

## A Appendix

### A.1 Limitations and Future Work

In our work, as in most previous approaches, we assume the target factors of variation are known in advance, and our goal is to learn how to disentangle them. However, in real-world scenarios, the factors of variation within a dataset may be unknown or considerably more complex than the idealized cases of attribute or object disentanglement. Consequently, an important future direction is to automatically identify underlying factors of variation and determine the appropriate mixing strategy (potentially through learning) without relying on prior knowledge beyond the dataset itself. Meanwhile, although our method shows strong performance in both attribute and object disentanglement, it does not provide theoretically grounded guarantees. Bridging the gap between methods that provide theoretical guarantees but only work on simple datasets, and methods like ours that demonstrate strong performance but lack such guarantees, is another important direction for future research. Moreover, although we focus on learning a disentangled representation in this work, exploring our framework on downstream tasks such as controllable image manipulation, *e.g.*, attribute- or object-level edits, and object-centric world-model training that leverages our object-disentangled representations, as exemplified by Jeong et al. [18].

### A.2 Broader Impact

Our method can extract attribute or object components from existing images and use this extracted information to generate new images. This capability may raise privacy issues if applied to deepfake generation or unauthorized copying of digital content.

### A.3 More Related Work

In this section, we discuss additional related work relevant to our method.

**Identifiable DRL** In object-level disentangled representation learning (or object-centric learning), a line of work [2, 25, 50] leverages identifiability theory to derive conditions that provide theoretical guarantees for disentangled representations aligned with underlying factors of variation. Specifically, Brady et al. [2] shows that, under certain assumptions in object-centric scenes, an invertible *compositional decoder* can recover the ground-truth object latents (up to permutation). Lachapelle et al. [25] provides theoretical conditions for identifiability (up to permutation and blockwise invertible transformations) when ground-truth latent variables are organized into specific blocks and an *additive decoder* is used. While these work can provide identifiability guarantees for object representations, it may not generally apply to factors of variation such as attributes, which globally affect the image and do not necessarily satisfy the assumptions of additive decoders and compositionality definitions. Moreover, these approaches often impose strong restrictions on model design and expressiveness [41], making them applicable only to relatively simple datasets. Our method aims in a different direction from these works, seeking a modular inductive bias that is decoupled from both learning objectives and architectural constraints, such as an additive decoder, thereby enabling disentangled representation learning for various factors under a single objective and architecture.

**Wiedemer et al. [50]** Wiedemer et al. [50] demonstrates that autoencoders satisfying encoder-decoder consistency, in combination with an additive decoder, can yield object-centric representations that provably generalize compositionally. While this approach shares the similar conceptual goal of enabling generalization to novel compositions of factors in disentangled representation learning as ours, there are fundamental differences in how we achieve valid generalization. As the common part, both methods recognize that valid generalization requires two key components: first, compositions of disentangled representations must yield valid data, *e.g.*, realistic images or representation, and second, the composite representation  $\mathbf{z}^c$  must properly encode the information from its corresponding composite image  $\mathbf{x}^c$  to satisfy the representation learning objective. To address the second requirement, both our method and Wiedemer et al. [50] employ a compositional consistency loss with minor differences. However, the approaches diverge significantly for the first component—how to ensure the validity of composite representations.

Wiedemer et al. [50] employs an additive decoder to ensure valid compositions of disentangled representations (Sec. 3.1 in [50]). However, additive decoders are known to be unscalable for complex scenes, as their local decoding mechanism cannot capture complex interactions between

Table 6: Comparison to Wiedemer et al. [50]. The additive decoder alone fails to achieve attribute, object, and joint disentanglement. While employing a slot attention module in the encoder leads to reasonable performance on object disentanglement, it still fails in joint disentanglement.

	Shapes3D		CLEVR				CLEVR-Style					
	FactorVAE	DCI	Shape	Color	Material	Position ( $\downarrow$ )	Acc	GRAM ( $\downarrow$ )	Shape	Color	Material	Position ( $\downarrow$ )
Ours	<b>0.975</b>	<b>0.837</b>	<b>87.04</b>	<b>93.93</b>	<b>94.81</b>	<b>0.032</b>	<b>96.50</b>	<b>5.05</b>	<b>83.56</b>	<b>90.48</b>	<b>93.74</b>	<b>0.053</b>
Additive Decoder w/o SA	0.000	0.031	33.87	15.24	51.87	0.765	23.30	18.59	34.20	13.54	50.90	0.517
Additive Decoder w/ SA	-	-	82.91	93.14	91.78	0.110	23.30	15.84	35.97	20.99	54.10	0.520

objects due to limited expressive power. More critically, additive decoders are designed based on the spatial exclusiveness bias (Def. 4 in [50]), which assumes that each pixel should be affected by only a single latent variable. This limits their applicability to object-centric learning scenarios, preventing generalization to other disentanglement tasks.

In contrast, our method ensures validity through a prior loss (Eq. 4) that leverages SDS loss [34] to encourage composite image  $\mathbf{x}^c$  to be realistic. As we do not require an additive decoder anymore, our approach allows us to utilize an expressive diffusion decoder for modeling complex scenes. Crucially, since our method does not embed factor-specific biases like spatial exclusiveness into the architecture, it can be applied beyond object-centric learning to attribute disentanglement and joint disentanglement scenarios. More importantly, to the best of our knowledge, we are the first to demonstrate that different underlying factors of variation can be disentangled simply by adjusting the factor-specific mixing strategy, distinguishing our approach from previous methods that introduce factor-specific biases through architectures or objective functions.

Finally, we provide a quantitative comparison of Wiedemer et al. [50] with our method in attribute-, object-, and joint disentanglement in Tab. 6. For a fair comparison, we used the same encoder for [50] as ours and only changed the decoder to an additive decoder. As expected, [50] cannot disentangle attribute factors at all (Shapes3D, CLEVR-Style), since they violate the spatial-exclusiveness assumptions. Moreover, in object disentanglement, we found that an additive decoder alone cannot disentangle objects (in fact, Wiedemer et al. [50] validated their method only on very simple 2D synthetic datasets). When we additionally use the slot-attention module in [50], it reasonably disentangles objects in the CLEVR dataset but is still significantly inferior to our method in CLEVR-Style, possibly due to the limited expressive power of the additive decoder. Additionally, we observed that object-wise manipulation with [50] always leads to unrealistic images with transparently overlapping objects due to the lack of interactions between latents inside the additive decoder.

**Group theory-based DRL** Disentangled representation learning using Group theory is an actively researched area and is related to our work. Group theory-based DRL typically define disentangled representation  $Z$  as follows: Given ground truth factors of variation  $W$  and decomposable group  $G$  (i.e.,  $G = G_1 \times G_2 \times \dots \times G_n$ ), the representation  $Z$  is disentangled w.r.t.  $G$  if (1) there exists a mapping  $f$  from  $W$  to  $Z$  such that  $f(g \cdot W) = g \cdot f(W)$  for all  $g \in G$  and  $w \in W$ , and (2) there is a decomposition  $Z = Z_1 \times \dots \times Z_n$  such that each  $Z_i$  is affected only by  $G_i$ . Despite this convincing principled definition, since the GT factors of variation  $W$  are infeasible to obtain in unsupervised DRL, existing unsupervised methods often utilize necessary conditions for group actions of  $G$  applied to  $Z$  and disentangled representation  $Z$  [46, 53]. For instance, Tao et al. [46] defines permutation group actions of element-wise addition on  $Z$  and introduces losses to enforce commutativity and cyclicity of group actions.

