HIERARCHICAL CROSS-MODAL REPRESENTATION ALIGNMENT FOR TEXT-TO-IMAGE DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern text-to-image diffusion models inherit rich compositional semantics from pretrained text encoders, yet their denoising objectives provide limited direct supervision for binding phrases to distinct image regions. Building on the success of representation alignment with external models, we harness the generator's own language representations as a source of hierarchical supervision. We introduce Hierarchical Intermodal Projection (HIP), a lightweight representation alignment method for fine-tuning text-to-image diffusion models. HIP promotes fine-grained cross-modal alignment, which is stabilized by anchoring local representations to their global semantic context. To preserve the distinctiveness of concepts sharing the same anchor, we introduce a shared hyperbolic space whose geometry supports local separation within a coherent scene neighborhood. On SD3.5-M and FLUX.1-dev, HIP improves GenEval macro accuracy by 14.48% and 8.24% over the respective baselines, and achieves the highest scores across all six categories in T2I-CompBench.

1 INTRODUCTION

Large-scale text-to-image generative models increasingly combine rectified-flow training with transformers that integrate pretrained language representations through joint text–image attention, enabling high-quality generation with improved prompt compliance (Saharia et al., 2022; Esser et al., 2024). However, compositional generation remains challenging: models often fail to faithfully translate the combinations of entities, attributes, and relations specified in a prompt into a coherent visual scene (Ghosh et al., 2023; Huang et al., 2023). One approach to improving semantic fidelity is to supplement the generative objective with explicit representation supervision. REPA (Yu et al., 2025) aligns denoiser features with representations from external self-supervised visual encoders, such as DINOv2 (Oquab et al., 2024), improving training efficiency and image quality. REPA-E (Leng et al., 2025) extends this principle to joint optimization of the visual tokenizer and diffusion transformer. For text-to-image generation, SoftREPA (Lee et al., 2026) improves semantic consistency by learning soft text tokens through contrastive training on matched and mismatched image–text pairs.

However, the semantic capabilities of the pretrained language models already embedded in large-scale text-to-image diffusion models remain underexploited as a source of supervision. For instance, the T5 language model serves as a text encoder in Stable Diffusion 3.5, converting prompts into contextual token embeddings for conditioning (Esser et al., 2024). Standard flow matching optimizes velocity prediction without an explicit objective for aligning visual and language representations. This motivates using the generator's existing language representations to directly supervise visual representation learning, encouraging semantic consistency and compositional correspondence.

Moreover, complex prompts and their paired training images contain hierarchical information that can serve as training signal. A prompt describes a complete scene through constituent noun phrases, while its paired image depicts that scene through corresponding entities and regions (Fig. 1(a)). For example, "a red cube beside a blue sphere" contains two object descriptions within a full scene description, mirrored by the two corresponding regions within the image. This intra-modal hierarchy enriches cross-modal representation alignment with finer-grained phrase–region correspondences. Furthermore, the whole–part relations can anchor these local representations to their shared scene context, supporting stable semantic alignment at the constituent level.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

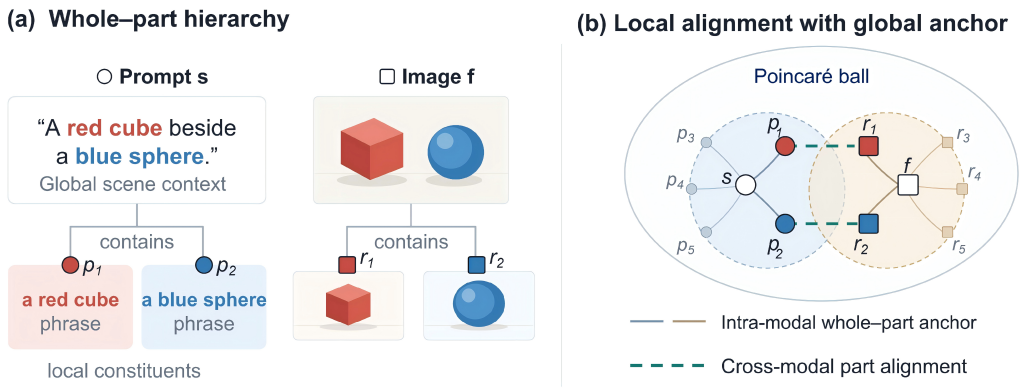

Figure 1: **Whole–part hierarchy and HIP alignment.** (a) Symbols label the representations associated with the depicted content: s for the full prompt, f for the full image, and p_i, r_i for a phrase and its corresponding region. The prompt text and image crops illustrate the content represented by these nodes. (b) HIP places these representations in a shared Poincaré ball, aligning matched phrase–region pairs (dashed links) while anchoring each constituent to its modality's global scene representation (solid links). The red cube and blue sphere illustrate distinct concepts with shared scene context and separate local correspondences.

We formulate this idea as *hierarchical cross-modal alignment with intra-modal whole–part anchoring*. Realizing this formulation presents two challenges. First, phrase–region correspondences have to be inferred without explicit annotations linking phrases to image regions. Second, keeping local representations close to their scene anchors can reduce their separation, creating a tension between contextual coherence and the semantic discrimination needed for phrase–region matching.

To address these challenges, we propose Hierarchical Intermodal Projection (**HIP**), a lightweight representation alignment method for diffusion fine-tuning. HIP leverages the generator's own joint attention to associate phrases with visual regions, yielding alignment supervision without external teachers or region-level training labels. To support contextual anchoring while preserving local distinctions, HIP embeds both modalities in a shared Poincar'e ball, where negative curvature supports greater constituent separation without requiring larger distances to the scene anchor (Nickel & Kiela, 2017; Desai et al., 2023; Pal et al., 2025), as illustrated in Fig. 1(b). The alignment loss introduced by HIP requires no extra training stage or additional computational cost at inference. Empirically, HIP improves compositional generation on both SD3.5-M and FLUX.1-dev, as measured by its leading performance on GenEval (Ghosh et al., 2023) and T2I-CompBench (Huang et al., 2023).

2 Related Work

We summarize the three threads our work intersects; an extended treatment with full citation coverage appears in App. A.

Diffusion transformers and compositional control. MMDiT (Esser et al., 2024), the rectified-flow (Liu et al., 2023; Lipman et al., 2023) transformer underpinning SD3/3.5 and FLUX (Black Forest Labs, 2024), descends from the DiT (Peebles & Xie, 2023) and PixArt (Chen et al., 2024b) line. Compositional failures of these backbones (Ghosh et al., 2023; Huang et al., 2023) are typically addressed at inference time by reshaping cross-modal attention (Feng et al., 2023; Chefer et al., 2023; Rassin et al., 2023) or at training time by reward fine-tuning and concept-matching auxiliaries (Jiang et al., 2024; Xu et al., 2023; Wallace et al., 2024).

Representation alignment for diffusion. Representation alignment supplements the generative objective with semantic training signals. REPA (Yu et al., 2025) aligns intermediate denoiser states with pretrained visual features such as DINOv2 (Oquab et al., 2024); REPA-E (Leng et al., 2025) extends alignment to joint tokenizer–diffusion training. SoftREPA (Lee et al., 2026) instead learns soft text tokens through contrastive score matching on positive and negative image–text pairs.

Hyperbolic vision–language learning. Hyperbolic geometry provides compact representations of hierarchical data (Sarkar, 2012; Nickel & Kiela, 2017; Ganea et al., 2018b; Sala et al., 2018).

MERU (Desai et al., 2023) and HyCoCLIP (Pal et al., 2025) combine hyperbolic distance and entailment objectives to learn general–specific relations in vision–language pretraining.

3 Alignment with Hierarchical Intermodal Projection

In this section, we present HIP, a representation alignment method for fine-tuning text-to-image diffusion models, which treats hierarchical organization within modalities and semantic correspondence across modalities as a joint geometric learning problem. We first review the necessary preliminaries in Section 3.1, then construct the whole–part hierarchies and their phrase–region correspondences in Section 3.2. We then present the HIP loss in Section 3.3, coupling cross-modal alignment with intra-modal whole–part anchoring to reconcile local semantic discrimination with global scene coherence. Finally, Section 3.4 provides theoretical analysis of HIP.

3.1 Preliminaries

Flow-based generation. Flow models learn a continuous transformation between a simple noise distribution and the data distribution (Lipman et al., 2023). We consider text-conditioned rectified flow in the latent space of a pretrained variational autoencoder (VAE) (Liu et al., 2023; Esser et al., 2024). Given a clean image latent x_0 and its caption y, training samples a time $t \in [0, 1]$ and independent Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ to form

$$x_t = (1 - t)x_0 + t\epsilon, \qquad \frac{dx_t}{dt} = \epsilon - x_0. \tag{1}$$

Here, $t = 0$ corresponds to data and $t = 1$ to noise. A neural network v_θ predicts the velocity along this path by minimizing

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{(x_0,y),t,\epsilon}\big[\|v_\theta(x_t, t, y) - (\epsilon - x_0)\|_2^2\big]. \tag{2}$$

The regression target yields an estimate of the conditional velocity $v(x, t, y) = \mathbb{E}[\epsilon - x_0 \mid x_t = x, y]$, without numerical integration during training.

At inference, a sample $\hat{x}_1 \sim \mathcal{N}(0, I)$ evolves from $t = 1$ to $t = 0$ according to the learned probability-flow ordinary differential equation (ODE),

$$\mathrm{d}\hat{x}_t = v_\theta(\hat{x}_t, t, y)\,\mathrm{d}t. \tag{3}$$

This deterministic formulation also connects to stochastic diffusion sampling (Ma et al., 2024; Yu et al., 2025). Let $p_t(x \mid y)$ be the conditional density induced by Eq. (1). With the exact velocity v, the same marginals can be generated by the reverse-time stochastic differential equation (SDE)

$$\mathrm{d}\hat{x}_t = \left[v(\hat{x}_t, t, y) - \frac{g(t)^2}{2}\nabla_x \log p_t(\hat{x}_t \mid y)\right]\mathrm{d}t + g(t)\,\mathrm{d}\bar{W}_t, \tag{4}$$

where $g(t) \geq 0$ controls the stochasticity and $\bar{W}_t$ is a reverse-time Wiener process. The score $\nabla_x \log p_t(x \mid y)$ is the gradient of the log density; for the linear path, it equals $-[x + (1 - t)v(x, t, y)]/t$ for $t > 0$, so it can be estimated from v_θ without a separate score network.

Multimodal diffusion transformer. The multimodal diffusion transformer (MMDiT) implements v_θ by jointly processing the caption and noisy image latent (Esser et al., 2024), as illustrated in Fig. 2. Text encoders produce a sequence of caption embeddings, while the latent is divided into image tokens. The two streams have separate transformation weights and interact through joint attention over all text and image tokens. This allows each image token to incorporate information from the caption and each text token to respond to the visual context. Time and pooled text embeddings condition the transformer blocks; the final image features are then mapped to the predicted velocity. FLUX extends this design with additional blocks that process the two modalities as a single sequence (Black Forest Labs, 2024).

HIP uses the text embeddings entering the transformer, $H^T = (h_a^T)_{a=1}^{n_T} \in \mathbb{R}^{n_T \times d_T}$, and the final image features, $H_t^I = (h_j^I)_{j=1}^{n_I} \in \mathbb{R}^{n_I \times d_I}$. Here n_T, n_I count text and image tokens, d_T, d_I are their feature dimensions, and a, j index the respective tokens. The text features supply language semantics, while the image features are the visual representations we aim to improve. We suppress the dependence of h_j^I and subsequent visual quantities on (x_t, t, y). Joint text–image attention associates the text tokens with spatial positions, providing the weights for phrase-conditioned visual pooling in Section 3.2.

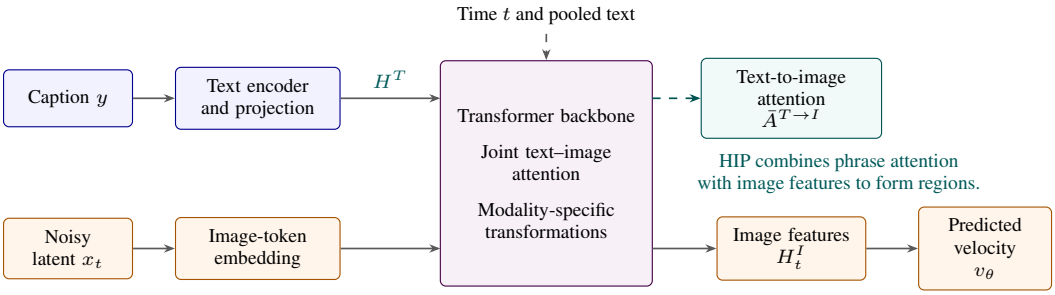

Figure 2: **MMDiT backbone and the features used by HIP.** Joint text–image attention supports velocity prediction v_θ. HIP combines text embeddings H^T, image features H_t^I, and attention weights $\bar{A}^{T\to I}$ to associate phrases with regions. Detailed structure is shown in Appendix C.1.

3.2 Whole–Part Hierarchies and Phrase–Region Correspondence

Whole–part structure. Compositional failures occur at the level of constituents: an attribute leaks to the wrong object, or an object is missing from the scene. We therefore represent each prompt–image pair at two levels. For a pair with K valid phrases, we reserve s, f, p_i, r_i for its final representations in the shared alignment space: s represents the whole caption, f the whole image, and p_i, r_i the ith phrase and its associated visual region. The input caption and image latent remain y and x_0; H^T, H_t^I denote their backbone feature sequences. Each modality forms a two-level whole–part hierarchy: s is the root, or *scene anchor*, for the phrase representations, and f anchors the region representations. The pairs (p_i, r_i) connect corresponding constituents across modalities. Below, we specify the representations of the constituents and their correspondence.

Attention-induced region construction. Supervising constituents requires knowing which image region each phrase describes, which is not usually provided in the training data. Joint attention offers an approach: during denoising, each text token already distributes attention over image tokens, and prior inference-time methods treat these attention maps as proxies for object layout (Feng et al., 2023; Rassin et al., 2023).

For each noun phrase extracted from the caption, let S_i contain its text-token indices, so that $\{h_a^T : a \in S_i\}$ supplies its textual features. Let $\bar{A}_{aj}^{T\to I}$ denote the attention from text token a to image token j, averaged across the selected heads and layers. We normalize the image-token subset of joint attention to obtain a spatial distribution: for phrase i, define

$$\widetilde{w}_{ij} = \frac{1}{|S_i|}\sum_{a\in S_i} \bar{A}_{aj}^{T\to I}, \qquad w_{ij} = \frac{\widetilde{w}_{ij}}{\sum_{j'=1}^{n_I}\widetilde{w}_{ij'}}. \tag{5}$$

For nonzero attention mass, $w_{ij} \geq 0$ and $\sum_j w_{ij} = 1$, so each phrase defines a normalized spatial distribution over image tokens. These soft weights give the local visual feature $\bar{h}_i^I$ for phrase i. We also define the global visual feature $\bar{h}^I$ by uniform pooling over the image tokens:

$$\bar{h}^I = \frac{1}{n_I}\sum_{j=1}^{n_I} h_j^I, \qquad \bar{h}_i^I = \sum_{j=1}^{n_I} w_{ij} h_j^I. \tag{6}$$

This construction associates each phrase's token span with a visual region in the current noisy image, without region-level training labels. Specifically, the phrase representation p_i is obtained by mapping each text-token embedding h_a^T, $a \in S_i$, to a shared alignment space and averaging the mapped embeddings. Its corresponding region representation r_i is obtained by mapping the pooled visual feature $\bar{h}_i^I$ to the same space. Section 3.3 specifies these mappings and gives the resulting representations in Eq. (8). Phrase extraction and attention-layer choices are detailed in App. G.

