RELVIT:

CONCEPT-GUIDED VISION TRANSFORMER FOR VISUAL RELATIONAL REASONING

By M. Hassan Shaikh(21d110016) & Pranav Singla(200040102)

Objectives of paper:

(RelViTs): The paper discusses the architecture of vision transformers, which tokenize images into patches for processing. This approach allows ViTs to capture complex relationships in visual data, making them suitable for tasks requiring relational reasoning, following are the key aspects paper is based on:

1. **Object centric learning:**

   highlights the role of concepts—defined as object entities and their relationships—in enhancing the reasoning capabilities of models. By leveraging these concepts, models can better understand and predict interactions within visual scenes.

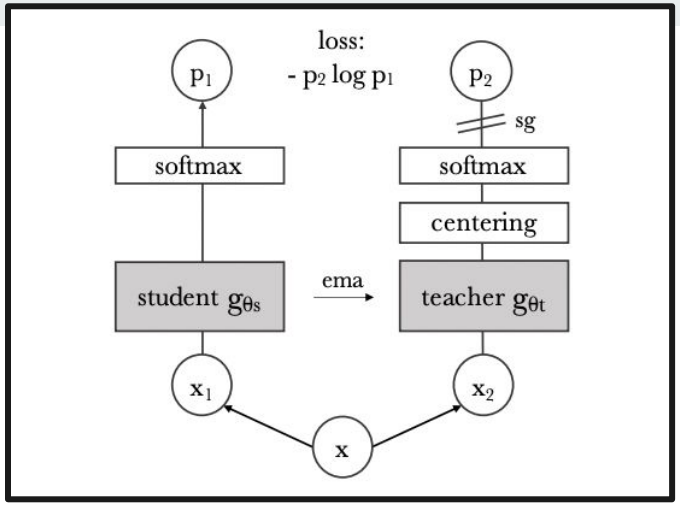
2. **Relational Reasoning**: -

   emphasize the importance of relational reasoning in visual tasks, where understanding the relationships between objects is crucial for accurate interpretation..

3. **Systematic Generalization**:

   introduces the idea of systematic generalization, which refers to a model's ability to apply learned knowledge to novel situations.

Model used EsViT and DINO



DINO (Self-Distillation with No Labels)

- **Overview:**
  - DINO is a self-supervised learning method that leverages self-distillation without requiring labeled data.
  - It trains a "student" and "teacher" network simultaneously, where the student network learns to mimic the teacher's predictions.
  - The teacher network's parameters are updated using an exponential moving average (EMA) of the student's parameters, allowing the teacher to maintain stability and consistency.
- **Architecture:**
  - **Input Image**: Multiple augmented views of the same image are generated.
  - **Teacher & Student Networks**: Both networks map these views to probability distributions using a shared projection head.
  - **Loss Function**: The student network's parameters are optimized by minimizing the cross-entropy loss between the teacher's and student's output distributions.
- **Why DINO?**
  - DINO's ability to learn without labels makes it suitable for large-scale image datasets where manual annotation is challenging.
  - It captures semantic features that are useful for downstream tasks, such as image classification and object detection.
  - The use of multiple views and self-distillation helps the model learn robust and invariant features.

Model used EsViT and DINO

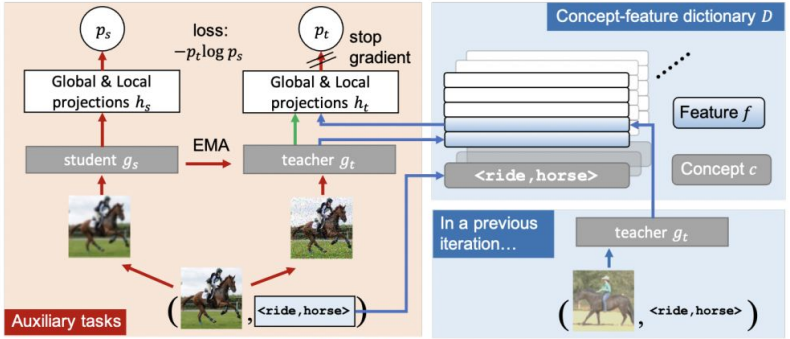


Figure 1: An overview of our method. Red+Green: the learning pipeline of DINO (Caron et al., 2021) and EsViT (Li et al., 2021); Red+Blue: our pipeline. We introduce a *concept-feature dictionary*, where the key is a concept c and its value is a queue of image features f with the same concept, to allow flexible feature retrieval with the concept keys. With the proposed dictionary, we further develop our concept-guided global and local tasks. EMA denotes the exponential moving average.