Our method takes a similar approach, but with important distinctions. From a Group theory perspective, unlike existing work, the group action in our method is defined on disentangled representation pairs  $(z^1, z^2)$  rather than a single latent  $z_i$ . By defining the group action on a pair  $(z^1, z^2)$ , we can impose additional necessary conditions for how each underlying factor combines to generate observations, which cannot be induced by the group action defined on a single latent  $z_i$ . For example, commutativity and cyclicity of group action are necessary conditions for both attributes and objects, but do not impose attribute or object-specific properties. Our main contribution here is that we define group action as factor-specific mixing, i.e., permutations, between two latents and demonstrate that this additional necessary condition imposes effective factor-specific inductive bias for attribute and object disentanglement without changing the overall learning objectives or model architectures. To the best of our knowledge, we are the first to study differently learned disentangled representations of

different factors of variation through a mixing strategy (or a form of group action) without employing factor-specific architectures or learning objectives.

**DRL with Factorized support** Roth et al. [39] leverages a similar idea of combining latents to promote factorized support. While ours and prior work both leverage a factorized support assumption, our main contributions are fundamentally different. First of all, we demonstrate that additional assumptions about the factorization structure of support, *e.g.*, product of  $K$  distinct supports in attributes or repetition of a shared support in objects, lead to disentanglement of different factors of variation. By combining these two assumptions, we can even achieve joint disentanglement of attributes and objects, which the factorized support assumption alone cannot handle. Secondly, we show that these factorization structures can be encoded via simple mixing strategies (Eqs. 1, 2, 3) replacing existing factor-specific biases. Those mixing strategies are decoupled from the architecture and objectives, so simply switching the mixing strategy enables us to disentangle attributes and objects within the single framework. Note that the prior method simply mixes the latent dimension-wise. Our ablations (Tab. 5) empirically show that such a strategy is insufficient for disentangling objects.

**Joint disentanglement of attribute and object** Recently, SysBinder [43] introduced the Block-Slot representation, which models each object as a slot formed by concatenating multiple multi-dimensional attribute (factor) representations called blocks, thereby enabling disentangled representations of both objects and their attributes. This approach differs from ours by fully redesigning the model architecture to handle object-centric scenes, making it inapplicable to broader factors. In contrast, our method aims to support disentangled representation learning for various factors of variation within a single framework with a modular inductive bias. Nonetheless, extending our approach to disentangle multiple factors of variation using a modular inductive bias remains an important direction for future work.

#### A.4 Equivalence between mixing two and multiple images

**Proof of equivalence** In this section, we explain why the random mixing between two images (*i.e.*,  $\mathbf{z}^c = \pi(\mathbf{z}^1, \mathbf{z}^2)$ ) can replace the random composition of  $\mathbf{z}_i$  from  $K$  images. Formally, we will show that:

$$\text{If } \mathcal{S}(p(\mathbf{z})) = \mathcal{S}(p(\mathbf{z}^c)) \text{ then } \mathcal{S}(p(\mathbf{z})) = \mathcal{S}^\times(p(\mathbf{z})), \quad (7)$$

where the factorized support  $\mathcal{S}^\times(p(\mathbf{z})) = \mathcal{S}(p(\mathbf{z}_1)) \times \mathcal{S}(p(\mathbf{z}_2)) \times \dots \times \mathcal{S}(p(\mathbf{z}_K))$  represents the random composition of each latent variable  $\mathbf{z}_i$  from  $K$  images.

*Proof.* Given  $\mathcal{S}(p(\mathbf{z})) = \mathcal{S}(p(\mathbf{z}^c))$ , we can prove the followings:

1. If  $p(\mathbf{z}_1)p(\mathbf{z}_2) > 0$  then  $p(\mathbf{z}_1, \mathbf{z}_2) > 0$ .  
Note that  $p(\mathbf{z}_1) > 0$  and  $p(\mathbf{z}_2) > 0$  ( $\Leftrightarrow p(\mathbf{z}_1)p(\mathbf{z}_2) > 0$ ) indicates the existence of  $\mathbf{z}^1, \mathbf{z}^2$  with  $\mathbf{z}_1^1 = \mathbf{z}_1, \mathbf{z}_2^2 = \mathbf{z}_2$ . By mixing  $\mathbf{z}^1$  and  $\mathbf{z}^2$ , we can compose  $\mathbf{z}^*$  where  $\mathbf{z}_1^* = \mathbf{z}_1, \mathbf{z}_2^* = \mathbf{z}_2$ . Then, by the definition of the support that  $\mathcal{S}(p(\mathbf{z})) = \{\mathbf{z} | p(\mathbf{z}) > 0\}$  and the given condition  $\mathbf{z}^* \in \mathcal{S}(p(\mathbf{z}^c)) = \mathcal{S}(p(\mathbf{z}))$ ,  $p(\mathbf{z}_1, \mathbf{z}_2) \geq p(\mathbf{z}^*) > 0$ .
2. Assume that for some  $k \geq 2$ , if  $\prod_{i=1}^k p(\mathbf{z}_i) > 0 \rightarrow p(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_k) > 0$  then  $\prod_{i=1}^{k+1} p(\mathbf{z}_i) > 0 \rightarrow p(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_k, \mathbf{z}_{k+1}) > 0$ .  
Note that  $\prod_{i=1}^{k+1} p(\mathbf{z}_i) > 0$  implies  $p(\mathbf{z}_{k+1}) > 0$  and  $\prod_{i=1}^k p(\mathbf{z}_i) > 0$ . By the given assumption,  $p(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_k) > 0$  and there exists  $\mathbf{z}^1, \mathbf{z}^2$  where  $\mathbf{z}_i^1 = \mathbf{z}_i$  for  $i \in \{1, \dots, k\}$  and  $\mathbf{z}_{k+1}^2 = \mathbf{z}_{k+1}$ . By mixing  $\mathbf{z}^1$  and  $\mathbf{z}^2$ , we can compose  $\mathbf{z}^*$  where  $\mathbf{z}_i^* = \mathbf{z}_i$  for  $i \in \{1, \dots, k+1\}$ . As a result, by the given condition  $\mathbf{z}^* \in \mathcal{S}(p(\mathbf{z}^c)) = \mathcal{S}(p(\mathbf{z}))$ ,  $p(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_k, \mathbf{z}_{k+1}) \geq p(\mathbf{z}^*) > 0$ .
3. By mathematical induction, we conclude that if  $\prod_{i=1}^K p(\mathbf{z}_i) > 0$  then  $p(\mathbf{z}) > 0$ .

Note that (3) implies  $\mathcal{S}(p(\mathbf{z})) = \mathcal{S}^\times(p(\mathbf{z}))$ , since  $\mathcal{S}^\times(p(\mathbf{z}))$  can be expressed as  $\{\mathbf{z} | p(\mathbf{z}_i) > 0\}$ . By using mathematical induction, we have proved that random mixing between two images can replace the random composition of multiple images to achieve disentanglement.

**Empirical results** In addition to proving equivalence, we compare our two image mixing strategies against mixing across multiple images (we use 64 here). We evaluate attribute disentanglement using

three random seeds and report the FactorVAE and DCI scores in Tab. 7. We observe no meaningful difference between mixing two images or 64 images, supporting our theoretical result.

Table 7: Effects of the number of samples used in mixing strategy.

# of samples for mixing	FactorVAE	DCI
2	0.975±0.040	0.837±0.105
64	0.966±0.032	0.802±0.088

## A.5 Evaluation metrics

In this section, we provide brief descriptions and definitions of the metrics we used in our experiments.

**FG-ARI** *Foreground Adjusted Rand Index* measures the agreement between predicted and ground-truth *instance* partitions, restricted to foreground pixels. Let  $\mathcal{I}$  be the set of foreground pixels in ground-truth labels, and let  $y \in \{1, \dots, K\}^{|\mathcal{I}|}$  and  $\hat{y} \in \{1, \dots, \hat{K}\}^{|\mathcal{I}|}$  denote ground-truth and predicted instance labels on  $\mathcal{I}$ , respectively. Then, we define the contingency table  $n_{ij} = |\{p \in \mathcal{I} : y_p = i, \hat{y}_p = j\}|$ , row sums  $a_i = \sum_j n_{ij}$ , and column sums  $b_j = \sum_i n_{ij}$ , with  $n = \sum_{ij} n_{ij} = |\mathcal{I}|$ . The FG-ARI is defined as

$$\text{FG-ARI} = \frac{\sum_{i,j} \binom{n_{ij}}{2} - \frac{\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}}{\binom{n}{2}}}{\frac{1}{2} \left( \sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2} \right) - \frac{\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}}{\binom{n}{2}}},$$

which lies in  $[-1, 1]$  and is higher when partitions agree better.

