Intrinsic Concept Extraction Based on Compositional Interpretability

Hanyu Shi[1*†] Hong Tao[2*] Guoheng Huang[1‡] Jianbin Jiang[2]
Xuhang Chen[3‡] Chi-Man Pun[4] Shanhu Wang[2‡] Pan Pan[2]
[1]Guangdong University of Technology [2]VIPSHOP [3]Huizhou University [4]University of Macau

Abstract

Unsupervised Concept Extraction aims to extract concepts from a single image; however, existing methods suffer from the inability to extract composable intrinsic concepts. To address this, this paper introduces a new task called Compositional and Interpretable Intrinsic Concept Extraction (CI-ICE). The CI-ICE task aims to leverage diffusion-based text-to-image models to extract composable object-level and attribute-level concepts from a single image, such that the original concept can be reconstructed through the combination of these concepts. To achieve this goal, we propose a method called HyperExpress, which addresses the CI-ICE task through two core aspects. Specifically, first, we propose a concept learning approach that leverages the inherent hierarchical modeling capability of hyperbolic space to achieve accurate concept disentanglement while preserving the hierarchical structure and relational dependencies among concepts; second, we introduce a concept-wise optimization method that maps the concept embedding space to maintain complex inter-concept relationships while ensuring concept composability. Our method demonstrates outstanding performance in extracting compositionally interpretable intrinsic concepts from a single image.

1. Introduction

Concept extraction [41] aims to extract symbols with human-interpretable meanings from visual images and is often used to explain the behaviors of models. With the development of multimodal models [47, 48, 50], in recent years, a variety of methods [2, 5, 15, 19, 41] have explored the concept extraction capabilities of diffusion-based Text-to-Image (T2I) models. Concept extraction can be categorized into supervised concept extraction [2, 19] and unsupervised concept extraction [5, 15, 41]. Supervised concept extraction [2, 19] relies on external human-labeled knowledge for supervised learning, and this limitation poses significant obstacles to its practical applications. In contrast, unsupervised concept extraction [5, 15, 41] aims to achieve concept extraction without relying on prior knowledge of concepts. Hao *et al*. [15] proposed the Unsupervised Concept Extraction (UCE) task. UCE methods such as Break-A-Scene [2], ConceptExpress [15], and AutoConcept [8]

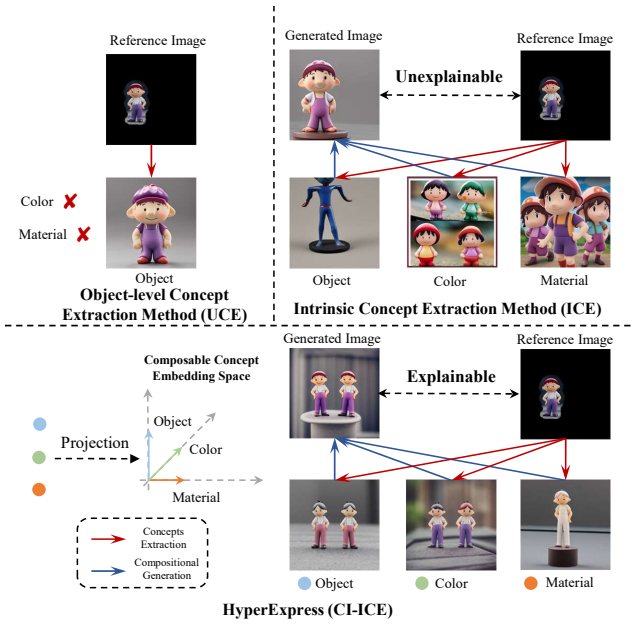

Figure 1. **The difference between unsupervised concept extraction (UCE) methods [5, 8, 41] and the composable and interpretable intrinsic concept extraction method**. Object-level concept extraction method [8, 15] can only extract object-level concepts and is unable to extract attribute-level concepts such as color and material. Although the intrinsic concept extraction method [5] can extract both object-level concepts and attribute-level concepts, the concepts it extracts are not sufficiently close to the original concepts, and it fails to consider the compositionality of the embedding space; thus, it has poor interpretability. In contrast, HyperExpress considers the relationships between concepts when learning them. As a result, the extracted concepts are more aligned with the objects in the image. Additionally, it imposes compositional constraints on the concept embedding space, thereby enabling concept extraction and combination capabilities that are understandable to humans.

*Equal contribution
†This work was completed during an internship at VIPSHOP
‡Corresponding authors

are designed to extract concepts from a single image; however, they can only extract object-level concepts. Cendra *et al*. [5] proposed an intrinsic concept extraction method, ICE [5], which is capable of extracting both object-level and attribute-level concepts from a single image. Although these methods all perform well in the UCE [15] task, none of them consider the composability of concepts. This leads to poor interpretability, and there is uncertainty when using these concepts, which lack composability, to explain the original image content. This limits the ability to control and trust the model [41]. Although Stein *et al*. [41] proposed the CCE [41] method for extracting composable concepts, it requires learning from multiple images containing the same concept, which also limits its application. Stein *et al*. [41] proposes that two composable concepts exhibit an approximately orthogonal relationship in the embedding space, and the effect of concepts is maximized when they possess composability. Current UCE methods [5, 8, 15] do not impose any constraints on the concept embedding space, which results in the concepts learned by these methods lacking composability and thus undermining their effectiveness.

To address the problem of extracting composable object-level concepts and attribute-level concepts from a single image, we propose a new task called Compositional and Interpretable Intrinsic Concept Extraction (CI-ICE). CI-ICE aims to leverage diffusion-based Text-to-Image (T2I) [36] models to extract object-level concepts and their corresponding attribute-level concepts from a single reference image. Moreover, these extracted concepts can be combined to form a complete sample, which is used to explain the complex concepts in the original reference image. In Fig. 1, we illustrate the distinctions between the method based on CI-ICE and existing concept extraction approaches. Current concept extraction methods place greater emphasis on concept disentanglement while overlooking their composability, thus failing to adequately accomplish the CI-ICE task. The CI-ICE task faces two challenges:

- The accurate disentanglement of object-level concepts and attribute-level concepts: there exist obvious hierarchical structures and associative relationships between these two types of concepts.
- The learned concept embedding space is required to have composability: only composable concepts can better explain the original complex concepts in the reference image.

