Intrinsic Concept Extraction Based on Compositional Interpretability

– Supplementary Material –

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*

1 Proof of Proposition 2

In this section, we prove that the horosphere projection (HP) module does not destroy the specific characteristics of the hierarchical structure and associative relationships between concepts.

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

$$d_{\mathbb{H}}(\pi^{\mathbb{H}}_{b,p_1,\ldots,p_n}(x), \pi^{\mathbb{H}}_{b,p_1,\ldots,p_n}(y)) = d_{\mathbb{H}}(x, y). \tag{1}$$

Proposition 2 demonstrates an advantage of the horosphere projection (HP) module: the distances between various concepts in the concept space remain unchanged before and after projection, and the distances between concepts truly reflect the relationships among them. This indicates that the HP will not destroy the hierarchical structure and associative relationships between concepts established by the concept learning method. Below, we present some preliminary knowledge for proving Proposition 2.

Lemma 1. Let $P = GH(p_1, p_2, \ldots, p_n)$, then for every $x \in \mathbb{H}^d$, the intersection of horosperes

$$S(x) = S(p_1, x) \cap S(p_2, x) \cap \cdots \cap S(p_n, x) \tag{2}$$

is precisely the orbit of x under the group G of rotations around P.

Chami et al. [3] provided a proof of this Lemma 1. Lemma 1 illustrates that the intersection of all $S(p_i, x)$is the rotational orbit of x around the $P = GH(p_1, \ldots, p_n)$; in other words, all points in the intersection can be obtained by rotating x around P.

Theorem 2. For any $x \in \mathbb{H}^d$, if $y \in GH(x, p)$, then:

$$d_{\mathbb{H}}(\pi^{\mathbb{H}}_{b,p}(x), \pi^{\mathbb{H}}_{b,p}(y)) = d_{\mathbb{H}}(x, y). \tag{3}$$

Proof. Since all geodesics with the ideal point p as an endpoint are orthogonal to all horospheres centered at p; moreover, for two horospheres sharing the same center, all orthogonal geodesic segments connecting them have equal lengths. The segment from point x to point y, and the segment from $\pi^H_{b,p}(x)$ to $\pi^H_{b,p}(y)$, are exactly two such orthogonal geodesic segments that connect the horospheres $S(p, x)$ and $S(p, y)$. Therefore, their distances are equal.

Corollary 1. If $x \in P = GH(p_1, p_2, \ldots, p_n)$, then $S(x) = x$. Otherwise, let $\pi^G_P = \pi^G_{p_1,p_2,\ldots,p_n}$ be the geodesic projection of x onto P, and $Q(x)$ be the geodesic submanifold that orthogonally complements P at π^G_P. Then $S(x) \subset Q(x)$ and is precisely the hypersphere in $Q(x)$ that is centred at π^G_P and passing through x.

Chami et al. [3] provided a proof of this Corollary 1. Corollary 1 shows that if $Q(x)$ is the orthogonal complement submanifold of P at $\pi^G_P(x)$, then $S(p_1, x) \cap \cdots \cap S(p_K, x)$ is a sphere in $Q(x)$ with $\pi^G_P(x)$ as the center and $d_{\mathbb{H}}(x, \pi^G_P(x))$ as the radius.

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

Next,we prove that Theorem 2 can be generalized to Proposition 2.

Proof. First, we construct an auxiliary submanifold $N = GH(x, y, p_1, p_2, \ldots, p_n)$. From $y \in GH(x, p_1, \ldots, p_n)$, we know that x and y belong to $GH(x, p_1, \ldots, p_n)$ together. Therefore, there exists a minimal totally geodesic submanifold $N = GH(x, y, p_1, \ldots, p_n)$that contains x,y, and all p_i. Since $P = GH(p_1, \ldots, p_n)$ is a $(n-1)$-dimensional submanifold and $x \notin P$, $GH(x, P)$ is a n-dimensional submanifold. Furthermore, as $y \in GH(x, P)$, we have $N = GH(x, P)$, which means $\dim(N) = n$.