3.3 Hyperbolic Alignment with Whole–Part Anchoring

Given phrase–region correspondences, a popular approach for supervision is through contrastive learning: each phrase should be closer to its corresponding region than to any other visual representation, including the other regions of the same image. This can discriminate constituents and reduce

attribute leakage and object mixing (Bica et al., 2024; Assouel et al., 2025). However, cross-modal discriminability alone does not ensure that the learned representations retain meaningful structure within each modality (Jiang et al., 2023; Swetha et al., 2023). In our setting, a purely local objective leaves each constituent's relation to its own sentence or image context underconstrained. We therefore introduce *whole–part anchoring* as an intra-modal regularizer, using sentence and image representations as shared semantic anchors for their respective constituents. By penalizing excessive local–global distance, this regularizer keeps local cross-modal alignment grounded in each modality's global context.

The two objectives place complementary demands on the same local representations: intra-modal anchoring ties constituents to their global context, while cross-modal alignment can cause them to drift apart. The geometric challenge is therefore to support constituent separation without increasing their distance from the shared anchor. We adopt a shared hyperbolic space (Nickel & Kiela, 2017; Sala et al., 2018), where negative curvature can support greater constituent separation at fixed anchor distances. This provides a geometric basis for jointly learning cross-modal correspondence and intra-modal contextual coherence (Fig. 3). See Section 3.4 for formal discussion.

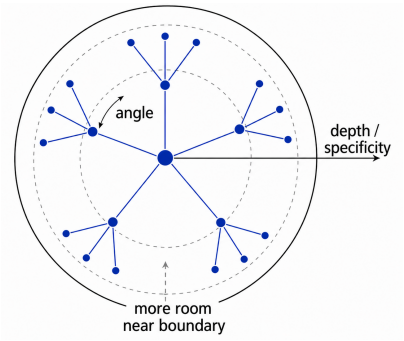

Figure 3: At fixed dimension and minimum separation, the capacity for constituents grows exponentially with the anchor radius in hyperbolic space, versus polynomially in Euclidean space.

Shared hyperbolic representations. We now map the features from Section 3.2 into a common space for phrase–region matching and whole–part anchoring. We use the Poincaré ball $\mathbb{B}_c^d = \{z \in \mathbb{R}^d : c\|z\|_2^2 < 1\}$, a coordinate model of d-dimensional hyperbolic space with curvature $-c$, where $c > 0$ controls the curvature magnitude. Its coordinates lie inside a Euclidean ball of radius $1/\sqrt{c}$, while distances are measured with the hyperbolic metric below. The common dimension d allows text and image representations to be compared even when their backbone dimensions d_T, d_I differ.

For each modality $M \in \{T, I\}$, a trainable multilayer perceptron $g_M : \mathbb{R}^{d_M} \to \mathbb{R}^d$ first produces a tangent vector u at the origin. The tangent space $T_0\mathbb{B}_c^d \simeq \mathbb{R}^d$ provides Euclidean coordinates in which this ordinary neural network operates. The exponential map $\exp_0$ then converts its output into a point in the ball (Ganea et al., 2018a). We denote the complete feature-to-ball mapping by Φ_M:

$$\begin{aligned} \Phi_M(h) &= \exp_0\big(g_M(h)\big), \quad M \in \{T, I\}, \\ \exp_0(u) &= \tanh(\sqrt{c}\|u\|_2)\frac{u}{\sqrt{c}\|u\|_2}, \end{aligned} \tag{7}$$

with $\exp_0(0) = 0$. Here h is an input feature vector and $\|\cdot\|_2$ is the Euclidean norm. The origin is the common reference for the mapping; each example has its own learned scene anchors s and f.

Let $z_a^T = \Phi_T(h_a^T)$ be the projected text-token representation and Ω_T the set of valid caption-token indices, with $S_i \subseteq \Omega_T$ for each phrase. Reusing the pooled visual features $\bar{h}^I, \bar{h}_i^I$ from Eq. (6), the four types of representations are

$$\begin{aligned} s &= \frac{1}{|\Omega_T|}\sum_{a\in\Omega_T} z_a^T, & p_i &= \frac{1}{|S_i|}\sum_{a\in S_i} z_a^T, \\ f &= \Phi_I(\bar{h}^I), & r_i &= \Phi_I(\bar{h}_i^I). \end{aligned} \tag{8}$$

Thus $s, f \in \mathbb{B}_c^d$ summarize the two modalities at scene level, while $p_i, r_i \in \mathbb{B}_c^d$ represent their corresponding constituents. Text features are projected before averaging, and visual features are pooled before projection. The text averages are arithmetic means of ball coordinates; they remain inside $\mathbb{B}_c^d$ because this coordinate domain is convex.

We compare all four types of representations using the same geodesic distance, the length of a shortest path under the hyperbolic metric. For two points $z, z' \in \mathbb{B}_c^d$,

$$d_{\mathbb{B}_c^d}(z, z') = \frac{1}{\sqrt{c}}\operatorname{arcosh}\left(1 + \frac{2c\|z - z'\|_2^2}{(1 - c\|z\|_2^2)(1 - c\|z'\|_2^2)}\right). \tag{9}$$

Here arcosh is the inverse hyperbolic cosine. This metric supplies both the phrase–region distances and the distances from constituents to scene anchors. Its equivalent logarithmic-map expression is given in App. B.

Cross-modal alignment. For a batch of image–caption pairs, we collect all valid phrase–region pairs as $\{(p_\nu, r_\nu)\}_{\nu=1}^{N}$, where N counts pairs across all images in the batch and ν, μ index these pairs. Each phrase p_ν is matched to its attention-induced region r_ν and compared against all N region candidates. Writing $D_{\nu\mu} = d_{\mathbb{B}_c^d}(p_\nu, r_\mu)$, the text-to-image InfoNCE loss (Oord et al., 2018) for phrase p_ν is

$$\ell_\nu = -\log \frac{\exp(-D_{\nu\nu}/\tau)}{\sum_{\mu=1}^{N} \exp(-D_{\nu\mu}/\tau)}, \tag{10}$$

where $\tau > 0$ is the temperature. The matched region is the positive candidate; all other regions serve as instance-level negatives, including those associated with other phrases from the same image. The contrastive objective $\mathcal{L}_{pr}$ averages these per-phrase losses; the exact training reduction and candidate-pool construction are specified in App. G.

Whole–part anchoring for local alignment. The contrastive objective $\mathcal{L}_{pr}$ constrains phrases and regions only through cross-modal comparisons. Complementing it, we instantiate the whole–part anchoring introduced above as two intra-modal penalties on each constituent's distance from its scene anchor. To avoid pulling every constituent all the way to its anchor and erasing the separation learned above, we penalize distances only beyond a margin m. Concretely, for a local batch of B examples padded to $K_{\max}$ phrase slots, with b indexing examples and i indexing phrase slots, the intra-modal anchoring loss terms are

$$\begin{aligned} \mathcal{L}_T &= \frac{1}{BK_{\max}} \sum_{b=1}^{B} \sum_{i=1}^{K_{\max}} [d_{\mathbb{B}_c^d}(s_b, p_{bi}) - m]_+, \\ \mathcal{L}_I &= \frac{1}{BK_{\max}} \sum_{b=1}^{B} \sum_{i=1}^{K_{\max}} [d_{\mathbb{B}_c^d}(f_b, r_{bi}) - m]_+, \end{aligned} \tag{11}$$

where $[x]_+ = \max(x, 0)$ and $m \geq 0$. The margin sets how tightly constituents are anchored to their scene. Constituents inside the margin receive no anchoring gradient, so the contrastive loss alone determines their relative positions. If the penalties for phrase representations p_i, p_j with sentence anchor s are both zero, the triangle inequality gives $d_{\mathbb{B}_c^d}(p_i, p_j) \leq d_{\mathbb{B}_c^d}(p_i, s) + d_{\mathbb{B}_c^d}(s, p_j) \leq 2m$. Anchoring therefore bounds how far siblings may drift apart without forcing them together. Whether a radius-m neighborhood leaves enough room for them to remain well separated depends on the geometry, which we analyze in Section 3.4.

HIP fine-tuning and inference. Together, these components constitute HIP as a representation alignment method for diffusion fine-tuning, which couples phrase–region alignment with intra-modal whole–part anchoring in a shared hyperbolic space. Specifically, HIP works by jointly optimizing the projection heads and backbone LoRA adapters with the combined objective:

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_{pr}\mathcal{L}_{pr} + \lambda_T\mathcal{L}_T + \lambda_I\mathcal{L}_I. \tag{12}$$

Here $\lambda_{pr}, \lambda_T, \lambda_I \geq 0$ weight the auxiliary losses. The projection heads and region construction provide supervision only during training, while the learned adaptation is retained in the backbone through the LoRA updates. At inference, we discard the projection heads and merge the LoRA updates into the backbone weights, introducing no additional parameters or computation relative to the original backbone. Optimization details are given in App. G.

3.4 Geometry of Scene Anchoring and Constituent Discrimination

We analyze how hyperbolic geometry supports constituent discrimination within a shared scene context, and how whole–part links transmit local cross-modal alignment to scene level. Proofs and extensions appear in App. B.

Separation within the anchoring margin. Consider two constituents at the same geodesic distance $0 < \rho \leq m$ from their scene anchor, with angle $\varphi \in [0, \pi]$ between the two geodesics. Writing

$D_{\mathbb{H}}$ and $D_{\mathbb{E}}$ for their hyperbolic and Euclidean separation at this radius and angle, Proposition 1 in App. B gives

$$D_{\mathbb{H}}(\rho, \varphi) \geq D_{\mathbb{E}}(\rho, \varphi) \tag{13}$$

with strict inequality for $0 < \varphi < \pi$ when curvature $-c < 0$. Thus negative curvature allows greater separation while both constituents remain within the anchoring margin, supporting local discrimination under the same scene-context constraint. This applies to both phrase and region representations.

Capacity for multiple constituents. For fixed dimension $d \geq 2$ and curvature $-c < 0$, hyperbolic packing capacity grows exponentially with the anchor radius R, whereas Euclidean capacity grows only polynomially (See Proposition 2 in App. B). This capacity growth supports larger sets of constituents around a shared scene anchor without reducing their minimum pairwise separation.

Implicit global cross-modal alignment. The following observation indicates that HIP provides a global cross-modal alignment without explicit supervision.

Observation 1 (Implicit sentence–image alignment)**.** *For the $K \geq 1$ valid phrase–region pairs of one example and margin $m \geq 0$, the triangle inequality gives*

$$d_{\mathbb{B}_c^d}(s, f) \leq 2m + \frac{1}{K} \sum_{i=1}^{K} \Big([d_{\mathbb{B}_c^d}(s, p_i) - m]_+ + [d_{\mathbb{B}_c^d}(f, r_i) - m]_+ + d_{\mathbb{B}_c^d}(p_i, r_i) \Big), \tag{14}$$

The left hand side of (14) is bounded by the auxiliary loss of HIP.

4 Experiments

In this section, we evaluate compositional generation across two diffusion backbones and examine how modules of HIP contribute to representation geometry.

4.1 Experimental setup

We fine-tune SD3.5-M and FLUX.1-dev on Flickr30k-Entities image–caption pairs (Plummer et al., 2015) using LoRA (Hu et al., 2022), with frozen VAE and text encoders and no phrase-box supervision. We compare against pretrained sampling, denoising-only **vanilla fine-tuning**, and **SoftREPA** (Lee et al., 2026). **Hierarchy + cosine** retains the whole–part anchoring terms and uses cosine similarity in the correspondence InfoNCE loss. REPA uses negative cosine similarity for visual-feature alignment (Yu et al., 2025), motivating this metric control; our formulation retains phrase–region contrastive learning and hierarchical supervision. GenEval (Ghosh et al., 2023) measures object, attribute, count, and position correctness; the official T2I-CompBench evaluator (Huang et al., 2023) assesses attribute binding and compositional relations. Image quality is evaluated on COCO-10k. Generation settings are fixed across methods within each backbone. The FLUX methods use the same effective epoch budget; on SD3.5-M, HIP uses one epoch and the evaluated SoftREPA checkpoint uses approximately 3.78 epochs. A single SD3.5-M HIP checkpoint is used across generation and representation evaluations. Training and evaluation details are in Apps. G and K.

4.2 Compositional generation

HIP achieves the highest GenEval macro accuracy and leads all six T2I-CompBench categories on both backbones (Table 1). The improvements span attribute binding and spatial organization, supporting HIP's ability to distinguish local concepts within a coherent scene. On SD3.5-M, where single-object accuracy is already saturated, the detailed GenEval results show particularly strong benefits for counting and position (App. E). On FLUX, HIP improves on the pretrained model even when denoising-only adaptation reduces compositional accuracy. This contrast highlights the value of hierarchical supervision for effective adaptation. HIP also outperforms SoftREPA in aggregate composition under the matched FLUX epoch budget and with fewer training epochs on SD3.5-M.

Figure 4 illustrates these gains: HIP correctly binds the requested colors to the cat and sheep and realizes multiple part-level shape constraints in the lamp. The extended qualitative comparison and additional model variants appear in App. D.

Table 1: **Compositional generation across backbones.** GenEval macro accuracy weights its six tasks equally; T2I-CompBench uses the official evaluator. Higher is better. Bold marks the best unrounded score within each backbone; a dash denotes an unreported score. Full task breakdowns, additional controls, and relative gains appear in App. E.