EsViT (Enhanced Self-supervised Vision Transformer)

- **Overview:**
  - EsViT builds on DINO, extending it to enhance performance on dense prediction tasks such as segmentation and object detection.
  - It introduces dense self-supervised learning, which allows the model to learn both global and local features from images.
- **Architecture:**
  - **Dense SSL**: In addition to the global feature alignment in DINO, EsViT focuses on local features by dividing the image into patches.
  - **Teacher & Student Networks**: Similar to DINO, but EsViT emphasizes the learning of correspondence between local patches of different views of the same image.
  - **Local Loss Function**: EsViT adds a patch-level loss that encourages the model to learn semantic correspondences across different image views, improving its ability to perform dense predictions.
- **Why EsViT?**
  - EsViT is specifically designed to handle tasks that require detailed, pixel-level understanding, such as segmentation.
  - The combination of global and local feature learning makes EsViT versatile for a wide range of vision tasks.
  - It outperforms standard ViT models in tasks that require fine-grained predictions.

Conceptual approach

Based on Features

- **Concept-Feature Dictionary**: Introduce a novel concept-feature dictionary where each key represents a concept (e.g., an object or action), and its value is a queue of image features sharing the same concept.
- **Dynamic Retrieval**: The dictionary allows dynamic retrieval of image features during training, enhancing the flexibility and effectiveness of the Vision Transformer (ViT) model.

Conceptual approach

Local

- **Object-Centric Learning**: The local task focuses on guiding the model to discover object-centric semantic correspondences across images, boosting the model's ability to recognize and reason about object relationships.
- **Correspondence Learning**: Leverages correspondence learning to facilitate object-level semantic understanding, thereby enhancing the model's capability in tasks requiring detailed object recognition.

  suppose we have two views $\{\mathbf{I}^{(1)}, \mathbf{I}^{(2)}\}$ of an image I, and we also tokenize the image feature into a sequence of N local image tokens. Then at the output of ViT, we obtain $g_t(\mathbf{I}^{(1)}) = [z^{(1)}_1, \cdots, z^{(1)}_N]$ and $g_s(\mathbf{I}^{(2)}) = [z^{(2)}_1, \cdots, z^{(2)}_N]$, where z denotes the local feature

$$\mathcal{L}_{\text{Local}} = -\frac{1}{N}\sum_{i=1}^{N} h_t(z^{(f)}_{j^\star}) \log h_s(z^{(2)}_i), \quad j^\star = \arg\max_j \text{CosineDistance}(z^{(f)}_j, z^{(2)}_i),$$

Conceptual approach

Global

- **Concept-Guided Global Task**: Augments the traditional ViT training by clustering images with the same concept, thereby producing semantically consistent relational representations.
- **Plug-and-Play**: This task is designed to be easily incorporated into existing ViT pipelines without additional input preprocessing, enhancing model compatibility.

Suppose we have two views $\{\mathbf{I}^{(1)}, \mathbf{I}^{(2)}\}$ of an image I, the main idea of our concept-guided global task is to replace I (1) in the DINO loss with the image feature f sampled from the concept-feature dictionary, and h_t and h_s are the projection head of the teacher and student network, respectively, and g_s is the student network.

$$\mathcal{L}_{\text{Global}} = -h_t(f) \log h_s(g_s(\mathbf{I}^{(2)})),$$

Learning algorithm:

Algorithm 1 RelViT: Concept-guided Vision Transformer

Input: A set of training images with concepts $\{(\mathbf{I}_1, C_1), \cdots\}$, an image augmentation function $\texttt{aug}(\cdot)$, momentum update factor λ, loss weight α, a concept-feature dictionary D, teacher and student ViT g_t and g_s, parameterized by θ_t and θ_s, respectively.