Second, we analyze the intersection of N and M as well as their projection properties. From $M = GH(b, p_1, \ldots, p_n)$ and $N = GH(x, p_1, \ldots, p_n)$, their intersection is $P = GH(p_1, \ldots, p_n)$. Since the definition of Horosphere Projection (HP) is as follows:

$$\pi^{\mathbb{H}}_{b,p_1,\ldots,p_n} : x \mapsto M \cap S(p_1, x) \cap \cdots \cap S(p_n, x), \tag{4}$$

we can obtain $\pi^{\mathbb{H}}_{b,p_1,\ldots,p_n}(x) \in N \cap M$ and $\pi^{\mathbb{H}}_{b,p_1,\ldots,p_n}(y) \in N \cap M$. Therefore,$\pi^{\mathbb{H}}_{b,p_1,\ldots,p_n}(x)$ and $\pi^{\mathbb{H}}_{b,p_1,\ldots,p_n}(y)$ lie in the submanifold that is the orthogonal complement of $M \cap N = P$ within M. In other words, $\pi^{\mathbb{H}}_{b,p_1,\ldots,p_n}(x)$ and $\pi^{\mathbb{H}}_{b,p_1,\ldots,p_n}(y)$ belong to the same n-dimensional subspace of M, and both are contained in N.

Third, according to Corollary 1, the Horosphere Projection (HP) is equivalent to rotating x around P to M: let the rotation transformation be $R : \mathbb{H}^d \to \mathbb{H}^d$, then $\pi^{\mathbb{H}}_{b,p_1,\ldots,p_n}(x) = R(x)$, and R is an isometric transformation of the hyperbolic space. We need to verify that $R(y) = \pi^{\mathbb{H}}_{b,p_1,\ldots,p_n}(y)$. From Corollary 1, since $y \in N = GH(x, P)$, when rotating around P,y and x belong to the same orthogonal complement submanifold $Q(x)$ of P; therefore, $R(y)$ remains in $Q(x)$. From $B_{p_i}(y) = B_{p_i}(x)$, we know that $y \in S(p_i, x)$; therefore, $R(y) \in S(p_i, R(x)) = S(p_i, \pi^{\mathbb{H}}_{b,p_1,\ldots,p_n}(x))$. Furthermore,$R(y) \in M$, so $R(y)$ is the intersection of M and $S(p_i, \pi^{\mathbb{H}}_{b,p_1,\ldots,p_n}(x))$, which means $R(y) = \pi^{\mathbb{H}}_{b,p_1,\ldots,p_n}(y)$.

Finally, since R is an isometric transformation, for any x and y, we have: $d_{\mathbb{H}}(\pi^{\mathbb{H}}_{b,p_1,\ldots,p_n}(x), \pi^{\mathbb{H}}_{b,p_1,\ldots,p_n}(y)) = d_{\mathbb{H}}(x, y)$. End proof.

2 Prove the Compositionality of HyperExpress

Following Stein et al. [7], we define the compositionality of concepts as:

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:

$$R([V_i] \cup [V_j]) = w_i R([V_i]) + w_j R([V_i]). \tag{5}$$

HyperExpress extracts concepts using a diffusion-based text-to-image (T2I) model [6], and employs CLIP [5] for text processing. Studies by Stein et al. [7] have shown that the concepts extracted by CLIP [5] exhibit compositionality, specifically: let R denote the concepts extracted by the CLIP [5] model, then Formula 5 can be satisfied. Since our horosphere projection (HP) module learns mapping relationships from the concept anchors, and the composability of these concept anchors has been verified by Stein et al. [7], we first need to prove that operations in the hyperbolic space will not destroy the composability of concepts.

Let R' denotes the hyperbolic TextEncoder:

$$R'(\cdot) = exp_0^c(R(\cdot)) \tag{6}$$

where he exponential operation can be defined as:

$$\exp_0^c(x) = \tanh(\sqrt{c}\|v\|)\frac{x}{\sqrt{c}\|x\|}. \tag{7}$$

We let $f(x) = \frac{\tanh(\sqrt{c}\cdot x)}{\sqrt{c}\cdot x}$, $R([V_i]) = v_i$, $R([V_j]) = v_j$ and $R([V_i] \cup [V_j]) = u$, then we have $u = w_i \cdot v_i + w_j \cdot v_j$, quivalent to proving that:

$$f(||u||) \cdot u = w_i' \cdot f(||v_i||) \cdot v_i + w_j' \cdot f(||v_j||) \cdot v_j. \tag{8}$$