**mIoU** *Mean Intersection-over-Union* averages class-wise IoU. For class  $c$ , let  $P_c$  and  $G_c$  be predicted and ground-truth masks. Then mIoU is defined as

$$\text{IoU}_c = \frac{|P_c \cap G_c|}{|P_c \cup G_c|}, \quad \text{mIoU} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \text{IoU}_c,$$

where  $\mathcal{C}$  is the set of evaluated classes.

**mBO** *Mean Best Overlap* measures, for each ground-truth object, how well it is covered by its single most overlapping prediction, and then averages these values. Let  $\mathcal{O}$  be the set of ground-truth instances with masks  $\{G_o\}_{o \in \mathcal{O}}$ , and let  $\{M_k\}_{k=1}^K$  be the set of predicted instance masks. For each  $o$ ,

$$\text{BO}(o) = \max_{1 \leq k \leq K} \text{IoU}(G_o, M_k), \quad \text{mBO} = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} \text{BO}(o).$$

**FactorVAE Score [22]** The *FactorVAE Score* quantifies disentanglement of a representation with respect to generative factors. We repeatedly sample mini-batches in which exactly one factor is fixed and all others vary. For each batch, we compute the empirical variance of each latent coordinate, normalize coordinate-wise, *e.g.*, by the mean variance across coordinates, and select the index of the lowest-variance coordinate as a feature. A simple classifier, such as a majority vote, is used to predict the fixed factor index from that feature. The classification accuracy on held-out batches is reported as the FactorVAE Score (higher is better).

**Disentanglement Score in DCI [10]** Let  $R \in \mathbb{R}^{d \times M}$  be an importance matrix obtained by training a predictive model from latent coordinates  $z \in \mathbb{R}^d$  to factors  $\{v_m\}$ , where  $R_{jm} \geq 0$  measures the contribution of latent dimension  $j$  to predicting factor  $m$ . Define normalized importances  $\rho_{jm} = R_{jm} / \sum_{m'} R_{jm'}$ . The per-dimension disentanglement is

$$D_j = 1 - H(\rho_{j:}) / \log M,$$

where  $H$  is the entropy. Weighting by total importance  $w_j = \sum_m R_{jm}$  and normalizing, the overall DCI Disentanglement is

$$\text{DCI-D} = \sum_{j=1}^d \tilde{w}_j D_j, \quad \tilde{w}_j = \frac{w_j}{\sum_{j'} w_{j'}}.$$

**GRAM Loss** A *Gram loss* measures feature correlations of a neural network  $\phi$  at selected layers  $\mathcal{L}$ . It measures second-order feature statistics (style) between two images  $x$  and  $y$ . For two images  $x$  and  $y$ , let  $F_\ell(x) \in \mathbb{R}^{C_\ell \times H_\ell W_\ell}$  be the feature at layer  $\ell$ , obtained by reshaping  $\phi_\ell(x)$ . The Gram matrix is  $G_\ell(x) = \frac{1}{C_\ell H_\ell W_\ell} F_\ell(x) F_\ell(x)^\top$ . The GRAM loss is measured as

$$\mathcal{L}_{\text{GRAM}}(x, y) = \sum_{\ell \in \mathcal{L}} \|G_\ell(x) - G_\ell(y)\|_F^2.$$

## A.6 Additional Implementation Details

When training our model, we use a fixed batch size of 64 and a learning rate of 0.0001 across all of the experiments. We use  $\lambda_{\text{Prior}} = 1$  and  $\lambda_{\text{Con}} = 0.01$  for all experiments. We set the number of latents to  $k = 10$  in attribute disentanglement and use  $K = 4, 11, 11, 12, 6$  for CLEVREasy, CLEVR, CLEVRTex, ClevrTex-Style, MSN-Style datasets in object disentanglement, respectively. When training the diffusion model, we use a v-prediction [40] loss to ensure reliable few-step generation.

Tab. 8–17 summarizes the hyper-parameters of our encoder and decoder architectures used in the experiments. Following the diffusion-based baselines: DisDiff [52] and LSD [19], we employ pre-trained VQ-VAE<sup>3</sup> and KL-regularized auto-encoder model<sup>4</sup> in attribute and object disentanglement, respectively. In attribute disentanglement, the encoder first maps the input  $\mathbf{x}$  into a  $KD$ -dimensional vector  $\mathbf{z} \in \mathbb{R}^{KD}$  and then uniformly divides it into  $K$  latents. In object disentanglement, we adopt the Q-former [27] of 4 transformer blocks with 8 attention heads and a hidden dimension of 256. Specifically, we have  $K$  learnable queries  $\{\mathbf{q}\}^K \in \mathbb{R}^{K \times D}$  and those queries are updated via multiple self attention layers and cross attention layers, where the keys and values are linearly projected from the U-net encoder feature of  $\mathbf{x}$ .

<sup>3</sup> <https://huggingface.co/stabilityai/sd-vae-ft-ema-original>

<sup>4</sup> <https://ommer-lab.com/files/latent-diffusion/celeba.zip>

---

Conv  $3 \times 3 \times 3 \times 128$ , stride=1  
BatchNorm2d  
ReLU  
Conv  $3 \times 3 \times 128 \times 128$ , stride=1  
BatchNorm2d  
ReLU  
Conv  $3 \times 3 \times 128 \times 128$ , stride=1  
BatchNorm2d  
ReLU  
Conv  $3 \times 3 \times 128 \times 128$ , stride=1  
BatchNorm2d  
ReLU  
ResBlock  $3 \times 3 \times 128 \times 128$ , stride=1  
BatchNorm2d  
ReLU  
ResBlock  $3 \times 3 \times 128 \times 128$ , stride=1  
BatchNorm2d  
ReLU  
FC  $4096 \times 10$

---

Table 8: Encoder Architecture used in attribute disentanglement.

Input Resolution	16
Input Channels	3
Output Resolution	16
Self Attention	Middle Layer
Base Channels	128
Channel Multipliers	[1,1,2,4]
# Heads	8
# Res Blocks / Layer	1

Table 11: Unet Encoder Architecture used in object disentanglement.

Input Resolution	16
Input Channels	4
$\beta$ scheduler	Linear
Mid Layer Attention	Yes
# Res Blocks / Layer	2
# Heads	8
Base Channels	192
Attention Resolution	[1,2,4,4]
Channel Multipliers	[1,2,4,4]

Table 14: Generative Prior Architecture used in object disentanglement.

---

ReLU  
Conv  $3 \times 3 \times 128 \times 128$ , stride=1  
BatchNorm2d  
ReLU  
Conv  $3 \times 3 \times 128 \times 128$ , stride=1

---

Table 9: ResBlock in the Encoder

Input Resolution	16
Input Channels	4
$\beta$ scheduler	Linear
Mid Layer Attention	Yes
# Res Blocks / Layer	2
# Heads	8
Base Channels	192
Attention Resolution	[1,2,4,4]
Channel Multipliers	[1,2,4,4]

Table 12: Decoder Architecture used in object disentanglement.

---

Conv  $3 \times 3 \times 3 \times 128$ , stride=1  
BatchNorm2d  
ReLU  
Conv  $3 \times 3 \times 128 \times 128$ , stride=1  
BatchNorm2d  
ReLU  
Conv  $3 \times 3 \times 128 \times 128$ , stride=1  
BatchNorm2d  
ReLU  
Conv  $3 \times 3 \times 128 \times 128$ , stride=1  
BatchNorm2d  
ReLU  
ResBlock  $3 \times 3 \times 128 \times 128$ , stride=1  
BatchNorm2d  
ReLU  
ResBlock  $3 \times 3 \times 128 \times 128$ , stride=1  
BatchNorm2d  
ReLU  
FC  $4096 \times 10$

---

Table 15: Encoder Architecture used in attribute disentanglement.