To address the two challenges of the CI-ICE task, we propose the HyperExpress method. The HyperExpress method tackles the two challenges of CI-ICE from two aspects: concept learning and concept optimization. In terms of concept learning, we propose a Hyperbolic Contrastive Learning module and a Hyperbolic Entailment Learning module. The Hyperbolic Contrastive Learning module leverages the inherent hierarchical modeling capability of hyperbolic space. In this design, object-level concepts and attribute-level concepts are positioned at different locations in the space, distinguishing object-level concepts from attribute-level concepts within complex concepts. The Hyperbolic Entailment Learning module, on the other hand, establishes relational dependencies between object-level and attribute-level concepts based on the hyperbolic entailment cone. In terms of concept optimization, we aim to achieve concept composability while preserving the hierarchical structure and relational dependencies among concepts through a Horosphere Projection module. As shown in Fig. 1, compared to existing concept extraction methods, our approach not only accurately disentangles object-level and attribute-level concepts but also ensures concept composability through the optimization of the concept embedding space. Experimental validation demonstrates that the HyperExpress method exhibits promising potential in mining composable visual concepts. Our contributions are summarized below:

- We propose the task of Compositionally Interpretable Intrinsic Concept Extraction (CI-ICE), aiming to address the issue that existing unsupervised concept extraction (UCE) methods fail to extract composable intrinsic concepts.
- We propose the HyperExpress method, which aims to extract composable intrinsic concepts from a single image to accomplish the CI-ICE task.
- We evaluated the effectiveness of the HyperExpress method on the UCE benchmark, and the experiments demonstrate that the HyperExpress method is a highly promising solution for extracting composable intrinsic concepts.

2. Related Work

2.1. Generative Concept Learning

Generative Concept Learning aims to decompose a complex visual concept into simple, basic elements. Some methods [7, 10, 11, 14, 18, 22, 25, 28, 33, 37, 39, 42, 45] extract concepts from multiple images that contain the same concept. For example, the Textual Inversion [10] method represents a certain concept by learning an embedding vector, but it can only learn a single concept; Liu *et al*. [27] extended the Textual Inversion method to extract multiple concepts. These methods all rely on multiple images representing the same concept, and this limitation undermines their practicality. Although methods such as Break-A-Scene [2], MCPL [19], and DisenDiff [51] can extract concepts from a single image, they depend on manually provided prior knowledge. ConceptExpress [15] proposed the task of unsupervised concept extraction, which aims to extract concepts from a single image without relying on manually provided prior knowledge, but it cannot distinguish between

object-level concepts and attribute-level concepts. Inspiration Tree [43] forces the model to learn different concepts by decomposing tokens, but it uses structured guidance. ICE [5] can structurally extract object-level concepts and attribute-level concepts from a single image, yet it fails to consider the associative relationships between object-level concepts and attribute-level concepts.

None of these methods take into account the compositionality and interpretability of the extracted concepts. This makes the process of decomposing a complex visual concept into multiple simple visual concepts irreversible, resulting in insufficient interpretability of these methods and thus weakening people's ability to control and trust the model. Although CCE [41] considers the compositionality of the concept embedding space, it still needs to extract concepts from multiple images containing the same concept. The HyperExpress method we propose not only extracts object-level concepts and attribute-level concepts in a structured manner and learns the associative relationships between concepts but also imposes compositional constraints on the embedding space, maximizing the realization of reversible decomposition from a complex visual concept to multiple simple visual concepts.

2.2. Hyperbolic learning

Euclidean space has been used for representation learning [49]. However, computer vision data often exhibits a highly non-Euclidean underlying geometric structure, in which case Euclidean embedding may not be the optimal choice [3, 13]. Hyperbolic space inherently has the ability to represent hierarchical structures with minimal distortion [4, 20, 21, 26], which makes it possible to learn hierarchical visual concept embeddings in hyperbolic space. Recent works [9, 17, 30] have demonstrated the potential of hyperbolic learning in learning cross-modal hierarchical embeddings in vision-language models, among which the study by Desai *et al*. [9] enforces the construction of entailment structures across modalities. Drawing inspiration from these methods, we establish hierarchical structures and entailment relationships between object-level concepts and attribute-level concepts, which can be used to realize the decomposition of complex visual concepts.

2.3. Compositionality in Concept Extraction

Existing studies [23, 44] have shown that compositionality can be used to control the behavior of generative models. Current concept extraction models focus on obtaining disentangled representations of concepts. "Disentanglement" focuses on how to distinguish different concepts, while "compositionality" focuses on the results produced when different concepts are combined with each other. Nevertheless, there is no inherent correlation between disentanglement and compositionality [46]. Stein *et al*. [41] discussed the compositionality of unsupervised concept extraction methods and proposed that the concepts extracted by current unsupervised concept extraction methods lack compositionality, as no constraints are imposed on the concept embedding space. Building on the work of CCE [41], we further consider emphasizing not only the disentanglement capability of concept extraction models but also their compositional generation capability.

3. Preliminaries

3.1. Diffusion Models

We use diffusion-based Text-to-Image (T2I) models [16, 29, 35, 36, 38, 40] for Compositional and Interpretable Intrinsic Concept Extraction (CI-ICE). A diffusion model gradually adds noise to data to convert the data into a random distribution and then reconstructs the original data from the noise through a reverse process. In the forward process, the diffusion model adds Gaussian noise to the original sample z_0 to obtain z_t:

$$q\left(z_t \mid z_0\right)=\mathcal{N}\left(z_t ; \sqrt{\bar{\alpha}_t} z_0,\left(1-\bar{\alpha}_t\right) \mathbf{I}\right), \tag{1}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, α_i represents the noise schedule. During the denoising process, the denoising network predicts the noise at the current time step based on the noisy image z_t at time step t. This process is achieved by minimizing the denoising loss:

$$\mathcal{L}_{\text{recon}}=\mathbb{E}_{\mathbf{x}_0, \epsilon, t}\left[\left\|c-c_\theta\left(\mathbf{x}_t, t, \mathcal{E}(\mathbf{p})\right)\right\|_2^2\right]. \tag{2}$$

