\text{gnr}} = \mathcal{G}(\mathbf{f}_{\text{gnr}})$, which is used only for VLM-based supervision. The training objective is

$$\hat{\Theta}_{\text{gen}} = \arg \min_{\Theta_{\text{gen}}} \mathbb{E} [\mathcal{L}_{\text{diff}} + \lambda \mathcal{L}_{\text{vlm}}], \quad (3)$$

where λ controls the strength of VLM-based supervision.

4.2 Improved Factor Alignment Accuracy (I-FAA)

To evaluate imaging factor disentanglement, we propose Improved Factor Alignment Accuracy (I-FAA) as a new metric. Given a set of N generated images $\{x_i\}_{i=1}^N$, where each x_i is conditioned on the factor tuple $\mathbf{f}_i = (f_i^{(1)}, \dots, f_i^{(\mathcal{K})})$, we use a set of factor classifiers $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_{\mathcal{K}}\}$, where $\mathcal{S}_k$ predicts factor values in $\mathcal{F}^{(k)}$. The per-category alignment accuracy is defined as

$$\text{I-FAA}_k = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[\mathcal{S}_k(x_i) == f_i^{(k)}]. \quad (4)$$

The overall I-FAA score is obtained by averaging over all factor categories

$$\text{I-FAA}_{\text{avg}} = \frac{1}{\mathcal{K}} \sum_{k=1}^{\mathcal{K}} \text{I-FAA}_k. \quad (5)$$

Classifier Architecture: Each factor classifier $\mathcal{S}_k$ consists of the shared frozen visual backbone ϕ with parameters Θ_{bb} , followed by a factor-specific trainable linear head h_k , with parameters Θ_{h_k} . Thus, $\mathcal{S}_k$ is parameterized by the shared frozen backbone parameters Θ_{bb} and the factor-specific head parameters Θ_{h_k} . During training, only Θ_{h_k} is optimized, while Θ_{bb} remains fixed.

Class-Balanced Sampling: To mitigate bias in highly skewed data, we adopt class-balanced under-sampling during training of the classifiers. Let $D_k = \{(x_i, f_i^{(k)})\}$ denote the training dataset for factor category k , where $f_i^{(k)} \in \mathcal{F}^{(k)}$ is the factor label associated with image x_i . Let $n_{\min}^{(k)}$ denote the number of the smallest class in category k . Samples from every class $c \in \mathcal{F}^{(k)}$ are randomly under-sampled to $n_{\min}^{(k)}$, ensuring balanced class representation.

Factor-Specific Augmentation: Beyond class imbalance, classifiers may acquire shortcuts by learning dataset-level features that co-occur with a target factor. To reduce such shortcut learning, we employ each classifier $\mathcal{S}_k$ with dedicated augmentations. Formally, for category k we define $\mathcal{A}^{(k)} = \{a_1^{(k)}, a_2^{(k)}, \dots, a_{M_k}^{(k)}\}$, where M_k denotes the number of augmentation operations assigned to category

Table 1: Factor-specific Prompts for VLM-based supervision. Our method uses structured prompts and JSON output constraints to independently query the zero-shot VLM for each imaging factor.

Factor	Prompt
Sensor	Classify the SENSOR type of this image (0-4): 0 = RGB: normal color camera 1 = Thermal: grayscale images with bright/warm colors/white 2 = Gated: high contrast black and white, harsh lighting, foggy/smoky 3 = Event: motion changes, red/blue motion capture 4 = RGB Thermal (Fusion): Color images containing an “overlay” or “glow” of thermal data. Look for heat signatures (bright orange/yellow/white highlights) mapped onto real-world objects and textures. Return only JSON: {"sensor": <int>}
Lens	Classify the LENS type of this image (0-1): 0 = Normal: standard camera lens 1 = Fisheye: distorted edges, wide field of view Return only JSON: {"lens": <int>}
Viewpoint	Classify the VIEWPOINT of this image (0-4): 0 = Drone: camera placed on a flying drone 1 = Pole: camera mounted on a pole 2 = Front: camera faces forward 3 = Back: camera faces backward 4 = Side: camera placed on left/right (typically on car) Return only JSON: {"viewpoint": <int>}
Domain	Classify the DOMAIN of this image (0-2): 0 = Real: real-world images 1 = Simulation: rendered or simulated images 2 = Video Game: captured from video-game Return only JSON: {"domain": <int>}

k . During training, augmentations are sampled uniformly from $\mathcal{A}^{(k)}$ and applied to the image x to produce the augmented image $\tilde{x}^{(k)}$.