Input Resolution	16
Input Channels	3
Input Channels	4
$\beta$ scheduler	Linear
Mid Layer Attention	Yes
# Res Blocks / Layer	2
# Heads	8
Base Channels	64
Attention Resolution	[1,2,4,4]
Channel Multipliers	[1,2,4,4]

Table 10: Decoder Architecture used in attribute disentanglement

Input Resolution	16
Input Channels	3
$\beta$ scheduler	Linear
Mid Layer Attention	Yes
# Res Blocks / Layer	2
# Heads	8
Base Channels	64
Attention Resolution	[1,2,4,4]
Channel Multipliers	[1,2,4,4]

Table 13: Generative Prior Architecture used in attribute disentanglement.

---

ReLU  
Conv  $3 \times 3 \times 128 \times 128$ , stride=1  
BatchNorm2d  
ReLU  
Conv  $3 \times 3 \times 128 \times 128$ , stride=1

---

Table 16: ResBlock in the Encoder

Input Resolution	16
Input Channels	4
$\beta$ scheduler	Linear
Mid Layer Attention	Yes
# Res Blocks / Layer	2
# Heads	8
Base Channels	192
Attention Resolution	[1,2,4,4]
Channel Multipliers	[1,2,4,4]

Table 17: Generative Prior Architecture used in object disentanglement.

### A.7 Details on Construction of CLEVR-Style and MSN-Style datasets

To construct the CLEVR-Style dataset, we first sample 25K images from the original CLEVR dataset and then augment each with three additional styles. Including the unmodified images, this produces a total of 80K/10K/10K images for the train/val/test splits, respectively. All augmentations combine simple color-space adjustments with diffusion-based translation. The first style (second column of Fig. 8) applies an HSV shift (hue=0, saturation=8, value=2) followed by image translation using Stable Diffusion [38] with the prompt “an oil painting.” The second style (third column of Fig. 8) applies an HSV shift (hue=0, saturation=1.5, value=2.5), converts to LAB color space, and reduces contrast by a factor of 0.15. The third style (fourth column of Fig. 8) applies an HSV shift (hue=5, saturation=3, value=1), converts to LAB color space for CLAHE (Contrast Limited Adaptive Histogram Equalization), and then segments the image into 480 superpixels. These combined transforms produce 4 different visual styles while preserving the underlying scene composition.

Similarly, we construct the MSN-Style dataset by first sampling 15k images from the original MSN dataset [45] and augmenting with three identical styles used in CLEVR-Style. It produces a total of 40k/10k/10k images for the train/val/test splits, respectively. Fig. 9 presents the sample in MSN-Style with four different styles.

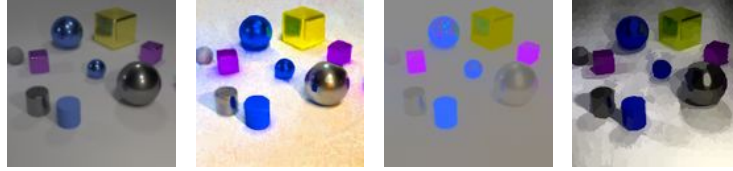


Figure 8: Example of CLEVR-Style dataset

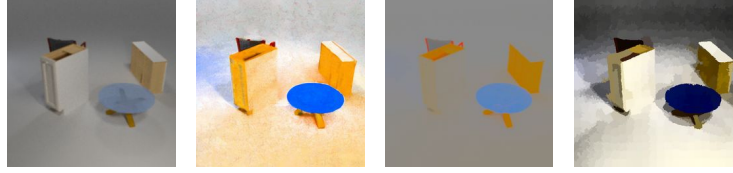


Figure 9: Example of MSN-Style dataset

### A.8 Experimental Details for Object Property Prediction

**Matching Technique** We developed a technique to identify the specific region corresponding to an object’s representation by analyzing images composed with that representation. For a given target object latent representation, we first randomly sample multiple images and encode each into an object representation. For each image, we then replace one object latent with the target latent, then decode the mixed representations. If the target object is properly encoded, it appears in the generated images. To determine the object region, we measure the RGB variance across these generated images and compare each generated image to the original one containing the target object representation. Finally, we combine these two metrics: variance and distance to the original image, to specify the object’s region. We provide the pseudocode in Algorithm 1.

**Evaluation protocol for object property prediction** For each property, we follow [21] to train a two-layer MLP classifier with hidden dimension 256 on the frozen object representations.

---

#### Algorithm 1 Matching Technique

---

**Require:**  $x$ : an image;  $z = \text{enc}(x)$ : object representation of  $x$ ;  $n$ : number of latent vectors;  $x_{\text{ref}}$ : randomly sampled reference  $B$  images

```

1: function MATCHING( $z, x, x_{\text{ref}}$ )
2:    $z_{\text{ref}} \leftarrow [\text{encode}(\_x) \text{ for } \_x \in x_{\text{ref}}]$ 
3:    $z_{\text{mixed}} \leftarrow [\text{replace\_ith\_latent}(z_{\text{ref}}, z, i) \text{ for } i = 1 \dots n]$ 
4:    $x_{\text{mixed}} \leftarrow [\text{decode}(\_z) \text{ for } \_z \in z_{\text{mixed}}]$ 
5:    $x_{\text{ref\_cm}} \leftarrow \text{mean}(x_{\text{ref}}, \text{dim} = 0)$ 
6:    $x_{\text{mixed\_cm}} \leftarrow \text{mean}(x_{\text{mixed}}, \text{dim} = 1)$ 
7:    $s \leftarrow \text{softmax}(1 - \text{distance}(x, x_{\text{mixed\_cm}}). \text{mean}(-1))$ 
8:    $d \leftarrow \text{softmax}(\text{distance}(x_{\text{mixed\_cm}}, x_{\text{ref\_cm}}). \text{mean}(-1))$ 
9:   return  $(s + d). \text{argmax}(\text{dim} = 0)$ 
10: end function
```

▷ Mask indicating matched region

---

### A.9 Comparison to Likelihood Maximization by [21]

We conducted experiments on the CLEVRText dataset by replacing our prior loss with the likelihood maximization loss proposed in L2C. Tab. 18 shows the results, which clearly demonstrate the superior performance of our likelihood maximization term compared to L2C.

Table 18: Object property prediction results with L2C’s likelihood maximization (prior) loss

	shape ( $\uparrow$ )	material ( $\uparrow$ )	position ( $\downarrow$ )
Ours	70.90	52.20	0.133
Ours + L2C prior loss	54.58	27.18	0.165

### A.10 Additional Results on Attribute Disentanglement

As in the Shapes3D dataset, our method captures three independent factors, *i.e.*, direction, axis, and appearance, enabling controlled manipulation of each factor in the Cars3D dataset and presents the result in Fig. 10.



Figure 10: Qualitative results on attribute disentanglement in Cars3D dataset.

### A.11 Additional Details and Results on Unsupervised Object Segmentation

As described in the main paper, our method does not include a built-in mechanism (*e.g.*, slot-attention) to explicitly cluster pixels. To address this, we followed [31] to train the Spatial Broadcast Decoder (SBD)[49] on the frozen latent representations to predict object masks for each latent. Specifically, each frozen object representation is decoded individually by the SBD into an RGB image and an alpha mask. We then normalize the alpha masks across all object representations using a softmax and use them as mixture weights to combine the RGB outputs into a single image. We treat the normalized alpha masks as object-mask proxies for evaluation. The decoder is trained with a reconstruction loss to recover the original image for 30k iterations with a learning rate of 0.001. Because the encoder is frozen and the SBD is shallow (see Tab. 19), this training process is relatively inexpensive.

