HIVE: HIERARCHICAL INDUCTION OF VISUAL EMBEDDINGS FOR PART–WHOLE REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual recognition benefits from representations that reveal how local image evidence is organized into higher-level semantic structure, yet existing hierarchy-aware visual backbones often provide limited quantitative access to the induced part–whole relations. To address this, we propose *HIVE*, hierarchical induction of visual embeddings for part–whole representation learning, whose cross-level assignments are directly readable as a part–whole composition graph. HIVE maps image patches to fine-level embeddings, recursively organizes them into higher-level structures through graph-based reasoning and hierarchical pooling, and reintegrates multi-level representations into the token space for recognition. We further introduce *part–whole contribution analysis* to quantify directional fine-to-mid and mid-to-coarse composition and visualize image-specific parsing trees. Experiments show that HIVE serves as a competitive general-purpose ViT backbone, while producing multilevel embeddings that are class-relevant and structurally coherent.

1 INTRODUCTION

A central challenge in computer vision is to learn representations that support structured visual understanding. Beyond recognizing objects or predicting image-level categories, visual perception requires understanding how local details compose into larger semantic units (Sabour et al., 2017; Hinton et al., 2018; Locatello et al., 2020). As illustrated in Fig. 1, a bedroom scene can be interpreted through multiple levels of organization: local regions form parts and objects, such as the bed, wall, furniture, and surrounding elements, which together define the overall scene. This part–whole organization is important for object recognition, fine-grained classification, scene understanding, and hierarchical category prediction (Sabour et al., 2017; Huang et al., 2016; Xu et al., 2017; Bertinetto et al., 2020; Krishna et al., 2017).

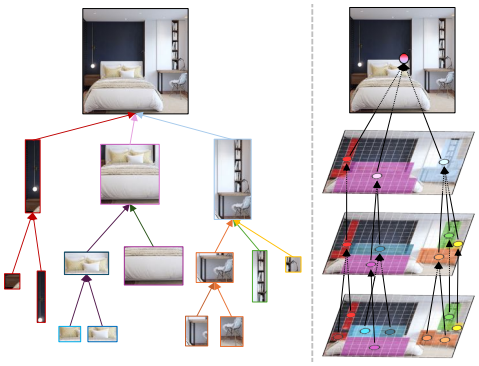

Figure 1: **(Left)** An illustrative hierarchical dependency structure. **(Right)** Decomposition of a visual scene into a part–whole hierarchy based on the semantic organization of its visual elements.

Modern deep networks have greatly improved visual representation learning. CNNs extract local spatial patterns (He et al., 2016), while Vision Transformers (ViTs) capture long-range interactions among image tokens through self-attention (Dosovitskiy et al., 2021; Touvron et al., 2021). Recent multi-scale and hierarchy-aware backbones further organize visual features across levels of abstraction (Chen et al., 2021; Wang et al., 2021; Liu et al., 2021; Chen et al., 2022; Tang et al., 2022). More explicit hierarchy-inducing approaches construct patch-level dependency trees (Ding et al., 2023) or decompose features through predefined probabilistic hierarchies (Kwon et al., 2024). Although these methods show that hierarchical structure can benefit recognition and dense prediction, the induced hierarchy is often assessed indirectly through downstream performance or illustrated with qualitative examples. They do not explicitly quantify how lower-level visual units support higher-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.

structures for each image, nor whether these relations form stable, directional, and semantically meaningful part–whole organization.

To address this gap, we propose *HIVE*, a visual hierarchy induction framework that learns multi-level visual embeddings directly from images. Without requiring relation annotations, HIVE encourages cross-level organization through self-supervised structural objectives while being trained for visual recognition. It constructs fine-level embeddings from image patches via competitive attention (Locatello et al., 2020), refines their relations with graph-attention-based reasoning (Veličković et al., 2018), and aggregates them through differentiable pooling (Ying et al., 2018) to obtain mid- and coarse-level embeddings for recognition. The novelty lies in an assignment-preserving recursive formulation that retains image-conditioned cross-level relations for both representation learning and part–whole analysis. HIVE further satisfies a structural consistency property under permutations of patches and learned embedding units, ensuring that the induced hierarchy is not an artifact of arbitrary indexing.

Our study makes two main contributions. First, we demonstrate that HIVE serves as a competitive general-purpose backbone across diverse visual recognition tasks, showing that the induced hierarchy supports recognition rather than functioning merely as an auxiliary interpretability mechanism. These results indicate that HIVE learns meaningful multilevel embeddings that are class-relevant, structurally coherent, and useful for both image-level and dense prediction settings. Second, we show that HIVE induces coherent part–whole organization and provides direct quantitative access to the learned cross-level relations. To examine these relations, we introduce *part–whole contribution analysis*, which measures directional contributions from lower-level embeddings to adjacent higher-level embeddings. This analysis reveals class-relevant fine-to-mid and mid-to-coarse compositions, enabling both example-specific parsing-tree visualization and quantitative profiling of the induced hierarchy.

2 Related Work

Learning Visual Hierarchies. Visual hierarchy has long been viewed as a core principle of visual understanding. Classical image parsing modeled scenes through Bayesian, stochastic grammar, and And–Or graph formulations (Tu et al., 2005; Zhu & Mumford, 2007), while capsule-based models revisited this idea by routing lower-level representations to higher-level entities (Sabour et al., 2017; Hinton et al., 2018). Modern architectures also encode hierarchy implicitly: CNNs build spatial feature hierarchies through convolution and pooling (He et al., 2016), whereas ViTs model long-range token interactions through self-attention (Dosovitskiy et al., 2020). Because such hierarchies are not explicitly organized, several methods have sought to reveal interlayer or feature-level structure after training (Kowal et al., 2024; Kim et al., 2026). Multi-scale and hybrid vision backbones extend this with cross-scale interaction, pyramid representations, token aggregation, or window-based attention (Chen et al., 2021; Wang et al., 2021; 2022; Liu et al., 2021), but they are often tied to resolution, token granularity, or downsampling. Recent methods more directly construct visual hierarchies through patch-level dependencies, hierarchy-aware attention, or predefined hierarchical templates (Ding et al., 2023; Geng et al., 2023; Kwon et al., 2024). In contrast, HIVE learns an emergent hierarchy over feature states rather than directly organizing image tokens or relying on a predefined hierarchy. Its distinguishing feature is an assignment-preserving recursive representation in which image-conditioned child–parent relations are retained across levels and remain directly accessible for both recognition and quantitative analysis of cross-level composition.

Structured Visual Representation Learning. Another line of work learns structured intermediate representations from visual features. Slot- and set-based methods organize perceptual inputs into compact latent units through attention, iterative refinement, or set-structured prediction (Locatello et al., 2020; Greff et al., 2019; Engelcke et al., 2020; Lee et al., 2019; Carion et al., 2020). Recent work on perceptual grouping further shows that contrastive vision–language models can preserve spatially coherent visual regions without dense supervision (Ranasinghe et al., 2023). Bidirectional interaction frameworks refine image embeddings by binding information from input features and projecting the refined representations back to the visual stream (Hong et al., 2026; 2024). While effective, these approaches typically operate at a single abstraction level.

Graph-based methods complement them by modeling relations among structured units. Message-passing networks learn node representations over structured inputs (Kipf & Welling, 2017; Hamilton et al., 2017), attention-based graph models adaptively weight neighboring nodes (Veličković

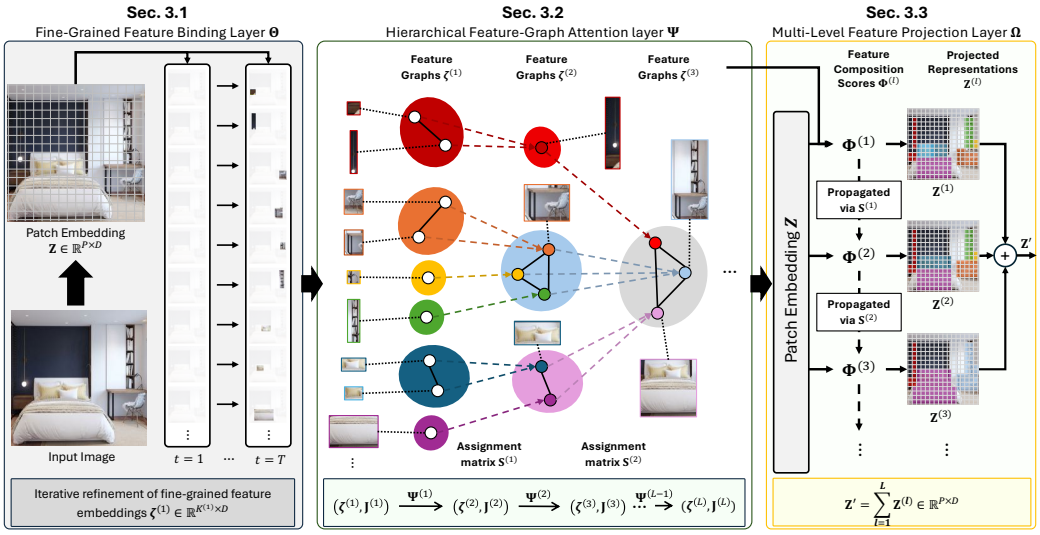

Figure 2: An architecture overview of HIVE, composed of: i) fine-grained feature binding $\boldsymbol{\Theta}$ (Sec. 3.1), ii) hierarchical feature-graph attention $\boldsymbol{\Psi}$ (Sec. 3.2), and iii) multi-level feature projection $\boldsymbol{\Omega}$ (Sec. 3.3).

et al., 2018), hierarchy-aware attention incorporates hierarchical similarity into attention computation (Tseng et al., 2023), and differentiable pooling constructs coarser graph representations through learned soft assignments (Ying et al., 2018). Building on these ideas, HIVE treats learned visual embeddings as nodes in a feature graph, performs graph-attention reasoning, and progressively coarsens the graph through differentiable pooling. This enables a structured multi-level visual representation rather than a single set of latent units, spatial groups, or a flat graph of image features.

3 Method

We propose *HIVE*, a hierarchy-inducing visual backbone framework consisting of three modules: fine-grained feature binding $\boldsymbol{\Theta}$ (Sec. 3.1), hierarchical feature-graph attention $\boldsymbol{\Psi}$ (Sec. 3.2), and multi-level feature projection $\boldsymbol{\Omega}$ (Sec. 3.3). HIVE captures hierarchical visual dependencies among patches without part–whole annotations, using self-supervised structural objectives during visual recognition training. Algorithm 1 summarizes the overall forward workflow of HIVE.

Notation. Throughout, $l \in \{1, \dots, L\}$ indexes the hierarchy level and $t \in \{0, \dots, T\}$ indexes the refinement step, with $l = 1$ denoting the lowest (finest) level. We use P for the number of patches, D for the embedding dimension, and $K^{(l)}$ for the number of feature states at level l. The input patch embeddings are denoted by $\mathbf{Z} \in \mathbb{R}^{P \times D}$, and the level-$l$ feature states by $\boldsymbol{\zeta}^{(l)} \in \mathbb{R}^{K^{(l)} \times D}$.

3.1 Fine-Grained Feature Binding Layer $\boldsymbol{\Theta}$

Following Locatello et al. (2020), the fine-grained feature binding layer initializes a compact set of input-dependent feature states from patch embeddings through competitive attention and recurrent refinement. Prior work (Hong et al., 2024; 2026) has shown that such feature extraction from image embeddings can be effective even without multiple objects or explicit object-centric structure.

At the lowest hierarchy level, we initialize a set of learnable fine-grained feature states $\boldsymbol{\zeta}_0^{(1)} = [\zeta_1, \dots, \zeta_{K^{(1)}}] \in \mathbb{R}^{K^{(1)} \times D}$, shared across images and refined conditioned on the input. At refinement step t, the current feature states $\boldsymbol{\zeta}_{t-1}^{(1)}$ attend competitively over the patch embeddings $\mathbf{Z}$:

$$\mathbf{A}^{\boldsymbol{\Theta}} = \mathrm{softmax}_{K^{(1)}}\left(q^{\boldsymbol{\Theta}}(\boldsymbol{\zeta}_{t-1}^{(1)})\, k^{\boldsymbol{\Theta}}(\mathbf{Z})^{\top} / \sqrt{D}\right) \in \mathbb{R}_{+}^{K^{(1)} \times P}.$$

We then normalize $\mathbf{A}^{\boldsymbol{\Theta}}$ along the patch dimension and aggregate value features:

$$\tilde{\mathbf{A}}^{\boldsymbol{\Theta}}_{k,p} \triangleq \mathbf{A}^{\boldsymbol{\Theta}}_{k,p} \Big/ \sum_{i=1}^{P} \mathbf{A}^{\boldsymbol{\Theta}}_{k,i}, \qquad \boldsymbol{u} = \tilde{\mathbf{A}}^{\boldsymbol{\Theta}} v^{\boldsymbol{\Theta}}(\mathbf{Z}) \in \mathbb{R}^{K^{(1)} \times D},$$

where $\tilde{\mathbf{A}}^{\mathbf{\Theta}} \in \mathbb{R}^{K^{(1)} \times P}$. The feature states are then updated with a Gated Recurrent Unit (GRU) (Cho et al., 2014):

$$\boldsymbol{\zeta}_t^{(1)} \triangleq \mathbf{\Theta}\Big(\boldsymbol{\zeta}_{t-1}^{(1)}, \boldsymbol{u}\Big) = \mathrm{GRU}\Big(\boldsymbol{\zeta}_{t-1}^{(1)}, \boldsymbol{u}\Big) \in \mathbb{R}^{K^{(1)} \times D}. \tag{1}$$

After T refinement steps, we denote the resulting fine-grained feature states by $\boldsymbol{\zeta}^{(1)} \triangleq \boldsymbol{\zeta}_T^{(1)}$.

3.2 Hierarchical Feature-Graph Attention Layer $\mathbf{\Psi}$

HIVE treats hierarchy as an internal computational mechanism over feature states. Starting from fine-grained feature states, the model constructs progressively coarser representations through graph-based relational updates (Veličković et al., 2018) and differentiable soft grouping (Ying et al., 2018).

Feature Graph Construction. For a transition from level l to level $l+1$, let $\boldsymbol{\zeta}^{(l)} \in \mathbb{R}^{K^{(l)} \times D}$ denote the level-l feature states and let $\mathbf{J}^{(l)} \in \mathbb{R}^{K^{(l)} \times K^{(l)}}$ denote the corresponding adjacency matrix. For the initial graph, $\mathbf{J}^{(1)}$ is constructed from $\boldsymbol{\zeta}^{(1)}$ using pairwise cosine similarity followed by min-max normalization. The neighborhood of node i is defined as $\mathcal{N}(i) = \{j \mid J_{ij}^{(l)} > 0\}$.

We first apply a shared linear transformation $\boldsymbol{W}_G^{\mathbf{\Psi}} \in \mathbb{R}^{D \times D}$ and compute masked attention coefficients:

$$e_{ij}^{(l)} \triangleq \mathrm{LeakyReLU}\Big(\boldsymbol{a}^\top \Big[\boldsymbol{W}_G^{\mathbf{\Psi}} \zeta_i^{(l)} \,\|\, \boldsymbol{W}_G^{\mathbf{\Psi}} \zeta_j^{(l)}\Big]\Big), \qquad j \in \mathcal{N}(i), \tag{2}$$

where $\boldsymbol{a} \in \mathbb{R}^{2D}$ is a learnable attention vector. The normalized coefficients are

$$\alpha_{ij}^{(l)} \triangleq \frac{\exp(e_{ij}^{(l)})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik}^{(l)})}, \qquad j \in \mathcal{N}(i), \tag{3}$$

which define the attention-induced adjacency matrix $\mathbf{A}^{(l)} \in \mathbb{R}^{K^{(l)} \times K^{(l)}}$, with entries $A_{ij}^{(l)} = \alpha_{ij}^{(l)}$.

With F attention heads, the updated node representation is

$$\bar{\zeta}_i^{(l)} \triangleq \sigma\left(\frac{1}{F}\sum_{f=1}^{F}\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(l),f} \boldsymbol{W}_U^{\mathbf{\Psi},f} \zeta_j^{(l)}\right), \tag{4}$$

where $\boldsymbol{W}_U^{\mathbf{\Psi},f} \in \mathbb{R}^{D \times D}$ is the learnable weight matrix of the f-th head. We denote the resulting updated node features by $\bar{\boldsymbol{\zeta}}^{(l)} \in \mathbb{R}^{K^{(l)} \times D}$.