for $(\mathbf{I}_i, C_i)$ in $\{(\mathbf{I}_1, C_1), \cdots\}$ **do**
  $\mathbf{I}_i^{(1)},\ \mathbf{I}_i^{(2)} = \texttt{aug}(\mathbf{I}_i),\ \texttt{aug}(\mathbf{I}_i)$
  Uniformly draw a concept code $c \sim C_i$.
  Retrieve Q from D with c.
  if Q is not empty **then**
    Sample feature $f \sim Q$, following some sampling tactics.
    $\mathcal{L}_{\text{aux}} = \mathcal{L}_{\text{Global}}(f,\ g_s(\mathbf{I}_i^{(2)})) + \mathcal{L}_{\text{Local}}(f,\ g_s(\mathbf{I}_i^{(2)}))$
    Insert feature $g_t(\mathbf{I}_i^{(1)})$ into Q; if it is full, remove the oldest feature.
  else
    $\mathcal{L}_{\text{aux}} = \mathcal{L}_{\text{Global}}(g_t(\mathbf{I}_i^{(1)}),\ g_s(\mathbf{I}_i^{(2)})) + \mathcal{L}_{\text{Local}}(g_t(\mathbf{I}_i^{(1)}),\ g_s(\mathbf{I}_i^{(2)}))$
  end if
  Update θ_s with the loss function $\mathcal{L} = \mathcal{L}_{\text{main}} + \alpha\mathcal{L}_{\text{aux}}$.
  Update θ_t using an EMA: $\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s$.
end for

Experiment(result on HICO dataset)

Method	Ext. superv.	Backbone	Orig.	Systematic-easy		Systematic-hard	
				Full cls.	Unseen cls.	Full cls.	Unseen cls.
Mallya & Lazebnik (2016)*		ResNet-101	33.8	-	-	-	-
Girdhar & Ramanan (2017)*	bbox	ResNet-101	34.6	-	-	-	-
Fang et al. (2018)*	pose	ResNet-101	39.9	-	-	-	-
Hou et al. (2020)†		ResNet-101	28.57	26.65	11.94	21.76	10.58
ViT-only		PVTv2-b2	35.48	31.06	11.14	19.03	18.85
EsViT (2021)		PVTv2-b2	38.23	35.15	11.53	22.55	21.84
RelViT (Ours)		PVTv2-b2	39.4	36.99	12.26	22.75	22.66
RelViT + EsViT (Ours)		PVTv2-b2	**40.12**	**37.21**	**12.51**	**23.06**	**22.89**

Result on GQA dataset

Method	Bbox feat.*	Backbone	Orig.	Sys.
BottomUp (2018)	✓	ResNet-101	53.21	-
MAC (2018b)	✓	ResNet-101	54.06	-
MCAN-Small (2019)	✓	ResNet-101	58.35	36.21
MCAN-Small (2019)		ResNet-101	51.1	30.12
ViT-only		PVTv2-b2	56.62	31.39
EsViT (2021)		PVTv2-b2	56.95	31.76
RelViT (Ours)		PVTv2-b2	57.87	35.48

Figure 3: Histogram of reasoning hops over GQA training questions.

Ablation study

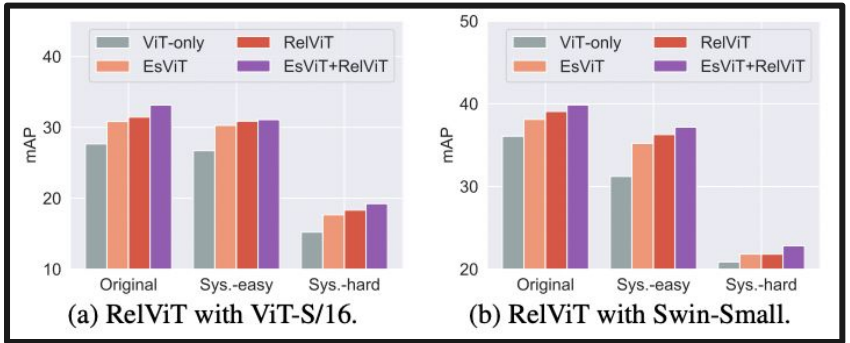


Figure 4: **Ablation study on HICO.** We investigate the impact of ViT architectures, implementation of concept-feature dictionary, auxiliary tasks, and the weight α on the performance of our method.

1. Impact of Different ViT Architectures

- **Testing on PVTv2-b2, ViT-S/16, and Swin-Small**:
  - RelViT shows consistent advantages across different ViT architectures.
  - Demonstrates compatibility with various ViT variants, indicating robustness and flexibility.
  - **Figures 4a & 4b** depict performance across systematic easy and hard tasks, highlighting the effectiveness of RelViT.