While conventional methods often extract object masks from the attention weights of slot-attentions, we evaluate the baselines using both these slot-attention masks and masks obtained by training an SBD on their frozen latent representations, for a fair comparison. The full results are reported in Tab. 20, Tab. 21. In the main paper, we present only the SBD-based results, since they outperform those derived from slot-attention masks. In Fig. 11, we also present object segmentation results on CLEVR and CLEVRText datasets.



Table 19: Decoder Architecture used in object segmentation

Deconv  $5 \times 5 \times 64 \times 64$ , stride=2, padding=2, output\_padding=1  
ReLU  
Deconv  $5 \times 5 \times 64 \times 64$ , stride=2, padding=2, output\_padding=1  
ReLU  
Deconv  $5 \times 5 \times 64 \times 64$ , stride=2, padding=2, output\_padding=1  
ReLU  
Deconv  $5 \times 5 \times 64 \times 64$ , stride=2, padding=2, output\_padding=1  
ReLU  
Deconv  $5 \times 5 \times 64 \times 64$ , stride=1, padding=1, output\_padding=1  
ReLU  
Deconv  $5 \times 5 \times 64 \times 4$ , stride=1, padding=1, output\_padding=1  
ReLU

Table 20: Quantitative Results of unsupervised segmentation in the CLEVR dataset.

Method	FG-ARI		mIoU		mBO	
	Slot-Attention	SBD Mask	Slot-Attention	SBD Mask	Slot-Attention	SBD Mask
LSD	82.00	<b>91.74</b>	22.69	25.59	22.98	25.84
L2C	54.01	80.05	19.30	<u>25.61</u>	20.36	<u>26.33</u>
Ours	-	<u>91.20</u>	-	<b>26.54</b>	-	<b>26.65</b>

Table 21: Quantitative Results of unsupervised segmentation in CLEVRTex dataset.

Method	FG-ARI		mIoU		mBO	
	Slot-Attention	SBD Mask	Slot-Attention	SBD Mask	Slot-Attention	SBD Mask
LSD	46.54	71.64	45.87	56.26	46.93	56.75
L2C	77.07	<u>82.55</u>	56.59	<u>58.33</u>	53.25	<u>58.68</u>
Ours	-	<b>87.68</b>	-	<b>58.88</b>	-	<b>59.12</b>

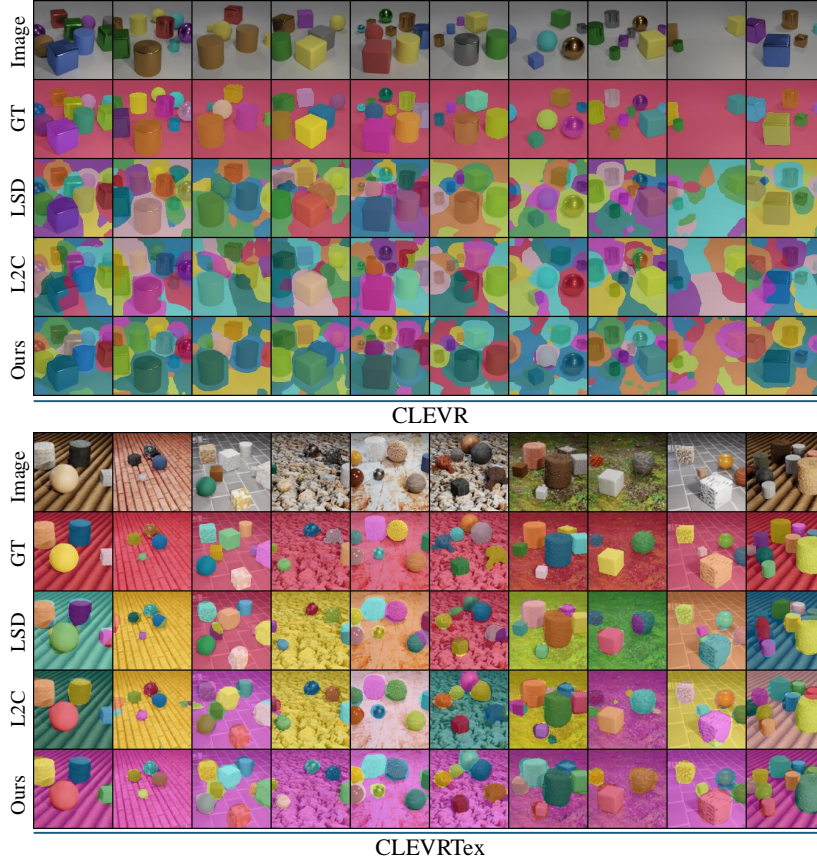


Figure 11: Unsupervised object segmentation results in the CLEVR and CLEVRTex dataset. In CLEVRTex, our method consistently encodes complete objects into distinct latents, resulting in clean object segmentation results. On CLEVR, the constant background appears in the object regions of the segmentation results, leading to relatively lower mIoU and mBO (Tab. 20). However, this does not affect the underlying quality of object representations, as our method still consistently captures each entire object in its segmentation region.

### A.12 Additional Results on Joint Disentanglement of Attribute and Object

We present additional qualitative results on joint disentanglement of attribute (style) and object in Fig. 12– 15 and Fig. 16, respectively. While all the baselines struggle to disentangle the style information into a single latent representation, our method successfully disentangles the style and transfers it from source to target images. In addition to style disentanglement, our method also disentangles individual objects and enables object-wise manipulation as shown in Fig. 16.

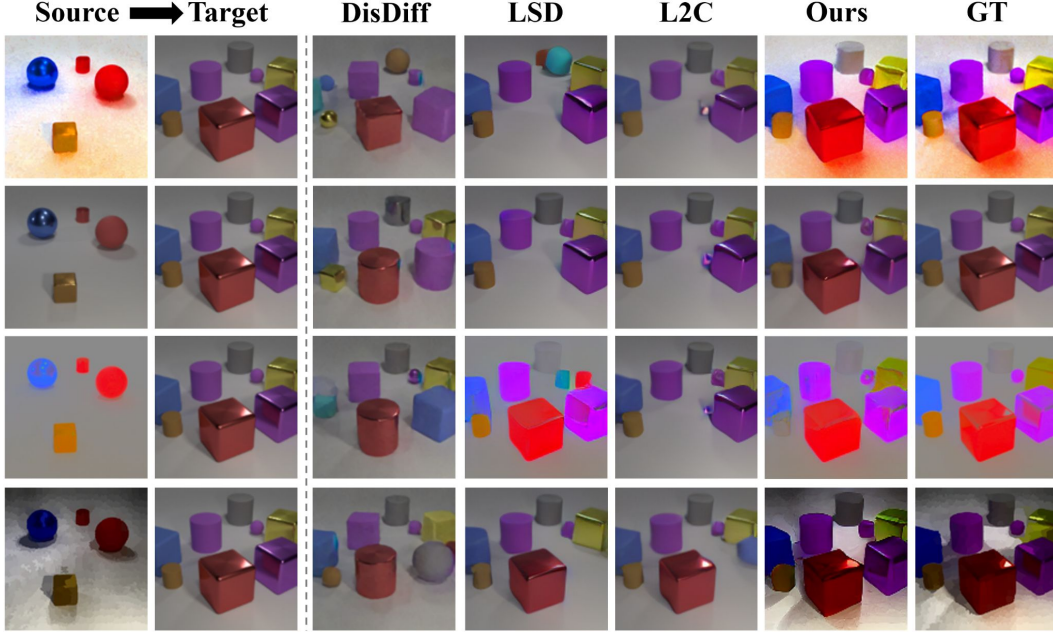


Figure 12: Style Transfer in CLEVR-Style dataset.