	GenEval	T2I-CompBench					
Method	Macro avg.	Color	Shape	Texture	2D spatial	Counting	Complex
SD3.5-M							
Pretrained baseline	0.64	0.80	0.57	0.77	0.31	0.60	0.37
Vanilla fine-tuning	0.67	0.82	0.57	0.76	0.31	0.59	0.38
Hierarchy + cosine	0.67	0.83	0.59	0.76	0.31	0.59	0.38
SoftREPA	0.70	0.82	0.64	0.78	0.36	0.60	0.38
HIP	**0.73**	**0.89**	**0.67**	**0.83**	**0.38**	**0.62**	**0.41**
FLUX.1-dev							
Pretrained baseline	0.64	0.75	0.50	0.64	0.27	0.60	0.37
Vanilla fine-tuning	0.57	0.71	0.49	0.62	0.23	0.53	0.36
Hierarchy + cosine	0.64	0.79	0.53	0.69	0.28	0.57	0.38
SoftREPA	0.67	0.81	0.57	0.73	0.29	0.59	0.39
HIP	**0.69**	**0.84**	**0.60**	**0.74**	**0.35**	**0.62**	**0.41**

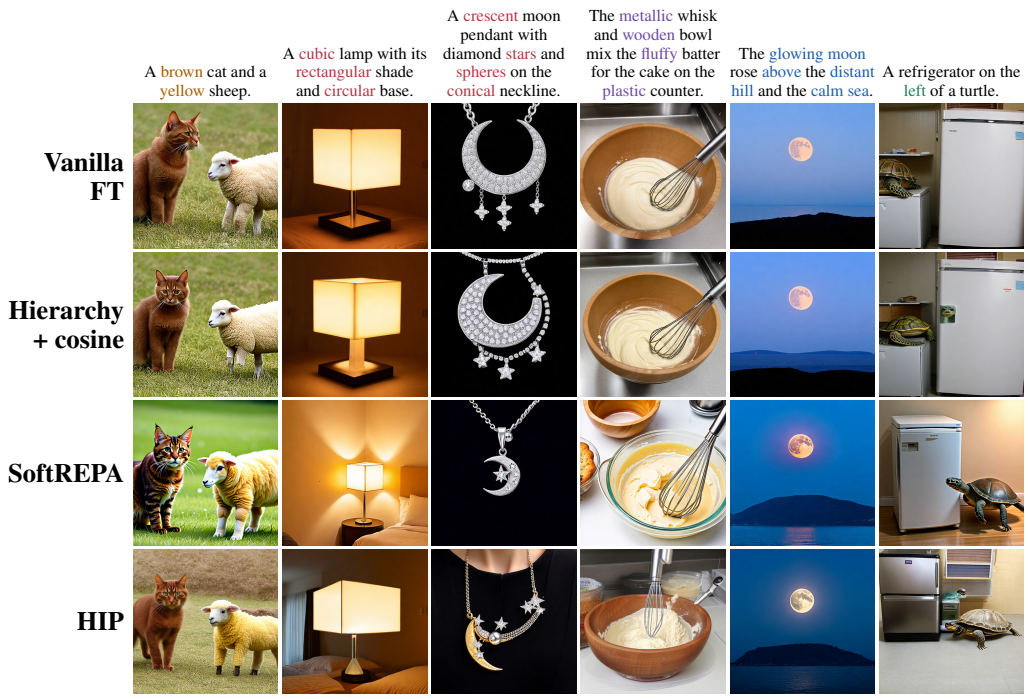

Figure 4: **Qualitative comparisons.** Rows show vanilla fine-tuning, hierarchy + cosine, SoftREPA and HIP. Each column fixes the prompt, seed, resolution, sampler, and guidance across methods. Highlighted phrases identify the color, shape, texture, complex, and spatial requirements.

4.3 Ablations and representation analysis

Complementary supervision. The component study in App. F identifies phrase–region correspondence as the strongest individual auxiliary, with the full objective performing best. This supports combining local discrimination with regularization of constituent distances to shared scene anchors. In the controlled FLUX comparison, replacing hyperbolic correspondence with cosine

Table 2: **SD3.5-M phrase–region correspondence on 128 unseen COCO-val pairs (414 phrases).**

Method	Nearest-neg. gap ↑	Top-1 ↑	Gromov gap ↑
Vanilla fine-tuning	−0.02	0.32	−0.01
Hierarchy + cosine	−0.04	0.39	0.01
Euclidean-L2	0.16	0.89	0.04
HIP	**0.18**	**0.93**	**0.09**

while fixing the whole–part anchoring terms and training configuration also reduces performance, supporting the geometric design.

Diagnostic measures. We test internal correspondence by retrieving each phrase's attention-pooled region among regions from the same image. **Top-1** measures correct retrieval frequency. The **nearest-negative gap** subtracts the matched distance from the closest incorrect distance. The **Gromov gap** subtracts the strongest incorrect affinity from the matched affinity, using the Gromov product $G_{ij} = [d_{\text{eval}}(0, p_i) + d_{\text{eval}}(0, r_j) - d_{\text{eval}}(p_i, r_j)]/2$. Here d_{eval} is the evaluation metric in both correspondence and anchoring losses. .

Discriminative cross-modal representations. HIP achieves the highest Top-1 accuracy, including over Euclidean-L2 (Table 2); retrieval rankings support comparison across geometries. Its positive Gromov gap, with a confidence interval above zero, further supports preferential origin-referenced affinity for matched concepts. Spatial attention analysis provides complementary evidence (App. I): HIP concentrates more attention within annotated regions and reduces overlap between phrases. Together, these diagnostics support accurate constituent correspondence in embedding space and distinct phrase localization within the image.

4.4 Image quality and efficiency

On 10,000 COCO-captioned images generated with SD3.5-M, HIP leads in text–image similarity (CLIP cosine), aesthetic appeal (Aesthetic Score), and predicted human preference (ImageReward). Fréchet Inception Distance (FID; lower is better) measures discrepancy from the reference image distribution and remains close to the pretrained model (See App. J). Thus HIP maintains image quality while improving compositional capabilities.

In our experiment, HIP forms contrastive comparisons from the features of matched image–caption pairs, with one denoiser evaluation per pair. By contrast, the SoftREPA configuration expands a batch of four pairs into 16 image–text conditions that each require a denoiser evaluation. In SD3.5-M runs, HIP processed 6.23 original image–caption pairs per second versus 2.07 for SoftREPA; HIP completed one epoch in 1.42 hours, while SoftREPA required 16.11 hours for approximately 3.78 epochs (App. K.1). At inference, HIP's LoRA weights can be merged into the backbone and its auxiliary heads discarded, leaving *zero additional parameters*, while SoftREPA retains soft tokens that add extra computation.

5 Conclusion

We introduced Hierarchical Intermodal Projection (HIP), a lightweight representation alignment method for compositional text-to-image generation. HIP uses the generator's own language representations and joint attention to align phrases with image regions, while anchoring these local representations to their global scene context. A shared hyperbolic space supports distinct local concepts within a coherent scene. Experiments on SD3.5-M and FLUX.1-dev show consistent improvements on GenEval and T2I-CompBench, without external teachers, region-level training annotations, or additional inference cost. These results highlight the value of the hierarchical structure already present in paired text and images as supervision for more faithful compositional generation.

Limitations. Our evaluation focuses on mmDiT based text-to-image generation. Extending HIP to other more diverse architectures such as unified multimodal models would be of interest, which we leave to future work.

AI USE STATEMENT

Generative AI tools were used to audit and migrate the LaTeX source code to the official ICLR 2027 template, recover and cross-check experimental records, and revise the experimental tables and text. The authors have reviewed all content in the paper before submission.

REPRODUCIBILITY STATEMENT

To support reproducibility, Sec. 3 defines the HIP architecture and training objectives, and Sec. 4.1 describes the datasets, baselines, and evaluation benchmarks. Appendix G documents image preprocessing, phrase extraction, feature and attention readouts, loss reductions, and training hyperparameters for both backbones. Appendix K specifies the evaluated checkpoints, training budgets, generation settings, and scoring procedures. The representation diagnostic and image-quality evaluation protocols are detailed in Apps. H and J, respectively. Assumptions and proofs supporting the theoretical analysis are provided in App. B.

REFERENCES

Rim Assouel, Pietro Astolfi, Florian Bordes, Michal Drozdzal, and Adriana Romero-Soriano. Object-centric binding in contrastive language-image pretraining. In *Advances in Neural Information Processing Systems*, pp. 145626–145666, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/be6047b64b8e38bc8af2d2c6aec9c4ae-Abstract-Conference.html.

Mina Ghadimi Atigh, Julian Schoep, Erman Acar, Nanne Van Noord, and Pascal Mettes. Hyperbolic image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4443–4452. IEEE, 2022.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A ViT backbone for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22669–22679, 2023.

Ioana Bica, Anastasija Ilic, Matthias Bauer, Goker Erdogan, Matko Bošnjak, Christos Kaplanis, Alexey A. Gritsenko, Matthias Minderer, Charles Blundell, Razvan Pascanu, and Jovana Mitrovic. Improving fine-grained understanding in image-text pre-training. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3974–3995. PMLR, 2024. URL https://proceedings.mlr.press/v235/bica24a.html.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, 2024.

Black Forest Labs. FLUX. https://github.com/black-forest-labs/flux, 2024.

Martin R. Bridson and André Haefliger. *Metric Spaces of Non-Positive Curvature*, volume 319 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 1999. doi: 10.1007/978-3-662-12494-9. URL https://doi.org/10.1007/978-3-662-12494-9.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. In *Advances in neural information processing systems*, volume 32, 2019.

Hila Chefer et al. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics*, 42(4):148, 2023.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024a.

Junsong Chen et al. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International Conference on Learning Representations*, 2024b.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/8b52c29d1a77e3fcc6c8f90fa0ac0e8e-Abstract-Conference.html.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/13e8be77982beb73d7ed0bbf122f9f3c-Abstract-Conference.html.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7694–7731. PMLR, 2023. URL https://proceedings.mlr.press/v202/desai23a.html.

Bhuwan Dhingra, Christopher Shallue, Mohammad Norouzi, Andrew Dai, and George Dahl. Embedding text in hyperbolic spaces. In *Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-12)*, pp. 59–69, 2018.

Aleksandr Ermolov, Leyla Mirvakhabova, Valentin Khrulkov, Nicu Sebe, and Ivan Oseledets. Hyperbolic vision transformers: Combining improvements in metric learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7399–7409. IEEE, 2022.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. In *Advances in neural information processing systems*, volume 36, pp. 79858–79885, 2023.

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *International Conference on Learning Representations*, 2023.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. In *Advances in Neural Information Processing Systems*, pp. 5345–5355, 2018a.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1646–1655. PMLR, 2018b. URL https://proceedings.mlr.press/v80/ganea18a.html.

Songwei Ge, Shlok Mishra, Simon Kornblith, Chun-Liang Li, and David Jacobs. Hyperbolic contrastive learning for visual representations beyond objects. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 6840–6849. IEEE, 2023.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, 2023.

Mikhael Gromov. Hyperbolic groups. In S. M. Gersten (ed.), *Essays in Group Theory*, volume 8 of *Mathematical Sciences Research Institute Publications*, pp. 75–263. Springer, New York, 1987. doi: 10.1007/978-1-4613-9586-7_3. URL https://doi.org/10.1007/978-1-4613-9586-7_3.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9726–9735. IEEE, 2020.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *Workshop on Deep Generative Models and Downstream Applications at the Conference on Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pp. 6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78723–78747, 2023.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *International Conference on Machine Learning*, 2024.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9307–9315. IEEE, 2024.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Dongzhi Jiang, Guanglu Song, Xiaoshi Wu, Renrui Zhang, Dazhong Shen, Zhuofan Zong, Yu Liu, and Hongsheng Li. Comat: Aligning text-to-image diffusion model with image-to-text concept matching. In *Advances in Neural Information Processing Systems*, volume 37, pp. 76177–76209, 2024.

Qian Jiang, Changyou Chen, Han Zhao, Liqun Chen, Qing Ping, Son Dinh Tran, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Understanding and constructing latent modality structures in multi-modal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7661–7671, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Jiang_Understanding_and_Constructing_Latent_Modality_Structures_in_Multi-Modal_Representation_Learning_CVPR_2023_paper.html.

Wonjun Kang, Kevin Galim, Hyung Il Koo, and Nam Ik Cho. Counting guidance for high fidelity text-to-image synthesis. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 899–908, 2025. URL https://openaccess.thecvf.com/content/WACV2025/html/Kang_Counting_Guidance_for_High_Fidelity_Text-to-Image_Synthesis_WACV_2025_paper.html.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, pp. 26565–26577, 2022.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5128–5137. IEEE, 2021.

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 6417–6427. IEEE, 2020.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/73aacd8b3b05b4b503d58310b523553c-Abstract-Conference.html.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

LAION-AI. LAION-Aesthetics Predictor V1. Software repository, 2022. URL https://github.com/LAION-AI/aesthetic-predictor.

Jaa-Yeon Lee, Byunghee Cha, Jeongsol Kim, and Jong Chul Ye. Aligning text to image in diffusion models is easier than you think. In *Advances in Neural Information Processing Systems*, volume 38, pp. 157106–157136, 2026.

Matthias Leimeister and Benjamin J. Wilson. Skip-gram word embeddings in hyperbolic space. *arXiv preprint arXiv:1809.01498*, 2018. URL https://arxiv.org/abs/1809.01498.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. REPA-E: Unlocking VAE for end-to-end tuning of latent diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18262–18272, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Leng_REPA-E_Unlocking_VAE_for_End-to-End_Tuning_of_Latent_Diffusion_Transformers_ICCV_2025_paper.html.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023a.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22511–22521. IEEE, 2023b.

Yaron Lipman et al. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European conference on computer vision*, pp. 423–439. Springer, 2022.

Qi Liu, Maximilian Nickel, and Douwe Kiela. Hyperbolic graph neural networks. In *Advances in Neural Information Processing Systems*, pp. 8230–8241, 2019.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Tuna Han Salih Meral, Enis Simsar, Federico Tombari, and Pinar Yanardag. Conform: Contrast is all you need for high-fidelity text-to-image diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9005–9014. IEEE, 2024.

Pascal Mettes, Mina Ghadimi Atigh, Martin Keller-Ressel, Jeffrey Gu, and Serena Yeung. Hyperbolic deep learning in computer vision: A survey. *International Journal of Computer Vision*, 132 (9):3484–3508, 2024.

Alex Nichol et al. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pp. 16784–16804, 2022.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, pp. 6338–6347, 2017.

Maximilian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International Conference on Machine Learning*, pp. 3779–3788, 2018.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3147–3157. IEEE, 2023.

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 *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d9dd691fa4eeb4741f16ac42ed5d197e-Abstract-Conference.html.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

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*, 44(12):10023–10044, 2021.

Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention refocusing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7932–7942, 2024.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

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

Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Ajanthan Thalaiyasingam. Accept the modality gap: An exploration in the hyperbolic space. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27253–27262. IEEE, 2024.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 3536–3559, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in neural information processing systems*, volume 35, pp. 36479–36494, 2022.

Frederic Sala, Chris De Sa, Albert Gu, and Christopher Ré. Representation tradeoffs for hyperbolic embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4460–4469. PMLR, 2018. URL https://proceedings.mlr.press/v80/sala18a.html.

Rik Sarkar. Low distortion Delaunay embedding of trees in hyperbolic plane. In *Graph Drawing*, volume 7034 of *Lecture Notes in Computer Science*, pp. 355–366. Springer, 2012. doi: 10.1007/978-3-642-25878-7_34. URL https://doi.org/10.1007/978-3-642-25878-7_34.

Ryohei Shimizu, Yusuke Mukuta, and Tatsuya Harada. Hyperbolic neural networks++. In *International Conference on Learning Representations*, 2021.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 3732–3784, 2023.

Sirnam Swetha, Mamshad Nayeem Rizve, Nina Shvetsova, Hilde Kuehne, and Mubarak Shah. Preserving modality structure improves multi-modal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21993–22003, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Swetha_Preserving_Modality_Structure_Improves_Multi-Modal_Learning_ICCV_2023_paper.html.

Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincaré GloVe: Hyperbolic word embeddings. In *International Conference on Learning Representations*, 2019.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238. IEEE, 2024.

Runqian Wang and Kaiming He. Diffuse and disperse: Image generation with representation regularization. *arXiv preprint arXiv:2506.09027*, 2025. URL https://arxiv.org/abs/2506.09027.

Chen Wei, Karttikeya Mangalam, Po-Yao Huang, Yanghao Li, Haoqi Fan, Hu Xu, Huiyu Wang, Cihang Xie, Alan Yuille, and Christoph Feichtenhofer. Diffusion models as masked autoencoders. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16238–16248. IEEE, 2023.