3.2. Hyperbolic Geometry

To learn the hierarchical relationships and associative relationships between concepts in greater detail, our method operates in hyperbolic space. Hyperbolic space is a Riemannian manifold with constant negative curvature, which ensures that in hyperbolic space, the embedding positions of concepts with greater differences are farther apart, while concepts with smaller differences are closer to each other—this is difficult to achieve in Euclidean space. Specifically, we select the Poincaré ball [13] for concept learning. The Poincaré ball [13] $(\mathbb{B}^n, g^{\mathbb{B}})$ is defined by the manifold $\mathbb{B}^n = \{x \in \mathbb{R}^n : \|x\| < 1\}$ equipped with the Riemannian metric $g_x^{\mathbb{B}} = \lambda_x^2 g^E$. Similar to concept extraction models based on Euclidean space, we use hyperbolic distance to measure the similarity between two concepts, which is defined as:

$$d_{\mathbb{D}}(x, y)=\cosh^{-1}\left(1+2 \frac{\|x-y\|^2}{(1-\|x\|^2)(1-\|y\|^2)}\right). \tag{3}$$

In Euclidean space, we often use orthogonal relationships to represent the composability of two concepts, and two mutually orthogonal concepts can be combined. This

first requires defining the directionality of the embedding space. In hyperbolic space, ideal points are used to represent directions, and they refer to the points on the unit sphere $\mathbb{S}_{\infty}^{d-1} = \{\|x\| = 1\}$. A geodesic is analogous to a straight line in Euclidean space. We can define a submanifold $M \subset \mathbb{H}^d$ such that: for any $x, y \in M$, the geodesic connecting x and y is contained in M; such a submanifold is called a geodesic submanifold. Furthermore, given a set of points $S \in \mathbb{H}^d$, the smallest geodesic submanifold in the hyperbolic space $\mathbb{H}^d$ that contains S is called the geodesic hull of S, denoted as $GH(S)$. In the Poincaré ball model, the level set of the Busemann function is called the horosphere centered at p. The intrinsic curvature of a horosphere is zero; therefore, it also possesses many properties of planes in Euclidean space.

3.3. Concepts and Compositionality

In our research, a concept is defined as a feature or attribute in an image that is clearly distinguishable and understandable to humans [41], such as objects, colors, and the materials of objects in the image. We use a set of tokens $\mathcal{T}$ to refer to these concepts, along with their corresponding embedding vectors $\mathcal{V}$. Furthermore, we define items, characters, and animals *etc*. in the image as object-level concepts, and attributes such as color, shape, and size as attribute-level concepts. The compositionality of concepts is defined in Proposition 1.

Proposition 1. *For concept tokens $[V_i], [V_j] \in \mathcal{T}$, the concept representation $R : \mathcal{T} \to \mathcal{V}$ is considered compositional if there exist positive weights $w_i, w_i \in \mathbb{R}^+$ such that:*

$$R([V_i] \cup [V_j]) = w_i R([V_i]) + w_j R([V_j]). \tag{4}$$

4. Method

We aim to learn a set of tokens that can refer to object-level concepts and attribute-level concepts from a single image, and these concepts need to be compositional. Specifically, given an image $\mathcal{I}$ containing N objects where each object has M attributes, we leverage diffusion-based Text-to-Image (T2I) models [16, 29, 35, 36, 38, 40] to mine a set of concept tokens $\mathcal{T} = \{[V_i]\}_{i=1}^{(\mathrm{M}+1)\cdot\mathrm{N}}$ and their embedding vectors $\mathcal{V} = \{v_i\}_{i=1}^{(\mathrm{M}+1)\cdot\mathrm{N}}$. These tokens $\mathcal{T}$ and embedding vectors $\mathcal{V}$ can capture specific concepts in the image, such as the category, color, and material of objects within the image $\mathcal{I}$. Furthermore, we impose constraints on the embedding space to ensure it satisfies Proposition 1, thereby enabling the extraction of concepts with compositional interpretability. To this end, we propose the HyperExpress method, which consists of a **Concept Learning** approach and a **Concept-wise Optimization** approach. Figure 2 illustrates the overall framework of the HyperExpress method.

4.1. Concept Learning

The first challenge of the CI-ICE task is how to accurately disentangle complex concepts, and this stage aims to learn a set of concept tokens $\mathcal{T} = \{\mathcal{T}^{obj}, \mathcal{T}^{att}\}$ and their embedding vectors $\mathcal{V} = \{\mathcal{V}^{obj}, \mathcal{V}^{att}\}$, so as to disentangle the complex concepts in the image. For an image containing N objects, where each object has M attributes, we have $\mathcal{T}^{obj} = \{[V_k^{obj}]\}_{k=1}^{\mathrm{N}}$, $\mathcal{V}^{obj} = \{v_k^{obj}\}_{k=1}^{\mathrm{N}}$, and $\mathcal{T}^{att} = \{[V_k^{att}]\}_{k=1}^{\mathrm{N}\cdot\mathrm{M}}$, $\mathcal{V}^{att} = \{v_k^{att}\}_{k=1}^{\mathrm{N}\cdot\mathrm{M}}$. There are obvious hierarchical and associative relationships between object-level concepts and attribute-level concepts; methods based on Euclidean space cannot accurately capture such relationships. However, hyperbolic space inherently possesses strong hierarchical modeling capabilities, which is why we conduct the learning process in hyperbolic space. Since our focus is on disentangling complex concepts in the embedding space, we first adopt the method from the first stage of ICE [5] to locate the main objects in the image and their semantic categories, and finally obtain the masks $\mathcal{M} = \{M_i\}_{i=1}^{\mathrm{N}}$ of the main objects and their corresponding text descriptions $\mathcal{T}^{anchor} = \{[V_k^{anchor}]\}_{k=1}^{\mathrm{N}}$. Subsequently, the Hyperbolic Contrastive Learning (HCL) module is used to learn the hierarchical structure between concepts, and the Hyperbolic Entailment Learning (HEL) module is used to learn the entailment relationships between concepts.