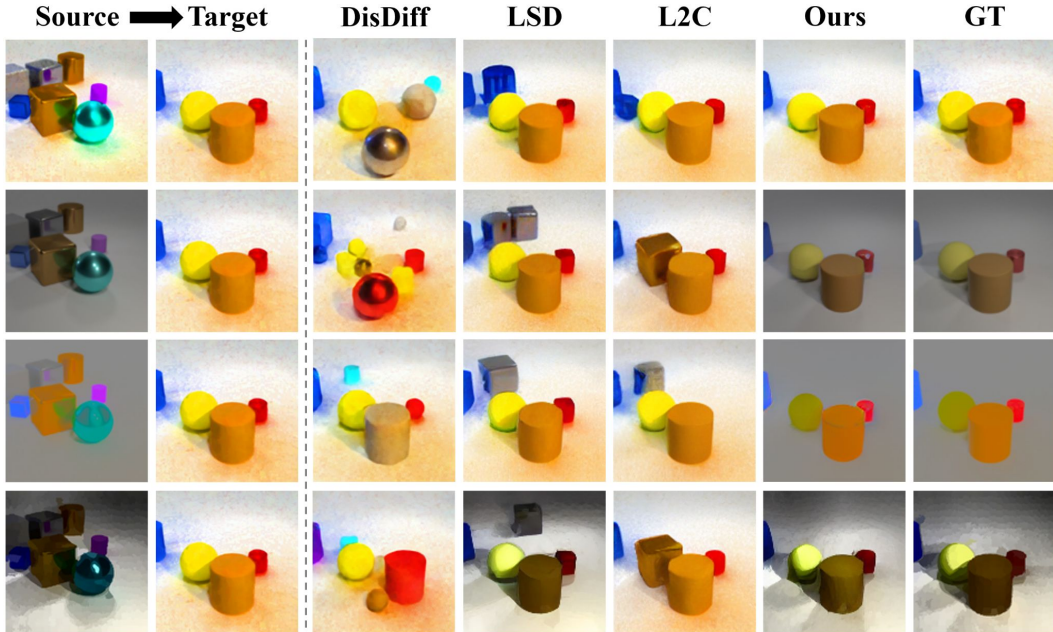


Figure 13: Style Transfer in CLEVR-Style dataset.

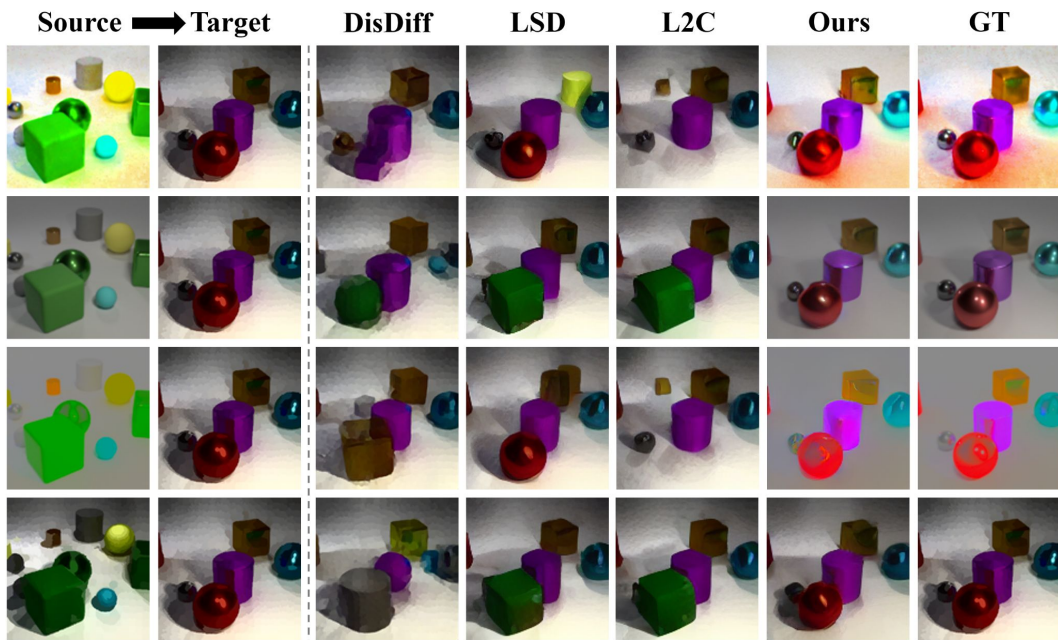


Figure 14: Style Transfer in CLEVR-Style dataset.

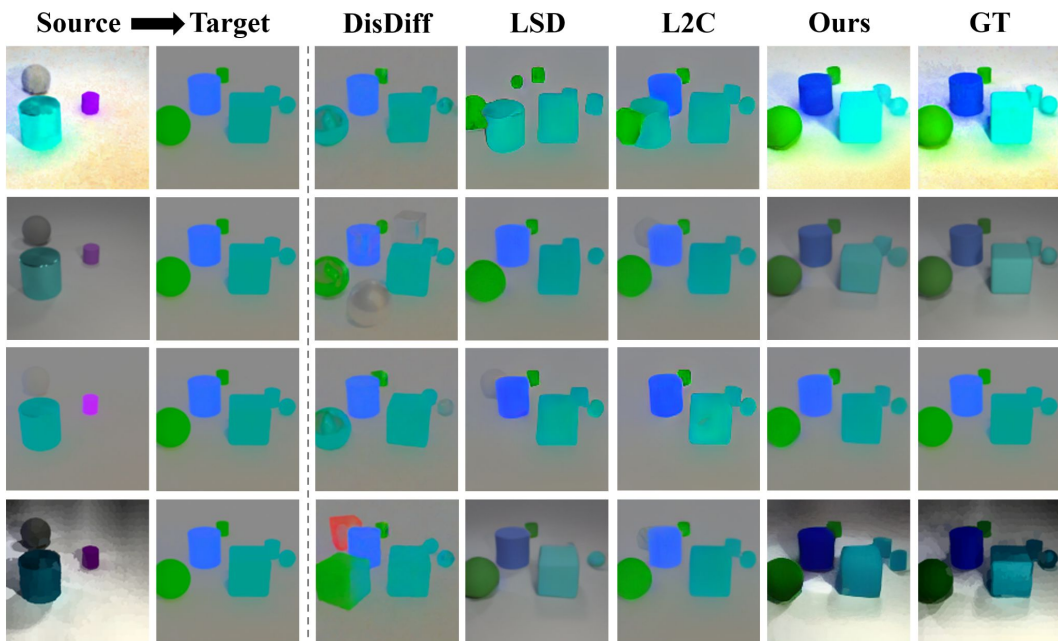


Figure 15: Style Transfer in CLEVR-Style dataset.