Method	Orig.	Sys.-easy	Sys.-hard
ViT-only	27.67	26.72	15.23
EsViT	30.83	30.28	17.67
RelViT	31.45	30.88	18.33
EsViT+RelViT	33.15	31.09	19.24

(a) RelViT with ViT-S/16

Method	Orig.	Sys.-easy	Sys.-hard
ViT-only	36.08	31.22	20.88
EsViT	38.11	35.22	21.82
RelViT	39.07	36.27	21.81
EsViT+RelViT	39.86	37.17	22.82

(b) RelViT with Siwn-Small

Ablation study

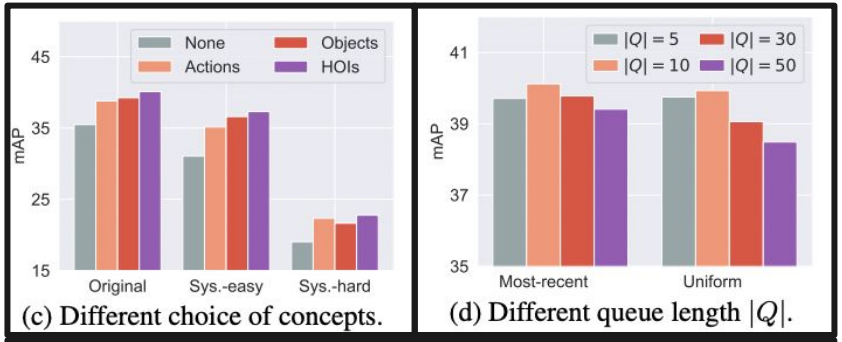


Figure 4: **Ablation study on HICO.** We investigate the impact of ViT architectures, implementation of concept-feature dictionary, auxiliary tasks, and the weight α on the performance of our method.

2. Concept-Feature Dictionary Implementation

- **Different Concepts**:
  - Tested with actions, objects, and Human-Object Interactions (HOIs).
  - **Findings**:
    - All concepts improve the baseline.
    - HOIs and objects yield more significant improvements, suggesting that finer-grained concepts are crucial.
- **Sampling Strategy & Queue Size (|Q|)**:
  - **Impact of |Q|**:
    - "Most-recent" sampling is less sensitive to |Q| size.
    - Uniform sampling, though diverse, may lead to unstable training, especially with large |Q|.

Ablation study

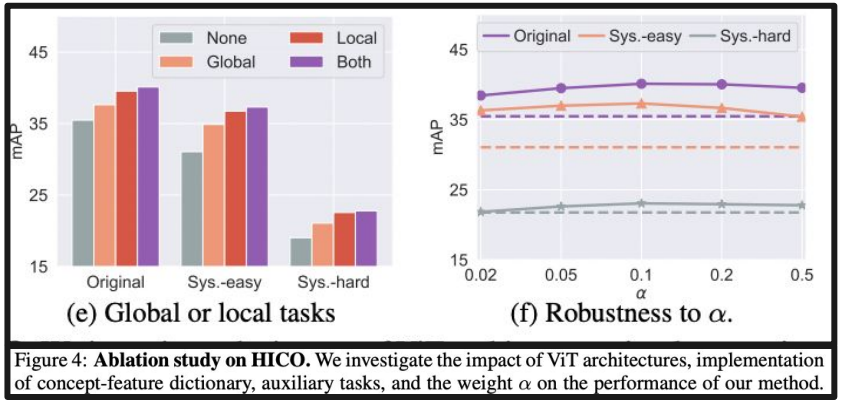


Figure 4: **Ablation study on HICO.** We investigate the impact of ViT architectures, implementation of concept-feature dictionary, auxiliary tasks, and the weight α on the performance of our method.

3. Auxiliary Tasks

- **Global vs. Local Tasks**:
  - Local tasks alone can achieve competitive results, emphasizing the importance of object-centric representations.
  - **Combined Tasks**:
    - Adding the global task enhances performance, validating the strength of concept-guided learning.
  - **Figure 4e**:
    - Displays the comparative results of global, local, and combined tasks, affirming the necessity of both.