Hyperbolic Contrastive Learning Module. The HCL module aims to distinguish between object-level concepts and attribute-level concepts by leveraging the hierarchical modeling capability of hyperbolic space. First, we define the hyperbolic text encoder. For the newly added tokens of each main object, we first encode them using the CLIP [34] model, and then map the text embeddings to the Poincaré ball [13, 31] through the exponential map [13, 31]. We add a weight W to the exponential map to learn the mapping relationship from the standard text encoder space to the tangent space, as follows:

$$\mathcal{E}_h(x) = \exp_0(W \cdot \mathcal{E}_s(x)), \tag{5}$$

where $\mathcal{E}_h(\cdot)$ is the hyperbolic text encoder and $\mathcal{E}_s(\cdot)$ is the CLIP [34] text encoder. Exponential map $\exp_0(\cdot)$ refers to the exponential operation in the Poincaré ball [13, 31], and its calculation is as follows:

$$\exp_0(\mathbf{x}) = \tanh\left(\frac{\|\mathbf{x}\|}{2}\right) \cdot \frac{\mathbf{x}}{\|\mathbf{x}\|}, \tag{6}$$

therefore, we have $\mathcal{V}^{obj} = \{\mathcal{E}_h([V_k^{obj}])\}_{k=1}^{\mathrm{N}}$ and $\mathcal{V}^{att} = \{\mathcal{E}_h([V_k^{att}])\}_{k=1}^{\mathrm{N}\cdot\mathrm{M}}$. Inspired by [5], we use the hyperbolic triplet loss to learn to distinguish between concepts.

The first step is to distinguish between object-level con-

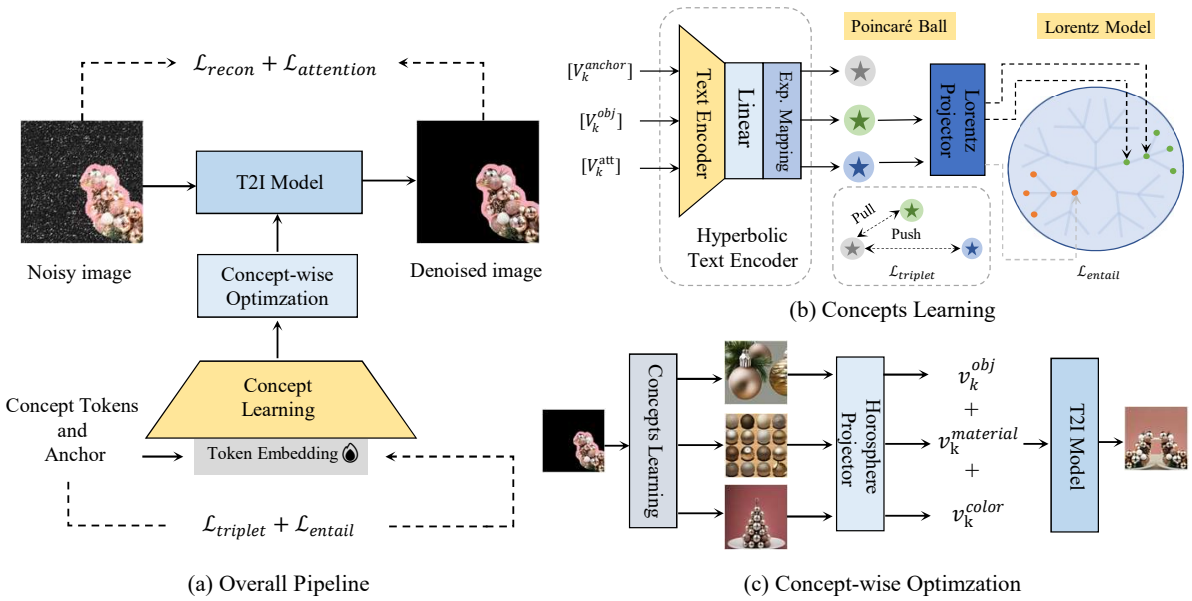

Figure 2. **The Proposed Method and Its Components**. **(a) The overall structure of HyperExpress**: It addresses the CI-ICE task from two aspects: concept learning and concept-wise optimization. **(b) Concept learning**: It leverages triplet loss $\mathcal{L}_{triplet}$ and hyperbolic entailment loss $\mathcal{L}_{entail}$ to learn the hierarchical structure and associative relationships between object-level concepts and attribute-level concepts. **(c) Concept-wise Optimization**: It uses Horosphere Projector (HP) to constrain the concept space, thereby ensuring the compositionality of concepts.

cepts and attribute-level concepts, as follows:

$$\mathcal{L}^{obj}_{\text{triplet,k}} = \max\left(0, d_{\mathbb{D}}(v_k^{\text{anchor}}, v_k^{\text{obj}}) - d_{\mathbb{D}}(v_k^{\text{anchor}}, v_k^{\text{att}}) + \gamma\right), \tag{7}$$

where $v_k^{\text{anchor}} = \mathcal{E}_h([V_k^{anchor}])$ and γ are the margin parameters. We use v_k^{anchor} to initialize v_k^{obj}.

Next is the distinction between different attribute-level concepts, as follows:

$$\mathcal{L}^{att}_{\text{triplet,k}} = \max\left(0, d_{\mathbb{D}}(v_k^{\text{intrinsic}}, v_k^{\text{att}}) - d_{\mathbb{D}}(v_k^{\text{intrinsic}}, v_j^{\text{att}}) + \gamma\right), \tag{8}$$

where $v_k^{\text{intrinsic}} = \mathcal{E}_h([V_k^{intrinsic}])$ and $[V_k^{intrinsic}]$ are set to $[V_k^{intrinsic}]$: "a $intrinsic_k$ concept" where $intrinsic_k$ represents a specific attribute such as color.

In contrast to ICE [5], we distinguish these concepts in hyperbolic space, which makes the concepts we have learned more hierarchical; however, the associative relationships between concepts are not considered. To address this issue, we propose the Hyperbolic Entailment Learning (HEL) module.

Hyperbolic Entailment Learning Module. Inspired by [12], we define the associative relationship between concepts as follows: if concept i entails concept j, then their tokens $[V_i], [V_j]$ and embeddings $v_i = \mathcal{E}_h([V_i]), v_j = \mathcal{E}_h([V_j])$ satisfy the following formula:

$$v_i \in \mathcal{S}_j \iff \theta(v_i, v_j) \leq \omega(v_i), \tag{9}$$

where $w(\cdot)$ denotes the entailment cone radius, $\theta(i,j)$ denotes the spatial angle, and $\mathcal{S}_i$ denotes the entailment cone of concept i. Since both w and θ have explicit formula definitions in the Lorentz model [31], and their calculation is simpler compared to the Poincaré model [13], we perform the computation in the Lorentz model [31]. That uses the following formula [31]:

$$x = (x_0, \ldots, x_n) \in \mathbb{B}^n \Leftrightarrow \left(\frac{1+\|x\|^2}{1-\|x\|^2}, \frac{2x_1}{1-\|x\|^2}, \ldots, \frac{2x_n}{1-\|x\|^2}\right) \in \mathbb{L}^n, \tag{10}$$

where, $\mathbb{B}^n$ and $\mathbb{L}^n$ are the n-dimensional Poincaré ball [13, 31] and Lorentz model [31] and $\|x\|^2 = x_0^2 + \cdots + x_n^2$.