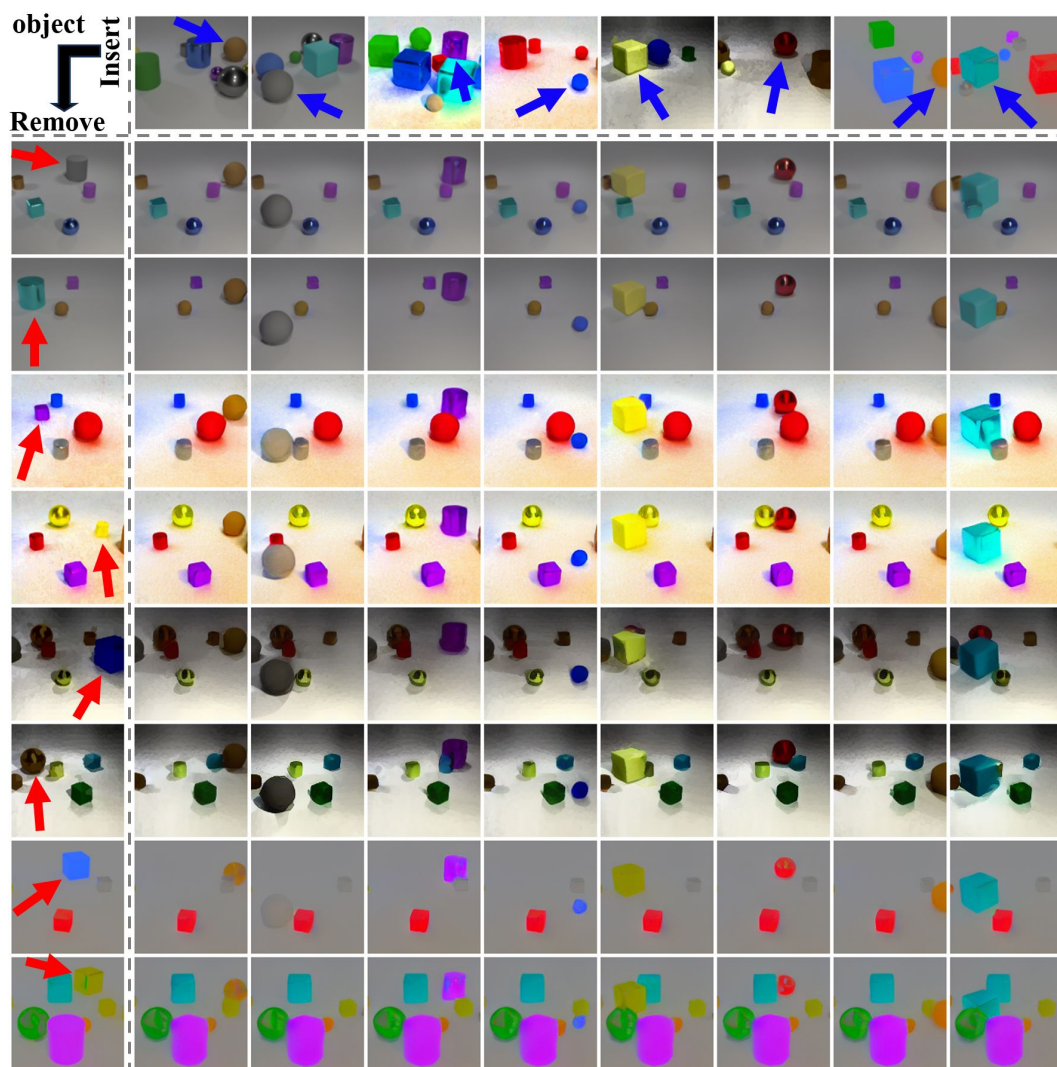


Figure 16: Additional qualitative results on Object Manipulation in CLEVR-Style dataset. Objects marked by red arrows are replaced with those marked by blue arrows. It demonstrates that our method effectively disentangles individual objects.

### A.13 Additional Results of Joint Disentanglement on Complex Dataset

To further validate our method on more complex and realistic datasets, we augment the MultiShapeNet (MSN) dataset [45] with four different global styles as in our previous CLEVR-Style experiments (See Appendix A.7 for details). MSN includes 11,733 unique furniture shapes with increased visual complexity compared to CLEVR. We compare our method with the strongest attribute-disentanglement baseline (DisDiff) and object-disentanglement baselines (LSD, L2C), reporting quantitative results in Tab. 22.

For style disentanglement, our method achieved the highest style prediction accuracy (Acc) and the lowest style loss (GRAM), demonstrating that it successfully isolates style information in a single latent and transfers it faithfully to the target image. Since baselines lack an internal mechanism to specify the latent representation for style information, we used a simple trick to identify the latent representation encoding style information. We decode all  $K \times K$  possible latent exchanges between two sets of  $K$  latents extracted from two images, and choose the pair yielding the lowest GRAM loss. Even with their best-case result, they fell short of our performance, confirming that they do not reliably capture a style factor. Although L2C achieves relatively high style prediction accuracy, style-swapping results in Fig. 17, 18 shows that swapping the single latent representation affects both the global style and objects, which indicates that L2C fails to isolate the global style but rather is entangled with object information. DisDiff also fails to produce reasonable compositions, and we conjecture that this is due to their disentanglement objective (variant/invariant loss), which is not well-suited for multi-object scenes and destabilizes the overall training.

For object disentanglement, Tab. 22 shows that our method produces the highest mIoU and mBO, indicating accurate localization and tight boundaries for each object mask, while FG-ARI was a bit lower than that of LSD and L2C. This trade-off arises because our broadcast decoder generated tight, object-internal masks. FG-ARI—which measures membership between two pixels within the same group—penalizes such masks when GT segmentations are looser, whereas mIoU and mBO reward the improved boundary precision. In contrast, slot attention modules in LSD and L2C often generate larger masks that bleed into the background or even include other objects, which leads to higher FG-ARI at the expense of mIoU and mBO. DisDiff, whose information-theoretic objective is not designed for a varying number of objects, fails to effectively separate either style or object factors.

Table 22: Quantitative results on Joint Disentanglement in MSN-Style

Method	Style		Object		
	Acc(↑)	GRAM(↑)	FG-ARI(↑)	mIoU (↑)	mBO (↑)
DisDiff	37.0	10.8	5.62	7.69	9.13
LSD	12.5	11.3	52.7	23.3	23.9
L2C	81.0	8.35	<b>53.3</b>	28.4	28.8
Ours	<b>97.5</b>	<b>7.16</b>	42.3	<b>44.1</b>	<b>44.2</b>

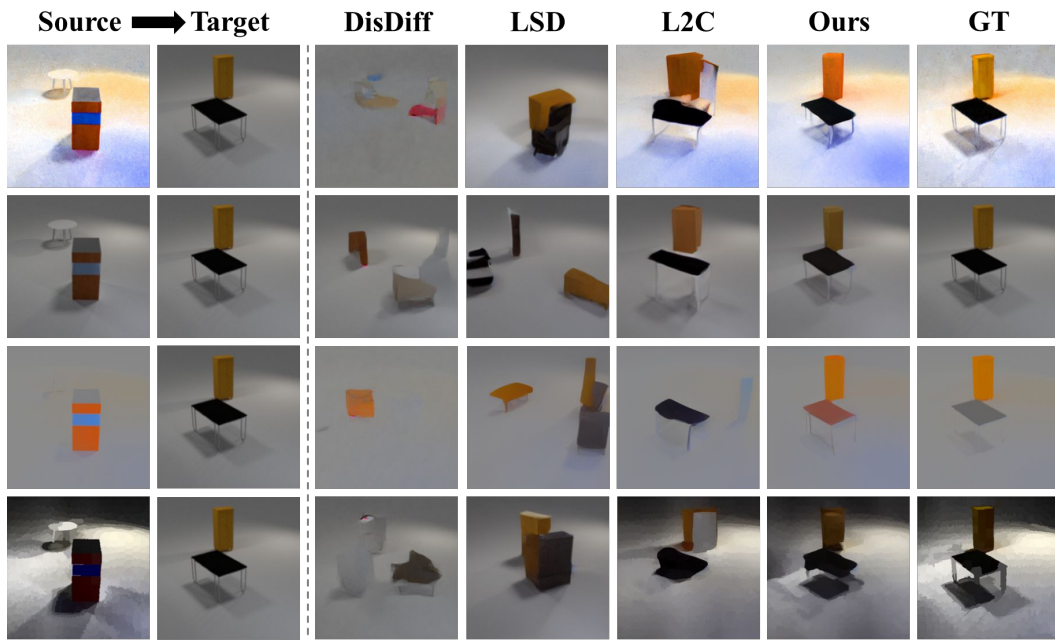


Figure 17: Style transfer results on MSN-Style dataset.