Let $\alpha = w_i' \cdot f(||v_i||)$, $\beta = w_j' \cdot f(||v_j||)$, this is equivalent to:

$$f(||u||) \cdot u = \alpha \cdot v_i + \beta \cdot v_j, \tag{9}$$

because $u = w_i \cdot v_i + w_j \cdot v_j$, and v_i is linearly independent of v_j, this is equivalent to:

$$f(||u||) \cdot (w_i \cdot v_i + w_j \cdot v_j) = \alpha \cdot v_i + \beta \cdot v_j, \tag{10}$$

to compare the coefficients, it is only necessary to set $w_i' = f(||u||) \cdot \frac{w_i'}{f(||v_i||)}$ and $w_j' = f(||u||) \cdot \frac{w_j'}{f(||v_j||)}$. Because the range of values of $f(||\cdot||)$ is greater than 0 and $w_i, w_j \in \mathbb{R}^+$, we have $w_i', w_j' \in \mathbb{R}^+$. Therefore, we have:

$$R'([V_i] \cup [V_j]) = w_i' R'([V_i]) + w_j' R'([V_i]). \tag{11}$$

This indicates that operations in the hyperbolic space do not destroy the composability of anchors. Our HP module, in turn, learns a mapping to such a composable space of anchors. After the HP module determines n geodesic directions, it can use an orthogonal matrix Q to map the concept embedding vector v to the composable space. This also shows that we can use the anchor concepts to map back to the concept embedding vector v through an inverse transformation. Since $Q^T Q = I$, the concepts extracted by HyperExpress also satisfy Formula 11, i.e., they possess the composability of concepts.

3 More Ablation Results

This section will elaborate on the relevant details of the margin parameter γ used in the triplet loss and the weight parameter λ_{entail} for the implicit loss term in the total loss.

First, there is the margin parameter γ, which is used to distinguish between object-level concepts, data concepts, and various attribute-level concepts. We set a larger parameter $\gamma = 1$ and a smaller parameter $\gamma = 10^{-6}$. The results show that when a larger parameter $\gamma = 1$ is used, the attribute-level concepts learned by the model become detached from the object-level concepts, disrupting the associative relationships between concepts. When a smaller parameter $\gamma = 10^{-6}$ is used, it becomes difficult to learn the attributes corresponding to the objects. When an appropriately chosen parameter $\gamma = 0.001$ is used, the model can distinguish between different concepts while maintaining their relationships.

Next is the weight of the implicit loss term. We also set a larger parameter $\lambda_{entail} = 10^{-2}$ and a smaller parameter $\lambda_{entail} = 10^{-8}$. The results show that when the smaller parameter $\lambda_{entail} = 10^{-8}$ is used, the model fails to learn the relationship between object-level concepts and attribute-level concepts, and the extracted attribute-level concepts are not sufficiently consistent with the original concepts. When a larger parameter $\lambda_{entail} = 10^{-2}$ is used, it affects the reconstruction performance and generation quality. When an appropriate parameter $\lambda_{entail} = 6 \times 10^{-5}$ is used, the model can not only learn the relationship between object-level concepts and attribute-level concepts but also achieve good reconstruction performance and generation quality.

4 More Qualitative Results

This section provides extensive experimental results illustrating HyperExpress's performance in intrinsic concept extraction and its application to compositional generation. Figure 4 presents additional qualitative results of the HyperExpress method to demonstrate its effectiveness in extracting composable visual intrinsic concepts. This method possesses the following capabilities: (1) Identifying object-level concepts and their corresponding attribute-level concepts from visual images. (2) Enabling the composability of object-level concepts and attribute-level concepts, which allows for the accurate restoration of original complex visual concepts through compositional generation.