Optimization: Given augmented inputs and balanced mini-batches, each classifier $\mathcal{S}_k$ is optimized via the standard cross-entropy objective

$$\hat{\Theta}_{h_k} = \arg \min_{\Theta_{h_k}} \mathbb{E}_{(x, f^{(k)}) \sim D_k} \left[\ell_{\text{CE}} \left(\mathcal{S}_k \left(\tilde{x}^{(k)} \right), f^{(k)} \right) \right]. \quad (6)$$

5 Experimental Results

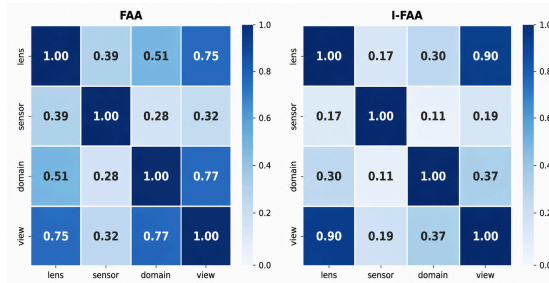
5.1 Implementation Details

We utilize the DF-RICO benchmark from MULTI [13], spanning 15 autonomous driving and surveillance datasets annotated with four factor categories: (1) camera lens (*normal*, *fisheye*), (2) source/domain (*real*, *simulation*, *video-game*), (3) viewpoint (*front*, *back*, *side*, *drone*, *pole*), and (4) sensor modality (*rgb*, *thermal*, *rgb-thermal*, *gated*, *event*).

X-MULTI: We follow the implementation setup of MULTI [13]. We use Stable Diffusion XL (SDXL) [30] as our generative backbone, where each factor is

Table 2: Data augmentations for I-FAA classifier training. Category-specific, non-overlapping image augmentations applied during I-FAA (ours) training.

Augmentation Category & Type	Lens	Sensor	Domain	Viewpoint
Target Classes	normal, fisheye	rgb, thermal, rgb-thermal, event, gated	real, simulation, video-game	front, back, side, pole, drone
Geometric				
Horizontal flip	✓	✓	✓	Mild
Vertical flip	✓	✓	✓	—
Random rotation & translation	Mild	Moderate	Moderate	Mild
Color / Appearance				
Color jitter, Grayscale, & Noise	✓	—	✓	✓
Gaussian blur	Moderate	Moderate	Strong	Mild
Occlusion / Region				
Cutout, Random erasing, & Hide-and-seek	Moderate	Mild	Moderate	—
Additional				
Radial distortion	—	Moderate	Moderate	Mild
Perspective warp	—	✓	✓	—

**Fig. 2: Cramer's V correlations.** Pairwise correlations between predicted factors for FAA (left) and I-FAA (ours) (right), showing that I-FAA reduces cross-factor information leakage. High values show strong correlations.

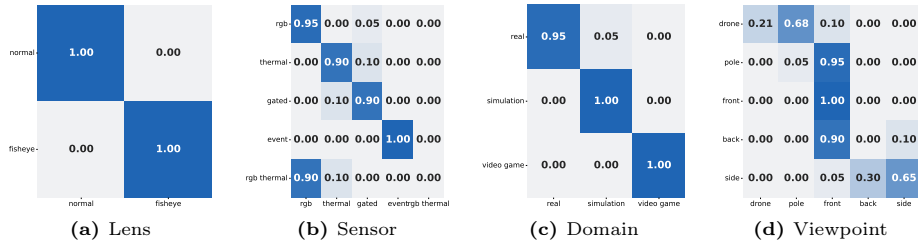
represented by $n = 15$ vectors optimized via AdamW [26] (weight decay: 10^{-2}) using a linear warmup and cosine annealing learning rate schedule (maximum: 10^{-4}). Training is conducted for 10 epochs with a total batch size of 4.