Differentiable Hierarchical Pooling. To construct hierarchy, the graph-refined feature states are softly assigned to a smaller set of higher-level groups. We learn an assignment matrix $\mathbf{S}^{(l)} \in \mathbb{R}^{K^{(l)} \times K^{(l+1)}}$ from the updated node features $\bar{\boldsymbol{\zeta}}^{(l)}$ in Eq. (4):

$$\mathbf{S}^{(l)} \triangleq \mathrm{softmax}_{\mathrm{row}}\Big(\bar{\boldsymbol{\zeta}}^{(l)} \boldsymbol{W}_P^{\mathbf{\Psi}}\Big), \tag{5}$$

where a weight matrix $\boldsymbol{W}_P^{\mathbf{\Psi}} \in \mathbb{R}^{D \times K^{(l+1)}}$. The coarsened feature graph at level $l+1$ is then given by

$$\boldsymbol{\zeta}^{(l+1)} \triangleq \mathbf{S}^{(l)\top} \bar{\boldsymbol{\zeta}}^{(l)} \in \mathbb{R}^{K^{(l+1)} \times D}, \qquad \mathbf{J}^{(l+1)} \triangleq \mathbf{S}^{(l)\top} \mathbf{A}^{(l)} \mathbf{S}^{(l)} \in \mathbb{R}^{K^{(l+1)} \times K^{(l+1)}}. \tag{6}$$

Hierarchical Feature Graph Stacking. Before training, we specify the hierarchy depth L and the number of feature clusters at each level, $\{K^{(1)}, \ldots, K^{(L)}\}$. Starting from $(\boldsymbol{\zeta}^{(1)}, \mathbf{J}^{(1)})$, each level-specific module $\mathbf{\Psi}^{(l)}$ maps the feature graph at level l to a coarser feature graph at level $l+1$:

$$(\boldsymbol{\zeta}^{(l+1)}, \mathbf{J}^{(l+1)}) = \mathbf{\Psi}^{(l)}\Big(\boldsymbol{\zeta}^{(l)}, \mathbf{J}^{(l)}\Big), \qquad l = 1, \ldots, L-1. \tag{7}$$

3.3 Multi-Level Feature Projection Layer $\mathbf{\Omega}$

Prior work (Hong et al., 2024; 2026) has shown that projecting extracted feature information back to the input embedding is effective. Building on this idea, we represent the input embedding as a

multi-level internal state. To this end, we introduce a cross-attention mechanism (Vaswani, 2017) that integrates refined feature states from all hierarchy levels into the original input embedding $\mathbf{Z}$.

The feature composition score at the lowest level is

$$\mathbf{\Phi}^{(1)} \triangleq \mathrm{softmax}_{K^{(1)}}\left(q^{\mathbf{\Omega}}(\mathbf{Z})\, k^{\mathbf{\Omega}}(\boldsymbol{\zeta}^{(1)})^{\top}/\sqrt{D}\right) \in \mathbb{R}_{+}^{P\times K^{(1)}}, \tag{8}$$

where each entry of $\mathbf{\Phi}^{(1)}$ measures the relevance of a fine-grained feature to an input patch.

Higher-level composition scores are induced recursively as

$$\mathbf{\Phi}^{(l+1)} \triangleq \mathbf{\Phi}^{(l)}\mathbf{S}^{(l)} \in \mathbb{R}_{+}^{P\times K^{(l+1)}}, \qquad l = 1,\dots,L-1, \tag{9}$$

using the assignment matrix $\mathbf{S}^{(l)}$ from Eq. (5). Thus, $\mathbf{\Phi}^{(l)}$ measures patch-to-feature alignment at level l, while multiplication by $\mathbf{S}^{(l)}$ propagates this relevance to level $l+1$. Given $\{\mathbf{\Phi}^{(l)}\}_{l=1}^{L}$, we define the level-specific projected representation as $\bar{\mathbf{Z}}^{(l)} \triangleq \mathbf{\Phi}^{(l)} v^{\mathbf{\Omega}}(\boldsymbol{\zeta}^{(l)}) \in \mathbb{R}^{P\times D}$, where $\boldsymbol{\zeta}^{(l)}$ denotes the feature states at level l. The final projected representation is

$$\mathbf{Z}' \triangleq \mathbf{\Omega}(\mathbf{Z}, \{\boldsymbol{\zeta}^{(l)}, \mathbf{J}^{(l)}\}_{l=1}^{L}) = \sum_{l=1}^{L} \bar{\mathbf{Z}}^{(l)} \in \mathbb{R}^{P\times D}. \tag{10}$$

3.4 Training Objectives

Given a task-specific objective $\mathcal{L}_{\mathrm{task}}$, we optimize the framework end-to-end with additional regularization terms that encourage coherent and informative hierarchical representations.

Task-specific Loss $\mathcal{L}_{\mathrm{task}}$. This term corresponds to the downstream objective of interest. For visual classification, $\mathcal{L}_{\mathrm{task}}$ can be instantiated as the standard cross-entropy loss on the final logits.

Hierarchy Consistency Loss $\mathcal{L}_{\mathrm{cons}}$. To encourage structurally related feature nodes to remain close in the embedding space, we define a Laplacian smoothness objective following the classical graph-Laplacian smoothness principle (Belkin et al., 2006). We first construct $\mathbf{M}^{(l)} \triangleq \mathbf{S}^{(l)}\mathbf{S}^{(l)\top} \in \mathbb{R}^{K^{(l)}\times K^{(l)}}$, where $M_{ij}^{(l)} = \sum_{k=1}^{K^{(l+1)}} S_{ik}^{(l)} S_{jk}^{(l)}$, define the diagonal degree matrix $\mathbf{D}^{(l)}$ by $D_{ii}^{(l)} \triangleq \sum_j M_{ij}^{(l)}$, and let $\mathbf{L}^{(l)} \triangleq \mathbf{D}^{(l)} - \mathbf{M}^{(l)}$. The resulting loss is $\mathcal{L}_{\mathrm{cons}} \triangleq \frac{1}{L-1}\sum_{l=1}^{L-1} \mathrm{Tr}\big(\boldsymbol{\zeta}^{(l)\top}\mathbf{L}^{(l)}\boldsymbol{\zeta}^{(l)}\big)$.

Hierarchy Alignment Loss $\mathcal{L}_{\mathrm{align}}$. To align higher-level feature embeddings with pooled lower-level representations, we define $\mathcal{L}_{\mathrm{align}} \triangleq \frac{1}{L-1}\sum_{l=1}^{L-1}\frac{1}{B}\sum_{b=1}^{B}\left\|\mathbf{S}_b^{(l)\top}\boldsymbol{\zeta}_b^{(l)} - \boldsymbol{\zeta}_b^{(l+1)}\right\|_F^2$, where $\mathbf{S}_b^{(l)} \in \mathbb{R}^{K^{(l)}\times K^{(l+1)}}$ and $\boldsymbol{\zeta}_b^{(l)} \in \mathbb{R}^{K^{(l)}\times D}$ denote the assignment matrix and feature embeddings for sample b.

Hierarchy Diversity Loss $\mathcal{L}_{\mathrm{div}}$. To discourage redundant hierarchical explanations, we encourage feature contribution maps within the same level to be distinct. Let $\mathbf{\Phi}^{(l)} = [\phi_1^{(l)},\dots,\phi_{K^{(l)}}^{(l)}] \in \mathbb{R}^{P\times K^{(l)}}$, where each $\phi_i^{(l)} \in \mathbb{R}^P$ denotes the contribution map of the i-th feature at level $l \geq 2$. We define $\mathcal{L}_{\mathrm{div}} \triangleq \frac{1}{L-1}\sum_{l=2}^{L}\frac{1}{K^{(l)}(K^{(l)}-1)}\sum_{i\neq j}\left(\frac{\phi_i^{(l)\top}\phi_j^{(l)}}{\|\phi_i^{(l)}\|_2\,\|\phi_j^{(l)}\|_2+\epsilon}\right)^2$, where $\epsilon > 0$ is a small constant.

Link Prediction Loss $\mathcal{L}_{\mathrm{link}}$. Following Ying et al. (2018), we encourage the learned assignments to preserve local graph structure by defining $\mathcal{L}_{\mathrm{link}} \triangleq \frac{1}{L-1}\sum_{l=1}^{L-1}\left\|\mathbf{J}^{(l)} - \mathbf{S}^{(l)}\mathbf{S}^{(l)\top}\right\|_F^2$.

Final Objective $\mathcal{L}_{\mathrm{final}}$. The overall training objective is $\mathcal{L}_{\mathrm{final}} \triangleq \mathcal{L}_{\mathrm{task}} + \lambda_{\mathrm{cons}}\mathcal{L}_{\mathrm{cons}} + \lambda_{\mathrm{align}}\mathcal{L}_{\mathrm{align}} + \lambda_{\mathrm{div}}\mathcal{L}_{\mathrm{div}} + \lambda_{\mathrm{link}}\mathcal{L}_{\mathrm{link}}$, where $\lambda_{\mathrm{cons}}, \lambda_{\mathrm{align}}, \lambda_{\mathrm{div}}, \lambda_{\mathrm{link}} \geq 0$ are hyperparameters.

3.5 Hierarchical Permutation Consistency Property

We identify a key structural property of HIVE: the learned hierarchical visual representation is *hierarchically permutation-consistent*. Intuitively, re-indexing patch embeddings or learned feature states does not create a different hierarchy; it only relabels the same hierarchy consistently.

Table 1: Classification results on ImageNet-1K. The Type column denotes standard CNN/ViT backbones and hierarchy-aware vision models. Bold indicates the best performance.

Type	Model	Params (M) ↓	FLOPs (G) ↓	Top-1 Acc. (%) ↑
Standard CNN/ViT	ResNet-50 (He et al., 2016)	25.6	4.1	78.5
	ResNet-101 (He et al., 2016)	44.7	7.9	79.8
	DeiT-T/16 (Touvron et al., 2021)	5.7	1.3	72.2
	DeiT-S/16 (Touvron et al., 2021)	22.2	4.5	79.8
	Swin-T (Liu et al., 2021)	28.3	4.6	81.3
	ViT-T/16 (Dosovitskiy et al., 2020)	5.8	1.3	72.4
	ViT-S/16 (Dosovitskiy et al., 2020)	22.1	4.2	76.5
Hierarchy-aware	CrossViT-S (Chen et al., 2021)	26.7	5.6	81.0
	PVT-Small (Wang et al., 2021)	24.5	3.8	79.8
	PVTv2-B2 (Wang et al., 2022)	25.4	4.0	82.0
	Hi-Mapper (DeiT-S) (Kwon et al., 2024)	23.3	4.8	82.6
	DependencyViT-S (Ding et al., 2023)	24.0	5.0	82.1
Ours	**HIVE-T**	6.4	1.7	$\mathbf{75.2 \pm 0.01}$
	HIVE-S	24.6	5.0	$\mathbf{82.8 \pm 0.02}$

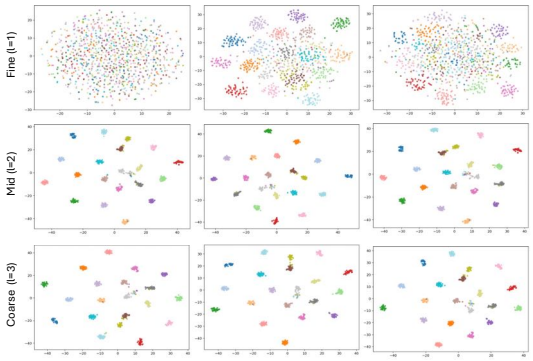

Figure 3: t-SNE projections of feature embeddings at each hierarchy level on a random 20-class ImageNet subset. Each panel visualizes the instantiations of one feature state, colored by class.

Proposition 1. *Using the notation, and equations (5), (6), (9), and (10), let* $\mathbf{P}$ *be a permutation matrix acting on patch embeddings and let* $\mathbf{Q}^{(l)}$ *be a permutation matrix acting on the level-*l *feature states. Then HIVE is hierarchically permutation-consistent: permuting the input patch embeddings and the learned feature states induces only the corresponding permutation of the hierarchical quantities* $\boldsymbol{\zeta}^{(l)}$, $\boldsymbol{\Phi}^{(l)}$, $\mathbf{J}^{(l)}$, $\mathbf{S}^{(l)}$, $\bar{\mathbf{Z}}^{(l)}$, $\mathbf{Z}'$, *without changing the underlying relational structure of the learned hierarchy.*

The proof is provided in Sec. A.2 in the Appendix. This property establishes that the learned visual hierarchy is independent of arbitrary indexing choices for patches and feature states. It does not by itself establish the semantic correctness of the induced relations; the structural and functional properties of the learned hierarchy are instead evaluated empirically in Sec. 4 and Appendix B.

4 Experiments

We organize our experiments around two questions: (i) *Does the induced hierarchy provide a competitive general-purpose backbone for visual recognition?* (Sec. 4.1); (ii) *Does the learned structure reflect part–whole organization rather than independent score maps across levels?* (Sec. 4.2).

4.1 Performance Experiments

We evaluate HIVE as a visual representation learner across complementary recognition tasks. First, we test whether the induced hierarchy supports standard image classification on ImageNet-1K (Deng et al., 2009). Second, we examine whether the learned representation transfers to part-level structure through unsupervised part discovery on PartImageNet (He et al., 2022). Additional semantic segmentation results on ADE20K (Zhou et al., 2017) are provided in Appendix B. For HIVE results implemented in our study, we train three random seeds and report the mean ± standard error unless otherwise stated; baseline results are taken from the corresponding papers.

4.1.1 Task 1: Image Classification

HIVE Configurations. We evaluate HIVE on ImageNet-1K classification (Deng et al., 2009) at two model scales. HIVE-T and HIVE-S are built on ViT-T/16 and ViT-S/16 backbones (Dosovitskiy et al., 2020), respectively. Both use a three-level hierarchy with $K^{(1)} = 128$, $K^{(2)} = 64$, and $K^{(3)} = 32$ feature states. This hierarchy configuration is selected based on the HIVE-S ablation in Appendix B.3. Unless otherwise specified, we use HIVE-S as the default configuration for subsequent analyses.

Setup. Each model is trained end-to-end using the standard classification loss as the task-specific objective $\mathcal{L}_{\text{task}}$, together with the hierarchy-inducing losses described in Sec. 3.4. Full training details are provided in Appendix B.2.1. We report top-1 accuracy, parameter count, and FLOPs [1].

Results. As shown in Table 1, HIVE achieves strong ImageNet-1K classification performance across both Tiny and Small model scales. HIVE-T reaches 75.2% top-1 accuracy with 6.4M parameters and 1.7G FLOPs, outperforming standard Tiny-scale ViT baselines under a similar computa-

[1] For Ding et al. (2023) and Kwon et al. (2024), we report the published results because reproducible implementations were unavailable.

Table 2: Part discovery results on PartImageNet-O and PartImageNet-S. P_{disc} denotes the number of discovered parts. We report results at $P_{\text{disc}} \in \{25, 50\}$ following prior work (Xia et al., 2025). For Xia et al. (2024), $4N_c$ denotes four parts per category. Parentheses denote backbone parameter counts. Bold and underline indicate the best and second-best performance within each P_{disc} setting, respectively.