5 Implementation Details

We adopt the Stable Diffusion v2.1 model [6] to implement the proposed framework. We fine-tuned the denoising network and text encoder: specifically, 600 training steps are conducted for object-level concept learning, and 600 training steps for attribute-level concept learning. After that, 400 steps of concept-wise optimization are performed to improve the quality of generated images. All experiments are carried out on a single NVIDIA A800 GPU. Next, we introduce the calculation methods of each metric. Hao et al. [4] proposed the calculation methods for SIM^I, SIM^C, and ACC^k. In this paper, we use CLIP [5] to calculate the similarities of SIM^I and SIM^C. For the calculation of

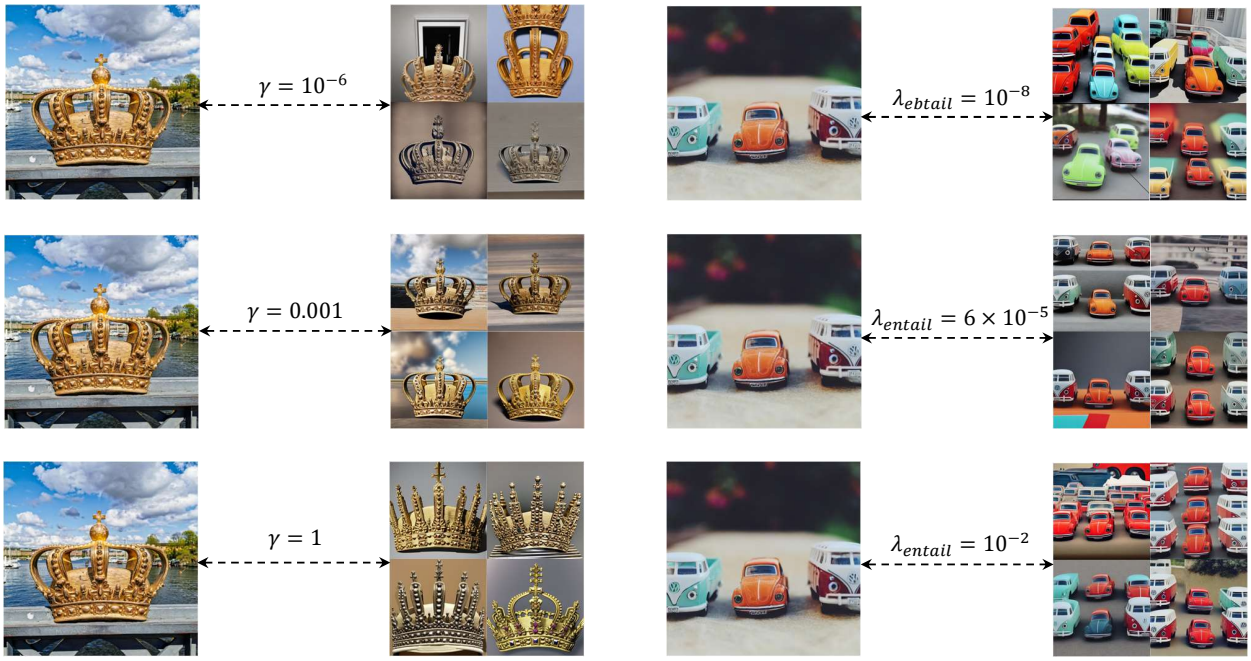

Figure 1: More ablation results.

Table 1: Ablation study on triplet loss weight $\lambda_{triplet}$.

$\lambda_{triplet}$	SIM^I (%)	SIM^C (%)	ACC^1 (%)	ACC^3 (%)
$1e-4$	0.433	0.521	0.145	0.299
$1e-2$	0.557	0.641	0.213	0.407
1.0 (default)	**0.625**	**0.769**	**0.348**	**0.509**

compositional similarity SIM^C, we adopt the prompt text "a photo of $[V_1][V_1] \ldots [V_n]$". Cendra et al. [2] put forward the calculation methods for SIM^{T-T} and SIM^{T-V}, and we extended the $D1$ dataset [2] following their approach. Specifically, we used GPT [1] to describe the objects, attributes, and colors of images with masks. These descriptions are only used for calculating the SIM^{T-T} and SIM^{T-V} metrics, not for model training. Examples of the extended part of the dataset are shown in Figure 3.