For the VLM-based supervision branch, we employ Qwen2-VL-7B-Instruct [40] as a zero-shot VLM classifier starting from epoch 3 with supervision strength of 10^{-6} . Each mini-batch consists of 3 real samples and 1 synthetic sample. We formulate structured factor-specific prompts as shown in Table 1 that independently query the VLM for each imaging factor: sensor, lens, domain, and viewpoint. To prevent incorrect supervision of unreliable VLM predictions which is discussed in Section 5.2, rgb-thermal sensor and non-front viewpoints are masked to 0.

I-FAA: We employ DINOv3 [37] as the vision backbone for all factor classifiers. We train the linear prediction heads using the AdamW optimizer [26] with a learning rate of 10^{-4} and a cosine annealing learning rate schedule [25], and a batch size of 32. Table 2 shows data augmentations carried out during training of each factor classifier.

Table 3: Qualitative comparison of novel combinations. Image generation with unseen factor combinations. Our method, X-MULTI shows better factor adherence.

Domain Sensor Viewpoint Lens	real thermal fisheye(loaf) front	video-game rgb fisheye drone(visdrone)	real gated normal pole	real rgb-thermal fisheye front	video-game rgb normal pole	simulation thermal normal drone(visdrone)	video-game thermal fisheye pole	simulation(shift) thermal(timo) normal pole(fisheye8k)
SDXL Zeroshot [30]								
DreamBooth [32]								
MULTI [13]								
X-MULTI								

**Fig. 3: Confusion matrices of the zero-shot VLM for factor classification:** (a) Lens, (b) Sensor, (c) Domain, and (d) Viewpoint. Rows correspond to ground-truth labels and columns to predicted labels.

Evaluation Metrics: We employ standard image generation metrics: Fréchet Inception Distance (FID) [17] to measure distributional similarity between real and generated images, CLIP Score [16] to assess semantic alignment with factor-based prompts (evaluated on three components: factors only, content description only, and full prompt), Inception Score (IS) [34] for image quality, and Diversity Score (DS) [46] for perceptual diversity. To measure factor disentanglement, we employ our introduced metric, I-FAA, and analyze its reliability in Sec. 5.3.

Baselines: We compare against: (1) SDXL Zeroshot [30], the pre-trained model without adaptation; (2) DreamBooth [32], fine-tuning-based personalization; (3) Inspiration Tree [39], hierarchical disentanglement of stylistic components (evaluated only on existing combinations, as its learned embeddings do not correspond to independent factor categories that can be recombined into novel combinations); and (4) MULTI [13], a prior factor disentanglement method.

Table 4: Qualitative comparison of novel factor combinations with ControlNets. Image-to-image generation using ControlNet guidance with altered factors, showing X-MULTI (ours) achieves best factor adherence.

Domain Sensor Lens Viewpoint	simulation rgb normal front	real rgb-thermal normal front	real rgb normal back	real event normal front	real rgb-thermal normal front	real rgb fisheye front	real event normal front	simulation rgb normal front
Original								
ControlNet Input								
Altered Factor	simulation ↓ real	rgb-thermal ↓ rgb	real ↓ video-game	event ↓ gated	rgb-thermal ↓ gated	rgb ↓ thermal	event ↓ event(desc)	rgb ↓ rgb(leaf)
SDXL Zeroshot [30]								
DreamBooth [32]								
MULTI [13]								
X-MULTI								

5.2 Imaging Factor Disentanglement Results

The primary goal is to evaluate the quality of the synthesis of images with unseen novel factor combinations. We show both qualitative and quantitative results along with the analysis of the choice of VLM and prompt designs for VLM-based supervision.