Qiucheng Wu, Yujian Liu, Handong Zhao, Trung Bui, Zhe Lin, Yang Zhang, and Shiyu Chang. Harnessing the spatial-temporal attention of diffusion models for high-fidelity text-to-image synthesis. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7732–7742. IEEE, 2023.

Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15756–15766. IEEE, 2023.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. SANA: Efficient high-resolution image synthesis with linear diffusion transformers. In *International Conference on Learning Representations*, 2025.

Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7418–7427. IEEE, 2023.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/33646ef0ed554145eab65f6250fab0c9-Abstract-Conference.html.

Jiexi Yan, Lei Luo, Cheng Deng, and Heng Huang. Unsupervised hyperbolic metric learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12460–12469. IEEE, 2021.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

A Extended Related Work

This appendix provides the full citation coverage and contextual narrative that the condensed Section 2 compresses into three short paragraphs. The structure mirrors the main-text version, with one subsection per thread.

A.1 Diffusion Transformers and Compositional Control

Modern text-to-image generation traces its lineage to denoising diffusion probabilistic models (Ho et al., 2020) and their continuous-time score-based formulation (Song et al., 2021), with latent diffusion (Rombach et al., 2022) establishing the now-standard recipe of running diffusion in a compressed VAE latent. Early text-conditioned systems such as GLIDE (Nichol et al., 2022), DALL-E 2 (Ramesh et al., 2022), Imagen (Saharia et al., 2022), and SDXL (Podell et al., 2024) relied on convolutional U-Net backbones with cross-attention to text features and were typically trained with the original DDPM objective and classifier-free guidance (Ho & Salimans, 2021); the EDM family (Karras et al., 2022) consolidated the design space around preconditioning and noise schedules. The shift to transformer denoisers began with DiT (Peebles & Xie, 2023) and U-ViT (Bao et al., 2023), which replaced the U-Net with a plain transformer operating on patchified latents, and was scaled to high-resolution text-to-image by PixArt-α/Σ (Chen et al., 2024b;a) and SANA (Xie et al., 2025). In parallel, flow-matching (Lipman et al., 2023) and rectified flow (Liu et al., 2023) replaced the DDPM training objective with a straight-line interpolant whose marginal velocity is regressed directly, and SiT (Ma et al., 2024) showed that scalable interpolant transformers match or exceed DiT under matched compute. MMDiT (Esser et al., 2024), the backbone we adopt, is the synthesis of these threads: a rectified-flow transformer in which text and image tokens are treated as co-equal streams that interact through joint self-attention rather than asymmetric cross-attention, with a CLIP encoder (Radford et al., 2021) supplying a pooled global vector and T5-XXL contributing per-token compositional structure. The same architectural pattern underpins FLUX-style pipelines (Black Forest Labs, 2024) that we also experiment with.

MMDiT's joint-attention block directly exposes text↔image weights, which motivates drawing candidate region tokens from the pooling already present inside the backbone rather than from an external detector; our phrase-box diagnostic tests whether this readout is spatially meaningful. Despite their scale, MMDiT-class models still falter on compositional prompts that specify multiple objects, attributes, counts, or spatial relations (Ghosh et al., 2023; Huang et al., 2023), prompting a wide range of training-time and inference-time remedies. Inference-time approaches re-weight or re-shape the existing attention maps to enforce attribute binding and spatial layout, including Structured Diffusion Guidance (Feng et al., 2023), Attend-and-Excite (Chefer et al., 2023), Linguistic Binding (Rassin et al., 2023), BoxDiff (Xie et al., 2023), attention refocusing (Phung et al., 2024), spatial-temporal attention harnessing (Wu et al., 2023), counting-aware guidance (Kang et al., 2025; Paiss et al., 2023), and CONFORM (Meral et al., 2024), which optimizes a contrastive attention objective at test time. Composable diffusion (Liu et al., 2022) combines conditional scores, while GLIGEN (Li et al., 2023b) uses grounding boxes. Training-time approaches add preference or concept-matching signals to denoising, including CoMat (Jiang et al., 2024), the ImageReward/ReFL framework (Xu et al., 2023), DPOK (Fan et al., 2023), DDPO (Black et al., 2024), Diffusion-DPO (Wallace et al., 2024), and direct reward backpropagation (DRaFT) (Clark et al., 2024). We leave the joint-attention transformer intact and adapt through LoRA (Hu et al., 2022), controlling for backbone and attention surgery while testing the added alignment signal.

A.2 Representation Alignment for Diffusion

A complementary line of work argues that the bottleneck of diffusion fine-tuning lies less in capacity than in the *representations* that the denoiser inhabits. The case for treating representation quality as a first-class training signal predates diffusion (SimCLR (Chen et al., 2020) and MoCo (He et al., 2020) for vision; CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), and SigLIP (Zhai et al., 2023) for vision–language; BLIP/BLIP-2 (Li et al., 2022; 2023a) for unified understanding/generation; and DINO/DINOv2 (Caron et al., 2021; Oquab et al., 2024) for self-distilled visual features), and an emerging "Platonic" view (Huh et al., 2024) suggests that strong models from different modalities converge to similar latent structure measurable via CKA (Kornblith et al., 2019). REPA (Yu et al., 2025) translated this intuition into the diffusion regime by aligning intermediate denoiser states with pretrained visual features, improving training efficiency and generation quality. REPA-E (Leng et al., 2025) uses representation alignment to enable joint training of the visual tokenizer and diffusion transformer. SoftREPA (Lee et al., 2026) addresses text–image compatibility through an InfoNCE-style (Oord et al., 2018) objective whose scores derive from denoising predictions under matched and mismatched text conditions, optimizing soft text tokens while keeping the pretrained model frozen. Related lines treat the diffusion process itself as a self-supervised feature learner (Chen et al., 2025; Wei et al., 2023; Xiang et al., 2023), and dispersive regularizers (Wang & He, 2025) aim to spread features apart without explicit positive pairs.

These approaches differ in their supervision: REPA transfers external visual features to denoiser states, REPA-E also adapts the visual tokenizer, and SoftREPA contrasts image–text compatibility through the generative model itself. HIP uses the language representations and joint attention already available in MMDiT to construct constituent correspondences. It regularizes each constituent's distance to its sentence or image context while learning phrase–region discrimination in a shared Poincaré space.

A.3 Hyperbolic Representation Learning

Hyperbolic geometry has long been advocated for hierarchically structured data on capacity grounds: a ball in constant negative curvature has exponentially growing volume and can represent branching structures compactly (Sarkar, 2012; Sala et al., 2018), whereas a fixed-dimensional Euclidean ball of comparable radius has only polynomial packing capacity. Nickel & Kiela (2017) introduced Poincaré embeddings of taxonomies, with a Lorentzian variant (Nickel & Kiela, 2018), and Ganea et al. (2018b) added explicit entailment cones for partial orders. Subsequent work generalized neural primitives to hyperbolic space: hyperbolic neural networks (Ganea et al., 2018a), HNN++ (Shimizu et al., 2021), graph convolutions (Chami et al., 2019; Liu et al., 2019), word embeddings (Tifrea et al., 2019; Dhingra et al., 2018; Leimeister & Wilson, 2018), and surveys (Peng et al., 2021; Mettes et al., 2024). Vision applications include image embeddings (Khrulkov et al., 2020), transformers for metric learning (Ermolov et al., 2022), segmentation (Atigh et al., 2022), contrastive learning (Ge et al., 2023), metric learning (Yan et al., 2021), and modality-gap analysis (Ramasinghe et al., 2024).

The vision–language thread is most directly relevant. MERU (Desai et al., 2023) introduced hyperbolic multimodal contrastive learning at CLIP scale and used entailment relations between images and captions. HyCoCLIP (Pal et al., 2025) explicitly exploits compositional relations between full captions and substrings. Together they establish that cross-modal correspondence and intra-modal entailment can organize multimodal hierarchies in a shared hyperbolic space. HIP adapts hierarchical representation learning to diffusion fine-tuning through attention-induced visual constituents. Its whole–part distance penalties provide scene anchors, and its phrase–region contrastive objective trains constituent discrimination. The finite-radius analysis characterizes how negative curvature supports these complementary relations within the shared representation space.

B Geometric Background and Metric Consistency

B.1 Geometric identities and conventions

Throughout this appendix, $c > 0$ denotes the magnitude of the negative sectional curvature: the Poincaré metric in Sec. 3.3 has curvature $-c$. The conformal metric is $g_x = \lambda_x^2 g^{\mathrm{E}}$ with $\lambda_x = 2/(1-$

$c\|x\|^2$). Unsubscripted vector norms are Euclidean, whereas $\|v\|_x = \lambda_x \|v\|$ is the Riemannian norm at x. For $x, y \in \mathbb{B}_c^d$, Möbius addition is

$$x \oplus_c y = \frac{(1 + 2c\langle x, y\rangle + c\|y\|^2)x + (1 - c\|x\|^2)y}{1 + 2c\langle x, y\rangle + c^2\|x\|^2\|y\|^2}.$$

For $v \in T_x\mathbb{B}_c^d \setminus \{0\}$ and $y \neq x$, the exponential and logarithmic maps admit the closed forms (Ganea et al., 2018a)

$$\exp_x(v) = x \oplus_c \left(\tanh\left(\frac{\sqrt{c}\,\lambda_x\,\|v\|}{2}\right)\frac{v}{\sqrt{c}\,\|v\|}\right),$$
$$\log_x(y) = \frac{2}{\sqrt{c}\,\lambda_x}\operatorname{arctanh}\big(\sqrt{c}\,\|(-x) \oplus_c y\|\big)\frac{(-x) \oplus_c y}{\|(-x) \oplus_c y\|}.$$

At the removable singularities, these expressions are defined by continuity as $\exp_x(0) = x$ and $\log_x(x) = 0$. In particular, $\lambda_0 = 2$, so the first formula reduces to Eq. (7), and

$$\log_0(y) = \operatorname{arctanh}(\sqrt{c}\|y\|)\frac{y}{\sqrt{c}\|y\|} \quad (y \neq 0), \qquad \log_0(0) = 0.$$

Taking the Riemannian norm of the logarithmic map gives

$$\|\log_x(y)\|_x = \lambda_x \|\log_x(y)\| = \frac{2}{\sqrt{c}}\operatorname{arctanh}\big(\sqrt{c}\|(-x) \oplus_c y\|\big) = d_{\mathbb{B}_c^d}(x, y),$$

where the last expression is the Möbius form of the distance in Sec. 3.3. The identity extends to $x = y$ by continuity. Thus the tangent-chart score uses the Riemannian norm, including the conformal factor λ_x, to recover the same geodesic distance. At the origin this convention gives $d_{\mathbb{B}_c^d}(0, \exp_0(u)) = 2\|u\|$.

B.2 Constituent separation around a shared anchor

Proposition 1 (Separation at a fixed anchor distance)**.** *Let $o, x, y \in \mathbb{H}_c^d$ with $d \geq 2$ and sectional curvature $-c < 0$. Write $d_{\mathbb{H}}$ for its geodesic distance, which is $d_{\mathbb{B}_c^d}$ in Poincaré coordinates. Suppose $d_{\mathbb{H}}(o, x) = d_{\mathbb{H}}(o, y) = \rho > 0$, and let $\varphi \in [0, \pi]$ be the Riemannian angle between the geodesics from o to x and y. Their hyperbolic distance $D_{\mathbb{H}}(\rho, \varphi) = d_{\mathbb{H}}(x, y)$ and the Euclidean distance $D_{\mathbb{E}}(\rho, \varphi)$ at the same radius and angle satisfy*

$$D_{\mathbb{H}}(\rho, \varphi) = \frac{2}{\sqrt{c}}\operatorname{arsinh}\left(\sinh(\sqrt{c}\,\rho)\sin\frac{\varphi}{2}\right),$$
$$D_{\mathbb{E}}(\rho, \varphi) = 2\rho\sin\frac{\varphi}{2}, \qquad D_{\mathbb{H}}(\rho, \varphi) \geq D_{\mathbb{E}}(\rho, \varphi). \tag{15}$$

The inequality is strict for $0 < \varphi < \pi$ and is an equality at $\varphi = 0$ or $\varphi = \pi$.

Proof. The hyperbolic law of cosines gives

$$\cosh\big(\sqrt{c}\,d_{\mathbb{H}}(x, y)\big) = \cosh^2(\sqrt{c}\,\rho) - \sinh^2(\sqrt{c}\,\rho)\cos\varphi$$
$$= 1 + 2\sinh^2(\sqrt{c}\,\rho)\sin^2\frac{\varphi}{2}.$$

Using $\cosh(2\operatorname{arsinh} u) = 1 + 2u^2$ for $u \geq 0$ yields the distance formula in Eq. (15). Set $\chi = \sqrt{c}\,\rho > 0$ and $\zeta = \sin(\varphi/2) \in [0, 1]$. Since arsinh is concave on $[0, \infty)$ and vanishes at zero,

$$\operatorname{arsinh}(\zeta \sinh\chi) \geq \zeta \operatorname{arsinh}(\sinh\chi) = \zeta\chi.$$

Multiplying by $2/\sqrt{c}$ proves the comparison with the Euclidean distance. Strict concavity gives strict inequality for $0 < \zeta < 1$; the two endpoint cases give equality. □

Interpretation for whole–part anchoring. Taking $o = f$ and $0 < \rho \leq m$ places both regions within the image neighborhood where their consistency penalties vanish. The proposition compares their separation with a Euclidean configuration at the same anchor distances and angle, isolating the contribution of curvature. Taking $o = s$ gives the corresponding interpretation for phrases. The comparison holds for any anchor in the space, so each scene representation supplies its own local context for constituent discrimination. Both geometries still obey the diameter bound $2m$; the separation advantage concerns fixed anchor distances and angle.

B.3 Fixed-dimensional branch capacity

Let $N_{\mathcal{M}}(R, \delta)$ be the largest number of points in a closed radius-R ball of a metric space $\mathcal{M}$ with pairwise distances at least δ. Accommodating K such constituents requires $K \leq N_{\mathcal{M}}(R, \delta)$. For fixed dimension $d \geq 2$, separation $\delta > 0$, and hyperbolic curvature $-c < 0$, volume comparison gives

$$\begin{aligned} N_{\mathbb{R}^d}(R, \delta) &\leq (1 + 2R/\delta)^d, \\ \log N_{\mathbb{H}^d_c}(R, \delta) &= (d-1)\sqrt{c}\, R + O(1), \end{aligned} \tag{16}$$

where the second relation holds as $R \to \infty$. At HIP's fixed training margin $R = m$, the exponential rate describes growth as the radius budget increases. The following proposition specializes this comparison to a hierarchy whose radius grows linearly with depth.

Proposition 2 (Fixed-dimensional branch capacity)**.** *Fix an integer dimension $d \geq 1$, an integer branching factor $\beta \geq 2$, and constants $\alpha, \delta > 0$, all independent of the integer depth $L \geq 1$. Consider β^L descendant embeddings within distance αL of a root, with all distinct descendants separated by at least δ. In $\mathbb{R}^d$, a necessary condition is*

$$\beta^L \;\leq\; \left(1 + \frac{2\alpha L}{\delta}\right)^d. \tag{17}$$

Thus a fixed-dimensional Euclidean ball with linearly growing radius has only polynomial capacity in L, and this condition fails for all sufficiently large L. For hyperbolic space $\mathcal{M} = \mathbb{H}^d_c$ with fixed $d \geq 2$ and sectional curvature $-c < 0$, its packing number satisfies, as $L \to \infty$,

$$N_{\mathbb{H}^d_c}(\alpha L, \delta) = \exp\big((d-1)\sqrt{c}\,\alpha L + O(1)\big)\,, \tag{18}$$

where the $O(1)$ term is bounded independently of L. Consequently, if $(d-1)\sqrt{c}\,\alpha > \log\beta$, then for every sufficiently large L there exist β^L pairwise δ-separated points within distance αL of a root.