6 User Study

To ensure that the concepts extracted by the model and the images generated using these concepts align with human preferences, we conducted a user study comparing HyperExpress and ICE [2]. We invited 12 volunteers to evaluate the two methods. Each volunteer was presented with generated images corresponding to 5 concepts from the DI [2] dataset. For each concept, we displayed 8 images generated through concept compositionality, 8 images generated at the object-level concept, and 16 images generated at the attribute-level concept. These images were produced by both HyperExpress and ICE [2]. Additionally, we provided the original concept images overlaid with masks. The volunteers were then asked to indicate which method generated images that more closely resembled the original concepts. Finally, we received a total of 240 pieces of voting data reflecting human preferences. Among all the votes, 39.58% votes were in favor of ICE [2], and 60.42% votes supported the HyperExpress method proposed in this paper. The detailed statistical results of the voting for each concept are shown in Figure 4. This user study further demonstrates that in terms of concept extraction and compositional generation, the results of HyperExpress align better with human preferences than ICE [2].

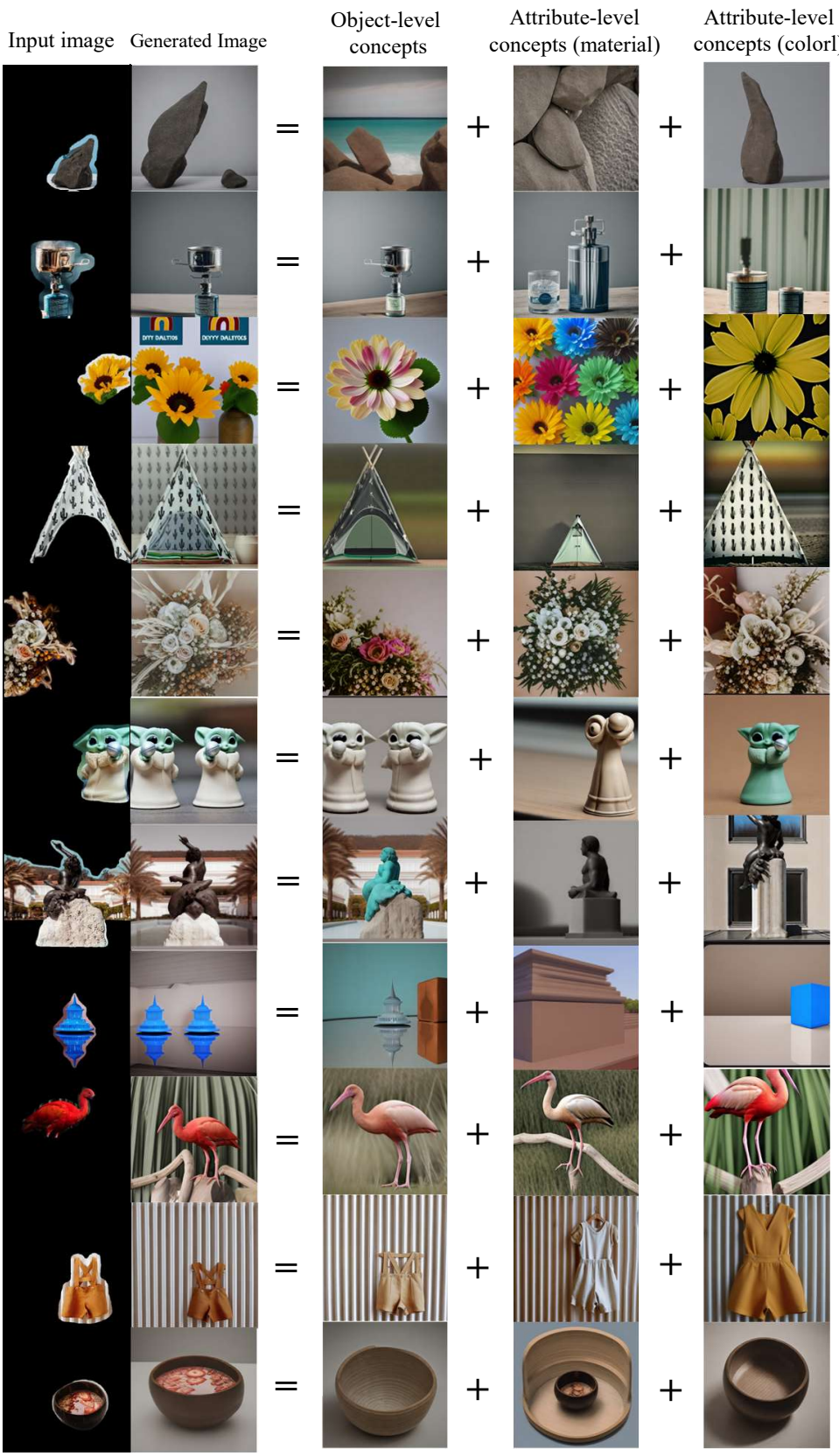

Figure 2: Additional qualitative results of the HyperExpress.