Novel Factor Combinations: Table 3 shows qualitative inspection of images generated with novel factor combinations. SDXL Zeroshot fails to reflect target factors, while MULTI, DreamBooth, and X-MULTI produce more plausible combinations. X-MULTI shows clearer factor adherence particularly for viewpoint, lens, and domain than other baselines. X-MULTI shows less proficiency with thermal sensor but handled other sensors competently.

Table 4 shows the usage of ControlNets [45] for guiding the image generation process, and altering a factor (to respect control maps we use sensor or domain factors only). In this setting, our approach demonstrates good precision compared to MULTI and DreamBooth, which outperforms SDXL Zeroshot. Specifically, when substituting the event sensor for a gated one, DreamBooth retained

Table 5: Quantitative comparison on novel factor combinations. Performance metrics across methods with and without ControlNet structural guidance, with best values highlighted in bold. X-MULTI (ours) achieves best overall disentanglement.

Method	Image Generation Metrics					Improved-Factor Alignment Accuracy (I-FAA)					
	IS [34] ↑	CLIP Score [16] ↑				DS [46] ↑	Lens ↑	Sensor ↑	Domain ↑	View ↑	Avg. ↑
		factor	context	average	full-prompt						
DreamBooth [32]	3.68	23.96	23.35	23.66	24.67	0.55	0.78	0.24	0.48	0.27	0.44
SDXL Zeroshot [30]	3.11	22.31	25.59	23.95	27.80	0.59	0.72	0.25	0.29	0.29	0.39
MULTI [13]	3.28	24.20	23.43	23.82	27.32	0.65	0.63	0.29	0.64	0.34	0.47
X-MULTI (ours)	3.23	24.62	23.57	24.10	28.22	0.59	0.76	0.30	0.66	0.40	0.53
DreamBooth-Canny	5.78	23.67	25.79	25.23	29.12	0.50	0.99	0.71	0.89	0.80	0.84
SDXL Zeroshot-Canny	5.74	23.88	26.75	25.31	28.95	0.51	0.99	0.63	0.84	0.75	0.78
MULTI-Canny	4.88	22.82	23.08	22.95	26.72	0.55	0.99	0.73	0.84	0.86	0.85
X-MULTI-Canny (ours)	4.64	23.91	23.12	23.51	29.60	0.56	0.99	0.81	0.89	0.87	0.89
DreamBooth-Depth	5.56	23.57	25.40	24.98	27.95	0.53	1.00	0.69	0.88	0.77	0.84
SDXL Zeroshot-Depth	5.49	23.62	26.57	25.10	29.82	0.52	1.00	0.66	0.81	0.71	0.77
MULTI-Depth	4.36	22.12	23.45	22.78	26.82	0.52	0.99	0.76	0.84	0.87	0.86
X-MULTI-Depth (ours)	4.87	23.63	24.79	23.21	29.81	0.53	1.00	0.71	0.88	0.88	0.87

event artifacts while our method executed the transformation cleanly. Similarly, shifting domains from real to video-game, our method completely eliminated real-world traces while MULTI and DreamBooth could not achieve this.

Table 5 reports quantitative metrics for novel factor combinations with and without using ControlNets [45]. Under I-FAA, X-MULTI achieves the best overall factor disentanglement, CLIP alignment with factors, and DS while other metrics remain mixed across methods.

Effect of VLM-based Supervision: We evaluate the contribution of VLM-based supervision by comparing MULTI (without VLM-based supervision) and X-MULTI (with supervision) on novel factor combinations. MULTI achieves an average I-FAA of 0.47, while X-MULTI achieves 0.53, representing a 0.06 (11% relative) improvement. X-MULTI shows a best improvements in lens and view-point factors with 21%, 18% relative improvement respectively, while both sensor and domain show a relative improvement of 0.01. This demonstrates that supervision enhances factor disentanglement, especially for lens, and viewpoint.