Type	Method + Backbone (Params)	P_{disc}	PartImage_O (%)		PartImage_S (%)	
			NMI ↑	ARI ↑	NMI ↑	ARI ↑
Supervised	Huang et al. (Huang & Li, 2020) + ResNet101 (45M)	25	7.6	1.3	–	–
		50	10.2	1.1	–	–
	PDiscoNet (Van Der Klis et al., 2023) + ViT-B/14 DINO (86M)	25	28.2	26.0	**46.8**	40.7
		50	29.5	28.7	34.1	39.4
	PDiscoFormer (Aniraj et al., 2024) + ViT-B/14 DINO (86M)	25	**44.7**	<u>59.3</u>	43.2	64.6
		50	<u>46.3</u>	62.2	<u>44.1</u>	60.1
Unsupervised	DINO ViT-S/8 (Amir et al., 2021) (21M)	25	31.5	14.2	–	–
		50	37.8	16.5	–	–
	Xia et al. (Xia et al., 2024) + DINO ViT-S/8 (21M)	$4N_c$	27.9	40.8	25.2	37.5
	MPAE (Xia et al., 2025) (21M)	25	37.3±2.0	57.3±2.5	41.7±1.0	<u>64.7±0.3</u>
		50	42.4±1.9	<u>63.1±2.5</u>	42.9±2.4	<u>64.3±2.9</u>
	MPAE + HIVE-S (22M)	25	<u>44.3±0.7</u>	**70.8±0.2**	<u>46.6±0.4</u>	**70.4±0.6**
		50	**46.5±0.3**	**70.3±0.2**	**47.3±0.4**	**70.1±0.1**



Figure 4: Qualitative part-discovery results on PartImageNet-O and -S. MPAE often produces fragmented or semantically mixed masks (e.g., fish and person, or bird body and background). In contrast, Ours (MPAE+HIVE-S) yields more coherent and spatially consistent parts, cleanly grouping object regions such as the animal head or body into single segments with improved boundary alignment.

tional budget. At the Small scale, HIVE-S achieves 82.8% top-1 accuracy with 24.6M parameters and 5.0G FLOPs, attaining the highest accuracy among all compared models. Notably, HIVE-S outperforms the hierarchy-aware models Hi-Mapper and DependencyViT-S, which achieve 82.6% and 82.1% top-1 accuracy, respectively. In addition to its recognition performance, HIVE exposes explicit cross-level assignments for analyzing the induced multi-level structure. Together, these results show that HIVE can induce structured hierarchical representations while maintaining strong recognition performance across model scales.

Feature States Separability. Beyond top-1 accuracy, we examine whether each learned feature state $\zeta_n^{(l)}$ from Eqs. (1) and (7) captures class-relevant visual structure. Fig. 3 shows t-SNE projections of feature states on a random 20-class subset: fine-level states range from diffuse scatter to partial class separation, whereas mid- and coarse-level states more consistently form compact, class-wise clusters. Both upper levels reach a nearest-centroid accuracy of 0.841 across all 1,000 classes, indicating that the upper hierarchy aggregates fine-level evidence into progressively more class-aligned representations. Full evaluation results are provided in Appendix B.5.

4.1.2 Task 2: Unsupervised Part Discovery

Unsupervised part discovery evaluates whether recurring parts can be recovered from learned visual features without part annotations during training. We use this task to test whether the hierarchy

induced by HIVE captures part-level visual structure, rather than only improving image-level recognition.

Setup. We integrate the default HIVE-S configuration from Task 1 in Sec. 4.1.1 into MPAE (Xia et al., 2025), which predicts part masks by matching patch features to learnable part descriptors. For a controlled comparison, we reimplement MPAE with a frozen ViT-S/16 backbone and compare two variants: MPAE, which uses raw ViT patch features, and MPAE+HIVE-S, which first reorganizes those features through the HIVE-S hierarchy. All other components and training settings are kept fixed.

We evaluate on PartImageNet-O and PartImageNet-S (He et al., 2022; Pan et al., 2023), using part annotations only for evaluation. Following Xia et al. (2025), we report Normalized Mutual Information (NMI) (Strehl & Ghosh, 2002) and Adjusted Rand Index (ARI) (Hubert & Arabie, 1985) at P_{disc}=25 and P_{disc}=50 discovered parts. MPAE and MPAE+HIVE-S use the smallest backbone listed in Table 2; Dataset statistics, metric definitions, and full evaluation details are provided in Appendix B.1 and Appendix B.2.2.

Results. As shown in Table 2, adding HIVE-S to the MPAE pipeline consistently improves part discovery across both PartImageNet variants and both part granularities ($P_{\text{disc}} = 25, 50$), where smaller P_{disc} corresponds to coarser partitions and larger P_{disc} yields finer-grained part decompositions. The gains appear in both NMI and ARI, indicating improved semantic agreement with ground-truth part labels and more spatially consistent discovered partitions. Despite using coarser 16×16 patch features, MPAE+HIVE-S outperforms the DINO ViT-S/8 baseline and remains competitive with methods using larger ViT-B/14 backbones trained with part supervision. Notably, although our method does not use part annotations during training, it achieves performance comparable to or exceeding several supervised baselines. These results suggest that HIVE makes part-relevant structure more accessible than raw ViT patch features.

Fig. 4 presents qualitative results on PartImageNet-O and -S. Compared to MPAE, which often produces fragmented or semantically mixed masks, MPAE+HIVE-S recovers more coherent and semantically meaningful part decompositions. For example, in the person-with-fish instances, MPAE tends to merge the fish and the person into a single region, whereas HIVE-S clearly separates them into distinct parts. In bird examples, HIVE-S further decomposes the object into finer structures such as beak, head, and legs, while MPAE typically captures only coarse or entangled regions. Similarly, for animals such as dogs and tigers, HIVE-S isolates meaningful components including the snout, mouth, and ears, whereas MPAE produces diffuse or partially mixed segments. In human-object interaction cases (e.g., the kart example), HIVE-S separates the human body, head, and different parts of the vehicle, while MPAE often groups them into a single or poorly localized region. Overall, these results indicate that HIVE-S yields both cleaner boundaries and more semantically consistent part decompositions across diverse object categories. Full qualitative results are provided in Appendix B.2.2.

4.2 Part–Whole Contribution Analysis

We next examine whether HIVE induces meaningful local part–whole organization. Unlike flat or level-wise hierarchy models, HIVE provides explicit cross-level composition maps, allowing us to quantify how lower-level visual structures contribute to higher-level embeddings. This analysis tests whether the hierarchy forms coherent directional part–whole relations rather than independent score maps across levels.

Setup. For each adjacent pair of hierarchy levels $(\ell, \ell+1)$, we compute a part–whole contribution matrix $\mathbf{R}^{(\ell\to\ell+1)}$ from the feature-composition score maps $\mathbf{\Phi}^{(\ell)}$ and $\mathbf{\Phi}^{(\ell+1)}$ from Eq. (9). After rectifying and patch-normalizing the score maps, we construct soft spatial footprints for each feature state. Each entry of $\mathbf{R}^{(\ell\to\ell+1)}$ combines an *influence* term, measuring spatial co-activation, with a directional *containment* term, measuring how much a lower-level footprint lies within a higher-level footprint. Thus, each column of $\mathbf{R}^{(\ell\to\ell+1)}$ represents a higher-level embedding as a distribution over lower-level contributors. Full experimental design and definitions are provided in Appendix B.4.

For analytical clarity, we select 20 ImageNet-1K classes and train a three-level hierarchical model with $K^{(1)}$=8 fine-, $K^{(2)}$=4 mid-, and $K^{(3)}$=2 coarse-level feature states per image. The two coarse-level states encourage a high-level foreground–background-like decomposition, while the finer lev-

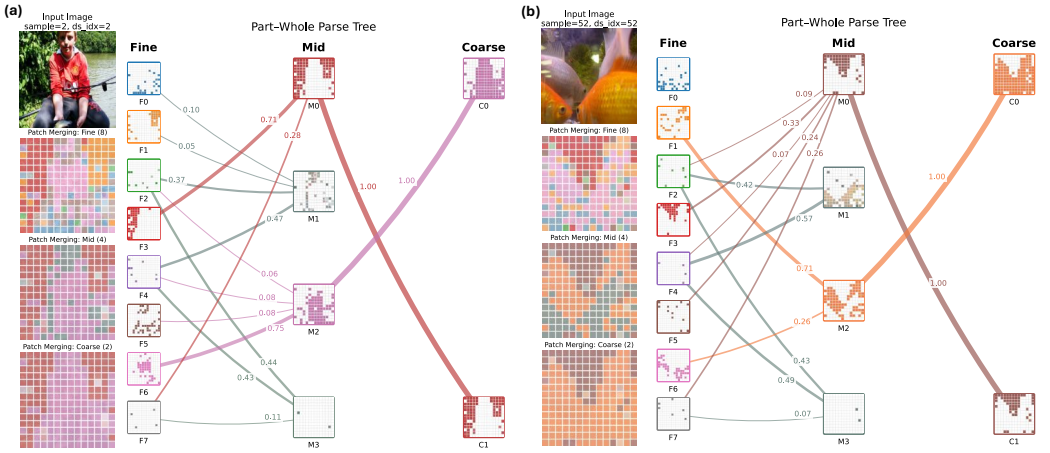

Figure 5: Part–whole visualizations for two ImageNet samples. Fine-level assignments are progressively consolidated into mid- and coarse-level groups, and edges with contribution values below 0.05 are omitted for clarity. **(a)** The hierarchy isolates semantically meaningful human-related structure, with the mid-level embedding $M1$ emphasizing the head/face region. **(b)** The mid-level hierarchy refines the scene into visually distinct fish groups, with $M1$ and $M2$ capturing different color-based fish regions. These examples illustrate that the learned hierarchy can represent interpretable intermediate structure beyond coarse scene-level grouping.

els allow this coarse split to branch into more detailed visual structures. Importantly, although this configuration has a binary-tree-like form, the proposed contribution score does not impose a hand-crafted tree. Instead, it reveals the local part–whole relations that emerge between adjacent hierarchy levels.

Results. Fig. 5 shows representative ImageNet examples of the learned part–whole organization. Each example contains two complementary visualizations. The patch-merging maps on the left show how image patches are grouped across hierarchy levels: at each level, a patch is colored according to the feature states with the largest composition score, so the fine-, mid-, and coarse-level maps visualize how local patch groups are progressively consolidated into larger visual regions. The part–whole parse tree on the right summarizes the same hierarchy at the feature level. Each node corresponds to a learned feature state at the fine, mid, or coarse level, and the small map inside each node shows its spatial footprint. Directed edges connect lower-level states to higher-level states, with edge thickness and edge labels indicating the part–whole contribution score $\mathbf{R}^{(\ell\to\ell+1)}$. The parse tree therefore shows which fine-level structures support each mid-level group and how those mid-level groups are further consolidated into coarse-level components.

The examples in Fig. 5 further show that the learned hierarchy captures interpretable intermediate structure beyond coarse grouping alone. In Fig. 5**(a)**, the mid-level feature state $M1$ emphasizes a localized human-related region, particularly around the head and face, indicating that HIVE can isolate interpretable body-part structure at an intermediate level. In Fig. 5**(b)**, the mid-level hierarchy separates visually distinct fish regions, with $M1$ and $M2$ capturing different color-based groupings. Together, these examples suggest that HIVE not only organizes local evidence into structured higher-level representations, but also preserves semantically meaningful intermediate decompositions. Additional examples are provided in Appendix B.4, further showing that HIVE learns flexible and visually coherent cross-level dependencies spanning contextual structures, human parts, object–context relations, and fine-grained intra-category variation.

5 Conclusion

We presented *HIVE*, a framework for inducing hierarchical visual embeddings for part–whole representation learning. HIVE organizes image patches into fine-, mid-, and coarse-level embeddings through feature binding, graph-based relational reasoning, hierarchical grouping, and projection back to the token space, without direct relation annotations. Experiments show that HIVE serves as a competitive general-purpose backbone while producing class-relevant and structurally coherent multilevel feature states. We also introduced part–whole contribution analysis to quantify directional cross-level composition and visualize image-specific parsing trees. Future work could evaluate HIVE on datasets with explicit object–part–subpart annotations or controlled compositional benchmarks. Extending HIVE to open-vocabulary recognition and compositional reasoning may further clarify how learned feature-level hierarchy supports visual understanding.

AI USE STATEMENT

Large language models were not used for any research tasks requiring disclosure under the ICLR AI policy, including hypothesis development, methodology or experiment design, method implementation, mathematical reasoning or proofs, data processing or analysis, or interpretation of results. Their use was limited to writing and editing assistance to improve the clarity and readability of the manuscript. All AI-assisted edits were reviewed by the authors, who take responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

We provide the full implementation of HIVE as an anonymized ZIP archive in the supplementary material. The archive includes code for model construction, training, evaluation, and the auxiliary hierarchy-inducing objectives described in Sec. 3. We also include configuration files and scripts covering the main experimental settings. Due to anonymity and file-size constraints, trained model weights are not included in the current submission. Upon acceptance, we will publicly release all trained checkpoints together with the codebase, enabling full reproduction of the reported results without retraining. We additionally provide Jupyter notebooks to support interactive inspection and analysis of the learned hierarchy. These notebooks include multiple analysis tools, such as visualization of feature-assignment maps, cross-level part–whole contribution structures, and image-specific parsing trees used in Sec. 4.2.

All experiments were conducted on 4×A100 (80GB) GPUs; For classification task, training takes about 12.5 hours. For unsupervised part discovery, training takes approximately 5–6 hours. For semantic segmentation, total training time takes approximately 10 hours. Reported times correspond to wall-clock training time under standard training settings.

REFERENCES

Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021.

Ananthu Aniraj, Cassio F Dantas, Dino Ienco, and Diego Marcos. Pdiscoformer: Relaxing part discovery constraints with vision transformers. In *European Conference on Computer Vision*, pp. 256–272. Springer, 2024.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7 (11), 2006.

Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12506–12515, 2020.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 357–366, 2021.

Richard Chen, Rameswar Panda, and Quanfu Fan. Regionvit: Regional-to-local attention for vision transformers. In *International Conference on Learning Representations*, 2022.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Mingyu Ding, Yikang Shen, Lijie Fan, Zhenfang Chen, Zitian Chen, Ping Luo, Joshua B Tenenbaum, and Chuang Gan. Visual dependency transformers: Dependency tree emerges from reversed attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14528–14539, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Martin Engelcke, Adam R Kosiorek, Oiwi Parker Jones, and Ingmar Posner. Genesis: Generative scene inference and sampling with object-centric latent representations. In *International Conference on Learning Representations*, 2020.

Michael D Ernst. Permutation methods: a basis for exact inference. *Statistical Science*, pp. 676–685, 2004.

Shijie Geng, Jianbo Yuan, Yu Tian, Yuxiao Chen, and Yongfeng Zhang. Hiclip: Contrastive language-image pretraining with hierarchy-aware attention. In *11th International Conference on Learning Representations, ICLR 2023*, 2023.

Phillip Good. *Permutation, parametric and bootstrap tests of hypotheses*. Springer, 2005.

Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *International Conference on Machine Learning*, pp. 2424–2433. PMLR, 2019.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts. In *European Conference on Computer Vision*, pp. 128–145. Springer, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Geoffrey E Hinton, Sara Sabour, and Nicholas Frosst. Matrix capsules with em routing. In *International conference on learning representations*, 2018.

Jinyung Hong, Keun Hee Park, and Theodore P Pavlic. Concept-centric transformers: Enhancing model interpretability through object-centric concept learning within a shared global workspace. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4880–4891, 2024.

Jinyung Hong, Yearim Kim, Keun Hee Park, Sangyu Han, Nojun Kwak, and Theodore P Pavlic. Bi-ice: An inner interpretable framework for image classification via bi-directional interactions between concept and input embeddings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3835–3845, 2026.

Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang. Part-stacked CNN for fine-grained visual categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1173–1182, 2016.

Zixuan Huang and Yin Li. Interpretable and accurate fine-grained recognition via region grouping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8662–8672, 2020.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

Yearim Kim, Sangyu Han, and Nojun Kwak. From local to global to mechanistic: An ierf-centered unified framework for interpreting vision models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2026. To appear.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6399–6408, 2019.

Matthew Kowal, Richard P Wildes, and Konstantinos G Derpanis. Visual concept connectome (vcc): Open world concept discovery and their interlayer connections in deep models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10895–10905, 2024.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Hyeongjun Kwon, Jinhyun Jang, Jin Kim, Kwonyoung Kim, and Kwanghoon Sohn. Improving visual recognition with hyperbolical visual hierarchy mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17364–17374, 2024.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pp. 3744–3753. PMLR, 2019.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V 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.

Tai-Yu Pan, Qing Liu, Wei-Lun Chao, and Brian Price. Towards open-world segmentation of parts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15392–15401, 2023.

Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual grouping in contrastive vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5571–5584, 2023.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. *Advances in neural information processing systems*, 30, 2017.

Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002.

Shitao Tang, Jiahui Zhang, Siyu Zhu, and Ping Tan. Quadtree attention for vision transformers. In *International Conference on Learning Representations*, 2022.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Albert Tseng, Tao Yu, Toni Liu, and Christopher M De Sa. Coneheads: Hierarchy aware attention. *Advances in Neural Information Processing Systems*, 36:51421–51433, 2023.

Zhuowen Tu, Xiangrong Chen, Alan L Yuille, and Song-Chun Zhu. Image parsing: Unifying segmentation, detection, and recognition. *International Journal of computer vision*, 63(2):113–140, 2005.

Robert Van Der Klis, Stephan Alaniz, Massimiliano Mancini, Cassio F Dantas, Dino Ienco, Zeynep Akata, and Diego Marcos. Pdisconet: Semantically consistent part discovery for fine-grained recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1866–1876, 2023.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 568–578, 2021.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational visual media*, 8(3):415–424, 2022.

Jiahao Xia, Wenjian Huang, Min Xu, Jianguo Zhang, Haimin Zhang, Ziyu Sheng, and Dong Xu. Unsupervised part discovery via dual representation alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10597–10613, 2024.

Jiahao Xia, Yike Wu, Wenjian Huang, Jianguo Zhang, and Jian Zhang. Unsupervised part discovery via descriptor-based masked image restoration with optimized constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8668–8677, 2025.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 418–434, 2018.

Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5410–5419, 2017.

Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. *Advances in neural information processing systems*, 31, 2018.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.

Song-Chun Zhu and David Mumford. A stochastic grammar of images. *Foundations and trends in computer graphics and vision*, 2(4):259–362, 2007.

A METHOD DETAILS

A.1 ALGORITHM.

Algorithm 1 describes the forward workflow of HIVE at a single epoch in Sec. 3 in the main text. Starting from patch embeddings $\mathbf{Z}$, the model first refines a set of learnable fine-grained feature embeddings through T iterations of the feature binding module $\boldsymbol{\Theta}$, yielding the lowest-level feature states $\boldsymbol{\zeta}^{(1)}$. An initial adjacency matrix $\boldsymbol{J}^{(1)}$ is then constructed from these features using pairwise cosine similarity with min-max normalization. Next, the hierarchical feature-graph attention modules $\{\boldsymbol{\Psi}^{(l)}\}_{l=1}^{L-1}$ recursively transform the current feature graph into progressively coarser feature graphs, producing $(\boldsymbol{\zeta}^{(l+1)}, \boldsymbol{J}^{(l+1)})$ at each level. Finally, the multi-level feature projection module $\boldsymbol{\Omega}$ projects the full hierarchy of feature graphs back to the patch space, generating the final representation $\mathbf{Z}'$, which integrates visual information across all hierarchy levels.

A.2 HIERARCHICAL PERMUTATION CONSISTENCY PROPERTY

We characterize a structural symmetry property of HIVE: the visual hierarchy induced by the model is permutation-consistent across hierarchy levels, while the final patch-space representation remains permutation equivariant with respect to patch re-indexing. Accordingly, the emergent hierarchy should be understood as a relational organization among patches and learned feature units, rather than as an artifact of arbitrary indexing choices.

We present this as an architectural property of HIVE. In particular, it supports interpreting HIVE as a hierarchy-inducing visual representation learner whose learned hierarchy constitutes a plausible reusable backbone. Its practical usefulness across tasks is then evaluated empirically. This property characterizes the joint consistency of HIVE's recursively constructed hierarchy rather than introducing a new notion of permutation equivariance, and it does not by itself imply semantic correctness of the learned hierarchy.

A.2.1 INTUITION

Permutation consistency means that the visual hierarchy learned by HIVE should not depend on arbitrary bookkeeping choices, such as the ordering of patch embeddings or the indexing of learned feature embeddings. If these elements are merely re-indexed, the model should produce the same hierarchy up to the corresponding re-indexing, rather than a fundamentally different representation. In this sense, permutation consistency formalizes the idea that the emergent hierarchy is determined by the relational structure among visual parts and feature units, rather than by the accidental ordering of rows in intermediate tensors.

A.2.2 DEFINITIONS

Before giving the proof, we introduce the definitions used below.

Algorithm 1 Overall Workflow of HIVE at a single epoch

Input: input patch embeddings $\mathbf{Z} \in \mathbb{R}^{P\times D}$, initial fine-grained feature states $\boldsymbol{\zeta}_0^{(1)} \in \mathbb{R}^{K^{(1)}\times D}$, refinement step T, hierarchy depth L, cluster sizes $\{K^{(l)}\}_{l=1}^{L}$

for $t = 1, \ldots, T$ **do**
  $\boldsymbol{\zeta}_t^{(1)} \leftarrow \boldsymbol{\Theta}(\mathbf{Z}, \boldsymbol{\zeta}_{t-1}^{(1)})$ ▷ Fine-grained feature binding; Sec. 3.1
end for
$\boldsymbol{\zeta}^{(1)} \leftarrow \boldsymbol{\zeta}_T^{(1)}$
$\boldsymbol{J}^{(1)} \leftarrow \mathrm{Adjacency}(\boldsymbol{\zeta}^{(1)})$ ▷ Pairwise cosine similarity with min-max normalization
for $l = 1, \ldots, L-1$ **do**
  $(\boldsymbol{\zeta}^{(l+1)}, \boldsymbol{J}^{(l+1)}) \leftarrow \boldsymbol{\Psi}^{(l)}(\boldsymbol{\zeta}^{(l)}, \boldsymbol{J}^{(l)})$ ▷ Hierarchical feature-graph attention; Sec. 3.2
end for
$\mathbf{Z}' \leftarrow \boldsymbol{\Omega}\big(\mathbf{Z}, \{\boldsymbol{\zeta}^{(l)}, \boldsymbol{J}^{(l)}\}_{l=1}^{L}\big)$ ▷ Multi-level feature projection; Sec. 3.3
return $\mathbf{Z}'$

Definition 1 (Permutation Matrix)**.** *A matrix $\boldsymbol{\pi} \in \mathbb{R}^{M\times M}$ is called a* permutation matrix *if each row and each column contains exactly one entry equal to* 1*, with all remaining entries equal to* 0*. Such a matrix represents a re-indexing of M elements.*

Definition 2 (Permutation Invariance)**.** *A function $f : \mathbb{R}^{M\times D_1} \to \mathbb{R}^{D_2}$ is called* permutation invariant *if, for every permutation matrix $\boldsymbol{\pi} \in \mathbb{R}^{M\times M}$,*

$$f(\boldsymbol{\pi}\mathbf{X}) = f(\mathbf{X}).$$

Definition 3 (Permutation Equivariance)**.** *A function $f : \mathbb{R}^{M\times D_1} \to \mathbb{R}^{M\times D_2}$ is called* permutation equivariant *if, for every permutation matrix $\boldsymbol{\pi} \in \mathbb{R}^{M\times M}$,*

$$f(\boldsymbol{\pi}\mathbf{X}) = \boldsymbol{\pi} f(\mathbf{X}).$$

Definition 4 (Patch Re-indexing)**.** *Let $\mathbf{Z} = [z_1, \ldots, z_P] \in \mathbb{R}^{P\times D}$ denote the input patch embeddings, and let $\mathbf{P} \in \mathbb{R}^{P\times P}$ be a permutation matrix. The* re-indexed patch embedding matrix *is defined by*

$$\widetilde{\mathbf{Z}} = \mathbf{P}\mathbf{Z}.$$

This operation permutes the ordering of the patch embeddings without changing their values.

Definition 5 (Level-wise Feature Re-indexing)**.** *Let*

$$\boldsymbol{\zeta}^{(l)} = [\zeta_1^{(l)}, \ldots, \zeta_{K^{(l)}}^{(l)}] \in \mathbb{R}^{K^{(l)}\times D}$$

denote the learned feature embeddings at hierarchy level l, and let $\mathbf{Q}^{(l)} \in \mathbb{R}^{K^{(l)}\times K^{(l)}}$ be a permutation matrix. The re-indexed level-l feature embedding matrix *is defined by*

$$\widetilde{\boldsymbol{\zeta}}^{(l)} = \mathbf{Q}^{(l)}\boldsymbol{\zeta}^{(l)}.$$

Definition 6 (Bi-index Permutation Equivariance)**.** *Let*

$$f : \mathbb{R}^{P\times D_1} \times \mathbb{R}^{K\times D_2} \to \mathbb{R}^{P\times K}.$$

We say that f is bi-index permutation equivariant *if, for every pair of permutation matrices $\mathbf{P} \in \mathbb{R}^{P\times P}$ and $\mathbf{Q} \in \mathbb{R}^{K\times K}$,*

$$f(\mathbf{P}\mathbf{X}, \mathbf{Q}\mathbf{Y}) = \mathbf{P}\, f(\mathbf{X}, \mathbf{Y})\, \mathbf{Q}^\top.$$

Definition 7 (Permutation Consistency of Assignment Matrices)**.** *Let*

$$\mathbf{S}^{(l)} \in \mathbb{R}^{K^{(l)}\times K^{(l+1)}}$$

denote the assignment matrix from level l to level $l+1$. We say that $\mathbf{S}^{(l)}$ is permutation-consistent *if, for every pair of permutation matrices*

$$\mathbf{Q}^{(l)} \in \mathbb{R}^{K^{(l)}\times K^{(l)}}, \qquad \mathbf{Q}^{(l+1)} \in \mathbb{R}^{K^{(l+1)}\times K^{(l+1)}},$$

the re-indexed assignment matrix satisfies

$$\widetilde{\mathbf{S}}^{(l)} = \mathbf{Q}^{(l)}\mathbf{S}^{(l)}\big(\mathbf{Q}^{(l+1)}\big)^\top.$$

Definition 8 (Hierarchical Permutation Consistency)**.** *A hierarchical representation*

$$\Big\{\boldsymbol{\zeta}^{(l)},\, \boldsymbol{\Phi}^{(l)},\, \mathbf{J}^{(l)},\, \mathbf{S}^{(l)},\, \bar{\mathbf{Z}}^{(l)}\Big\}_l$$

is said to be hierarchically permutation-consistent *if arbitrary re-indexing of patch embeddings and level-wise feature embeddings induces only the corresponding re-indexing of all derived quantities, without changing the underlying relational structure.*

A.2.3 PROOF

The proof for Proposition 1 in the main text is straightforward and is provided for completeness. The key idea is that each component of HIVE is permutation-consistent with respect to the relevant indexing of its inputs, and therefore their composition remains permutation-consistent. In particular, we repeatedly use the facts that shared row-wise maps are permutation equivariant and that matrix multiplication preserves compatible re-indexing.

Linear Projections. The linear projections used in HIVE, such as $q^{\mathbf{\Omega}}(\cdot)$, $k^{\mathbf{\Omega}}(\cdot)$, and $v^{\mathbf{\Omega}}(\cdot)$, are applied independently to each row of their input with shared parameters. Hence, for any permutation matrix $\boldsymbol{\pi}$ and any row-indexed input matrix, the corresponding projection is permutation equivariant. In particular, for the patch embedding matrix $\mathbf{Z}$,

$$q^{\mathbf{\Omega}}(\boldsymbol{\pi}\mathbf{Z}) = \boldsymbol{\pi} q^{\mathbf{\Omega}}(\mathbf{Z}), \qquad k^{\mathbf{\Omega}}(\boldsymbol{\pi}\mathbf{Z}) = \boldsymbol{\pi} k^{\mathbf{\Omega}}(\mathbf{Z}), \qquad v^{\mathbf{\Omega}}(\boldsymbol{\pi}\mathbf{Z}) = \boldsymbol{\pi} v^{\mathbf{\Omega}}(\mathbf{Z}).$$

Therefore, such linear projections are permutation equivariant.

Fine-grained Feature Binding Layer $\mathbf{\Theta}$. The fine-grained feature binding layer $\mathbf{\Theta}$ follows a Slot-Attention-style binding mechanism (Locatello et al., 2020), which is permutation invariant with respect to the ordering of the input embeddings and permutation equivariant with respect to the ordering of the output feature embeddings. Thus, re-indexing the input patch embeddings induces only the corresponding re-indexing of the learned level-wise feature embeddings. This permutation-consistent binding property is established in (Locatello et al., 2020), and the same reasoning applies to the binding mechanism used here.

Feature Composition Score $\mathbf{\Phi}^{(1)}$. The first-level feature composition score is defined by

$$\mathbf{\Phi}^{(1)} = \mathrm{softmax}_{K^{(1)}}\left(q^{\mathbf{\Omega}}(\mathbf{Z})\, k^{\mathbf{\Omega}}(\boldsymbol{\zeta}^{(1)})^{\top}/\sqrt{D}\right).$$

Let $\mathbf{P} \in \mathbb{R}^{P\times P}$ be a permutation matrix acting on the patch embeddings, and let $\mathbf{Q}^{(1)} \in \mathbb{R}^{K^{(1)}\times K^{(1)}}$ be a permutation matrix acting on the level-1 feature embeddings. Since the linear projections are permutation equivariant,

$$q^{\mathbf{\Omega}}(\mathbf{P}\mathbf{Z}) = \mathbf{P} q^{\mathbf{\Omega}}(\mathbf{Z}), \qquad k^{\mathbf{\Omega}}(\mathbf{Q}^{(1)}\boldsymbol{\zeta}^{(1)}) = \mathbf{Q}^{(1)} k^{\mathbf{\Omega}}(\boldsymbol{\zeta}^{(1)}).$$

It follows that the attention score matrix transforms as

$$q^{\mathbf{\Omega}}(\mathbf{P}\mathbf{Z})\, k^{\mathbf{\Omega}}(\mathbf{Q}^{(1)}\boldsymbol{\zeta}^{(1)})^{\top} = \mathbf{P}\, q^{\mathbf{\Omega}}(\mathbf{Z})\, k^{\mathbf{\Omega}}(\boldsymbol{\zeta}^{(1)})^{\top}\big(\mathbf{Q}^{(1)}\big)^{\top}.$$

Since the softmax is applied row-wise over the feature dimension, row permutations act on the left and column permutations act on the right. Hence,

$$\widetilde{\mathbf{\Phi}}^{(1)} = \mathbf{P}\, \mathbf{\Phi}^{(1)}\, \big(\mathbf{Q}^{(1)}\big)^{\top}.$$

Thus, $\mathbf{\Phi}^{(1)}$ is bi-index permutation equivariant.

Hierarchical Feature Composition Scores $\mathbf{\Phi}^{(l)}$. For $l = 1, \ldots, L-1$, the higher-level feature composition scores are defined recursively by

$$\mathbf{\Phi}^{(l+1)} = \mathbf{\Phi}^{(l)}\mathbf{S}^{(l)}.$$

Assume that

$$\widetilde{\mathbf{\Phi}}^{(l)} = \mathbf{P}\, \mathbf{\Phi}^{(l)}\, \big(\mathbf{Q}^{(l)}\big)^{\top}$$

and that the assignment matrix is permutation-consistent:

$$\widetilde{\mathbf{S}}^{(l)} = \mathbf{Q}^{(l)}\mathbf{S}^{(l)}\big(\mathbf{Q}^{(l+1)}\big)^{\top}.$$

Then

$$\begin{aligned}
\widetilde{\mathbf{\Phi}}^{(l+1)} &= \widetilde{\mathbf{\Phi}}^{(l)}\widetilde{\mathbf{S}}^{(l)} \\
&= \mathbf{P}\, \mathbf{\Phi}^{(l)}\, \big(\mathbf{Q}^{(l)}\big)^{\top}\mathbf{Q}^{(l)}\mathbf{S}^{(l)}\big(\mathbf{Q}^{(l+1)}\big)^{\top} \\
&= \mathbf{P}\, \mathbf{\Phi}^{(l)}\mathbf{S}^{(l)}\big(\mathbf{Q}^{(l+1)}\big)^{\top} \\
&= \mathbf{P}\, \mathbf{\Phi}^{(l+1)}\big(\mathbf{Q}^{(l+1)}\big)^{\top}.
\end{aligned}$$

Therefore, every hierarchical feature composition score $\mathbf{\Phi}^{(l)}$ is bi-index permutation equivariant.