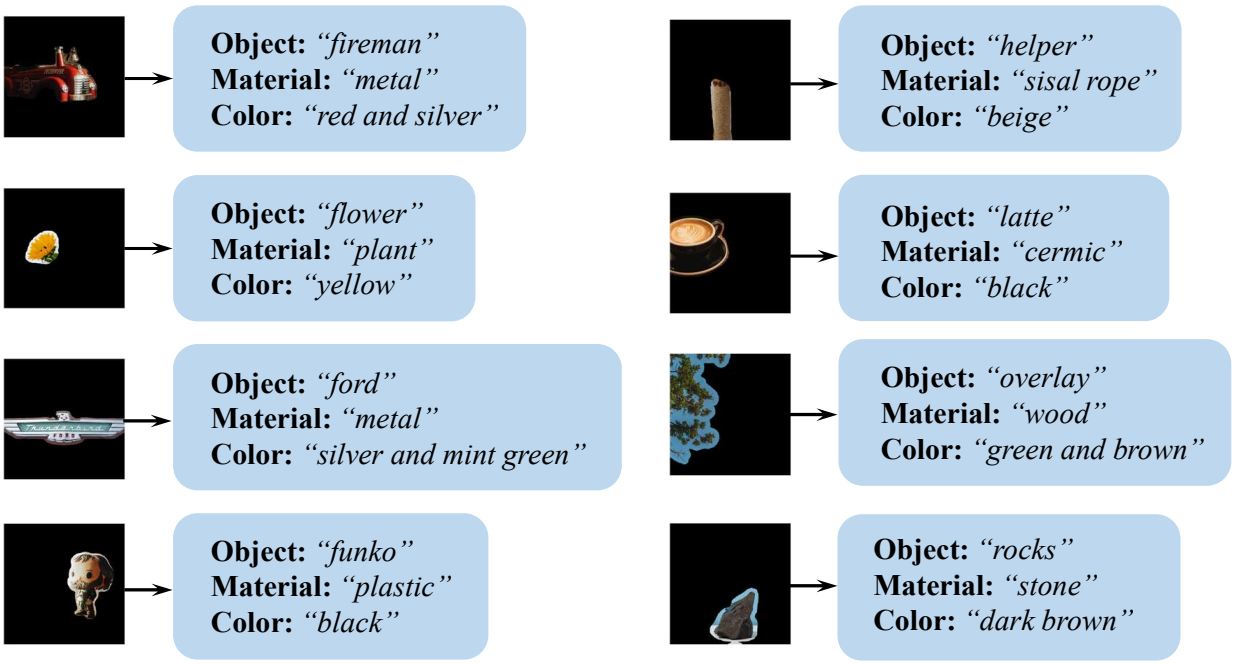

Figure 3: For the extended $D1$ dataset [2], GPT [1] is used to describe the object (**Object**), material (**Material**), and color (**Color**).