In the Lorentz model [31], the definitions of the entailment cone radius $\omega(\cdot)$ are as follows [9, 24, 32]:

$$\omega(x) = \sin^{-1}\left(\frac{2K}{\sqrt{\kappa}\|x\|}\right), \tag{11}$$

and the spatial angle $\theta(x,y)$ is as follows:

$$\theta(x,y) = \cos^{-1}\left(\frac{x_0 + y_0 \cdot \kappa \cdot \langle x, y\rangle_{\mathcal{L}}}{\|\tilde{y}\| \cdot \sqrt{(\kappa \cdot \langle x, y\rangle_{\mathcal{L}})^2 - 1}}\right), \tag{12}$$

where κ denotes the curvature, and $\langle x, y\rangle_{\mathcal{L}}$ represents the Lorentz inner product. Next, based on Eq. (9), we set up an

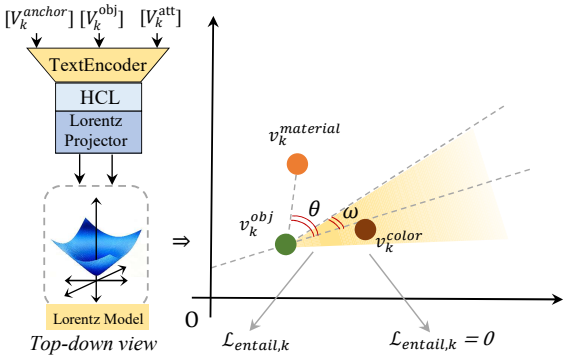

Figure 3. **Explanation of the proposed HEL module**. Unlike the HCL module, the hyperbolic entailment loss is computed within the Lorentz model. If the object-level concept (v_k^{obj}) and attribute-level concept (v_k^{color} and $v_k^{material}$) satisfy the condition in Eq. (9), the entailment loss will be 0; otherwise, the corresponding entailment loss will be calculated.

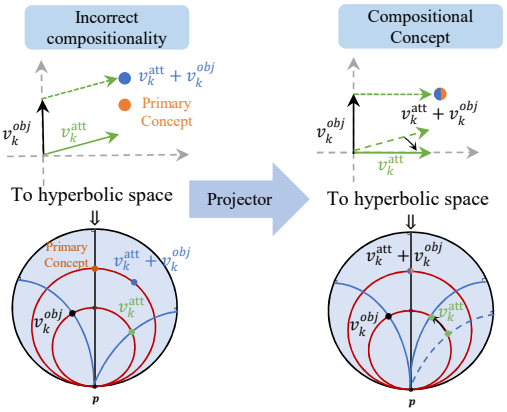

Figure 4. **Horosphere Projection Module and an illustration of concept compositionality**. If no constraints are applied to the concept embedding space, it may lead to incorrect composition of concepts. For instance, using $v_k^{att} + v_k^{obj}$ alone cannot reconstruct the primary concept. To address this, we reproject the embedding space to ensure composability between concepts.

entailment loss. The entailment loss between object-level concepts and attribute-level concepts is shown as follows:

$$\mathcal{L}_{\text{entail,k}} = \max\Big(0, \cos(\omega(v_k^{obj})) - \cos(\theta(v_k^{obj}, v_k^{att}))\Big), \tag{13}$$

where $\mathcal{L}_{\text{entail,k}}$ denotes the entailment loss of the k-th object concept in the image. As shown in Fig. 3, this means that the attributes of this object should fall within the entailment cone of the object. By constraining the embedding positions of object-level concepts and attribute-level concepts, we establish the associative relationship between them and achieve accurate disentanglement of complex concepts. Next, we impose further constraints on the concept embedding vectors to ensure their composability.

4.2. Concept-wise Optimization

Another core challenge faced by CI-ICE lies in the composability of the learned concept embedding vectors. Currently, mainstream unsupervised concept extraction methods [5, 15] have failed to address this issue effectively, and the fundamental reason is that these methods lack an explicit constraint mechanism on the concept embedding space. Although the CCE [41] method proposed by Stein *et al*. incorporates constraints, its modeling in Euclidean space makes it difficult for the model to effectively learn the hierarchical structures and associative relationships between concepts. In view of this, inspired by [6], this study proposes a concept composability Horosphere Projection (HP) module suitable for hyperbolic space. The core goal of HP is to not only maintain the complex intrinsic relationships between concepts in hyperbolic space but also to achieve the composability of concept embedding vectors, thereby solving the aforementioned key issue in a targeted manner.

Horosphere Projection Module. Given n ideal points $\{p_1, p_2, \ldots, p_n\}$ and one base point b, our goal is to learn a mapping from the concept embedding space to the composable submanifold.

$$\pi^{\mathbb{H}}_{b,p_1,\ldots,p_n} : v \mapsto M \cap S(p_1, v) \cap \cdots \cap S(p_n, v), \tag{14}$$

where $M := GH(b, p_1, \ldots, p_n)$ and $S(p_i, v)$ denote the horosphere centered at p_i and passing through the concept embedding point v.

We train our HP on the anchors of concepts. Since anchors have been proven to be composable in the embedding space [41], our goal is to enable the newly added anchors to possess the composability of the original semantic concepts. The training objective is to find n geodesic directions such that the variance of the data after horosphere projection is maximized, as shown in the following formula:

$$\begin{aligned} p_1 &= \underset{p\in\mathbb{S}^{d-1}_{\infty}}{\text{argmax}} \sigma^2_{\mathbb{H}}\left(\pi^{\mathbb{H}}_{b,p}(S)\right), \\ p_{k+1} &= \underset{p\in\mathbb{S}^{d-1}_{\infty}}{\text{argmax}} \sigma^2_{\mathbb{H}}\left(\pi^{\mathbb{H}}_{b,p_1,\ldots,p_n,p}(S)\right), \\ \sigma^2_{\mathbb{H}}(S) &= \frac{1}{n^2}\sum_{x,y\in S} d_{\mathbb{H}}(x,y)^2. \end{aligned} \tag{15}$$

Algorithm 1 in Appendix 7 illustrates the calculation process of the HP module.