Hierarchical Feature-Graph Attention Layer $\mathbf{\Psi}$. The hierarchical feature graph attention layer $\mathbf{\Psi}$ follows the permutation-equivariant graph processing principle used in (Veličković et al., 2018), where shared attention and aggregation operations are applied over graph nodes and their neighbors. Under a re-indexing of the level-l feature embeddings by $\mathbf{Q}^{(l)}$, the hierarchy-level adjacency matrix transforms as

$$\widetilde{\mathbf{J}}^{(l)} = \mathbf{Q}^{(l)}\mathbf{J}^{(l)}\big(\mathbf{Q}^{(l)}\big)^{\top}.$$

Since graph attention is applied with shared parameters over the re-indexed nodes and edges, the updated level-wise feature representations transform equivariantly under the same permutation. Thus, the same structural principle as in (Veličković et al., 2018) applies here.

Assignment Matrices $\mathbf{S}^{(l)}$. The assignment matrix $\mathbf{S}^{(l)}$ maps the level-l feature embeddings to the coarser level-$(l+1)$ feature embeddings. In HIVE, this step follows the same permutation-consistent coarsening principle as in (Ying et al., 2018): if the level-l and level-$(l+1)$ feature embeddings are re-indexed by permutation matrices $\mathbf{Q}^{(l)}$ and $\mathbf{Q}^{(l+1)}$, respectively, then the corresponding assignment matrix transforms as

$$\widetilde{\mathbf{S}}^{(l)} = \mathbf{Q}^{(l)}\mathbf{S}^{(l)}\big(\mathbf{Q}^{(l+1)}\big)^{\top}.$$

Thus, the soft assignment relation between fine-level and coarse-level features is preserved under arbitrary re-indexing. Consequently, the pooling structure induced by $\mathbf{S}^{(l)}$ depends only on the underlying feature relations, not on the particular ordering of feature embeddings at either level.

Multi-level Feature Projection Layer $\mathbf{\Omega}$. The multi-level feature projection layer $\mathbf{\Omega}$ maps the hierarchical feature embeddings back to the patch space through the feature composition scores. For level l, let

$$\bar{\mathbf{Z}}^{(l)} = \mathbf{\Phi}^{(l)} v^{\mathbf{\Omega}}(\boldsymbol{\zeta}^{(l)}),$$

where $v^{\mathbf{\Omega}}(\cdot)$ is a shared row-wise value projection. Since $v^{\mathbf{\Omega}}(\cdot)$ is permutation equivariant,

$$v^{\mathbf{\Omega}}(\mathbf{Q}^{(l)}\boldsymbol{\zeta}^{(l)}) = \mathbf{Q}^{(l)} v^{\mathbf{\Omega}}(\boldsymbol{\zeta}^{(l)}),$$

and since $\mathbf{\Phi}^{(l)}$ is bi-index permutation equivariant,

$$\widetilde{\mathbf{\Phi}}^{(l)} = \mathbf{P}\,\mathbf{\Phi}^{(l)}\,\big(\mathbf{Q}^{(l)}\big)^{\top},$$

it follows that

$$\widetilde{\bar{\mathbf{Z}}}^{(l)} = \widetilde{\mathbf{\Phi}}^{(l)}\, v^{\mathbf{\Omega}}(\widetilde{\boldsymbol{\zeta}}^{(l)}) = \mathbf{P}\,\mathbf{\Phi}^{(l)}\,\big(\mathbf{Q}^{(l)}\big)^{\top}\mathbf{Q}^{(l)} v^{\mathbf{\Omega}}(\boldsymbol{\zeta}^{(l)}) = \mathbf{P}\,\bar{\mathbf{Z}}^{(l)}.$$

Therefore, the projected patch-level representation induced by level l is permutation equivariant with respect to patch re-indexing and insensitive to the arbitrary ordering of the level-wise feature embeddings.

If the final projected representation is obtained by aggregating the level-specific projections, namely

$$\mathbf{Z}' = \sum_{l=1}^{L} \bar{\mathbf{Z}}^{(l)},$$

then

$$\widetilde{\mathbf{Z}}' = \mathbf{P}\mathbf{Z}',$$

since summation preserves compatible permutation equivariance.

Combining All Modules. The fine-grained feature binding layer $\mathbf{\Theta}$ yields permutation-consistent level-wise feature embeddings. The linear projections are permutation equivariant. The first-level feature composition score $\mathbf{\Phi}^{(1)}$ is bi-index permutation equivariant, and the recursive relation

$$\mathbf{\Phi}^{(l+1)} = \mathbf{\Phi}^{(l)}\mathbf{S}^{(l)}$$

preserves this property across all hierarchy levels. The hierarchical feature-graph attention layer $\mathbf{\Psi}$ is permutation equivariant with respect to level-wise feature re-indexing, and the assignment matrices $\mathbf{S}^{(l)}$ are permutation-consistent across consecutive levels. Finally, the multi-level feature projection layer $\mathbf{\Omega}$ maps the hierarchy back to patch space while preserving compatible re-indexing.

Table 3: Dataset statistics. "#Categories" refers to image-level classes for ImageNet-1K, object/super-classes for PartImageNet, and semantic categories for ADE20K. PartImageNet provides pixel-level part masks, which are used only for evaluation; the part-discovery model is trained without part annotations.

Task	Dataset	#Train	#Eval	#Categories	Annotation	Input Resolution
Image Classification	ImageNet-1K (Deng et al., 2009)	$\sim$1.28M	50,000 (val)	1,000	Image-level class label	224×224
Unsupervised Part Discovery	PartImageNet-O (OOD) (Pan et al., 2023)	14,865	1,658 (test)	110 / 11 super	Pixel-level part masks (eval only)	224×224
	PartImageNet-S (Seg.) (He et al., 2022)	20,457	2,405 (test)	158	Pixel-level part masks (eval only)	224×224
Semantic Segmentation	ADE20K (Zhou et al., 2017)	20,210	2,000 (val)	150	Pixel-level semantic labels	512×512

Therefore, arbitrary re-indexing of patch embeddings or level-wise feature embeddings induces only the corresponding re-indexing of all derived quantities:

$$\boldsymbol{\zeta}^{(l)}, \quad \boldsymbol{\Phi}^{(l)}, \quad \mathbf{J}^{(l)}, \quad \mathbf{S}^{(l)}, \quad \bar{\mathbf{Z}}^{(l)}, \quad \mathbf{Z}'.$$

Consequently, the hierarchical visual representation learned by HIVE is hierarchically permutation-consistent: its emergent visual hierarchy is determined by the underlying relational structure rather than by arbitrary indexing choices.

B Further Experimental Results

This section provides the experimental details and additional results that support the main performance evaluation in Sec. 4. We first summarize the datasets used for image classification, unsupervised part discovery, and semantic segmentation. We then describe the task-specific configurations for ImageNet-1K classification, PartImageNet part discovery, and ADE20K semantic segmentation, including model settings, optimization details, data processing, and evaluation protocols. Finally, we provide additional qualitative results for part discovery and ablation studies that analyze the effect of hierarchy size, auxiliary objectives, and multi-level projected representations.

B.1 Dataset Statistics

Table 3 summarizes the datasets used in our experiments. We evaluate HIVE-S across three complementary settings: ImageNet-1K classification for image-level recognition, PartImageNet part discovery for part-level structure, and ADE20K semantic segmentation for dense pixel-level prediction. The table reports the number of training and evaluation images, the number of categories, the annotation type, and the input resolution used for each task.

ImageNet-1K Classification. ImageNet-1K (Deng et al., 2009) is used to evaluate image-level recognition. It contains $1{,}000$ object categories, approximately 1.28M training images, and $50{,}000$ validation images. We use the official train/validation split and report top-1 validation accuracy.

Unsupervised Part Discovery. PartImageNet (He et al., 2022) provides pixel-level part annotations for ImageNet object categories. We evaluate on two variants: PartImageNet-O (Pan et al., 2023), which contains 110 classes grouped into 11 super-classes and is used for out-of-distribution cross-category part discovery, and PartImageNet-S (He et al., 2022), which contains 158 classes and provides a larger segmentation-oriented benchmark with greater appearance variation. Part annotations are not used during training; they are used only for evaluation. We report Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI) between predicted part assignments and ground-truth part masks at $P_{\text{disc}}{=}25$ and $P_{\text{disc}}{=}50$ discovered parts.

Semantic Segmentation. ADE20K (Zhou et al., 2017) is used to evaluate dense pixel-level transfer. It contains $20{,}210$ training images and $2{,}000$ validation images annotated over 150 semantic categories. We use the standard train/validation split and report mean Intersection-over-Union (mIoU) on the validation set. Unless otherwise stated, our ADE20K experiments use single-scale inference without multi-scale or flip testing.

B.2 Performance Experiments

B.2.1 Task 1: ImageNet-1K Classification

ImageNet-1K classification evaluates HIVE at two model scales for image-level recognition. HIVE-T and HIVE-S are built on ViT-T/16 and ViT-S/16 backbones, respectively. We use this task as the primary general-purpose recognition benchmark in Sec. 4. Dataset statistics are summarized in Sec. B.1.

Architecture. We train HIVE-T and HIVE-S using ViT-T/16 and ViT-S/16 backbones (Dosovitskiy et al., 2020), respectively. Unless otherwise stated, both models use a three-level hierarchy (L=3) with cluster sizes $(K^{(1)}, K^{(2)}, K^{(3)}) = (128, 64, 32)$. This configuration is selected by the HIVE-S hierarchy-size ablation in Sec. B.3 and Table 5, and is used as the default hierarchy configuration throughout the experiments.

Implementation Details. The model is trained end-to-end with the standard image-classification loss and the auxiliary hierarchy regularizers introduced in Sec. 3.4. The final objective is

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{task}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}} + \lambda_{\text{link}}\mathcal{L}_{\text{link}}.$$

We set λ_{cons}=1×10^{-7}, λ_{align}=1×10^{-2}, λ_{div}=5×10^{-5}, and λ_{link}=1×10^{-3}. The contribution of each auxiliary objective is evaluated in Appendix B.3 and Table 7.

We follow the supervised ViT training protocol of DeiT (Touvron et al., 2021) unless otherwise specified. We use AdamW (Loshchilov & Hutter, 2017) with weight decay 1×10^{-3} and initial learning rate 5×10^{-5}. The learning rate is linearly warmed up for the first 10 epochs and then decayed with a cosine schedule. All ImageNet-1K models are trained for 50 epochs with batch size 128 in 16-bit precision.

Data Augmentation. Training images are resized to 256×256 with bilinear interpolation, followed by a random 224×224 crop and random horizontal flip. We use RandAugment (Cubuk et al., 2020) with magnitude 9, standard deviation 0.5, and increasing severity (`rand-m9-mstd0.5-inc1`). We also use random erasing with probability 0.25, scale $(0.02, 0.33)$, aspect ratio $(0.3, 3.3)$, and random pixel values. Inputs are normalized with ImageNet mean $[0.485, 0.456, 0.406]$ and standard deviation $[0.229, 0.224, 0.225]$. Mixup (Zhang et al., 2018) with α=0.8 and CutMix (Yun et al., 2019) with α=1.0 are applied at the batch level with mixing probability 1.0 and switching probability 0.5. We use label smoothing with coefficient 0.1.

Evaluation Protocol. At test time, images are resized to 256×256 and center-cropped to 224×224. We report top-1 accuracy on the ImageNet-1K validation set. For HIVE-T and HIVE-S, we run three random seeds and report mean $\pm$ standard error. Parameter counts and FLOPs are computed for the same 224×224 input resolution used at evaluation time. Baseline numbers in Table 1 are taken from the corresponding papers unless otherwise stated.

B.2.2 Task 2: Unsupervised Part Discovery

Unsupervised part discovery evaluates whether HIVE-S exposes part-level visual structure without using part annotations during training. We use this task to test whether the hierarchy learned for ImageNet-1K classification transfers to object-part discovery, rather than only improving image-level recognition. Dataset statistics are summarized in Sec. B.1.

Architecture. We evaluate part discovery by plugging the HIVE-S representation into MPAE (Xia et al., 2025). For a controlled comparison, we compare two variants under the same MPAE pipeline: MPAE, which directly uses raw ViT-S/16 patch features, and MPAE+HIVE-S, which first passes the patch features through the ImageNet-trained HIVE-S hierarchy.

Implementation Details. We follow the part-discovery objective and training procedure of MPAE (Xia et al., 2025). The original MPAE setup uses a frozen ViT-B/14 DINOv2 backbone (Oquab et al., 2024) with register tokens. For a controlled comparison with HIVE-S, we reimplement MPAE using a frozen supervised ViT-S/16 backbone (Dosovitskiy et al., 2020). The model

is trained without ground-truth part annotations, and PartImageNet pixel-level part masks are used only for evaluation.

Because replacing ViT-B/14 with ViT-S/16 changes the spatial resolution and channel dimension of the feature map, we modify only the shape-dependent components of the MPAE implementation, such as the 1×1 projection layer and the spatial dimensions used in the matching block. All other algorithmic components and hyperparameters are kept the same as in MPAE (Xia et al., 2025). This ensures that the downstream part-discovery pipeline remains fixed, so that differences between MPAE and MPAE+HIVE-S reflect the representation supplied to MPAE.

Evaluation Protocol. We evaluate predicted part assignments against ground-truth part masks using Normalized Mutual Information (NMI) (Strehl & Ghosh, 2002) and Adjusted Rand Index (ARI) (Hubert & Arabie, 1985). Let $\mathcal{A}$ denote the partition induced by the predicted part labels and $\mathcal{B}$ denote the partition induced by the ground-truth part labels over image pixels. NMI is defined as

$$\mathrm{NMI}(\mathcal{A},\mathcal{B}) = \frac{2I(\mathcal{A};\mathcal{B})}{H(\mathcal{A}) + H(\mathcal{B})},$$

where $I(\mathcal{A};\mathcal{B})$ is the mutual information between the two partitions and $H(\cdot)$ denotes entropy. NMI measures agreement between predicted and ground-truth partitions, with larger values indicating better alignment.

ARI measures pairwise agreement between two partitions after correcting for chance. Let m_{rs} be the number of pixels assigned to predicted part group r and ground-truth part group s. Let $p_r = \sum_s m_{rs}$ and $q_s = \sum_r m_{rs}$ denote the corresponding marginal counts, and let N_{pix} denote the total number of evaluated pixels. ARI is defined as

$$\mathrm{ARI} = \frac{\sum_{rs}\binom{m_{rs}}{2} - \frac{\sum_r \binom{p_r}{2}\sum_s \binom{q_s}{2}}{\binom{N_{\mathrm{pix}}}{2}}}{\frac{1}{2}\left[\sum_r \binom{p_r}{2} + \sum_s \binom{q_s}{2}\right] - \frac{\sum_r \binom{p_r}{2}\sum_s \binom{q_s}{2}}{\binom{N_{\mathrm{pix}}}{2}}}.$$

ARI is permutation-invariant with respect to part indices and corrects for random clustering agreement, making it suitable for unsupervised part discovery where predicted part labels have no fixed semantic ordering.

We compute NMI and ARI at $P_{\mathrm{disc}}{=}25$ and $P_{\mathrm{disc}}{=}50$ discovered parts. Higher NMI and ARI indicate better agreement with annotated object parts. For MPAE and MPAE+HIVE-S, results in Table 2 are averaged over three random seeds and reported as mean $\pm$ standard error. Baseline numbers are taken from the corresponding papers unless otherwise stated.

Qualitative Results. Fig. 6 and Fig. 7 present qualitative comparisons on PartImageNet-O and PartImageNet-S. Across both splits, MPAE frequently produces fragmented or semantically mixed part assignments, whereas our method recovers more coherent and object-aligned decompositions.

In PartImageNet-O (Fig. 6), our method consistently separates meaningful object components. For instance, in animal examples, it isolates regions such as the head and body, while MPAE often merges these into a single diffuse segment or produces scattered masks. In scenes with background context, our method more cleanly separates foreground objects from surrounding regions, whereas MPAE frequently mixes object and background pixels. It also preserves spatially contiguous parts instead of breaking them into disconnected fragments.