B.4 Proof of Proposition 2

Write $\overline{B}_{\mathcal{M}}(o, R) = \{x : d_{\mathcal{M}}(o, x) \leq R\}$ for a closed metric ball and $B^{\circ}_{\mathcal{M}}(o, R)$ for its open counterpart. A set X is δ-separated if $d_{\mathcal{M}}(x, y) \geq \delta$ for every two distinct points in X. The packing number used in the proposition is

$$N_{\mathcal{M}}(R, \delta) = \max\big\{|X| : X \subseteq \overline{B}_{\mathcal{M}}(o, R),\ X \text{ is } \delta\text{-separated}\big\}.$$

For the Euclidean and hyperbolic spaces considered here, the choice of o does not affect this number because isometries act transitively. Their closed balls of finite radius are compact. Hence such a ball admits a finite cover by open balls of radius $\delta/3$, each of which contains at most one point of a δ-separated set. The possible cardinalities are therefore bounded integers, so the maximum exists. A set is *maximal by inclusion* if no further point of the ball can be added while preserving separation; this is weaker than attaining the maximum cardinality, but suffices for the covering argument below.

Proof. **Step 1: Euclidean upper bound.** Let $X_L \subseteq \overline{B}_{\mathbb{R}^d}(o, \alpha L)$ consist of the β^L descendant embeddings. The open balls $B^{\circ}_{\mathbb{R}^d}(x, \delta/2)$, $x \in X_L$, are pairwise disjoint: if a point z belonged to the balls around two different centers x, y, the triangle inequality would imply

$$\|x - y\| \leq \|x - z\| + \|z - y\| < \delta,$$

contradicting the assumed separation. The use of open balls also handles centers whose separation is exactly δ. Furthermore, every such ball is contained in $B^{\circ}_{\mathbb{R}^d}(o, \alpha L + \delta/2)$, since

$$\|z - o\| \leq \|z - x\| + \|x - o\| < \delta/2 + \alpha L.$$

Let v_d be the volume of the Euclidean unit ball. Translation invariance and additivity of volume on disjoint sets give

$$\beta^L v_d (\delta/2)^d = \sum_{x \in X_L} \mathrm{Vol}\left(B^{\circ}_{\mathbb{R}^d}(x, \delta/2)\right) \leq v_d(\alpha L + \delta/2)^d.$$

Dividing by the positive quantity $v_d(\delta/2)^d$ proves Eq. (17). Its logarithmic form is

$$\log\beta \leq \frac{d}{L}\log\left(1 + \frac{2\alpha L}{\delta}\right).$$

For fixed d, α, δ, the right-hand side tends to zero as $L \to \infty$, whereas $\log \beta > 0$. Thus such Euclidean configurations cannot exist for all sufficiently large depths under the stated constraints.

Step 2: Hyperbolic packing and covering bounds. Set

$$V_{c,d}(R) = \mathrm{Vol}_{\mathbb{H}^d_c}\big(\overline{B}_{\mathbb{H}^d_c}(o, R)\big).$$

Homogeneity makes this volume independent of the center, and open and closed balls of the same radius have the same volume. For any δ-separated $X \subseteq \overline{B}_{\mathbb{H}^d_c}(o, R)$, the same triangle-inequality argument as in Step 1 shows that the open radius-$\delta/2$ balls around X are disjoint and lie inside the radius-$(R + \delta/2)$ ball. Consequently,

$$|X| V_{c,d}(\delta/2) \le V_{c,d}(R + \delta/2).$$

Taking a set of maximum cardinality yields the upper bound on $N_{\mathbb{H}^d_c}(R, \delta)$ below.

For the lower bound, choose a maximal-by-inclusion δ-separated set Y in $\overline{B}_{\mathbb{H}^d_c}(o, R)$; for example, add points greedily until no point can be added, which terminates by the finite cardinality bound above. Every point of the ball is at distance strictly less than δ from some point of Y. Otherwise that point could be added to Y, contradicting maximality. Thus

$$\overline{B}_{\mathbb{H}^d_c}(o, R) \subseteq \bigcup_{y \in Y} B^{\circ}_{\mathbb{H}^d_c}(y, \delta).$$

Subadditivity of volume, without any assumption that these covering balls are disjoint, gives $V_{c,d}(R) \le |Y| V_{c,d}(\delta)$. Since $|Y| \le N_{\mathbb{H}^d_c}(R, \delta)$, the two bounds combine to give

$$\frac{V_{c,d}(R)}{V_{c,d}(\delta)} \le N_{\mathbb{H}^d_c}(R, \delta) \le \frac{V_{c,d}(R + \delta/2)}{V_{c,d}(\delta/2)}. \tag{19}$$

The denominators are strictly positive constants because c, d, δ are fixed and $\delta > 0$.

Step 3: Derivation and bounds for hyperbolic ball volume. Let $\kappa = \sqrt{c}$. At the origin of the Poincaré model, write $\varrho = \|x\|$ for the Euclidean radial coordinate and ℓ for geodesic radius. The metric gives

$$\ell = \int_0^{\varrho} \frac{2\,\mathrm{d}u}{1 - \kappa^2 u^2} = \frac{2}{\kappa} \operatorname{arctanh}(\kappa \varrho), \qquad \varrho = \frac{1}{\kappa} \tanh\left(\frac{\kappa \ell}{2}\right).$$

Writing $\lambda(\varrho) = 2/(1 - c\varrho^2)$, in Euclidean polar coordinates, $g^{\mathrm{E}} = \mathrm{d}\varrho^2 + \varrho^2 g_{\mathbb{S}^{d-1}}$. Substitution into $g = \lambda(\varrho)^2 g^{\mathrm{E}}$, using $\lambda(\varrho)\,\mathrm{d}\varrho = \mathrm{d}\ell$ and $\lambda(\varrho)\varrho = \sinh(\kappa\ell)/\kappa$, yields

$$g = \mathrm{d}\ell^2 + \left(\frac{\sinh(\kappa \ell)}{\kappa}\right)^2 g_{\mathbb{S}^{d-1}}.$$

Here $g_{\mathbb{S}^{d-1}}$ is the standard metric on the Euclidean unit sphere. Taking the volume element and integrating over the angular coordinates therefore gives

$$V_{c,d}(R) = \omega_{d-1} \int_0^R \left(\frac{\sinh(\kappa \ell)}{\kappa}\right)^{d-1} \mathrm{d}\ell, \tag{20}$$

where ω_{d-1} is the area of $\mathbb{S}^{d-1}$.

To obtain bounds with constants independent of R, define $\gamma_{\mathrm{vol}} = (d-1)\kappa > 0$. Since $\sinh(\kappa\ell) \le e^{\kappa\ell}/2$ for $\ell \ge 0$,

$$V_{c,d}(R) \le \frac{\omega_{d-1}}{(2\kappa)^{d-1}} \int_0^R e^{\gamma_{\mathrm{vol}} \ell}\,\mathrm{d}\ell \le A_+ e^{\gamma_{\mathrm{vol}} R}, \qquad A_+ = \frac{\omega_{d-1}}{\gamma_{\mathrm{vol}}(2\kappa)^{d-1}}.$$

For a matching lower bound, let $\ell_0 = 1/\kappa$ and $q = (1 - e^{-2})/2 > 0$. For $\ell \ge \ell_0$,

$$\sinh(\kappa \ell) = \frac{e^{\kappa \ell}}{2}(1 - e^{-2\kappa \ell}) \ge q e^{\kappa \ell}.$$

If $R \geq 2\ell_0$, the interval $[R-\ell_0, R]$ lies in $[\ell_0, \infty)$, so

$$V_{c,d}(R) \geq \frac{\omega_{d-1} q^{d-1}}{\kappa^{d-1}} \int_{R-\ell_0}^{R} e^{\gamma_{\mathrm{vol}}\ell}\, \mathrm{d}\ell = A_- e^{\gamma_{\mathrm{vol}} R},$$
$$A_- = \frac{\omega_{d-1} q^{d-1}}{\gamma_{\mathrm{vol}} \kappa^{d-1}} (1 - e^{-\gamma_{\mathrm{vol}}\ell_0}) > 0.$$

Thus $A_- e^{\gamma_{\mathrm{vol}} R} \leq V_{c,d}(R) \leq A_+ e^{\gamma_{\mathrm{vol}} R}$ for $R \geq 2\ell_0$, with both constants depending only on c and d. Substituting into Eq. (19) gives

$$C_- e^{\gamma_{\mathrm{vol}} R} \leq N_{\mathbb{H}^d_c}(R, \delta) \leq C_+ e^{\gamma_{\mathrm{vol}} R}, \qquad R \geq 2\ell_0, \tag{21}$$

where

$$C_- = \frac{A_-}{V_{c,d}(\delta)}, \qquad C_+ = \frac{A_+ e^{\gamma_{\mathrm{vol}}\delta/2}}{V_{c,d}(\delta/2)}.$$

These positive constants depend only on c, d, δ, not on R or L. Setting $R = \alpha L$ and taking logarithms gives

$$\log C_- \leq \log N_{\mathbb{H}^d_c}(\alpha L, \delta) - \gamma_{\mathrm{vol}} \alpha L \leq \log C_+,$$

which proves Eq. (18), including its bounded $O(1)$ remainder.

Step 4: Comparison with the branching rate. If $\gamma_{\mathrm{vol}}\alpha > \log \beta$, then

$$\frac{N_{\mathbb{H}^d_c}(\alpha L, \delta)}{\beta^L} \geq C_- \exp\big((\gamma_{\mathrm{vol}}\alpha - \log\beta) L\big) \longrightarrow \infty.$$

It follows that $N_{\mathbb{H}^d_c}(\alpha L, \delta) \geq \beta^L$ for every sufficiently large integer L. Selecting β^L points from such a packing proves the stated existence claim. □

Implications for hierarchical representation. Proposition 2 quantifies the capacity to separate descendants under a radius budget growing linearly with depth. With fixed $d \geq 2$, $c > 0$, and $\delta > 0$, the hyperbolic packing rate is $(d-1)\sqrt{c}\,\alpha$; choosing this rate above $\log \beta$ accommodates exponential branching. Proposition 1 characterizes separation between individual constituents at a fixed anchor distance, while the packing result describes how capacity grows with the represented hierarchy. These geometric properties support the joint organization specified by HIP's correspondence and whole–part objectives.

B.5 Proof of Observation 1

We prove the bound in a common metric space $(\mathcal{M}, d_{\mathcal{M}})$; setting $d_{\mathcal{M}} = d_{\mathbb{B}^d_c}$ gives the hyperbolic statement in the main text. For the $K \geq 1$ valid correspondences, define

$$e_i = d_{\mathcal{M}}(p_i, r_i), \qquad \xi_i^T = [d_{\mathcal{M}}(s, p_i) - m]_+, \qquad \xi_i^I = [d_{\mathcal{M}}(f, r_i) - m]_+,$$

and write

$$\bar{e} = \frac{1}{K}\sum_{i=1}^{K} e_i, \qquad \bar{\xi}^T = \frac{1}{K}\sum_{i=1}^{K} \xi_i^T, \qquad \bar{\xi}^I = \frac{1}{K}\sum_{i=1}^{K} \xi_i^I.$$

These means are taken over valid correspondences of one example. The training losses $\mathcal{L}_T, \mathcal{L}_I$ in Eq. (11) average the same hinge residuals over all phrase slots and batch examples, including padding.

Proof. For every real $q \geq 0$, the definition of the positive part implies

$$q = m + (q - m) \leq m + [q - m]_+.$$

Applying this to the two parent–child distances gives

$$d_{\mathcal{M}}(s, p_i) \leq m + \xi_i^T, \qquad d_{\mathcal{M}}(f, r_i) \leq m + \xi_i^I.$$

Following the path $s \to p_i \to r_i \to f$ and applying the triangle inequality twice, then symmetry of $d_{\mathcal{M}}$, yields

$$\begin{aligned} d_{\mathcal{M}}(s, f) &\le d_{\mathcal{M}}(s, p_i) + d_{\mathcal{M}}(p_i, r_i) + d_{\mathcal{M}}(r_i, f) \\ &\le (m + \xi_i^T) + e_i + (m + \xi_i^I) = 2m + \xi_i^T + \xi_i^I + e_i. \end{aligned}$$

The left-hand side is independent of i. Summing these K inequalities and dividing by K therefore proves

$$d_{\mathcal{M}}(s, f) \le 2m + \bar{\xi}^T + \bar{\xi}^I + \bar{e},$$

which, with $d_{\mathcal{M}} = d_{\mathbb{B}_c^d}$ and the residual means expanded, is Eq. (14). □

Pairwise leaf discrepancy. For $K \ge 2$, the same correspondences also satisfy

$$\Delta_{\text{pair}} := \frac{2}{K(K-1)} \sum_{i<j} |d_{\mathcal{M}}(p_i, p_j) - d_{\mathcal{M}}(r_i, r_j)| \le 2\bar{e}. \tag{22}$$

Proof. For any pair $i < j$, the path $p_i \to r_i \to r_j \to p_j$ gives

$$d_{\mathcal{M}}(p_i, p_j) \le e_i + d_{\mathcal{M}}(r_i, r_j) + e_j.$$

Interchanging the two modalities, the path $r_i \to p_i \to p_j \to r_j$ gives

$$d_{\mathcal{M}}(r_i, r_j) \le e_i + d_{\mathcal{M}}(p_i, p_j) + e_j.$$

Rearranging these two inequalities bounds the signed discrepancy from both sides and hence gives

$$|d_{\mathcal{M}}(p_i, p_j) - d_{\mathcal{M}}(r_i, r_j)| \le e_i + e_j. \tag{23}$$

There are $K(K-1)/2$ unordered pairs. In the sum of their upper bounds, each e_i occurs once with each of the other $K - 1$ indices, so

$$\sum_{i<j} (e_i + e_j) = (K-1) \sum_{i=1}^{K} e_i.$$

Using $K \ge 2$ to divide by the number of pairs gives

$$\Delta_{\text{pair}} \le \frac{2}{K(K-1)} (K-1) \sum_{i=1}^{K} e_i = \frac{2}{K} \sum_{i=1}^{K} e_i = 2\bar{e},$$

proving Eq. (22). □

Sharpness and interpretation. Both constants are sharp under the observation's assumptions. On the real line, set $p_i = 0$, $r_i = \eta$, $s = -m$, and $f = \eta + m$ for every i, with $\eta \ge 0$. Repeated leaf locations are allowed by the observation. Both hinge residuals vanish, $\bar{e} = \eta$, and $d_{\mathcal{M}}(s, f) = 2m + \eta$, attaining the root bound. At $\eta = 0$, this construction attains the $2m$ margin budget exactly. For the pairwise bound, take $K = 2$, $p_1 = 0$, $p_2 = D$, $r_1 = -\eta$, and $r_2 = D + \eta$, with $D, \eta > 0$. Then $e_1 = e_2 = \eta$ and $\Delta_{\text{pair}} = |(D) - (D + 2\eta)| = 2\eta = 2\bar{e}$. Both examples can also be placed along a single hyperbolic geodesic parametrized by arc length.