VLM Analysis: Since VLM predictions provide the supervision in X-MULTI, we evaluate their reliability in a zero-shot setting using the prompts in Table 1. Figure 3 presents the confusion matrices for Lens, Sensor, Domain, and Viewpoint classification. Across all factors, the VLM achieves an average classification accuracy of ~ 0.78 , demonstrating that the model captures meaningful semantic representations of imaging factors. However, the analysis also reveals that certain factor categories are less reliably recognized. In particular, the rgb-thermal sensor exhibits strong confusion with rgb images, and several viewpoint categories other than front show inconsistent predictions. Since these inaccurate classification signals can negatively affect disentanglement learning in X-MULTI, we disable the VLM supervision loss for these categories during training. Specifically, the supervision strength is set to zero for the rgb-thermal sensor and for all viewpoints except front.

Table 6: Comparison of attention maps for FAA and I-FAA. Grad-CAM visualizations showing that I-FAA (ours) learns more semantically meaningful attention regions.

Image	Metric	Lens	Sensor	Domain	View
 normal/rgb/video-game/front	FAA [13]				
					
	I-FAA				
 fisheye/rgb/real/drone	FAA [13]				
					
	I-FAA				
 normal/thermal/real/front	FAA [13]				
					
	I-FAA				

Table 7: FAA and I-FAA on ground-truth DF-RICO images. Comparison of FAA and I-FAA (ours) metric on ground-truth images.

Metric	Lens	Sensor	Domain	View	Avg.
FAA [13]	0.23	0.83	0.50	0.17	0.43
I-FAA	1.00	0.97	0.98	0.84	0.95

Ablation Studies: We carry out different ablation studies to evaluate the design of core components of X-MULTI.

- **(1) VLM Selection and Prompt Design:** Since factor disentanglement depends on VLM feedback quality, we evaluate the zero-shot classification performance of VLM on ground-truth data. First, comparing VLM backbones shows that Qwen2-VL-7B-Instruct [40] gives 0.78 average accuracy and 0.51 for LLaVA-1.6 [24], indicating Qwen provides much stronger visual grounding. Second, comparing prompt designs using Qwen2-VL shows that detailed factor-specific prompts with per-class descriptions achieve 0.78 accuracy, simple prompts without descriptions achieve 0.70, and a single unified prompt classifying all factors jointly drops to 0.63, demonstrating that detailed, isolated prompts provide the best signal for X-MULTI.
- **(2) Supervision strength:** We vary the supervision strength (λ) and evaluate I-FAA on novel factor combinations. Moderate (10^{-6}) and weak (10^{-8}) supervision perform best, both reaching an I-FAA of 0.53. In contrast, overly

Table 8: Factor prediction comparison. Per-factor classification outputs compared against ground-truth labels for ground truth images. I-FAA (ours) predicts factor correctly.

Image	Factor	GT	FAA [13]	I-FAA	Image	Factor	GT	FAA [13]	I-FAA
	Lens	normal	fisheye	normal		Lens	normal	normal	normal
	Sensor	rgb	rgb	rgb		Sensor	rgb	rgb	rgb
	Domain	real	simulation	real		Domain	real	real	real
	View	front	front	front		View	back	back	back
	Lens	normal	fisheye	normal		Lens	normal	normal	normal
	Sensor	event	rgb	event		Sensor	rgb	rgb	rgb
	Domain	real	real	real		Domain	simulation	simulation	simulation
	View	front	front	front		View	front	back	front
	Lens	fisheye	fisheye	fisheye		Lens	normal	normal	normal
	Sensor	rgb	rgb	rgb		Sensor	rgb	rgb	rgb
	Domain	real	real	real		Domain	video-game	simulation	video-game
	View	pole	red	pole		View	front	front	front
	Lens	normal	normal	normal		Lens	normal	normal	normal
	Sensor	thermal	thermal	thermal		Sensor	rgb	rgb	rgb
	Domain	real	real	real		Domain	real	simulation	real
	View	front	pole	front		View	drone	drone	drone

strong supervision (10^{-4}) degrades performance to 0.47, indicating that excessive supervision over-constrains factor adaptation.