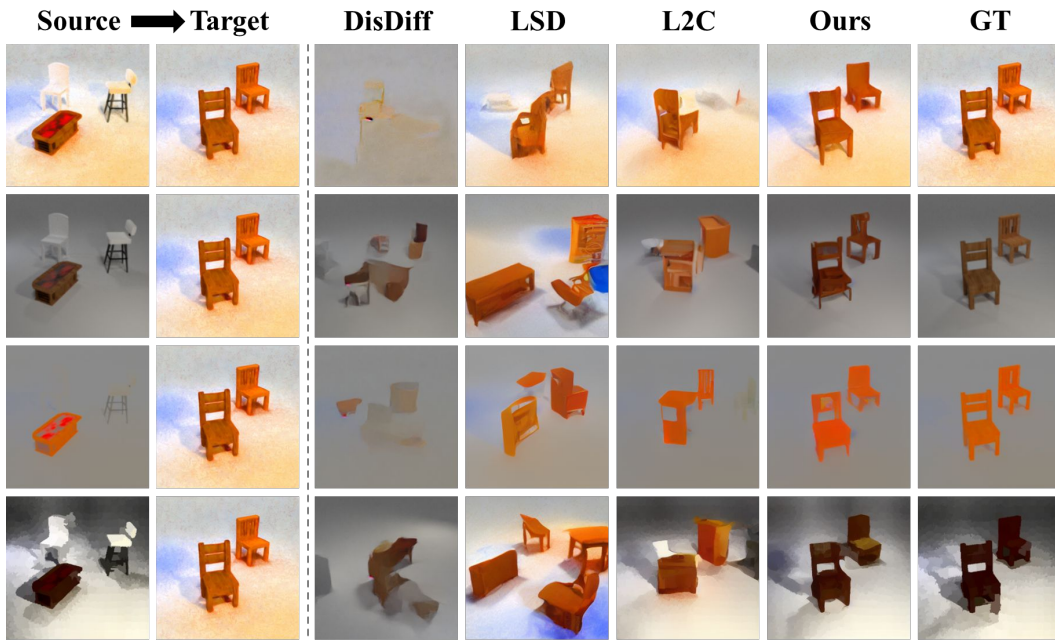


Figure 18: Style transfer results on MSN-Style dataset.

#### A.14 Effect of random seeds on performance

We repeat our attribute/object disentanglement experiments with ten/three different random seeds and present the results in Tab. 23, Tab. 24, respectively, showing that our method achieves competitive performance in both tasks.

Table 23: Quantitative results on scene-level disentanglement. Our method achieves state-of-the-art performance in almost all of the datasets, except FactorVAE score in Cars3D.

Method	Cars3D		Shapes3D		MPI3D	
	FactorVAE	DCI	FactorVAE	DCI	FactorVAE	DCI
FactorVAE [22]	0.906±0.052	0.161±0.019	0.840±0.066	0.611±0.082	0.152±0.025	0.240±0.051
$\beta$ -TCVAE [5]	0.855±0.082	0.140±0.019	0.873±0.074	0.613±0.114	0.179±0.017	0.237±0.056
InfoGAN-CR [29]	0.411±0.013	0.020±0.011	0.587±0.058	0.478±0.055	0.439±0.061	0.241±0.056
LD [47]	0.852±0.039	0.216±0.072	0.805±0.064	0.380±0.064	0.391±0.039	0.196±0.038
GS [15]	0.932±0.018	0.209±0.031	0.788±0.091	0.284±0.034	0.464±0.036	0.229±0.042
DisCo [37]	0.855±0.074	0.271±0.037	0.877±0.031	0.708±0.048	0.371±0.030	0.292±0.024
DisDiff-VQ [52]	<b>0.976±0.018</b>	0.232±0.019	0.902±0.043	0.723±0.013	0.617±0.070	0.337±0.057
<b>Ours</b>	0.877±0.089	<b>0.365±0.073</b>	<b>0.975±0.059</b>	<b>0.837±0.105</b>	<b>0.708±0.060</b>	<b>0.458±0.052</b>

Table 24: Object property prediction results with 3 different runs for our model.

Method	CLEVREasy			CLEVR				CLEVRTex		
	Shape (↑)	Color (↑)	Position* (↑)	Shape (↑)	Color (↑)	Material (↑)	Position (↓)	Shape (↑)	Material (↑)	Position (↓)
SA	72.25	72.33	44.08	79.4	91.30	93.18	0.064	30.44	7.890	0.482
SLASH	86.06	89.23	46.97	83.28	92.26	93.16	0.078	53.13	37.49	0.148
LSD	<b>96.03</b>	<b>98.05</b>	50.29	<b>87.66</b>	91.46	<b>94.87</b>	0.062	68.25	<u>51.54</u>	0.197
L2C	92.78	93.57	47.62	73.61	74.03	86.93	0.168	<b>71.54</b>	<b>51.62</b>	<b>0.116</b>
<b>Ours</b>	<u>93.74±2.10</u>	<u>94.29±0.97</u>	<u>49.42±1.15</u>	<u>85.72±0.37</u>	<b>93.79±0.22</b>	<b>94.93±0.07</b>	<b>0.058±0.006</b>	<u>68.29±2.55</u>	47.89±4.89	<u>0.143±0.009</u>

#### A.15 Experiments on Correlated Factors

The datasets used in our main experiments handle only simple scenarios with statistically independent factors of variation (FoVs), unlike real-world cases. To verify our method on more challenging scenarios with correlated factors, we followed Roth et al. [39] to generate correlated benchmarks. We introduced 0.1 correlation between 1/2/3 pairs of GT factors in Shapes3D, as in [39], then trained and evaluated our models using FactorVAE and DCI metrics across 3 seeds. Tab. 25 reports the results.

Our method showed only slight metric drops despite correlations. Unlike prior work that relies on statistical independence between latent variables, *e.g.*, Total Correlation, our compositional objective only encourages discovering atomic factors whose compositions are valid under predefined mixing strategies. This allows successful disentanglement even with correlated factors, as the compositional requirement doesn’t penalize correlations.

Table 25: Quantitative results on the dataset with correlated factors. Our method robustly disentangles the underlying ground-truth factors even with correlated factors.

Amount of Correlation	Factor VAE	DCI
No Correlation	0.975±0.059	0.837±0.105
Pairs: 1, $\sigma = 0.1$	0.930±0.061	0.798±0.082
Pairs: 2, $\sigma = 0.1$	0.997±0.002	0.806±0.060
Pairs: 3, $\sigma = 0.1$	0.965±0.030	0.801±0.024

#### A.16 Computing Resources

We conduct all our experiments on a GPU Server that consists of an Intel Xeon Gold 6230 CPU, 256GB RAM, and 8 NVIDIA RTX 3090 GPUs (with 24GB VRAM), or 8 NVIDIA RTX 6000 GPUs (with 48GB VRAM). It takes about 24 GPU hours and from 36 to 48 GPU hours for the attribute and object disentanglement experiment, respectively.

## NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [NA].
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

**The checklist answers are an integral part of your paper submission.** They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]" provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No]" or "[NA]" is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading “NeurIPS Paper Checklist”,**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers.**

### 1. Claims

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

Answer: [Yes]

Justification: We states our motivation, contributions, scope of our work in abstract and introduction.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

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

Answer: [Yes]



Justification: We provide limitation of our work in Appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 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: We do not claim for theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

### 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: We provide all the information needed to reproduce the experimental results.

Guidelines:

- The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 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: Our code is not cleaned and prepared enough for sharing. Our dataset and codes will be released in future.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

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

Answer: [\[Yes\]](#)

Justification: We specify all the details for experimental setting.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 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: We report standard deviation of our experiments for multiple runs with different seeds.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 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: We state information about computing resources in Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

#### 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: We followed the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

#### 10. Broader impacts

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

Answer: [Yes]

Justification: We provide Broader impacts in Appendix.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

#### 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., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: Our paper possess no risk.

Guidelines:

- The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 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: We cite all the codes, data, papers, and pretrained models in our paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

## 13. New assets

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

Answer: [\[Yes\]](#)

Justification: We provide a detailed information about generating our new synthetic dataset.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

## 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: [\[NA\]](#)

Justification: We do not conduct crowdsourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

**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: [NA]

Justification: We do not conduct crowdsourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

**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 rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: We do not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.