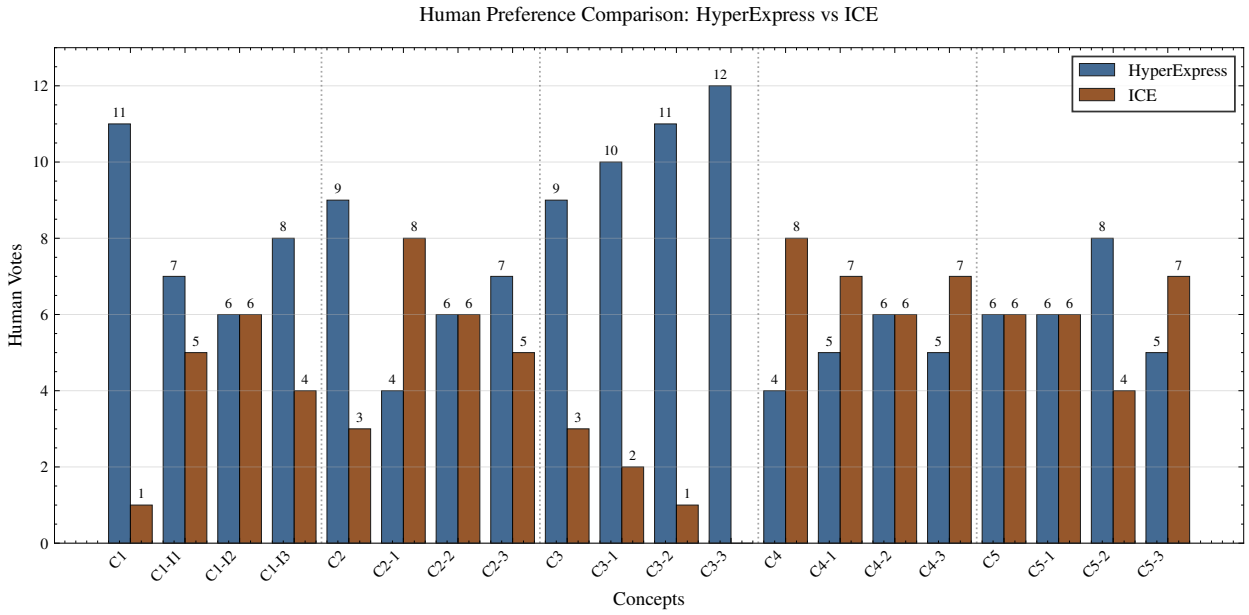

Figure 4: **User study statistics**. This figure presents a comparative analysis of human voting results between the ICE method and HyperExpress method for 15 intrinsic concepts extracted from 5 original concepts and 5 compositional concepts. On the horizontal axis, "Cn" denotes the n-th compositional concept, while "Cn-Im" represents the m-th intrinsic concept (object, material, color) of the n-th concept. The vertical axis displays the number of human votes received by each method, with different colors corresponding to the two methods respectively.

Table 2: Ablation study on the number of ideal points n in the Horizon Projection module.

n	SIM^I (%)	SIM^C (%)	ACC^1 (%)	ACC^3 (%)
50	0.691	0.775	0.382	0.583
100	0.695	0.780	0.438	0.651
150 (default)	**0.699**	**0.786**	**0.504**	**0.736**

7 Algorithm of Horosphere Projection

We present the calculation process of Horosphere Projection in Algorithm 1.

Algorithm 1 Horosphere Projection

Input: $x \in \mathbb{H}^d$, anchor $b \in \mathbb{H}^d$, ideal points $\{p_1, \ldots, p_n\}$
Output: $y = \pi^{\mathbb{H}}_{b,p_1,\ldots,p_n}(x)$
$P \leftarrow \text{GH}(p_1, \ldots, p_n), \quad M \leftarrow \text{GH}(b, p_1, \ldots, p_n)$
$y \leftarrow \pi^{\text{G}}_{P}(x)$
Find geodesic $\alpha \subset M$ with $\alpha \perp P$ at y
Find $y_1, y_2 \in \alpha$ with $d_{\text{H}}(y_i, y) = d_{\text{H}}(x, y)$
if $d_{\text{H}}(y_1, b) < d_{\text{H}}(y_2, b)$ **then**
    $y \leftarrow y_1$
else
    $y \leftarrow y_2$
end if
return y

8 More Ablation Study.

We supplemented ablation experiments for the weight of the hyperbolic triplet loss (Table 1) and the number of ideal points n in the Horosphere projector module (Table 2). The hyperbolic text encoder differs from conventional ones by an additional projection onto hyperbolic space, an indispensable component that cannot be replaced for ablation verification. The Lorentz projector, designed to compute the entailment cone radius in hyperbolic entailment loss, has a rigorous mathematical formulation in Lorentz space—such calculation is intractable in other geometric spaces. Moreover, this module is parameter-free and thus not suitable for parameter-based ablation analysis.

References

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

[2] Fernando Julio Cendra and Kai Han. Ice: Intrinsic concept extraction from a single image via diffusion models. In *CVPR*, 2025.

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

[4] 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.

[5] 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.

[6] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2021.

[7] Adam Stein, Aaditya Naik, Yinjun Wu, Mayur Naik, and Eric Wong. Towards compositionality in concept learning. In *ICML*, 2024.