In PartImageNet-S (Fig. 7), which contains more diverse and segmentation-oriented examples, these differences are more pronounced. For example, in vehicle scenes, our method separates the main body from attached components, while MPAE tends to group them together or produce noisy assignments. In animal and object-centric images, our method captures consistent part regions aligned with object structure, whereas MPAE produces over-segmented or entangled masks. In cluttered scenes, our method groups visually similar regions into coherent parts, while MPAE yields highly fragmented outputs.

Overall, these results show that our method improves both spatial coherence and semantic consistency of part assignments, making part-level structure more accessible than the raw ViT features used in the MPAE baseline.

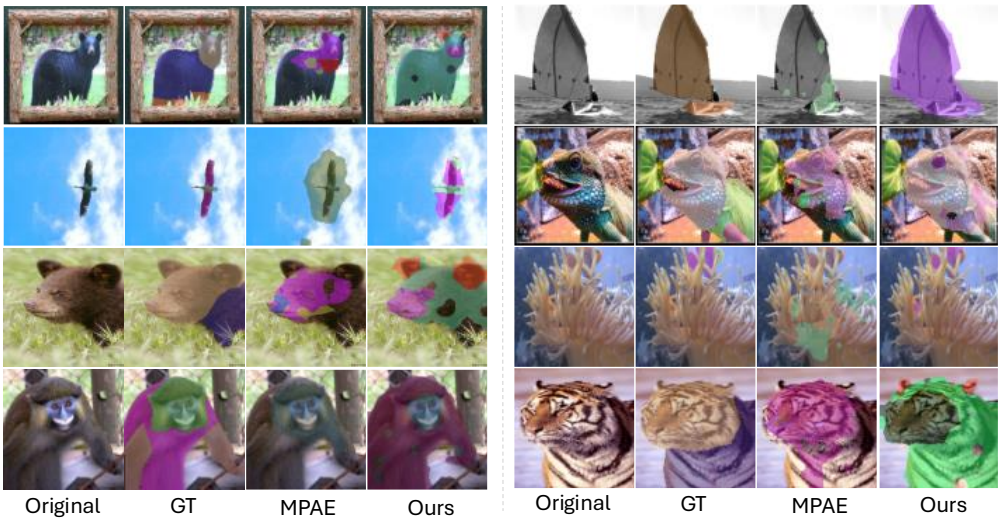

Figure 6: Qualitative unsupervised part-discovery results on PartImageNet-O. We show the original image, ground-truth part annotation, MPAE prediction, and our method (MPAE+HIVE-S). MPAE often produces fragmented or semantically mixed masks, whereas our method yields more coherent and object-aligned part decompositions. For example, our method separates distinct object components such as animal heads and bodies, isolates foreground objects from background, and avoids mixing unrelated regions, while MPAE frequently merges these regions or produces scattered assignments.

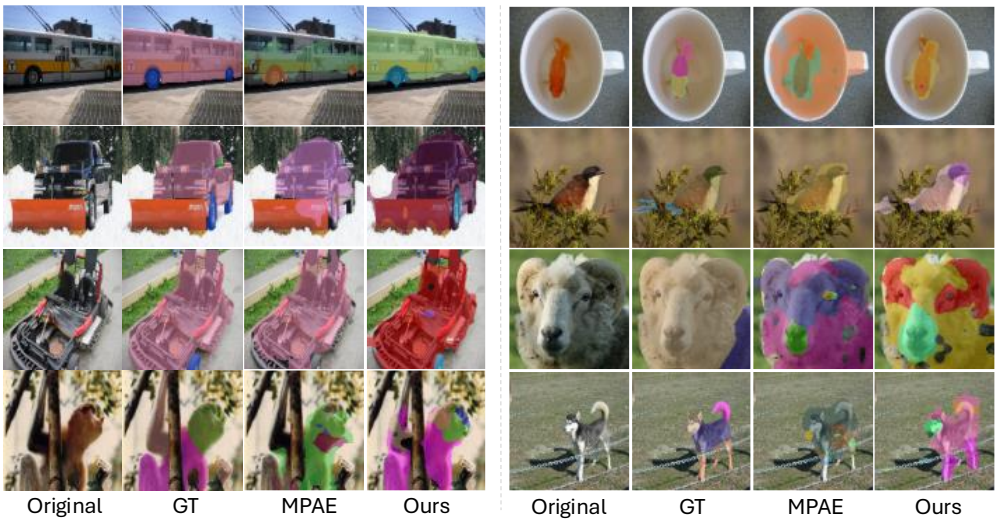

Figure 7: Qualitative unsupervised part-discovery results on PartImageNet-S. We show the original image, ground-truth part annotation, MPAE prediction, and our method (MPAE+HIVE-S). On this more diverse and segmentation-oriented split, our method produces more spatially consistent and semantically meaningful part masks. In particular, it better separates object components (e.g., vehicle body vs. attachments or animal parts) and reduces fragmentation compared to MPAE, which often yields noisy or entangled regions.

B.2.3 Task 3: Semantic Segmentation

Semantic segmentation evaluates whether the hierarchical representations learned by HIVE transfer to dense prediction, where the model must assign semantic labels at the pixel level rather than only producing image-level predictions. This setting complements our image-classification and part-

discovery evaluations by testing whether the learned hierarchy remains useful for spatially detailed visual understanding. Dataset statistics are provided in Sec. B.1.

Architecture. We evaluate HIVE under two complementary semantic-segmentation settings on ADE20K (Zhou et al., 2017).

First, we conduct a controlled comparison between ViT-S/16 and HIVE-S using the same Semantic FPN decoder (Kirillov et al., 2019), training recipe, data-augmentation pipeline, and evaluation protocol. This comparison isolates the effect of replacing the vanilla ViT-S/16 representation with the hierarchical HIVE-S representation.

Second, we evaluate HIVE-T and HIVE-S using UperNet (Xiao et al., 2018), following the setting used by the hierarchy-aware baselines in Table 4. The HIVE backbones are initialized from their corresponding ImageNet-1K classification models and fine-tuned end-to-end for semantic segmentation. This setting enables comparisons with DependencyViT and Hi-Mapper at corresponding model scales.

Implementation Details. For our semantic-segmentation experiments, the backbone and segmentation head are optimized end-to-end using the standard pixel-wise semantic-segmentation objective. No additional segmentation-specific hierarchy loss is introduced during fine-tuning.

Models are fine-tuned for $80{,}000$ iterations with AdamW (Loshchilov & Hutter, 2017). We use an initial learning rate of $6{\times}10^{-5}$, weight decay of $1{\times}10^{-4}$, polynomial learning-rate decay with power 0.9, and batch size 16. During training, images are randomly resized with scale range $[0.5, 2.0]$, horizontally flipped, cropped to $512{\times}512$, and augmented with photometric distortion.

Data Augmentation. During training, images are randomly resized with scale range $[0.5, 2.0]$, randomly horizontally flipped, randomly cropped to $512{\times}512$, and augmented with photometric distortion. The input resolution used for training and evaluation is summarized in Table 3.

Evaluation Protocol. We report mean Intersection-over-Union (mIoU) on the ADE20K validation set. Let $\mathcal{C}_{\mathrm{seg}}$ denote the semantic label set and let $\mathcal{P}_{\mathrm{eval}}$ denote the set of evaluated pixel locations. For each semantic label $c \in \mathcal{C}_{\mathrm{seg}}$, we define

$$\mathrm{IoU}_c = \frac{\left|\left\{\rho \in \mathcal{P}_{\mathrm{eval}} : \hat{\ell}(\rho) = c \ \wedge \ \ell^\star(\rho) = c\right\}\right|}{\left|\left\{\rho \in \mathcal{P}_{\mathrm{eval}} : \hat{\ell}(\rho) = c \ \vee \ \ell^\star(\rho) = c\right\}\right|},$$

where $\hat{\ell}(\rho)$ and $\ell^\star(\rho)$ denote the predicted and ground-truth semantic labels at pixel location ρ, respectively. The mean Intersection-over-Union is then computed as

$$\mathrm{mIoU} = \frac{1}{|\mathcal{C}_{\mathrm{seg}}|} \sum_{c \in \mathcal{C}_{\mathrm{seg}}} \mathrm{IoU}_c.$$

Higher mIoU indicates better agreement between predicted and ground-truth semantic segmentation masks.

At inference time, we use single-scale evaluation without multi-scale or flip testing. Results for ViT-S/16, HIVE-T, and HIVE-S implemented in our study are averaged over three random seeds and reported as mean $\pm$ standard error. Rows marked with $*$ in Table 4 use UperNet (Xiao et al., 2018); other baseline results are taken from the corresponding papers unless otherwise stated.

Quantitative Results. As shown in Table 4, HIVE transfers effectively to dense semantic prediction under both evaluation settings.

In the controlled Semantic FPN comparison, HIVE-S improves the reproduced ViT-S/16 baseline from 39.9 $\pm$ 0.2 to 41.1 $\pm$ 0.2 mIoU under the same decoder, training recipe, data-augmentation pipeline, and evaluation protocol. The resulting +1.2-point gain isolates the effect of replacing the vanilla ViT-S/16 representation with the hierarchical HIVE-S representation while keeping the downstream segmentation pipeline fixed.

Table 4: Semantic segmentation results on the ADE20K validation set. ViT-S/16 and HIVE-S (Semantic FPN) are evaluated using the same Semantic FPN decoder (Kirillov et al., 2019), training recipe, and single-scale inference protocol. Rows marked with $*$ use UperNet (Xiao et al., 2018). For the UperNet setting, HIVE-T and HIVE-S are compared with hierarchy-aware baselines at corresponding model scales. Results for our reproduced ViT-S/16 and HIVE models are averaged over three random seeds and reported as mean $\pm$ standard error; other baseline results are taken from the corresponding papers unless otherwise stated.

Backbone	#Param (M) ↓	FLOPs (G) ↓	mIoU (%) ↑
Standard CNN/ViT backbones			
ResNet-50 (He et al., 2016)	28.5	45.4	36.7
PVT-Small (Wang et al., 2021)	28.2	42.9	39.8
Swin-T (Liu et al., 2021)	31.9	46.0	41.5
ViT-S/16 (Dosovitskiy et al., 2020)	25.9	26.1	39.9 ± 0.2
HIVE-S (Ours)	**28.1**	**34.4**	$\mathbf{41.1 \pm 0.2}$
Hierarchy-aware / structured vision models			
DependencyViT-T* (Ding et al., 2023)	11.1	14.5	40.3
Hi-Mapper (DeiT-T)* (Kwon et al., 2024)	11.6	14.4	39.8
HIVE-T* (Ours)	**10.9**	**14.3**	$\mathbf{43.6 \pm 0.2}$
DependencyViT-S* (Ding et al., 2023)	43.1	57.4	45.7
Hi-Mapper (Swin-T)* (Kwon et al., 2024)	62.1	94.9	46.8
HIVE-S* (Ours)	**41.7**	**57.0**	$\mathbf{47.2 \pm 0.2}$

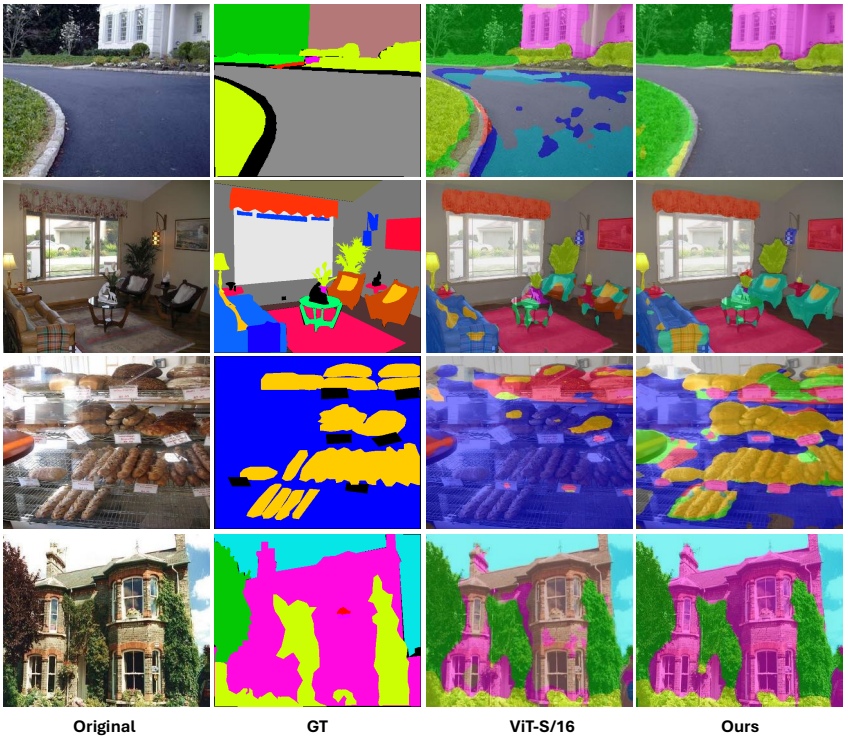

Figure 8: Qualitative semantic segmentation results on ADE20K. Compared to ViT-S/16, which produces noisy and fragmented predictions, HIVE-S yields more coherent regions with improved boundary alignment. For example, ours (HIVE-S) better separates road, vegetation, and building regions in outdoor scenes and produces cleaner object-level segmentation in indoor layouts (e.g., chairs, tables, and plants).

Under the UperNet setting, the gains are also consistent across model scales. At the Tiny scale, HIVE-T achieves 43.6 ± 0.2 mIoU, compared with 40.3 for DependencyViT-T and 39.8 for Hi-Mapper (DeiT-T), while using a comparable parameter and FLOP budget. At the Small scale, HIVE-S achieves 47.2 ± 0.2 mIoU, compared with 45.7 for DependencyViT-S and 46.8 for Hi-

Table 5: Hierarchy-size ablation for HIVE-S on ImageNet-1K. We compare two-level and three-level hierarchies while varying the number of feature embeddings at each level. The default configuration used in the main experiments is $(128, 64, 32)$.

Hierarchy	Feature-embedding configuration	Top-1 Acc. (%) ↑
Two-level	$(512, 256)$	82.5
	$(256, 128)$	82.6
	$(128, 64)$	82.5
	$(64, 32)$	82.4
Three-level	$(512, 256, 128)$	82.7
	$(256, 128, 64)$	82.6
	$(128, 64, 32)$	82.8
	$(64, 32, 16)$	82.5

Table 6: One-module-at-a-time ablation for HIVE-S on ImageNet-1K. Each variant modifies only the indicated component while using the same training and evaluation protocol. ✓ denotes that the corresponding component is retained.

Variant	Binding	Graph refinement	Learned pooling	Multi-level projection	Top-1 Acc. (%) ↑
Full HIVE	✓	✓	✓	✓	**82.8**
No binding	–	✓	✓	✓	79.9
No graph attention	✓	–	✓	✓	80.7
Fixed pooling	✓	✓	–	✓	80.4
Final level only	✓	✓	✓	–	80.1

Mapper (Swin-T). These results indicate that the hierarchical representations learned by HIVE transfer beyond image-level classification and part discovery to dense semantic prediction across multiple model scales.

Qualitative Results. Fig. 8 shows qualitative comparisons on ADE20K. Compared to ViT-S/16, which often produces noisy, over-fragmented predictions and mixes semantically distinct regions, HIVE-S yields more coherent and spatially consistent segmentation maps.

In outdoor scenes, HIVE-S better separates large-scale regions such as road, vegetation, sky, and buildings, while ViT-S/16 exhibits boundary leakage and scattered predictions. In indoor scenes, HIVE-S produces cleaner object-level segmentation, clearly delineating furniture such as chairs, tables, and plants, whereas ViT-S/16 tends to merge or fragment these regions. For cluttered scenes such as the bakery display, HIVE-S groups visually similar objects into consistent regions, while ViT-S/16 produces highly fragmented assignments. Similarly, in building scenes, HIVE-S more accurately separates structural components such as walls and vegetation with sharper boundaries.

These results indicate that the hierarchical representation learned by HIVE-S improves both spatial consistency and semantic grouping in dense prediction tasks.