As shown in Fig. 4, HP essentially uses anchors to find n geodesic directions and adjusts the newly added Token vectors to these n geodesic directions. The learned submanifold inherits Euclidean properties from the zero-curvature horosphere, allowing native operations like vector addition [6]. Therefore, after finding the n geodesic directions, we can complete the rotation operation using an orthogonal matrix Q. The HP module is designed based on the isometry of

horospherical projection, which endows it with two advantages. The isometry of horospherical projection is defined as Proposition 2.

Proposition 2. *For any* $x \in \mathbb{H}^d$, *if* $y \in \mathrm{GH}(x, , p_1, \ldots, p_n)$, *then:*

$$d_{\mathbb{H}}(\pi^{\mathbb{H}}_{b,p_1,\ldots,p_n}(x), \pi^{\mathbb{H}}_{b,,p_1,\ldots,p_n}(y)) = d_{\mathbb{H}}(x, y). \quad (16)$$

The one advantage is that the HP module does not disrupt the hierarchical structures and associative relationships between concepts. The complex relationships between concepts are mainly determined by their distances: the stronger the association between two concepts, the closer their distance should be. Since the HP module satisfies the Proposition 2, it does not alter the distance between two concepts before projection; thus, it does not affect concept disentanglement. We have proven the Proposition 2 in Appendix 1. Another advantage of the HP module is that, through its rotation operation, we project the newly added concepts into a composable concept space. This space can satisfy Proposition 1, thereby enabling composability. We have provided the corresponding proof process in Appendix 2.

4.3. Overall training objective

The method we proposed includes multiple loss terms, and the total loss is as follows:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_{\text{triplet}} \cdot \mathcal{L}_{\text{triplet}} + \lambda_{attention} \cdot \mathcal{L}_{attention} + \lambda_{\text{entail}} \cdot \mathcal{L}_{entail}. \quad (17)$$

Since the concept learning process involves learning the embeddings of object-level concepts and the embeddings of attribute-level concepts, we have two types of triplet losses:

$$\mathcal{L}^{\text{obj}}_{\text{triplet}} = \sum_k \mathcal{L}^{\text{obj}}_{\text{triplet,k}}, \quad \mathcal{L}^{\text{att}}_{\text{triplet}} = \sum_k \mathcal{L}^{\text{att}}_{\text{triplet,k}}. \quad (18)$$

Similarly, we also have entailment loss:

$$\mathcal{L}_{\text{entail}} = \sum_k \mathcal{L}_{\text{entail,k}}. \quad (19)$$

The attention loss $\mathcal{L}_{attention}$ adopts the Wasserstein loss [15], which can align the attention of the T2I model with the masked regions, thereby reducing the impact of object-irrelevant features:

$$\mathcal{L}_{\text{attention}} = \mathcal{W}(A_i, M_i), \quad (20)$$

where $\mathcal{W}$ is Wasserstein distance and A_i, M_i are the attention region and mask of the i-th concept, respectively.

5. Experiments

5.1. Baseline and Evaluation Metrics

We evaluated HyperExpress on the Unsupervised Concept Extraction Benchmark (UCEBench) [15] and the Intrinsic Concept Benchmark (ICBench) [5]. Specifically, on the UCEBench [15], we compared it with state-of-the-art UCE models [2, 5, 8, 15] using the identity similarity SIM^I, compositional similarity SIM^C, and Top-k accuracy ACC^k metrics. On the ICBench [5], we conducted comparisons using the SIM^{T-T} and SIM^{T-V} metrics. We conducted training and testing on the $D1$ dataset [5] by referring to the approach of ICE [5].

Among these metrics, SIM^I measures the reconstruction accuracy of individual concepts; SIM^Cevaluates the overall consistency of generated images based on extracted concepts; ACC^k assesses the concept disentanglement capability of models; SIM^{T-T} calculates the similarity between concepts described by GPT [1] and learned tokens; SIM^{T-V} computes the similarity between concepts described by GPT [1] and images generated using the learned concepts. We present more experimental details and the calculation methods of evaluation metrics in Appendix 5.

5.2. Quantitative results

Table 1. Performance comparison on UCEBench [15].

Method	SIM^I (%)	SIM^C (%)	ACC^1 (%)	ACC^3 (%)
Break-A-scene [2]	0.627	0.773	0.174	0.282
ConceptExpress [15]	0.689	0.784	0.263	0.385
AutoConcept [8]	0.690	0.770	0.350	0.520
ICE [5]	**0.738**	**0.822**	0.325	0.518
HyperExpress (Ours)	0.699	0.786	**0.504**	**0.736**

Table 2. Performance comparison on ICBench [5].

Method	$\text{SIM}^{T-T}_{object}$	$\text{SIM}^{T-T}_{material}$	SIM^{T-T}_{color}	$\text{SIM}^{T-V}_{object}$	$\text{SIM}^{T-V}_{material}$	SIM^{V-T}_{color}
ICE [5]	0.249	0.101	0.093	0.264	0.208	0.215
HyperExpress (Ours)	**0.280**	**0.115**	**0.098**	**0.305**	**0.211**	**0.222**

As shown in Tabs. 1 and 2, HyperExpress demonstrates competitive performance compared with existing UCE models [2, 5, 8, 15]. However, it is worth noting that ICE [5] sacrifices the interpretability of concept composition. The paths generated by its concept composition are difficult to understand, which is not conducive to people's control over the model. Although the HyperExpress model we proposed loses some performance, it gains easily interpretable concept composition paths and a composable concept space.

5.3. Qualitative results

The CI-ICE task requires extracting intrinsic concepts with compositionality from images. Among existing UCE methods, only ICE [5] can extract intrinsic concepts; therefore, we conduct a qualitative comparison between HyperExpress and ICE [5] to evaluate the models' concept extraction and combination capabilities, with the results shown in Fig. 5.