5.3 Analysis of Metrics

To validate the reliability of FAA and I-FAA, we test whether their underlying classifiers isolate imaging factors independently rather than exploiting dataset-specific shortcuts. Specifically, we analyze cross-factor alignment using Grad-CAM attention maps to visually verify semantic localization, Cramer’s V correlation [8] to statistically detect information leakage, and metric scores alongside factor predictions on ground-truth DF-RICO [13] images to establish performance.

Attention Map Visualization: Table 6 visualizes Grad-CAM [35] attention maps for FAA and I-FAA across imaging factors. It shows that I-FAA learns more localized and semantically meaningful attention regions for each factor category, whereas FAA often exhibits entangled attention patterns. For example, in the fisheye lens samples, I-FAA strongly attends to image boundary curvature and geometric distortions, while FAA focuses on unrelated scene regions. Similarly, for viewpoint prediction, I-FAA concentrates on perspective-specific regions, whereas FAA frequently attends to the image center irrespective of viewpoint semantics. For sensor-related factors, I-FAA captures modality-specific characteristics scattered across regions more distinctly than FAA. Domain attention maps also demonstrate that I-FAA focuses on scene appearance cues by attending on objects, while FAA exhibits more inconsistent activations.

Cramér’s V Correlation Analysis: Figure 2 presents the pairwise Cramér’s V correlations [8] between predicted imaging factors for FAA and I-FAA. Ideally, a factor-isolated evaluation metric should produce low inter-factor correlations, since each factor prediction should depend primarily on the target factor rather than on co-occurring factors. FAA exhibits strong coupling between multiple factor pairs, particularly lens-viewpoint, domain-viewpoint, and lens-domain. In

contrast, I-FAA reduces most of these correlations, indicating lower cross-factor leakage. However, residual correlations remain, especially for factor pairs that are strongly coupled or sparsely represented in DF-RICO. This highlights that augmentation in I-FAA classifier training mitigates shortcut learning but cannot fully compensate for correlations in the underlying data.

Metric and Factor Prediction Comparison: Since FAA was originally proposed as an evaluation metric without independently validating the accuracy of its underlying classifiers on ground-truth images, we first evaluate both FAA and I-FAA classifiers on DF-RICO. Table 7 shows that I-FAA improves the average classification accuracy from 0.43 to 0.95. This improvement stems from the class-balanced training and factor-specific augmentations introduced in I-FAA, which reduce shortcut learning from correlated imaging factors. The largest gains are observed for viewpoint and domain, where FAA frequently confuses factor predictions. Table 8 further illustrates that I-FAA produces more reliable factor-wise predictions, while FAA is more affected by inter-factor interference.

6 Discussion

Our numerical results lead to the following main findings:

VLM-Guided Supervision: Zero-shot VLM supervision provides additional training signals by predicting factors from synthetically generated images during training. This significantly enhances factor alignment, particularly for geometric attributes. ► *Zero-shot supervision enhances disentanglement.*

VLM Reliability and Prompt Context: X-MULTI relies on a VLM to provide supervision. Our VLM analysis shows that some factor values are less reliably recognized, so we mask unreliable predictions during training. This indicates that performance depends on both the VLM choice and the design of factor-specific prompts. Providing richer visual context in the prompts may further improve supervision. ► *VLM-based supervision is useful, but its reliability depends on prompt design and VLM factor recognition.*

Reliability of I-FAA: FAA exhibits cross-factor correlations consistent with shared representations entangling factor predictions, while I-FAA reduces this leakage through factor-specific augmentations. However, residual correlations remain and could be addressed with stronger augmentations or correlation-reduction methods [6, 14, 21, 44, 48]. ► *I-FAA reduces correlations for imaging factor disentanglement, but they could be reduced further.*