B.3 Ablation

We conduct four ablation studies on ImageNet-1K to examine how HIVE-S depends on hierarchy size, individual architectural components, auxiliary objectives, and multi-level projected representations. Unless otherwise stated, all ablations use HIVE-S with the same ImageNet-1K training protocol described in Sec. B.2.1. Detailed results for the four ablation studies are reported in Tables 5, 6, 7, and 8, respectively.

Hierarchy Size. We examine whether HIVE-S benefits from a deeper hierarchy or simply from using more feature states. As shown in Table 5, performance remains stable across a broad range of two- and three-level configurations, with top-1 accuracy varying only slightly. The default three-level configuration $(K^{(1)}, K^{(2)}, K^{(3)}) = (128, 64, 32)$ achieves the best result of 82.8% and is used throughout the main experiments. Increasing the hierarchy size to $(256, 128, 64)$ or $(512, 256, 128)$ does not yield further gains, while reducing it to $(64, 32, 16)$ slightly degrades performance. The

Table 7: Ablation over auxiliary objectives on ImageNet-1K. The task loss $\mathcal{L}_{\text{task}}$ is always used. ✓ denotes inclusion of the corresponding auxiliary objective. Accuracy is reported as top-1 validation accuracy.

Case	$\mathcal{L}_{\text{cons}}$	$\mathcal{L}_{\text{align}}$	$\mathcal{L}_{\text{div}}$	$\mathcal{L}_{\text{link}}$	Top-1 Acc. (%) ↑
1	–	–	–	–	80.6
2	✓	–	–	–	81.5
3	–	✓	–	–	81.2
4	–	–	✓	–	81.9
5	–	–	–	✓	81.5
6	✓	✓	–	–	82.3
7	✓	–	✓	–	82.4
8	✓	–	–	✓	82.4
9	–	✓	✓	–	82.4
10	–	✓	–	✓	82.4
11	–	–	✓	✓	82.4
12	✓	✓	✓	–	82.5
13	✓	✓	–	✓	82.5
14	✓	–	✓	✓	82.5
15	–	✓	✓	✓	82.5
16	✓	✓	✓	✓	**82.8**

Table 8: Inference-time ablation of multi-level projected representations on ImageNet-1K. Each row retains a different subset of hierarchy levels without retraining. Backbone: ViT-S/16; hierarchy configuration: $(K^{(1)}, K^{(2)}, K^{(3)}) = (128, 64, 32)$.

Levels used at inference	Top-1 Acc. (%) ↑
Low only ($\bar{\mathbf{Z}}^{(1)}$)	75.3
Low + Mid ($\bar{\mathbf{Z}}^{(1)} + \bar{\mathbf{Z}}^{(2)}$)	79.2
Full hierarchy ($\bar{\mathbf{Z}}^{(1)} + \bar{\mathbf{Z}}^{(2)} + \bar{\mathbf{Z}}^{(3)}$)	**82.8**

two-level hierarchy shows a similar trend, with its best result at $(256, 128)$. Overall, these results indicate that HIVE-S does not benefit from simply increasing the number of feature states and instead performs well with a compact multi-level hierarchy.

One-Module-at-a-Time Ablation. We next examine whether HIVE-S benefits from the coordinated use of its individual architectural components. Table 6 reports controlled one-module-at-a-time ablations, where only the indicated component is modified and all variants use the same ImageNet-1K training and evaluation protocol.

Removing the feature-binding module and applying the hierarchy directly to ViT patch tokens produces the largest degradation, reducing top-1 accuracy from 82.8% to 79.9%. This indicates that constructing compact, input-conditioned feature states before hierarchical reasoning is important. Removing graph attention and applying hierarchical pooling to the unrefined feature states reduces accuracy to 80.7%, showing that relational refinement provides additional benefit beyond direct aggregation. Replacing the learned image-conditioned assignments $\mathbf{S}^{(l)}$ with fixed pooling assignments reduces accuracy to 80.4%, supporting the importance of learned cross-level grouping. Finally, using only the coarsest hierarchy level instead of projecting and combining representations from all levels yields 80.1%. Overall, these ablations show that no single architectural component accounts for the full gain: feature binding, relational refinement, learned cross-level assignment, and multi-level projection provide complementary contributions to the final representation.

Auxiliary Objectives. We next examine whether HIVE-S depends on a particular auxiliary objective or benefits from their combined structural regularization. Table 7 reports all $2^4 = 16$ combinations of the auxiliary objectives while keeping $\mathcal{L}_{\text{task}}$ active. Using only $\mathcal{L}_{\text{task}}$ yields 80.6% top-1 accuracy, whereas the full objective achieves the best result of 82.8%, corresponding to a $+2.2$-point improvement with no additional inference cost.

Each single-loss variant improves over the task-only baseline, while several two-loss combinations reach 82.4% and the three-loss variants achieve approximately 82.5%. Thus, most of the performance gain can already be obtained with a subset of the auxiliary objectives, while their joint use provides the strongest result. Overall, these results suggest that $\mathcal{L}_{\text{cons}}$, $\mathcal{L}_{\text{align}}$, $\mathcal{L}_{\text{div}}$, and $\mathcal{L}_{\text{link}}$ provide complementary structural regularization rather than the improvement being driven by a single isolated term.

Multi-level Projected Representations. Finally, we examine whether all hierarchy levels contribute to prediction or whether the final representation is dominated by the lowest-level embeddings. Since the final patch-space representation is defined as

$$\mathbf{Z}' = \sum_{l=1}^{L} \bar{\mathbf{Z}}^{(l)},$$

we can ablate hierarchy levels at inference time without retraining.

As shown in Table 8, using only the low-level projection $\bar{\mathbf{Z}}^{(1)}$ gives 75.3% top-1 accuracy. Adding the mid-level projection increases accuracy to 79.2%, while using the full hierarchy further improves performance to 82.8%. The monotonic improvement as additional levels are retained indicates that the level-specific projected representations provide complementary information. In particular, the gain from low to low+mid shows that intermediate representations contribute substantial predictive evidence, while the additional gain from the high-level projection demonstrates that the coarsest representation remains useful. These results support the interpretation of HIVE-S as a genuinely multi-level representation rather than a model whose predictions are driven primarily by the lowest-level feature embeddings.

B.4 Part-whole Contribution Analysis

Definition Detail. We define the *part–whole contribution score* to quantify how much each lower-level feature state contributes to an adjacent higher-level feature state for a given image. Let $\mathbf{\Phi}^{(\ell)} \in \mathbb{R}^{P \times K^{(\ell)}}$ denote the patch-level feature-composition score map at hierarchy level ℓ from Eq. (9) in the main text, where P is the number of patches and $K^{(\ell)}$ is the number of feature states at that level. In a multi-level hierarchy, the score is computed recursively only between adjacent levels, e.g., $(\ell, \ell+1)$, such as low-to-mid and mid-to-high.

We first rectify the score maps to be nonnegative and normalize each feature channel over patches. To focus on the strongest spatial support of each feature state, we apply feature-wise quantile thresholding using feature-specific thresholds τ. The retained values are then renormalized over patches, yielding the thresholded soft footprints $\mathbf{F}_{\text{thr}}^{(\ell)}$ and $\mathbf{F}_{\text{thr}}^{(\ell+1)}$.

We define the *influence* from lower-level embedding i at level ℓ to higher-level embedding j at level $\ell+1$ as

$$\mathbf{I}^{(\ell \to \ell+1)} = \big(\mathbf{F}_{\text{thr}}^{(\ell)}\big)^{\top} \mathbf{F}_{\text{thr}}^{(\ell+1)},$$

whose (i,j)-th entry is

$$I_{ij}^{(\ell \to \ell+1)} = \sum_{p=1}^{P} F_{\text{thr}}^{(\ell,i)}(p)\, F_{\text{thr}}^{(\ell+1,j)}(p).$$

This term measures the degree of spatial co-activation between the two feature states across adjacent levels.

To capture directional part-inside-whole structure, we further define the *containment* score as

$$C_{ij}^{(\ell \to \ell+1)} = \frac{\sum_{p=1}^{P} \min\big(F_{\text{thr}}^{(\ell,i)}(p), F_{\text{thr}}^{(\ell+1,j)}(p)\big)}{\sum_{p=1}^{P} F_{\text{thr}}^{(\ell,i)}(p) + \varepsilon},$$

where $\varepsilon > 0$ is a small constant for numerical stability. This term measures the extent to which the activated support of the lower-level embedding is spatially contained within that of the adjacent higher-level states.

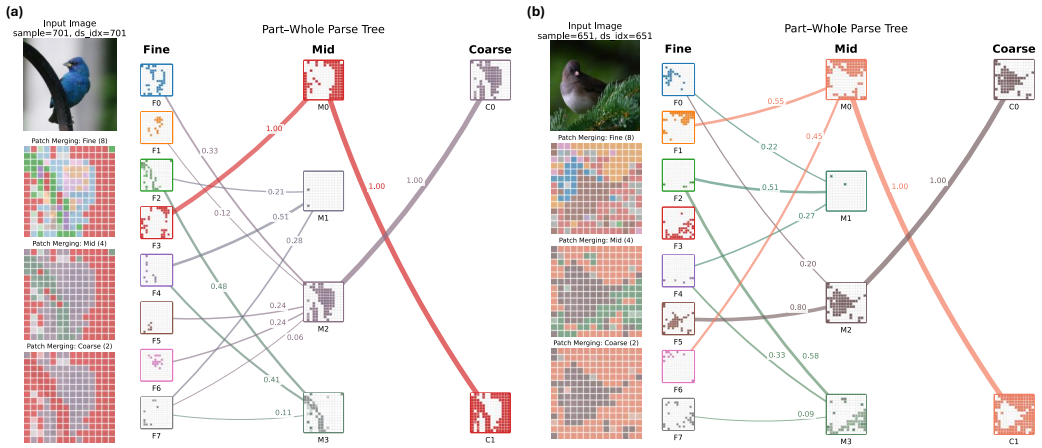

Figure 9: Additional part–whole visualizations for two ImageNet bird samples. Fine-level assignments are progressively consolidated into mid- and coarse-level groups, and edges with contribution values ≤ 0.05 are omitted for clarity. In both examples, the mid-level embedding $M3$ captures semantically meaningful contextual structure, corresponding to a branch in **(a)** and surrounding leaves in **(b)**. These examples suggest that the learned hierarchy can represent not only coarse object-level grouping but also coherent intermediate scene context.

The final part–whole contribution score is obtained by combining influence and containment multiplicatively:

$$\widetilde{R}_{ij}^{(\ell\to\ell+1)} = I_{ij}^{(\ell\to\ell+1)}\, C_{ij}^{(\ell\to\ell+1)}.$$

Intuitively, this favors embedding pairs that are both strongly co-active and spatially nested. To facilitate interpretation, we normalize each higher-level column so that it forms a distribution over lower-level contributors:

$$R_{ij}^{(\ell\to\ell+1)} = \frac{\widetilde{R}_{ij}^{(\ell\to\ell+1)}}{\sum_{i'} \widetilde{R}_{i'j}^{(\ell\to\ell+1)} + \varepsilon}.$$

Under this normalization, each column of $\mathbf{R}^{(\ell\to\ell+1)} \in \mathbb{R}^{K^{(\ell)}\times K^{(\ell+1)}}$ represents the compositional structure of one higher-level feature state in terms of lower-level parts. Thus, $R_{ij}^{(\ell\to\ell+1)}$ can be interpreted as the relative contribution of lower-level states i at level ℓ to higher-level states j at level $\ell+1$. In hierarchies with more than two levels, the same computation is applied recursively between adjacent levels, yielding a sequence of local part–whole relations such as low-to-mid and mid-to-high.

Results. Figs. 9–11 provide additional ImageNet examples that further illustrate the interpretability of the learned part–whole hierarchy. Across all cases, fine-level feature states are progressively consolidated into mid- and coarse-level groups, while the parse trees summarize the dominant adjacent-level relations using cross-level contribution scores, with weak edges omitted for clarity. Together, these examples show that the learned hierarchy captures not only coarse object-level grouping, but also semantically meaningful intermediate structure at the mid level.

In Fig. 9, the mid-level state $M3$ consistently captures meaningful contextual structure across two bird images, corresponding to a branch in **(a)** and surrounding leaves in **(b)**. This suggests that the learned hierarchy does not treat all non-object regions as undifferentiated background, but can instead isolate coherent scene context as an intermediate component. In Fig. 10, the mid-level state $M1$ captures localized object-part structure in both examples, highlighting the human head/face region in **(a)** and the rooster's comb in **(b)**. This indicates that HIVE can recover semantically meaningful intermediate parts across distinct object categories. Finally, Fig. 11 shows that the hierarchy can also disentangle both object–context and fine-grained part-level structure. In **(a)**, $M2$ captures the steel cage as a coherent contextual element, while in **(b)**, $M1$ emphasizes the fish fins as a meaningful object part. Taken together, these additional examples suggest that HIVE learns flexible and visu-

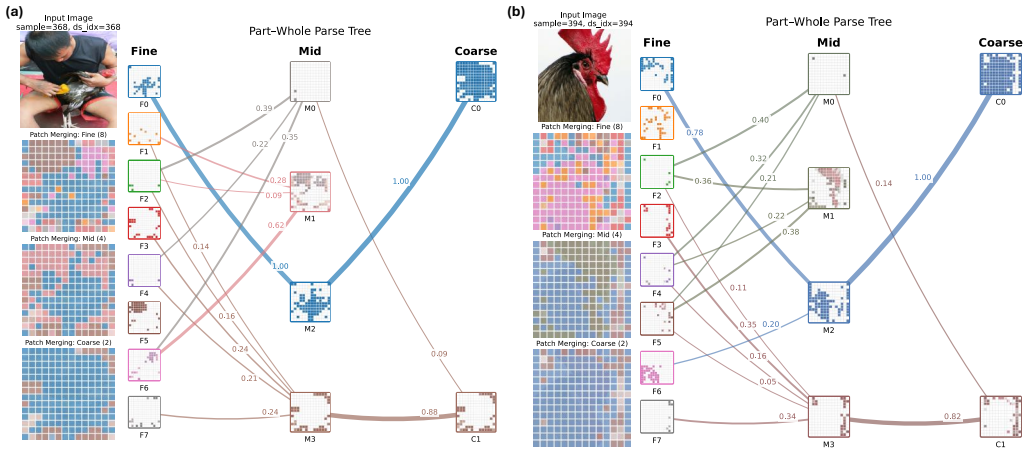

Figure 10: Additional part–whole visualizations for two ImageNet samples. Fine-level assignments are progressively consolidated into mid- and coarse-level groups, and edges with contribution values ≤ 0.05 are omitted for clarity. In both examples, the mid-level embedding $M1$ captures semantically meaningful localized structure: in **(a)** it highlights the human head/face region, while in **(b)** it appears to localize the rooster's comb. These examples suggest that the learned hierarchy can represent not only coarse scene-level groupings but also interpretable intermediate object-part structure.

ally meaningful part–whole dependencies spanning contextual structure, localized object parts, and object–context decomposition.

To verify that the observed cross-level organization is not specific to the compact 8–4–2 hierarchy used for visualization, we additionally analyzed three independently trained full-scale HIVE-S models with hierarchy $(128, 64, 32)$ on 1,000 ImageNet validation images. For each adjacent transition, each child state was assigned to its strongest-contributing parent, and the child and parent activation maps were converted to binary spatial supports using fixed top-activation fractions. We quantified spatial nesting using child-in-parent containment, defined as the fraction of the child support overlapping its assigned parent support.

We compared the learned correspondence with a degree-preserving permutation null (Good, 2005; Ernst, 2004). For each image, model, and hierarchy transition, we randomly permuted the observed parent labels across child states while preserving their multiset. This keeps the child and parent spatial supports and the exact number of children assigned to each parent unchanged, while disrupting only the learned child–parent correspondence. We sampled 100 permissible permutations per image.

For each independently trained model, containment values were averaged across images, and we report mean $\pm$ SD across the three models. Learned containment was substantially greater than the permutation null at both $128 \to 64$ (0.428 ± 0.002 vs. 0.280 ± 0.001) and $64 \to 32$ (0.842 ± 0.027 vs. 0.427 ± 0.012). These results indicate that the cross-level spatial organization observed in the compact visualization model also emerges reliably in the full-scale HIVE-S hierarchy.