The mean bound accommodates heterogeneous correspondence errors, and Eq. (23) gives the guarantee for each individual pair. Exact local matches yield identical cross-modal leaf-distance matrices; the two intra-modal constraints additionally control root agreement through the $2m$ margin budget. The bounds are expressed in absolute metric distances and hinge residuals, while InfoNCE trains relative discrimination between matched and mismatched candidates. These complementary roles explain HIP's joint objective. The attention-specificity analysis in App. I provides spatial evidence that phrase attention concentrates on constituents within the broader scene context.

Connection to hierarchy diagnostics. For a rooted tree metric, let z be the lowest common ancestor of x and y. The unique paths satisfy $d_{\mathcal{M}}(o,x) = d_{\mathcal{M}}(o,z) + d_{\mathcal{M}}(z,x)$, $d_{\mathcal{M}}(o,y) = d_{\mathcal{M}}(o,z) + d_{\mathcal{M}}(z,y)$, and $d_{\mathcal{M}}(x,y) = d_{\mathcal{M}}(z,x) + d_{\mathcal{M}}(z,y)$. Substitution into the standard Gromov product $(x \mid y)_o = [d_{\mathcal{M}}(o,x) + d_{\mathcal{M}}(o,y) - d_{\mathcal{M}}(x,y)]/2$ (Gromov, 1987; Bridson & Haefliger, 1999) gives $d_{\mathcal{M}}(o,z)$, the length of the shared root path.

The raw mean absolute sibling-distance discrepancy, together with the mean absolute correspondence error in the same metric, checks Eq. (22). Checking Eq. (14) additionally requires the root distance, margin and mean whole–part consistency residuals. These bounds describe how the supervised relations fit together; they hold for any configuration in the exact metric. Evidence for learned correspondence comes from retrieval and matched-versus-mismatched gaps relative to the controls in App. H.

C Backbone and HIP Architectures

C.1 MMDiT blocks and feature representations

Figure 5 expands the main-text architecture in Fig. 2. MMDiT processes text and image streams with separate modality weights and joint attention (Esser et al., 2024). For one attention head of width d_h, let Q_M, K_M, V_M be the query, key, and value matrices of modality $M \in \{T, I\}$. Concatenating each pair along the token dimension gives Q, K, V. The row-wise softmax yields the attention matrix $A = \mathrm{softmax}(QK^\mathsf{T}/\sqrt{d_h})$ and output $O = AV$, where T denotes matrix transpose. The text-query/image-key submatrix of A supplies the weights used in Sec. 3.2; the softmax runs over both text and image keys.

The text embeddings H^T enter the first joint-attention block, and the attention weights are averaged over the selected blocks specified in App. G. The image features H_t^I are obtained after adaptive normalization and before output projection on SD3.5-M. On FLUX, they are the image slice after the single-stream blocks and before final normalization. Thus, the attention maps and image features come from different points in the backbone: the former associate phrases with spatial positions, while the latter supply the visual features pooled at those positions.

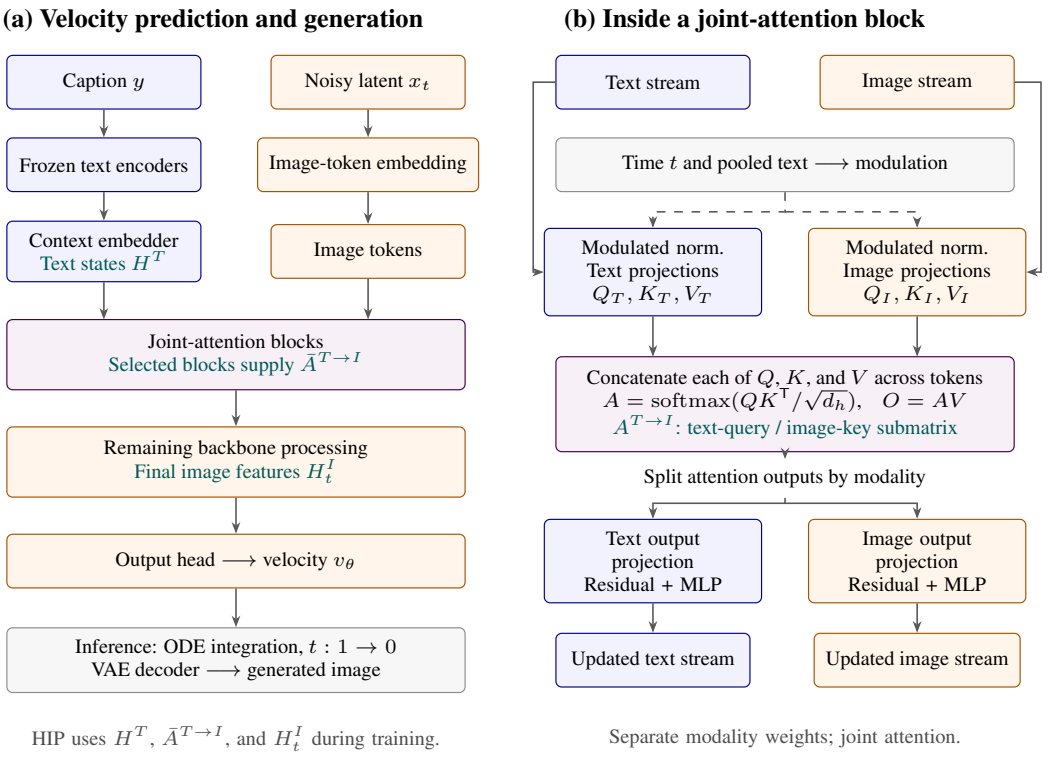

Figure 5: **Detailed MMDiT computation.** (a) Text and noisy image tokens interact to predict the flow velocity; repeated predictions drive the ODE solver before VAE decoding. Teal labels identify the features and attention used by HIP. Remaining backbone processing includes FLUX's single-stream blocks. (b) A representative joint-attention block uses separate modality projections and attention over the concatenated token sequences. Modulation, residual paths, and head-specific normalization are summarized for clarity.

D Extended Qualitative Comparison

Figure 6 includes the four methods in Fig. 4 and adds the hierarchy-free cosine identity-projector control and HIP-lite, a separately trained variant using rank-1 LoRA adapters in all 24 MMDiT blocks. HIP-lite illustrates generation with a smaller adaptation capacity. As with full HIP, the projection heads are discarded after training and the LoRA updates can be merged into the backbone without increasing the final model's parameter count.

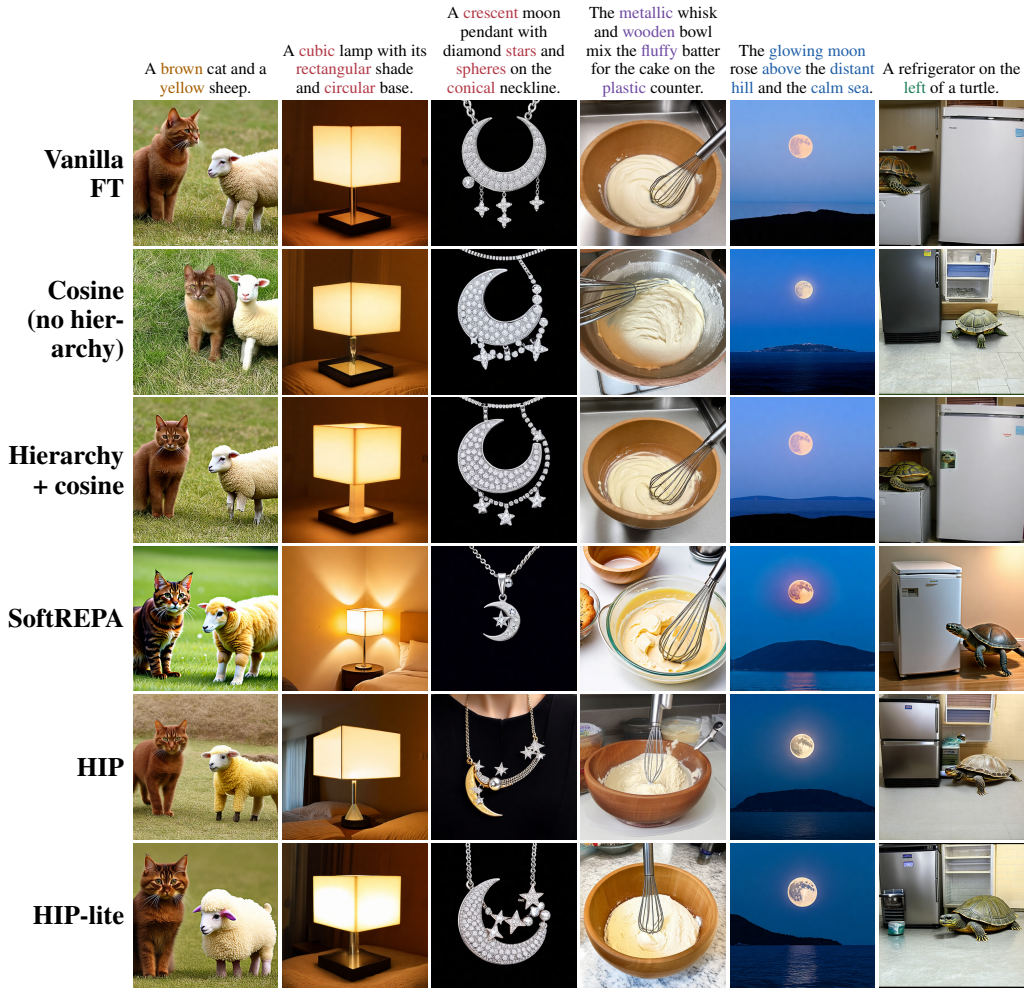

Figure 6: **Extended qualitative comparison.** Rows show vanilla fine-tuning, cosine with an identity projector, hierarchy + cosine, SoftREPA, HIP, and full-depth HIP-lite (r=1).

E Full Compositional Benchmark Results

Tables 3–6 expand the main comparison in Table 1 with full task breakdowns, relative gains over pretrained sampling, and additional geometry and hierarchy controls. Macro accuracy weights the six GenEval tasks equally; prompt accuracy weights all prompts equally. The original comparison scores are averaged over models trained with three random seeds: 111, 666, 777. Hierarchy + Euclidean is included as a supplementary evaluation.

Additional SD3.5-M controls. **Cosine identity (no hierarchy)** uses cosine alignment with an identity projector and omits the whole–part anchoring terms. **Hierarchy + Euclidean** retains correspondence and anchoring but replaces the exponential map with the identity and uses Euclidean L2 distance in both objectives. This is the Euclidean-L2 geometry substitution described in App. H; its supplementary generation results are reported in Tables 3 and 4.

Table 3: **SD3.5-M on GenEval.** All six task scores and both aggregations are reported on 553 prompts. Macro avg. equally weights tasks; Prompt avg. equally weights prompts. Bold marks the best score in each column. The shaded HIP row is followed by relative gains over the pretrained baseline.

Method	Macro avg.	Prompt avg.	Single	Two	Counting	Colors	Position	Color attr.
Pretrained baseline	0.64	0.63	**1.00**	0.79	0.53	0.80	0.23	0.51
Vanilla fine-tuning	0.67	0.66	**1.00**	0.79	0.63	0.84	0.26	0.53
Hierarchy + cosine	0.67	0.66	0.99	0.81	0.61	0.87	0.22	0.54
Hierarchy + Euclidean	0.69	0.68	0.96	**0.93**	0.56	0.86	0.26	0.56
SoftREPA	0.70	0.70	**1.00**	0.91	0.53	**0.89**	0.28	0.61
Cosine identity (no hierarchy)	0.65	0.64	0.99	0.88	0.49	0.81	0.25	0.50
HIP	**0.73**	**0.73**	**1.00**	0.92	**0.69**	0.86	**0.32**	**0.62**
HIP gain	+14.48%	+14.89%	0.00%	+16.66%	+30.95%	+8.00%	+39.13%	+21.57%

Table 4: **SD3.5-M on T2I-CompBench with the official evaluator.** All six evaluated categories are reported; higher is better. Bold marks the best score in each column. The shaded HIP row is followed by relative gains over the pretrained baseline.

Method	Color	Shape	Texture	2D spatial	Counting	Complex
Pretrained baseline	0.80	0.57	0.77	0.31	0.60	0.37
Vanilla fine-tuning	0.82	0.57	0.76	0.31	0.59	0.38
Hierarchy + cosine	0.83	0.59	0.76	0.31	0.59	0.38
Hierarchy + Euclidean	0.86	0.65	0.80	0.36	0.61	0.40
SoftREPA	0.82	0.64	0.78	0.36	0.60	0.38
Cosine identity (no hierarchy)	0.83	0.60	0.77	0.32	0.61	0.39
HIP	**0.89**	**0.67**	**0.83**	**0.38**	**0.62**	**0.41**
HIP gain	+11.21%	+17.46%	+7.89%	+19.98%	+4.77%	+9.19%

Euclidean generation results. Hierarchy + Euclidean improves over hierarchy + cosine in GenEval macro accuracy and all six T2I-CompBench categories, but remains below HIP on these measures (Tables 3 and 4). The full GenEval breakdown shows a task-specific exception: Hierarchy + Euclidean achieves the highest two-object accuracy, 0.9297 before rounding.

Hierarchy-free FLUX control. **InfoNCE (no hierarchy)** uses learned MLP projections followed by L2 normalization and cosine-similarity logits, with the exponential map and hierarchy terms removed. It retains the matched global and phrase–region contrastive objectives, training budget, and projection-head capacity. Tables 5 and 6 report this additional control alongside the main comparison.

Table 5: **FLUX.1-dev on GenEval.** All methods use the same generation settings, and adapted models are compared at the same epoch budget. Macro avg. and Prompt avg. follow Table 3. Bold marks the best score in each column; HIP gains are relative to the pretrained baseline.

Method	Macro avg.	Prompt avg.	Single	Two	Counting	Colors	Position	Color attr.
Pretrained baseline	0.64	0.62	0.99	0.81	**0.74**	0.79	0.17	0.36
Vanilla fine-tuning	0.57	0.55	0.96	0.64	0.65	0.70	0.16	0.30
Hierarchy + cosine	0.64	0.63	0.99	0.78	0.66	0.80	0.16	0.46
SoftREPA	0.67	0.66	0.98	0.87	0.60	0.79	0.20	**0.57**
InfoNCE (no hierarchy)	0.68	0.67	0.99	0.86	0.68	0.79	**0.26**	0.50
HIP	**0.69**	**0.68**	**1.00**	**0.89**	0.70	**0.81**	0.22	0.55
HIP gain	+8.24%	+9.26%	+1.27%	+10.00%	−5.08%	+2.71%	+29.41%	+52.78%

Table 6: **FLUX.1-dev on T2I-CompBench with the official evaluator.** All six evaluated categories are reported; higher is better. Bold marks the best score in each column. The shaded HIP row is followed by relative gains over the pretrained baseline.