Ablation study

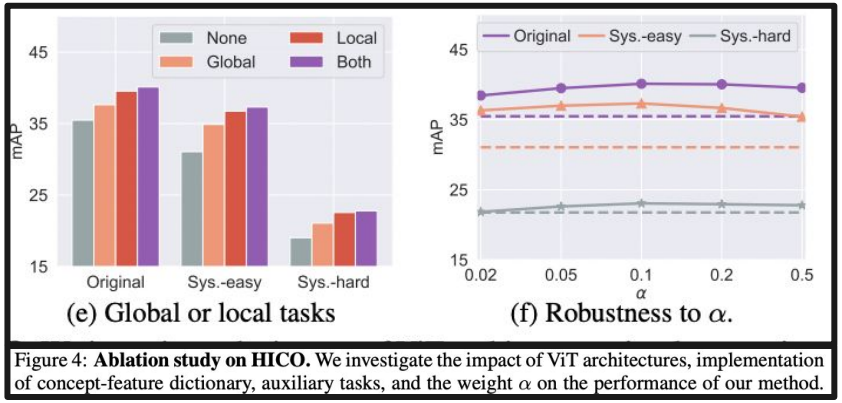


Figure 4: **Ablation study on HICO.** We investigate the impact of ViT architectures, implementation of concept-feature dictionary, auxiliary tasks, and the weight α on the performance of our method.

4. Robustness to Trade-off Weight (α)

- **Range of α (0.02 to 0.5)**:
  - Consistent performance improvements over the baseline, showcasing robustness to hyper-parameter changes.
  - **Optimal Value**:
    - Peak performance observed around α = 0.1.
  - **Figure 4f**:
    - Illustrates the steady performance gains across different α values, indicating stable optimization.

Quantitative inspection

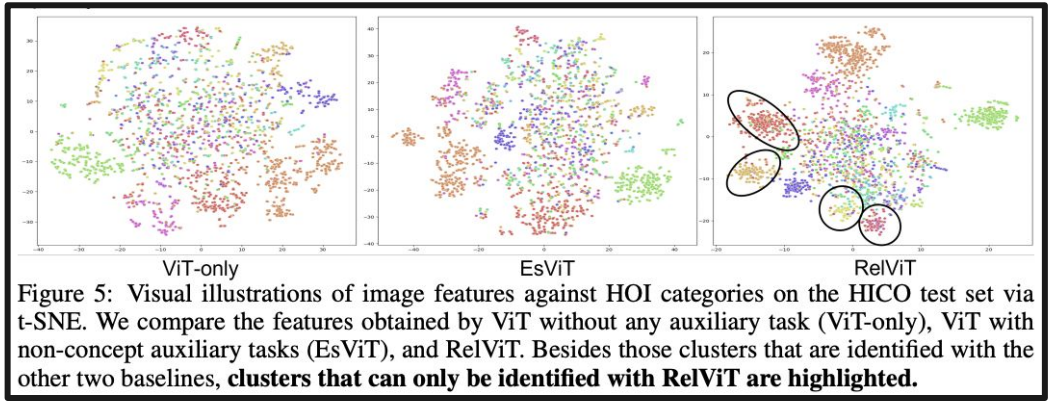


Figure 5: Visual illustrations of image features against HOI categories on the HICO test set via t-SNE. We compare the features obtained by ViT without any auxiliary task (ViT-only), ViT with non-concept auxiliary tasks (EsViT), and RelViT. Besides those clusters that are identified with the other two baselines, **clusters that can only be identified with RelViT are highlighted.**

1. Visualization of HOI Categories with t-SNE

- **Purpose**:
  - Illustrates how different models handle Human-Object Interaction (HOI) categories in the HICO dataset.
- **Comparison of Models**:
  - **ViT-only**: No auxiliary tasks; fewer identifiable clusters.
  - **EsViT**: Non-concept auxiliary tasks; better than ViT-only but still limited.
  - **RelViT**: Incorporates concept-guided auxiliary tasks; reveals more distinct clusters, indicating superior relational reasoning and feature discrimination.
- **Outcome**:
  - **Cluster Identification**:
    - RelViT identifies clusters that are not distinguishable by the other models, showcasing its ability to learn more discriminative and relational features.