7 Conclusion

In this work, we addressed imaging factor disentanglement for controllable image generation of unseen novel combinations of imaging factors. We first identify that MULTI’s pixel-level reconstruction focuses only on available training data leading to no factor-level supervision for novel combinations. We therefore proposed X-MULTI, which augments MULTI with zero-shot VLM supervision for

unseen factor combinations to enforce semantic factor identity. We also identified a limitation in the FAA metric, i.e., it produces strongly entangled factor predictions, evidenced by high cross-factor correlation, and attention maps. To address this, we introduced **I-FAA**, an evaluation framework with targeted augmentation strategies that reduces these correlations and provides improved disentanglement assessment. Experiments demonstrated that X-MULTI achieves stronger factor disentanglement than baselines, particularly for novel factor combinations, and I-FAA substantially reduces correlations and establishes a more robust metric for factor disentanglement evaluation. Despite promising results, imaging factor disentanglement remains challenging.

References

1. Avrahami, O., Aberman, K., Fried, O., Cohen-Or, D., Lischinski, D.: Break-a-scene: Extracting multiple concepts from a single image. In: SIGGRAPH Asia. pp. 1–12 (2023)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Butt, M.A., Wang, K., Vazquez-Corral, J., van de Weijer, J.: Colorpeel: Color prompt learning with diffusion models via color and shape disentanglement. In: ECCV. pp. 456–472. Springer (2024)
4. Cai, M., Liu, H., Mustikovela, S.K., Meyer, G.P., Chai, Y., Park, D., Lee, Y.J.: Vip-llava: Making large multimodal models understand arbitrary visual prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12914–12923 (2024)
5. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)* **42**(4), 1–10 (2023)
6. Chew, O., Lin, H.T., Chang, K.W., Huang, K.H.: Understanding and mitigating spurious correlations in text classification with neighborhood analysis. In: Findings of the association for computational linguistics (EACL). pp. 1013–1025 (2024)
7. Cho, J., Zala, A., Bansal, M.: Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3043–3054 (2023)
8. Cramér, H.: *Mathematical methods of statistics*, vol. 9. Princeton university press (1999)
9. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML. pp. 12606–12633. PMLR (2024)
10. Frenkel, Y., Vinker, Y., Shamir, A., Cohen-Or, D.: Implicit style-content separation using b-lora. In: ECCV. pp. 181–198. Springer (2024)
11. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: ICLR (2022)
12. Garibi, D., Yadin, S., Paiss, R., Tov, O., Zada, S., Ephrat, A., Michaeli, T., Mosseri, I., Dekel, T.: Tokenverse: Versatile multi-concept personalization in token modulation space. *ACM Transactions On Graphics (TOG)* **44**(4), 1–11 (2025)