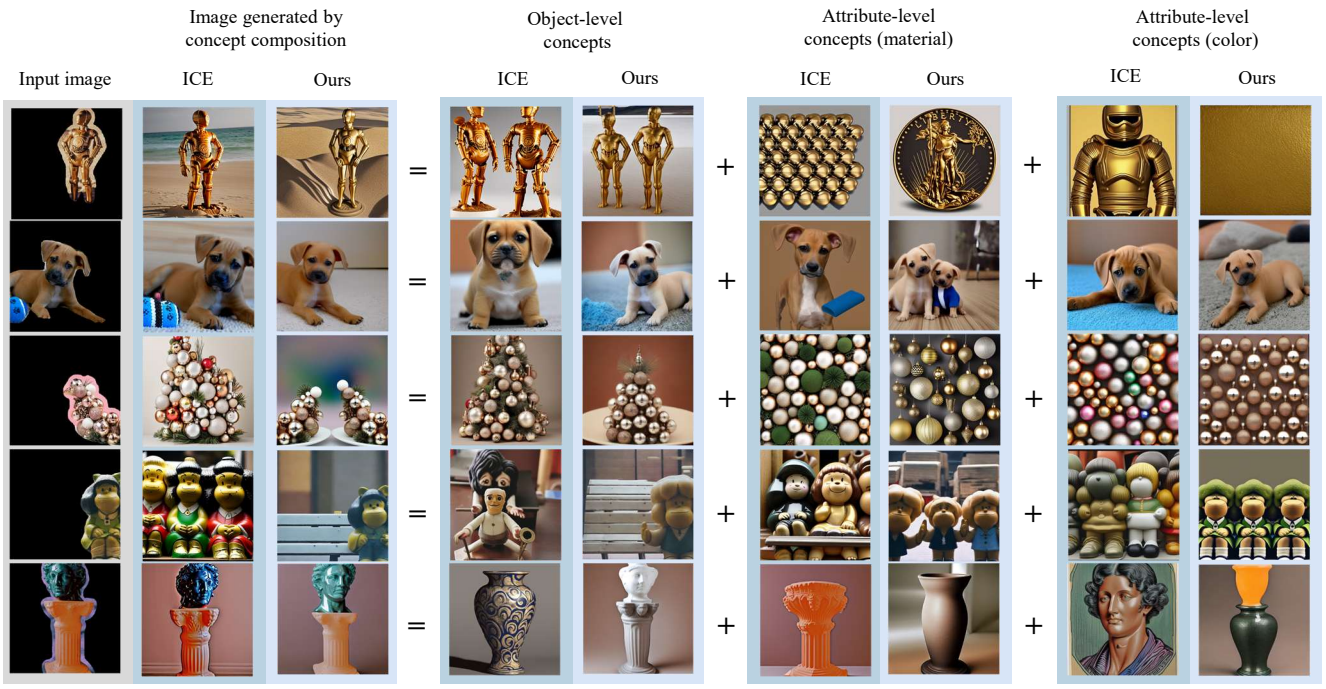

Figure 5. **Comparison of Qualitative Results Between HyperExpress and ICE [5].** This comparison includes two processes: one is extracting object-level concepts and further attribute-level concepts from a single image, and the other is performing compositional reconstruction using the extracted concepts.

In concept extraction, HyperExpress distinguishes itself from ICE [5] by learning associative relationships between concepts, leading to the extraction of specific, concrete object concepts from images. This specificity enhances the interpretability of the compositional process. Consequently, in compositional reconstruction, HyperExpress generates more interpretable pathways than ICE. For instance, as shown in Fig. 5, it logically combines the concepts of "robot," "metal," and "gold" to form the complex concept "a golden robot made of metal." This clarity stems from learning concept associations and applying constraints to the concept embedding space, whereas ICE [5]'s compositional results remain difficult to interpret. We also provide the user study in Appendix 6.

5.4. Model component analysis

To verify the effectiveness of HyperExpress, we conducted ablation experiments on the three modules of HyperExpress. As shown in Tab. 3, each module improves the model performance to a certain extent. Additionally, we present the ablation results of the margin parameter γ and the weights λ of each loss term in Appendix 3.

Table 3. Ablation study on HyperExpress model components on the $D1$ dataset [5].

HCL Module	HEL Module	HP Module	SIM^I (%)	SIM^C (%)	ACC^1 (%)	ACC^3 (%)
✔	✗	✗	0.625	0.769	0.326	0.509
✔	✔	✗	0.688	0.771	0.330	0.518
✔	✗	✔	0.621	0.765	0.348	0.522
✔	✔	✔	**0.699**	**0.786**	**0.504**	**0.736**

6. Conclusion

This paper proposes the task of Compositionally Interpretable Intrinsic Concept Extraction (CI-ICE), which aims to extract composable object-level concepts and attribute-level concepts from a single image using Text-to-Image (T2I) models. We design the HyperExpress method, which addresses the CI-ICE problem from two core aspects. In terms of concept disentanglement, we have devised a concept learning approach that can decompose complex visual concepts into object-level concepts and attribute-level concepts while preserving the hierarchical structure and associative relationships between concepts. Subsequently, regarding the compositionality of concepts, we have designed a concept optimization method to achieve concept compositionality by constraining the concept embedding space. Experimental results demonstrate that HyperExpress is an effective solution for the CI-ICE task.

Acknowledgements

This work was supported by Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515140010, Key Areas Research and Development Program of Guangzhou under Grant 2023B01J0029, the Guangdong Provincial Key Laboratory of Cyber-Physical System under Grant 2020B1212060069, the Science and Technology Development Fund, Macau SAR, under Grant 0079/2025/AFJ and 0193/2023/RIA3, and the University of Macau under Grant MYRG-GRG2024-00065-FST-UMDF.