Method	Color	Shape	Texture	2D spatial	Counting	Complex
Pretrained baseline	0.75	0.50	0.64	0.27	0.60	0.37
Vanilla fine-tuning	0.71	0.49	0.62	0.23	0.53	0.36
Hierarchy + cosine	0.79	0.53	0.69	0.28	0.57	0.38
SoftREPA	0.81	0.57	0.73	0.29	0.59	0.39
InfoNCE (no hierarchy)	0.81	0.57	0.73	0.28	0.60	0.39
HIP	**0.84**	**0.60**	**0.74**	**0.35**	**0.62**	**0.41**
HIP gain	+11.66%	+20.51%	+16.70%	+31.12%	+3.69%	+8.32%

F Extended Ablation Study

We examine how the three auxiliary objectives contribute to compositional generation. Table 7 compares denoising-only adaptation, each auxiliary term used individually, and the full HIP objective on SD3.5-M. Training details are collected in App. G.

Table 7: **SD3.5-M auxiliary-objective ablations on GenEval.** Each single-term variant adds the indicated auxiliary to the denoising loss; HIP uses all three. Macro avg. and Prompt avg. follow the definitions in Table 3.

Method	Macro avg.	Prompt avg.
Vanilla fine-tuning	0.67	0.66
Text consistency only	0.65	0.65
Image consistency only	0.68	0.67
Phrase–region correspondence only	0.70	0.69
HIP	**0.73**	**0.73**

Local correspondence supplies the most effective individual signal. Phrase–region alignment gives the strongest single-term result, supporting direct supervision of the visual constituents associated with different phrases. The anchoring terms have complementary gradient paths: image anchoring updates visual representations through the denoiser, while text anchoring shapes the phrase embeddings that supervise visual learning through the cross-modal objective.

The full objective combines discrimination with context. HIP outperforms every single-term variant, supporting the joint use of constituent discrimination and scene anchoring. Each constituent

participates in a cross-modal correspondence and a within-modality relation to its scene representation. Their combination provides the most effective supervision in this component study.

Geometry complements the objective design. The component study is complemented by the controlled FLUX metric comparison in Sec. 4.2 and the SD3.5-M Euclidean-L2 control, evaluated for generation as Hierarchy + Euclidean in App. E and for representations in Sec. 4.3. The former holds the whole–part anchoring terms fixed and tests the correspondence metric; the latter substitutes the geometry in both correspondence and anchoring losses. Together, these comparisons provide empirical support for the combined supervision and the geometry used to organize its representations.

G Implementation Details of HIP

This appendix specifies the training configuration, feature construction, and loss reductions. We use the main-text notation: $d_{\mathbb{B}_c^d}$ is the Poincaré distance, $\Phi_M = \exp_0 \circ g_M$ is the mapping for modality $M \in \{T, I\}$, and s, f, p_i, r_i are the resulting scene and constituent representations.

Backbone adaptation. Both backbones use 512×512 center crops and rank-64 LoRA adapters, with frozen VAE and text encoders. SD3.5-M is trained for one epoch on the 31,783-pair Flickr shard, using learning rate 10^{-4}, per-device batch size 1, and gradient accumulation over eight microsteps. FLUX uses LoRA in all 19 double-stream blocks and the attention projections of all 38 single-stream blocks, with learning rate 5×10^{-5}, two data-parallel workers, batch size 1 per worker, and accumulation 8. Its main comparison uses the step-750 checkpoint, selected before evaluation, corresponding to 12,000 original pair presentations. All FLUX runs use the native timestep normalization and attention output path.

Features and projection order. Text features are read after the frozen context embedder, before joint attention. SD3.5-M combines its CLIP and T5 conditioning streams; FLUX uses its T5 sequence. The image features are outputs of the denoiser processing the noisy latent: SD3.5-M reads them after final adaptive normalization and before the output projection, while FLUX reads the image slice after the single-stream blocks and before final normalization. The feature dimensions are $d_T = d_I = 1536$ on SD3.5-M and $d_T = d_I = 3072$ on FLUX. Each modality head has two linear layers with a GELU between them, followed by $\exp_0$; the hidden width equals the input width and the common output dimension is $d = 256$. Text tokens are projected before sentence/phrase averaging. Image features are pooled before projection:

$$\begin{aligned} s &= \frac{1}{|\Omega_T|} \sum_{a \in \Omega_T} z_a^T, \quad & p_i &= \frac{1}{|S_i|} \sum_{a \in S_i} z_a^T, \\ f &= \Phi_I(\bar{h}^I), & r_i &= \Phi_I(\bar{h}_i^I). \end{aligned} \tag{24}$$

Here $z_a^T = \Phi_T(h_a^T)$, and $\bar{h}^I, \bar{h}_i^I$ are the global and phrase-conditioned feature means in Eq. (6). The index set Ω_T contains the valid prompt-token positions, including valid special-token positions; its indicator gives the sentence mask. Ordinary phrase masks select only T5 positions; sentence masks additionally include the valid CLIP positions on SD3.5-M.

Phrase extraction and masking. The configured parser is spaCy. The loader tries `en_core_web_sm`, then `en_core_web_md`; noun chunks are preferred, with sentence spans and a rule-based splitter as fallbacks. Duplicate entries with identical lowercased text and character offsets are removed, while repeated phrases at different offsets remain distinct. The phrases are retained and restored to sentence order. Character spans are matched to overlapping, nonempty T5 tokenizer offsets, with sequence limits of 256 tokens on SD3.5-M and 512 on FLUX. Phrases with no surviving tokens are invalid for contrastive learning; if none survive, the first slot uses the valid full-prompt mask. The local verification environment uses spaCy 3.8.14 and `en_core_web_sm` 3.8.0.

Attention aggregation and normalization. Attention uses text queries and image keys from the full joint text–image softmax. All heads are averaged within each selected block, followed by averaging across blocks and over phrase tokens. The fixed block sets are 19–22 for SD3.5-M and

15–18 for FLUX (zero-based); the latter are double-stream blocks. Using $\bar{A}^{T\to I}_{aj}$ for this head/block average, as in Section 3.2, define

$$\widetilde{w}_{ij} = \frac{1}{|S_i|}\sum_{a\in S_i}\bar{A}^{T\to I}_{aj}, \qquad w_{ij} = \frac{\widetilde{w}_{ij}}{\sum_{j'=1}^{n_I}\widetilde{w}_{ij'}}. \tag{25}$$

For positive attention mass, these weights sum to one, and Eq. (24) forms each region as a weighted mean of image features. Invalid phrase slots and zero-mass rows receive zero weights.

Contrastive direction and distributed reduction. The phrase–region InfoNCE uses one text-to-image direction. The batch in Eq. (10) is the image–caption microbatch across participating workers. Each worker retains its local text anchors and gathers visual candidates from all workers using differentiable `all_gather`. Variable-length candidate tensors are padded for the collective and then unpadded. The per-phrase losses ℓ_ν are averaged over local anchors, and data-parallel training averages worker losses to obtain $\mathcal{L}_{pr}$. This equal-worker reduction gives each participating worker the same weight, including when valid phrase counts differ. Regions associated with other phrases from the same image also serve as instance-level negatives; candidate identities follow phrase instances, including repeated or synonymous phrases. A pool with fewer than two pairs contributes zero contrastive loss. Gradient accumulation sets the optimizer batch size, while each microstep has its own cross-worker candidate pool.

Objectives for both backbones. Using the losses defined in Sec. 3.3, the full training objectives are

$$\mathcal{L}_{\text{SD3.5}} = \mathcal{L}_{\text{diff}} + \omega_{\text{warm}}(k)\big(0.05\mathcal{L}_{pr} + 0.02\mathcal{L}_T + 0.02\mathcal{L}_I\big), \tag{26}$$

$$\mathcal{L}_{\text{FLUX}} = \mathcal{L}_{\text{diff}} + \omega_{\text{warm}}(k)\big(0.05\mathcal{L}_{pr} + 0.02\mathcal{L}_T + 0.02\mathcal{L}_I\big), \tag{27}$$

Here k indexes training microsteps, before gradient accumulation; it is distinct from the diffusion time t and the optimizer-update count. The warmup factor $\omega_{\text{warm}}(k)$ increases linearly from zero to one over the first 200 microsteps and remains one thereafter. Equation (12) presents the objective with the full auxiliary weights. SD3.5-M uses $\tau = 0.05$; FLUX uses $\tau = 0.07$. Both use $c = 1$ and $m = 0.1$.

Trainable parameters. We train the two projection heads together with LoRA adapters (Hu et al., 2022), keeping the VAE and text-encoder parameters frozen. Gradients flow through the MLP projectors, visual features, and captured MMDiT attention. Cross-modal alignment and image whole–part anchoring can update LoRA through final image features; phrase-conditioned regions additionally provide a path through the differentiable attention weights. The pre-MMDiT text states are frozen under the configured adapter targets, so text whole–part anchoring updates the text head, shaping the phrase representations used in the joint contrastive objective.

Inference model. For each adapted linear map, LoRA learns low-rank factors $A_{\text{L}}, B_{\text{L}}$ with update $\Delta W = B_{\text{L}}A_{\text{L}}$, which can be merged as $W_{\text{final}} = W + B_{\text{L}}A_{\text{L}}$ before deployment. The merged matrix has the same shape as W. Discarding the projection heads and attention readout then leaves the original generator architecture and parameter count, with no auxiliary losses, region construction, or hyperbolic operations during sampling.

Numerical safeguards. Training is configured for BF16 mixed precision with Accelerate. Attention aggregation and weighted visual pooling explicitly accumulate in float32; the pooled visual vectors are cast back to the feature dtype before projection. The geometry routines use $\varepsilon = 10^{-5}$: the exponential-map output is radially clipped toward $(1-\varepsilon)/\sqrt{c}$, the distance denominator is clamped below at ε, and the arcosh argument is clamped below at $1+\varepsilon$. Geometry operations follow their active operation dtypes within mixed-precision training. In float32 the self-distance floor is approximately $\operatorname{arcosh}(1+\varepsilon)/\sqrt{c}$. The analysis uses the exact Poincaré metric; training evaluates its numerically stabilized form.

H PHRASE–REGION CORRESPONDENCE EVALUATION

This appendix specifies the representation diagnostic in Sec. 4.3: whether a phrase is closer to its corresponding attention-pooled region than to other regions of the same image. It evaluates representations of real image–caption pairs, independently of the generation benchmarks.

Examples and candidate regions. We use 128 COCO-val image–caption pairs, disjoint from Flickr30k-Entities training, with 414 valid phrases in total. Each image has $K \in \{2, 3, 4\}$ phrases. All checkpoints receive the same images, captions, phrase extraction, deterministic noise, and native diffusion timestep index 500. For phrase i, this attention pools the image features into $\bar{h}_i^I$, which is then mapped to the region representation r_i as in Eq. (8). The candidate set for p_i consists of all K regions of that image: r_i is the matched candidate and each r_j, $j \neq i$, is an incorrect candidate. No regions from other images enter this diagnostic. Since phrase attention defines the region correspondence, the diagnostic tests internal consistency on unseen pairs; the separate spatial evaluation in App. I uses annotated phrase boxes.

Training conditions and evaluation distances. The four SD3.5-M conditions in Tables 2 and 8 are defined as follows.

- **Vanilla fine-tuning**: denoising-only LoRA adaptation with HIP's projection-head capacity and a zero auxiliary-loss weight.
- **Hierarchy + cosine**: the SD3.5-M checkpoint trained with cosine correspondence and whole–part anchoring, introduced in Sec. 4.1. The diagnostic evaluates its ball embeddings with Poincaré distance; "cosine" identifies its training objective. Figures 4 and 6 also show generations from this checkpoint; the appendix additionally includes the separately identified, hierarchy-free cosine-identity control.
- **Euclidean-L2**: the geometry substitution for HIP's $\tau = 0.05$ configuration. It preserves the data, backbone, optimizer settings, projection-head dimensions, and auxiliary weights, replaces $\exp_0$ by the identity, and uses $d_{\text{eval}}(z, z') = \|z - z'\|_2$ for both phrase–region correspondence and intra-modal anchoring. Evaluation uses the same Euclidean distance, directly testing the effect of the ambient geometry on learned representations. For this control, s, f, p_i, r_i denote final representations in $\mathbb{R}^d$.
- **HIP**: the full objective in Eq. (12), with $\tau = 0.05$ and Poincaré distance for training and evaluation. The no-auxiliary control also uses Poincaré distance in the diagnostic.

Top-1 retrieval and the nearest incorrect region. Write $d_{\text{eval}} = d_{\mathbb{B}_c^d}$ for the ball-based conditions and $d_{\text{eval}}(z, z') = \|z - z'\|_2$ for Euclidean-L2. For one image, define $D_{ij} = d_{\text{eval}}(p_i, r_j)$ using the evaluation distance above. The closest candidate to p_i has index $\hat{j}_i = \arg\min_j D_{ij}$. The image's retrieval accuracy and nearest-negative gap are

$$\begin{aligned} \text{Top1} &= \frac{1}{K} \sum_{i=1}^{K} \mathbf{1}\{\hat{j}_i = i\}, \\ \Delta_D &= \frac{1}{K} \sum_{i=1}^{K} \left[\min_{j \neq i} D_{ij} - D_{ii} \right]. \end{aligned} \tag{28}$$

The nearest incorrect region is the *hardest negative*: the competitor most likely to be confused with the matched region. A positive gap for a phrase means that its matched region is closer than every incorrect candidate. Larger gaps indicate stronger separation: matched distance 0.2 and nearest-incorrect distance 0.7 give a margin of $+0.5$. Top-1 measures the frequency of correct retrieval, and the average gap measures the mean separation from the strongest competing regions, providing complementary views of correspondence quality.

As an additional mismatch check, the audit pairs each phrase with the next phrase's region, wrapping the last phrase back to the first: $(p_1, r_2), \ldots, (p_{K-1}, r_K), (p_K, r_1)$. This fixed one-position rotation is the *cyclic shuffle*. It supplies a mismatched reference without searching for the nearest incorrect region; Tables 2 and 8 report the more demanding nearest-negative comparison.

Table 8: **Phrase–region correspondence with 95% bootstrap confidence intervals.** Entries are image-averaged means over 128 examples, followed by intervals in brackets, rounded to two decimal places. HIP Top-1 is 0.926 before rounding. HIP improves retrieval over Euclidean-L2 and has a positive Gromov gap with its interval above zero. Sentence–image distance is $d_{\text{eval}}(s, f)$.