13. Godavarthy, S., Neuwirth-Trapp, M., Faasch, T.F., Bieshaar, M., Moeller, M., Paudel, D.P.: Multi: Disentangling camera lens, sensor, view, and domain for novel image generation. arXiv preprint arXiv:2605.12134 (2026)
14. Gui, S., Ji, S.: Mitigating spurious correlations in llms via causality-aware post-training. arXiv preprint arXiv:2506.09433 (2025)
15. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: ICLR (2022)
16. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 7514–7528 (2021)
17. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **30** (2017)
18. Jin, C., Tanno, R., Saseendran, A., Diethe, T., Teare, P.A.: An image is worth multiple words: Discovering object level concepts using multi-concept prompt learning. In: International Conference on Machine Learning (ICML). PMLR (2024)
19. Kwon, G., Jenni, S., Li, D., Lee, J.Y., Ye, J.C., Heilbron, F.C.: Concept weaver: Enabling multi-concept fusion in text-to-image models. In: CVPR. pp. 8880–8889 (2024)
20. Laurençon, H., Tronchon, L., Cord, M., Sanh, V.: What matters when building vision-language models? *Advances in Neural Information Processing Systems* **37**, 87874–87907 (2024)
21. Lee, S., Payani, A., Chau, D.H.P.: Towards mitigating spurious correlations in image classifiers with simple yes-no feedback. In: AI and HCI Workshop at the International Conference on Machine Learning (ICML Workshop) (2023)
22. Li, P., Huang, Q., Ding, Y., Li, Z.: Layerdiffusion: Layered controlled image editing with diffusion models. In: SIGGRAPH Asia 2023 Technical Communications (SIGGRAPH Asia), pp. 1–4 (2023)
23. Liu, C., Shah, V., Cui, A., Lazebnik, S.: Unziplora: Separating content and style from a single image. In: ICCV. pp. 16776–16785 (2025)
24. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 26296–26306 (2024)
25. Liu, Z.: Super convergence cosine annealing with warm-up learning rate. In: CAIBDA 2022; 2nd International Conference on Artificial Intelligence, Big Data and Algorithms. pp. 1–7. VDE (2022)
26. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2017)
27. Luo, G., Granskog, J., Holynski, A., Darrell, T.: Dual-process image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17972–17983 (2025)
28. Motamed, S., Paudel, D.P., Van Gool, L.: Lego: Learning to disentangle and invert personalized concepts beyond object appearance in text-to-image diffusion models. In: European Conference on Computer Vision (ECCV). pp. 116–133. Springer (2024)
29. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: ICML. pp. 16784–16804. PMLR (2022)
30. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. In: ICLR (2023)

31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
32. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dream-booth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR. pp. 22500–22510 (2023)
33. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* **35**, 36479–36494 (2022)
34. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. *Advances in neural information processing systems* **29** (2016)
35. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE international conference on computer vision*. pp. 618–626 (2017)
36. Shentu, J., Watson, M., Al Moubayed, N.: Attencraft: Attention-guided disentanglement of multiple concepts for text-to-image customization. *CoRR* (2024)
37. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. *arXiv preprint arXiv:2508.10104* (2025)
38. Sohn, K., Ruiz, N., Lee, K., Chin, D.C., Blok, I., Chang, H., Barber, J., Jiang, L., Entis, G., Li, Y., et al.: Styledrop: text-to-image generation in any style. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. pp. 66860–66889 (2023)
39. Vinker, Y., Voynov, A., Cohen-Or, D., Shamir, A.: Concept decomposition for visual exploration and inspiration. *ACM Transactions on Graphics (TOG)* **42**(6), 1–13 (2023)
40. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* (2024)
41. Wen, S., Fang, G., Zhang, R., Gao, P., Dong, H., Metaxas, D.: Improving compositional text-to-image generation with large vision-language models. *arXiv preprint arXiv:2310.06311* (2023)
42. Wu, P., Xie, S.: V?: Guided visual search as a core mechanism in multimodal llms. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 13084–13094 (2024)
43. Xu, C., Xu, Y., Zhang, H., Xu, X., He, S.: Dreamanime: Learning style-identity textual disentanglement for anime and beyond. *IEEE Transactions on Visualization and Computer Graphics* (2024)
44. Yang, Y., Nushi, B., Palangi, H., Mirzasoileiman, B.: Mitigating spurious correlations in multi-modal models during fine-tuning. In: *International Conference on Machine Learning (ICML)*. pp. 39365–39379. PMLR (2023)
45. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: *ICCV*. pp. 3836–3847 (2023)
46. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 586–595 (2018)
47. Zhang, Y., Yang, M., Zhou, Q., Wang, Z.: Attention calibration for disentangled text-to-image personalization. In: *CVPR*. pp. 4764–4774 (2024)

48. Zheng, G., Ye, W., Zhang, A.: Learning robust classifiers with self-guided spurious correlation mitigation. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). pp. 5599–5607 (2024)
49. Zhong, W., Yang, H., Liu, Z., He, H., He, Z., Niu, X., Zhang, D., Li, G.: Mod-adapter: Tuning-free and versatile multi-concept personalization via modulation adapter. arXiv preprint arXiv:2505.18612 (2025)