B.5 Feature States Separability Analysis

Here, we examine whether each learned feature state $\zeta_n^{(l)}$ captures class-relevant visual structure, rather than only spatial location or weak activation patterns. If a feature state encodes a meaningful attribute, its instances from images of the same class should form coherent groups, while instances from different classes should be separated.

Setup. We analyze the feature states produced by the best HIVE-S configuration from Task 1 in Sec. 4.1.1 of the main text. For qualitative visualization, we randomly sample 20 ImageNet-1K classes and project the corresponding feature states into two dimensions with t-SNE. This reduced class count is used only for readability, as it keeps individual classes visually distinguishable in the projection. Also, we randomly choose 5 feature states for the figures.

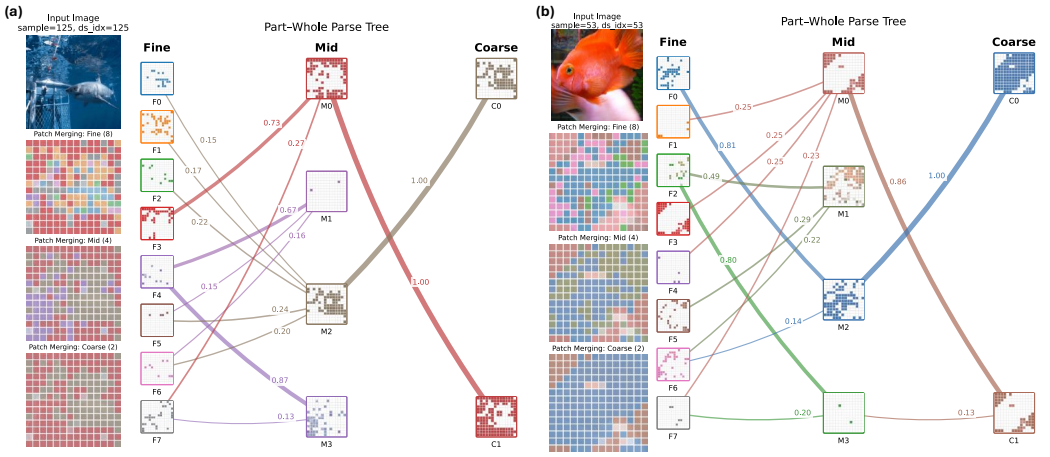

Figure 11: Additional part–whole visualizations for two ImageNet samples. Fine-level assignments are progressively consolidated into mid- and coarse-level groups, and edges with contribution values ≤ 0.05 are omitted for clarity. In **(a)**, the learned hierarchy separates semantically distinct scene elements, with the mid-level embedding $M2$ capturing the steel cage. In **(b)**, the hierarchy also identifies meaningful object-part structure, with $M1$ highlighting the fish fins. Together, these examples illustrate that the learned hierarchy can capture both object–context relations and fine-grained part-level structure.

Results. Across random subsets, we observe the same qualitative pattern as in Fig. 12, 13, 14. Fine-level feature states remain heterogeneous: some embeddings show weak or diffuse class structure, while others separate a subset of classes more clearly. By contrast, mid- and coarse-level feature states more consistently produce compact class-wise clusters. This supports the interpretation that the upper hierarchy aggregates fine-level evidence into more uniformly class-aligned feature representations.

Quantitative Separability Metrics We further quantify the class-wise separability of feature states on the full ImageNet-1K validation set, using all $1{,}000$ classes and 50 images per class. Unlike the t-SNE visualizations, which use 20 classes for visual clarity, this quantitative evaluation is performed on the full validation set.

We use four metrics. The intra-/inter-class ratio measures the average within-class distance divided by the average between-class distance; lower values indicate tighter within-class clustering relative to between-class separation. Nearest-centroid accuracy measures whether each embedding instance is closest to the centroid of its own class. The silhouette score measures cluster separation, with larger values indicating clearer class structure. The centroid distance reports the average pairwise distance between class centroids. For each hierarchy level, we compute these metrics for all feature states at that level and report the mean and standard deviation across embeddings.

Table 9 shows that the upper hierarchy yields more class-aligned feature states than the fine level. The intra-/inter-class ratio decreases from 0.926 at the fine level to approximately 0.61 at the mid and coarse levels, indicating tighter within-class grouping relative to between-class spread. The silhouette score also changes from negative at the fine level to positive at the upper levels, suggesting clearer class-wise organization after hierarchical aggregation. Nearest-centroid accuracy follows the same trend, reaching 0.841 at both the mid and coarse levels. Therefore, these results support the qualitative observation that fine-level feature states provide a heterogeneous pool of local evidence, whereas upper-level feature states form more uniformly class-aligned representations.

B.5.1 Hierarchical Causal Mediation

We next test whether this hierarchical organization is causally used by the model. Concretely, we ask whether perturbing a class-relevant fine-level feature states affects the final prediction only through its direct local contribution, or whether its effect is amplified once the upper hierarchy is recomputed.

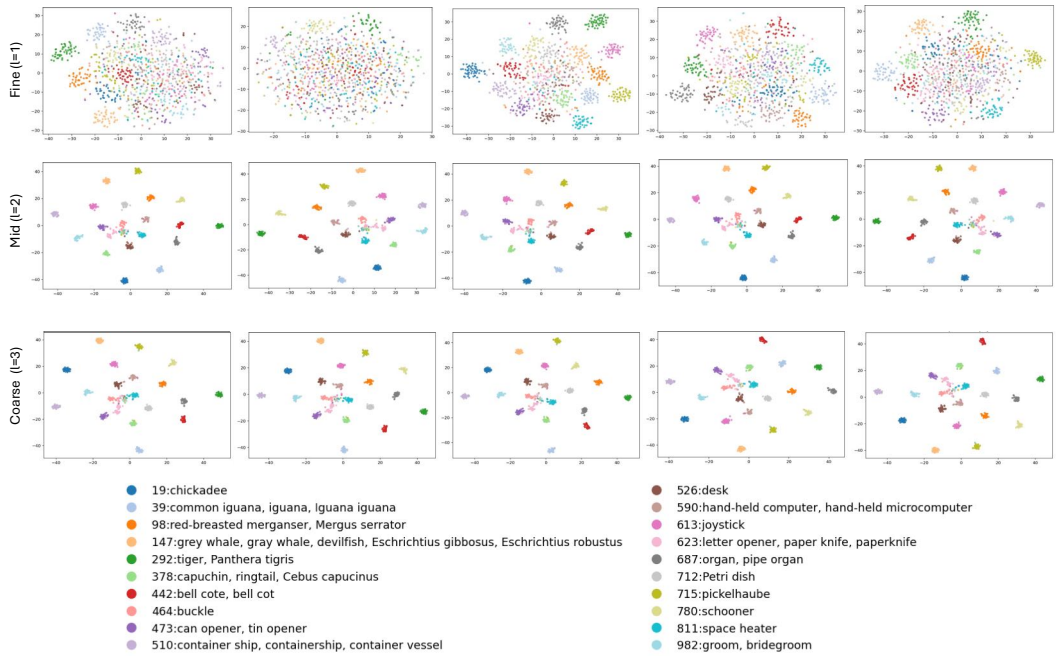

Figure 12: Additional t-SNE visualizations of feature states across multiple random 20-class subsets of ImageNet-1K (**seed 43**). Each panel visualizes the instantiations of one feature state, colored by class. The qualitative pattern is consistent across seeds: fine-level feature states are more heterogeneous, whereas mid- and coarse-level feature states form more compact class-wise clusters.

Table 9: Intra-/inter-class consistency of feature states on the ImageNet-1K validation set (1,000 classes, 50 images per class). Metrics are computed in the raw 384-dimensional embedding space and averaged over all feature states at each level. We report mean ± standard deviation across embeddings.

Level	$K^{(l)}$	Intra/Inter ↓	NC Acc. ↑	Silhouette ↑	Centroid Dist. ↑
Fine	128	0.926 ± 0.033	0.763 ± 0.112	-0.066 ± 0.043	5.799 ± 1.461
Mid	64	0.611 ± 0.001	0.841 ± 0.000	0.205 ± 0.001	6.253 ± 0.810
Coarse	32	0.607 ± 0.002	0.841 ± 0.000	0.205 ± 0.002	9.921 ± 1.480

Setup. For each sample, we identify class-relevant fine-level feature states using a class-conditional importance score. Let $\mathrm{Score}_{p,n}$ denote the assignment strength of fine-level feature states n at patch p, and let $\mathrm{contrib}_{p,y}$ denote the contribution of patch p to the ground-truth class y. We define

$$I(n; y) = \sum_{p=1}^{P} \mathrm{Score}_{p,n} \cdot \mathrm{contrib}_{p,y}. \tag{11}$$

For each sample, the top-3 fine-level embeddings under $I(n; y)$ are targeted to be intervened.

We use three reference vectors that do not introduce a counterfactual class direction. The *global-mean* reference replaces the selected feature vector with the dataset-wide mean vector for the same embedding index. The *random-init* reference samples a vector from the empirical mean and standard deviation of the same embedding index, preserving scale while removing image-specific semantic content. The *zero* reference removes the feature vector entirely.

We compare two intervention regimes. In the *direct* regime, the selected fine-level feature states are replaced while the upper-level feature states are held fixed at their pre-intervention values. In the *total* regime, the selected fine-level embedding is replaced and the upper hierarchy is recomputed

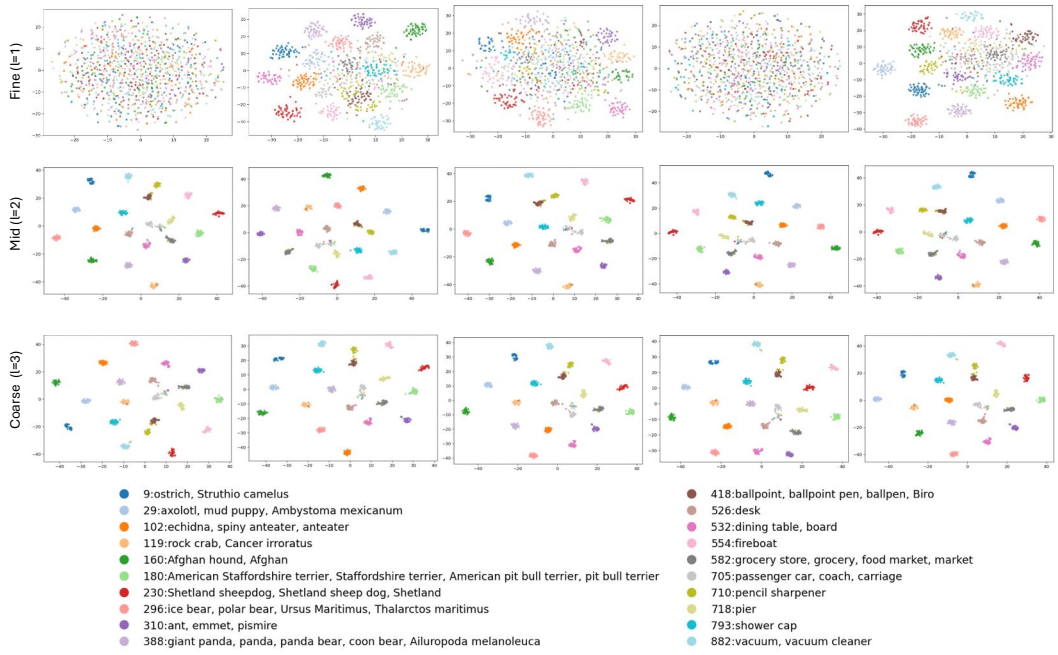

Figure 13: Additional t-SNE visualizations of feature states across multiple random 20-class subsets of ImageNet-1K (**seed 44**). Each panel visualizes the instantiations of one feature state, colored by class. The qualitative pattern is consistent across seeds: fine-level feature states are more heterogeneous, whereas mid- and coarse-level feature states form more compact class-wise clusters.

Table 10: Hierarchical causal mediation on target fine-level feature embeddings (top-3 per sample under $I(n; y)$).

Reference	Direct Δp_y	Indirect Δp_y
Global-mean	-0.00289	$\mathbf{-0.00832}$
Random-init	-0.00322	$\mathbf{-0.00831}$
Zero	-0.00401	$\mathbf{-0.01058}$

end-to-end through Ψ and Ω. The indirect effect is then defined as

$$\Delta p_y^{\text{indirect}} \triangleq \Delta p_y^{\text{total}} - \Delta p_y^{\text{direct}}, \tag{12}$$

which isolates the portion of the prediction change that is mediated by the learned hierarchy.

Results. Table 10 shows that a substantial fraction of the effect of perturbing a target fine-level feature states is mediated by the upper hierarchy. Under global-mean replacement, target embeddings reduce the ground-truth probability by only -0.00289 in the direct regime, but by -0.01121 once the upper hierarchy is recomputed, yielding an indirect effect of -0.00832. Random-init replacement gives a nearly identical indirect effect of -0.00831, indicating that the mediated effect does not depend on the semantic content of the substitute vector. Zero replacement produces the largest total effect, as expected from a stronger ablation. Across all three references, the indirect effect is several times larger in magnitude than the corresponding direct effect, supporting a causal reading of the learned hierarchy: class-relevant fine-level feature states are not merely local descriptors, but inputs whose perturbations propagate through upper-level feature formation and substantially affect the final prediction.

C LIMITATIONS

While HIVE induces coherent hierarchical visual representations, its hierarchy is model-internal and is not directly supervised to align with human-defined semantic parts; accordingly, the resulting

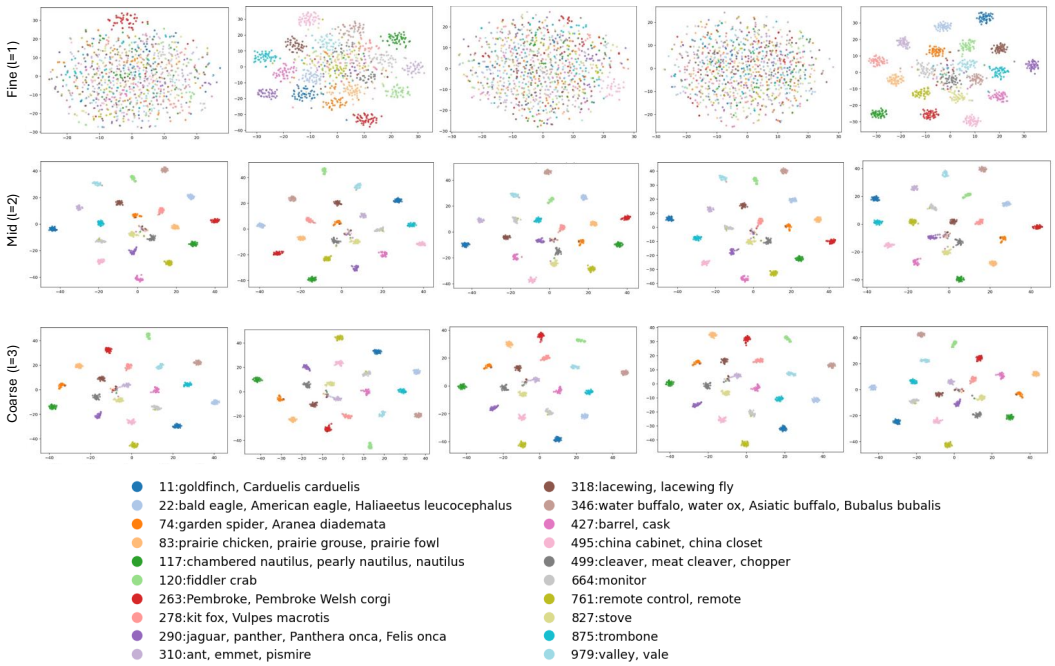

Figure 14: Additional t-SNE visualizations of feature states across multiple random 20-class subsets of ImageNet-1K (**seed 45**). Each panel visualizes the instantiations of one feature state, colored by class. The qualitative pattern is consistent across seeds: fine-level feature states are more heterogeneous, whereas mid- and coarse-level feature states form more compact class-wise clusters.

parse trees should be interpreted as cross-level feature compositions. HIVE also relies on a fixed hierarchy shape, which may be suboptimal for images with varying structural complexity. Future work could explore image-adaptive hierarchies and vision-language grounding of the learned components.