Method	Nearest-neg. gap ↑	Top-1 ↑	Gromov gap ↑	Sentence–image distance ↓
Vanilla fine-tuning	-0.02 $[-0.02, -0.01]$	0.32 $[0.30, 0.34]$	-0.01 $[-0.01, -0.01]$	1.88 $[1.86, 1.90]$
Hierarchy + cosine	-0.04 $[-0.04, -0.03]$	0.39 $[0.36, 0.42]$	0.01 $[0.01, 0.01]$	0.86 $[0.85, 0.87]$
Euclidean-L2	0.16 $[0.14, 0.17]$	0.89 $[0.86, 0.91]$	0.04 $[0.04, 0.05]$	0.23 $[0.23, 0.24]$
HIP	0.18 $[0.16, 0.19]$	0.93 $[0.90, 0.95]$	0.09 $[0.08, 0.10]$	0.45 $[0.44, 0.46]$

Gromov-product gap. Using the standard Gromov product (Gromov, 1987; Bridson & Haefliger, 1999) with the embedding origin 0 as basepoint, we define the following matched-versus-incorrect diagnostic gap:

$$G_{ij} = (p_i \mid r_j)_0 = \frac{d_{\text{eval}}(0, p_i) + d_{\text{eval}}(0, r_j) - d_{\text{eval}}(p_i, r_j)}{2},$$
$$\Delta_G = \frac{1}{K} \sum_{i=1}^{K} \left[G_{ii} - \max_{j \neq i} G_{ij} \right]. \tag{29}$$

In a rooted tree metric, the Gromov product equals the length of the path shared by two nodes from the root, as derived in App. B. We use this identity to motivate an origin-referenced geometric measure of shared hierarchical affinity. The origin is the zero embedding, while s and f denote the learned scene-level representations. For any incorrect candidate $j \neq i$, the affinity difference satisfies the exact identity

$$G_{ii} - G_{ij} = \frac{D_{ij} - D_{ii} + d_{\text{eval}}(0, r_i) - d_{\text{eval}}(0, r_j)}{2}. \tag{30}$$

Thus the Gromov comparison combines phrase–region discrimination with the candidates' radial organization. A positive per-phrase gap means that its matched region has greater shared-root affinity than every incorrect candidate. HIP's mean gap of 0.09 with confidence interval $[0.08, 0.10]$ supports this preference on the evaluated images. Together with its improved Top-1 accuracy over Euclidean-L2 (0.93 versus 0.89), this provides evidence that the hyperbolic objective improves correspondence while organizing matched concepts coherently in the shared space. Table 8 reports the associated metric values and uncertainty.

Sentence–image distance and consistency bounds. The final column of Table 8 reports $d_{\text{eval}}(s, f)$, the distance between the two scene-level representations. It is the root distance in Observation 1. The whole–part consistency residuals $[d_{\text{eval}}(s, p_i) - m]_+$ and $[d_{\text{eval}}(f, r_i) - m]_+$, together with matched distances D_{ii}, determine the right-hand side of Eq. (14). The sibling-distance discrepancy in Eq. (22) compares distances between phrases with those between their corresponding regions. A bound's *slack* is its right-hand side minus its left-hand side; nonnegative slack verifies consistency with the metric inequality. Retrieval and matched-versus-incorrect gaps assess the learned correspondence.

Averaging and uncertainty. Each metric is first computed for one image and then averaged over the 128 images, giving every image equal weight regardless of its phrase count. The reported 95% confidence intervals use 2,000 bootstrap resamples: each resample draws 128 images with replacement and recomputes the mean; the interval endpoints are the 2.5th and 97.5th percentiles of these means. These intervals quantify evaluation-sample uncertainty for fixed checkpoints. Table 9 gives the image-level averages by phrase count K for the three Poincaré-space conditions. Each group consists of the images with that phrase count, characterizing correspondence at different candidate-set sizes.

I Spatial Evaluation of Phrase Attention

This diagnostic asks whether phrase-conditioned attention is spatially concentrated on the corresponding annotated objects. It uses 128 real Flickr30k-Entities images with 442 phrase instances and

Table 9: **Correspondence by phrase count for the Poincaré-space conditions.** K is the number of phrases and candidate regions in an image, so each phrase has $K-1$ incorrect candidates. Values average images within each group. HIP maintains high retrieval accuracy and positive nearest-negative gaps across all three candidate-set sizes.

	Top-1 ↑			Nearest-negative gap ↑		
Method	K=2	K=3	K=4	K=2	K=3	K=4
Vanilla fine-tuning	0.50	0.33	0.24	0.00	−0.02	−0.02
Hierarchy + cosine	0.61	0.38	0.30	0.03	−0.03	−0.07
HIP	1.00	0.96	0.86	0.28	0.18	0.13

their box annotations, at native diffusion timestep index 500. The images come from the training-data manifest, so the results characterize the learned attention readout on this subset. The unseen COCO-val evaluation in App. H separately measures internal correspondence without box labels. Phrase-box annotations from dataset are reserved for this spatial evaluation; training regions and attention-block selection follow the annotation-free procedure in Sec. 3.2.

Diagnostic checkpoints. This evaluation uses a separate set of SD3.5-M checkpoints: a denoising-only **No auxiliary** control, a hierarchy-free **Cosine identity** control, and **HIP** checkpoint. Cosine identity uses cosine alignment with an identity projector. This spatial study uses its own checkpoint set, while Table 2 evaluates the current representation controls. The two studies provide complementary evidence on attention localization and embedding organization.

Box mass and pointing accuracy. For phrase i, use the normalized spatial weights w_{ij} from Eq. (5), with $\sum_j w_{ij} = 1$. Let $\mathcal{B}_i$ be the set of image-token locations covered by that phrase's annotated box mask. Its *box mass* is $\sum_{j\in\mathcal{B}_i} w_{ij}$, and the area-normalized score is

$$\mathrm{BoxMass/Area}_i = \frac{\sum_{j\in\mathcal{B}_i} w_{ij}}{|\mathcal{B}_i|/n_I}.$$

A uniform attention map scores one; a larger value indicates greater concentration inside the box than its area alone would predict. *Pointing accuracy* is the fraction of phrases whose highest-attention image token lies inside the corresponding box mask.

From sentence attention to phrase attention. Write $\mathbf{w}_i = (w_{ij})_{j=1}^{n_I}$ for phrase i's spatial map and $\mathbf{w}_s$ for the normalized sentence-level attention map. For any normalized spatial map $\mathbf{w} = (w_j)_{j=1}^{n_I}$, define normalized entropy as

$$\mathcal{H}(\mathbf{w}) = -\frac{\sum_{j=1}^{n_I} w_j \log w_j}{\log n_I}, \qquad 0\log 0 = 0.$$

Higher entropy means that attention is more spatially spread out. Whole–part attention specificity is $\mathcal{H}(\mathbf{w}_s) - K^{-1}\sum_{i=1}^{K}\mathcal{H}(\mathbf{w}_i)$: it is positive when phrase maps are, on average, more concentrated than the sentence map. This measures the progression from broad scene attention to spatially specific phrase attention.

Overlap between phrases. For each phrase i, let $\mathcal{T}_i$ contain its $n_{\mathrm{top}} = \max(1, \mathrm{round}(0.1 n_I))$ highest-attention image-token locations. The *sibling overlap* is the mean Jaccard overlap $|\mathcal{T}_i \cap \mathcal{T}_j|/|\mathcal{T}_i \cup \mathcal{T}_j|$ over pairs of distinct phrases from the same image. "Sibling" refers to phrases sharing a caption, and the retained evaluation uses the top 10% of locations. Lower overlap indicates that different phrases concentrate on more distinct locations. Box mass and pointing accuracy complement this separation measure by evaluating spatial agreement with annotated objects.

Mismatched-box control. The shuffled-box column reassigns box masks by a one-position circular shift of the phrase-box list: each attention map is scored against the next entry's box, with the final entry wrapping back to the first. The same box-mass/area calculation is then repeated with the reassigned box. This breaks the intended phrase–box correspondence while preserving the set of box areas, providing a reference for concentration explained by box size and spatial priors.

Method	Box mass / area ↑	Pointing ↑	Specificity ↑	Sibling overlap ↓	Shuffled box mass / area
No auxiliary	3.47	0.83	0.04	0.16	1.58
Cosine identity	3.68	0.82	0.04	0.17	1.69
HIP	5.11	0.82	0.09	0.12	1.76



Figure 7: **Attention-induced phrase–region readout.** Flickr30k-Entities diagnostic examples showing the image, sentence-level attention, and phrase-level attention maps used to construct region tokens. Boxes are the corresponding phrase annotations from the same real image.

J Image Quality Metrics

Table 10 provides the quality means for the SD3.5-M checkpoints compared in the main experiments. clean-fid uses Inception features and the clean preprocessing mode on 10,000 generated images and 10,000 matched COCO Karpathy validation/test reference images. CLIP uses normalized image and caption embeddings from ViT-L/14, with captions truncated to the evaluator's 77-token limit; its reported value is the raw cosine similarity. MUSIQ (Ke et al., 2021) uses the retained `pyiqa` evaluator. FID measures distance to the reference distribution, whereas no-reference image scores and learned image–text preferences assess different properties of individual samples. Together, these metrics characterize distributional fidelity, perceptual quality, and semantic agreement (Stein et al., 2023; Jayasumana et al., 2024).

Table 10: **SD3.5-M image quality and text–image agreement on COCO-10k.** All methods use the same 10,000 captions and matched reference set. Bold marks the best unrounded score; displayed values are rounded to two decimal places. Metric definitions and uncertainty estimates are given in App. J.

Method	FID10k ↓	CLIP cosine ↑	MUSIQ ↑	Aesthetic ↑	PickScore ↑	ImageReward ↑
Pretrained baseline	24.93	0.27	73.29	6.08	**22.38**	1.00
Vanilla fine-tuning	**23.67**	0.28	72.17	6.02	22.21	1.02
Hierarchy + cosine	23.68	0.28	72.88	6.02	22.21	1.02
SoftREPA	31.28	0.27	**74.28**	6.13	22.38	1.11
HIP	24.73	**0.28**	72.60	**6.13**	22.23	**1.17**

Aesthetic and preference scoring. Aesthetic Score uses the original LAION V1 linear predictor on L2-normalized OpenAI CLIP ViT-L/14 image embeddings (LAION-AI, 2022), specifically `sa_0_4_vit_l_14_linear.pth`. PickScore (Kirstain et al., 2023) uses `PickScore_v1` with its ViT-H/14 processor and reports $\exp(\beta_{\mathrm{PS}})\langle \hat{e}_I^{\mathrm{PS}}, \hat{e}_T^{\mathrm{PS}} \rangle$, where β_{PS} is the checkpoint's learned logit scale and $\hat{e}_I^{\mathrm{PS}}, \hat{e}_T^{\mathrm{PS}}$ are its L2-normalized image and text embeddings. We report the mean raw paired score across the evaluation set. ImageReward (Xu et al., 2023) uses the official `ImageReward-v1.0` scorer and its native standardized reward for each caption–image pair. All three scorers run in float32, use their prescribed image preprocessing, and evaluate the same saved images and COCO captions across all five checkpoints. Per-image indices, caption strings and image hashes are retained for paired comparisons.

Calibration and uncertainty. Single-image versus six-image batch checks give maximum score differences of 9.54×10^{-6} for Aesthetic and 3.82×10^{-6} for PickScore. ImageReward reproduces the ranking of the four official example images, with a maximum absolute deviation of approximately 0.0029 from their published scores across the checked CPU/GPU and dependency configurations. Table 11 distinguishes the retained CLIP/MUSIQ bootstrap standard errors (1,000 resamples, seed 42) from the added scorers' sample standard errors, computed as sample standard deviation divided by $\sqrt{10{,}000}$.

Table 11: **Standard errors for the COCO-10k quality and preference means.** Means appear in Table 10; each estimate uses 10,000 images. All displayed entries are in units of 10^{-3}. CLIP and MUSIQ use retained bootstrap errors, while Aesthetic, PickScore and ImageReward use sample standard deviation divided by $\sqrt{n}$ with $n = 10{,}000$.

	Bootstrap SE		Sample SE		
Method	CLIP	MUSIQ	Aesthetic	PickScore	ImageReward
Pretrained baseline	0.34	51.88	8.31	10.90	6.81
Vanilla fine-tuning	0.33	51.60	8.59	10.64	6.45
Hierarchy + cosine	0.34	50.10	8.56	10.61	6.47
SoftREPA	0.34	43.37	8.16	10.62	6.31
HIP	0.34	43.40	8.47	10.53	5.66

K Evaluation Reproducibility

Training and checkpoint identity. The training configurations are given in App. G. The SD3.5-M generation and representation results use the same HIP checkpoint with $\tau = 0.05$ and attention-induced regions. Its vanilla control has the same projection-head capacity but applies no auxiliary gradient. The SD3.5-M SoftREPA results use step 30,000 with four original pairs per update, totaling 120,000 pair presentations, or approximately 3.78 epochs on the 31,783-pair training set; HIP

uses one epoch. For FLUX, vanilla fine-tuning, hierarchy + cosine, and HIP share the data snapshot, seed, LoRA rank and targets, and optimizer settings. Hierarchy + cosine changes the correspondence metric to cosine after the logarithmic map and normalization. For FLUX, we evaluate SoftREPA at 3,000 updates and the LoRA-based methods at 750 updates. With effective batch sizes of 4 and 16, respectively, each evaluated checkpoint has processed 12,000 original image–caption pairs, equivalent to approximately 0.378 epochs on the 31,783-pair training set. It showns that HIP has better data efficiency than SoftREPA.

Cross-worker contrastive candidates are gathered within each microstep, independently of gradient accumulation. In the two-worker FLUX configuration, each microstep contains two image–caption examples and up to eight phrase–region pairs. The 200-microstep auxiliary warmup therefore spans 25 optimizer updates with accumulation 8.

Generation and scoring. The original GenEval evaluations use 553 prompts with four images per prompt; task sizes are 80 single-object, 99 two-object, 80 counting, 94 color, 100 position and 100 color-attribute prompts. SD3.5-M uses 20 inference steps and guidance 7.0, and FLUX uses 50 steps and guidance 3.5, all at 512^2 with seed base 42. GenEval v2 uses the corrected class-name mapping, a Mask2Former COCO-instance detector and CLIP color classification. The corrected scorer is used for the six original SD3.5-M checkpoint rows in Table 3, including vanilla, hierarchy + cosine, cosine identity, and HIP ($\tau = 0.05$); rescoring reuses the retained 553 images for each checkpoint. The table additionally includes the supplementary Hierarchy + Euclidean results. Macro accuracy weights the six tasks equally, while prompt accuracy weights all prompts equally. Benchmark gain rows use $100(S_{\mathrm{HIP}}/S_{\mathrm{base}} - 1)\%$, where S_{HIP} and S_{base} are the unrounded benchmark scores of HIP and its baseline.

K.1 Training and inference cost

Denoiser workload. For a raw batch of B image–caption pairs, the SoftREPA training implementation constructs all B^2 image–text combinations. With $B = 4$, this gives 16 denoiser condition instances for four original pairs. HIP processes each matched pair once and computes contrastive comparisons between the resulting projected features. Thus SoftREPA evaluates four times as many denoiser condition instances per original pair in this configuration.

Recorded training throughput. SD3.5-M runs provide a practical timing reference. A one-epoch HIP run processed 31,782 pairs in approximately 1.42 hours of training, or 6.23 original pairs per second; including initialization, the run took 1.58 hours. SoftREPA processed 120,000 pairs in approximately 16.11 hours, or 2.07 pairs per second. Normalizing SoftREPA's throughput to one epoch yields an estimated training time of 4.27 hours, which is much slower than 1.42 hours for HIP under the recorded training configurations.

Inference computation and parameter count. The SD3.5-M SoftREPA checkpoint retains 849,920 learned parameters: four soft tokens in each of five layers, together with timestep-dependent embeddings for 100 time buckets. These tokens remain active during sampling and introduce token preparation and additional attention computation at each denoising step. The FLUX SoftREPA configuration has 1.618M learned parameters. HIP's LoRA and projection parameters are training-time quantities: the FLUX configuration trains 104.595M LoRA and 20.454M projection parameters, but merging the former and discarding the latter leaves zero additional parameters in the final generator. HIP therefore avoids the persistent soft-token overhead and preserves the backbone's inference workload. This architectural comparison supports an expected inference-speed advantage.