References

[1] Hurst A, Lerer A, and et al. Goucher A P. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 7

[2] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia*, 2023. 1, 2, 7

[3] Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 2017. 3

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 3

[5] Fernando Julio Cendra and Kai Han. Ice: Intrinsic concept extraction from a single image via diffusion models. In *CVPR*, 2025. 1, 2, 3, 4, 5, 6, 7, 8

[6] Ines Chami, Albert Gu, Dat Nguyen, and Christopher Ré. Horopca: Hyperbolic dimensionality reduction via horospherical projections. In *ICML*, 2021. 6

[7] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W. Cohen. Subject-driven text-to-image generation via apprenticeship learning. In *NeurIPS*, 2023. 2

[8] Pranav Singh Chib, Kirtankumar Vijaykumar Patel, Mudit Gupta, Pise Ashutosh Kalidas, and Pravendra Singh. Autoconcept: Unsupervised extraction of constituent concepts from single image. In *ICCV Workshops*, 2025. 1, 2, 7

[9] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Ramakrishna Vedantam. Hyperbolic image-text representations. In *ICML*, 2023. 3, 5

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 2

[11] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)*, 2023. 2

[12] Octavian Ganea, Gary Becigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *ICML*, 2018. 5

[13] Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. In *NeurIPS*, 2018. 3, 4, 5

[14] Shaozhe Hao, Kai Han, Shihao Zhao, and Kwan-Yee K. Wong. Vico: Plug-and-play visual condition for personalized text-to-image generation. *arXiv preprint arXiv:2306.00971*, 2023. 2

[15] Shaozhe Hao, Kai Han, Zhengyao Lv, Shihao Zhao, and Kwan-Yee K. Wong. ConceptExpress: Harnessing diffusion models for single-image unsupervised concept extraction. In *ECCV*, 2024. 1, 2, 6, 7

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3, 4

[17] Sarah Ibrahimi, Mina Ghadimi Atigh, Pascal Mettes, and Marcel Worring. Intriguing properties of hyperbolic embeddings in vision-language models. *Transactions on Machine Learning Research*, 2024. 3

[18] Xuhui Jia, Yang Zhao, Kelvin C. K. Chan, Yandong Li, Han-Ying Zhang, Boqing Gong, Tingbo Hou, H. Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*, 2023. 2

[19] Chen Jin, Ryutaro Tanno, Amrutha Saseendran, Tom Diethe, and Philip Teare. An image is worth multiplewords: discovering object level concepts using multi-concept prompt learning. In *ICML*, 2024. 1, 2

[20] Sungnyun Kim, Gihun Lee, Sangmin Bae, and Seyoung Yun. Mixco: Mix-up contrastive learning for visual representation. *arXiv preprint arXiv:2010.06300*, 2020. 3

[21] Wonsik Kim, Bhavya Goyal, Kunal Chawla, Jungmin Lee, and Keunjoo Kwon. Attention-based ensemble for deep metric learning. In *ECCV*, 2018. 3

[22] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 2

[23] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *ICLR*, 2023. 3

[24] Matthew Le, Stephen Roller, Laetitia Papaxanthos, Douwe Kiela, and Maximilian Nickel. Inferring concept hierarchies from text corpora via hyperbolic embeddings. In *ACL*, 2019. 5

[25] Dongxu Li, Junnan Li, and Steven C. H. Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *NeurIPS*, 2023. 2

[26] Jiahong Liu, Menglin Yang, Min Zhou, Shanshan Feng, and Philippe Fournier-Viger. Enhancing hyperbolic graph embeddings via contrastive learning. *arXiv preprint arXiv:2201.08554*, 2022. 3

[27] Nan Liu, Yilun Du, Shuang Li, Joshua B. Tenenbaum, and Antonio Torralba. Unsupervised compositional concepts discovery with text-to-image generative models. In *ICCV*, 2023. 2

[28] Y. Ma, Huan Yang, Wenjing Wang, Jianlong Fu, and Jiaying Liu. Unified multi-modal latent diffusion for joint subject and text conditional image generation. *arXiv preprint arXiv:2303.09319*, 2023. 2

[29] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2021. 3, 4

[30] Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. In *ICLR*, 2025. 3

[31] Wei Peng, Tuomas Varanka, Abdelrahman Mostafa, Henglin Shi, and Guoying Zhao. Hyperbolic deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 4, 5

[32] Tobia Poppi, Tejaswi Kasarla, Pascal Mettes, Lorenzo Baraldi, and Rita Cucchiara. Hyperbolic safety-aware vision-language models. In *CVPR*, 2025. 5

[33] Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. In *NeurIPS*, 2023. 2

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4

[35] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3, 4

[36] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 4

[37] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 2

[38] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Lit, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 3, 4

[39] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *CVPR*, 2024. 2

[40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3, 4

[41] Adam Stein, Aaditya Naik, Yinjun Wu, Mayur Naik, and Eric Wong. Towards compositionality in concept learning. In *ICML*, 2024. 1, 2, 3, 4, 6

[42] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *SIGGRAPH Asia*, 2023. 2

[43] Yael Vinker, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. Concept decomposition for visual exploration and inspiration. *ACM Transactions on Graphics (TOG)*, 2023. 3

[44] Zihao Wang, Lin Gui, Jeffrey Negrea, and Victor Veitch. Concept algebra for (score-based) text-controlled generative models. In *NeurIPS*, 2023. 3

[45] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *ICCV*, 2023. 2

[46] Zhenlin Xu, Marc Niethammer, and Colin Raffel. Compositional generalization in unsupervised compositional representation learning: a study on disentanglement and emergent language. In *NeurIPS*, 2022. 3

[47] Guanting Ye, Qiyan Zhao, Wenhao Yu, Liangyu Yuan, Mingkai Li, Xiaofeng Zhang, Jianmin Ji, Yanyong Zhang, Qing Jiang, and Ka-Veng Yuen. Sope: Spherical coordinate-based positional embedding for enhancing spatial perception of 3d lvlms. *arXiv preprint arXiv:2602.22716*, 2026. 1

[48] Guanting Ye, Qiyan Zhao, Wenhao Yu, Xiaofeng Zhang, Jianmin Ji, Yanyong Zhang, and Ka-Veng Yuen. C^2rope: Causal continuous rotary positional encoding for 3d large multimodal-models reasoning. In *ICRA*, 2026. 1

[49] Yun Yue, Fangzhou Lin, Guanyi Mou, and Ziming Zhang. Understanding hyperbolic metric learning through hard negative sampling. In *WACV*, 2024. 3

[50] Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Xiaosong Yuan, Qiyan Zhao, Jiawei Cao, Feilong Tang, Sinan Fan, Yaomin Shen, Chen Shen, and Hao Tang. Hallucination begins where saliency drops. In *ICLR*, 2026. 1

[51] Yanbing Zhang, Mengping Yang, Qin Zhou, and Zhe Wang. Attention calibration for disentangled text-to-image personalization. In *CVPR*, 2024. 2