Quantitative inspection

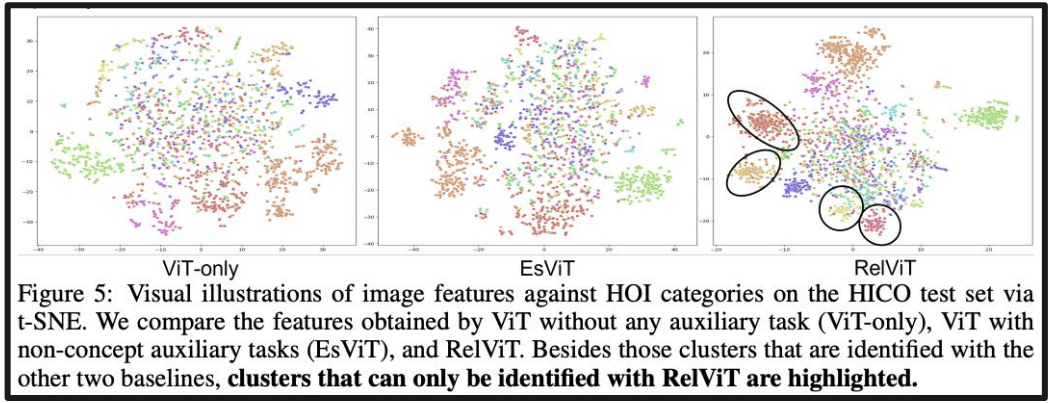


Figure 5: Visual illustrations of image features against HOI categories on the HICO test set via t-SNE. We compare the features obtained by ViT without any auxiliary task (ViT-only), ViT with non-concept auxiliary tasks (EsViT), and RelViT. Besides those clusters that are identified with the other two baselines, **clusters that can only be identified with RelViT are highlighted.**

2. Features vs. Concepts

- **RelViT's Strength**:
  - Demonstrates that the global task in RelViT makes learned features more relational, aligning well with the intended HOI categories.
- **Concept-Guided Learning**:
  - Encourages the learned representations to be more discriminative, aiding in better relational reasoning compared to baseline models.

Quantitative inspection

Figure 6: Visualization of correspondence. The correspondence is extracted between two views of the same image (upper) and two images that belong to the same concept (lower), using the learned model on HICO. **RelViT can extract correspondence on more objects in the two images (semantic correspondence) setting**. Best viewed on screen.

3. Semantic Correspondence

- **Objective**:
  - Assess how well models identify semantic correspondence between images.
- **Evaluation Settings**:
  - **Semantic Setting**: Two images belonging to the same concept.
  - **Non-Semantic Setting**: Two views of the same image.
- **Model Performance**:
  - **ViT-only**: Struggles to maintain meaningful spatial information; poor at semantic correspondence.
  - **EsViT**: Handles non-semantic tasks well but lacks in semantic correspondence.
  - **RelViT**: Excels in both settings, particularly in semantic correspondence, due to concept-guided auxiliary tasks.

Conclusion

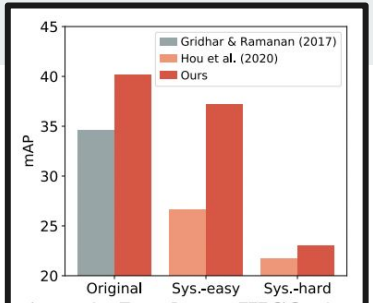


Figure 2: **Results on HICO**. Our method improves the best baseline by 16%, 43%, and 7% on the original non-systematic and two new systematic splits. **Sys.**: systematic.

ViTs for Visual Relational Reasoning:

- Vision Transformers (ViTs) have proven to be effective in relational reasoning, object-centric learning, and systematic generalization.
- ViTs show great promise in handling real-world visual relational reasoning tasks.

Introduction of RelViT:

- **RelViT**: A method designed to enhance ViTs' performance in visual relational reasoning by leveraging auxiliary tasks.
- **Global Task**: Focuses on achieving semantically consistent relational representations.
- **Local Task**: Aims at learning object-centric semantic correspondence.

Concept-Feature Dictionary:

- The concept-feature dictionary is pivotal in implementing the auxiliary tasks, aiding in improving ViTs' ability to handle complex visual reasoning.

Performance:

- RelViT significantly outperforms existing methods on challenging visual relational reasoning benchmarks, demonstrating its effectiveness.

Future Directions:

- While RelViT mainly focuses on auxiliary tasks, future work could explore architectural modifications to ViTs for even better generalization in visual reasoning tasks.

References

- RelViT: Concept-guided Vision Transformer for Visual Relational Reasoning, Xiaojian Ma and Weili Nie and Zhiding Yu and Huaizu Jiang and Chaowei Xiao and Yuke Zhu and Song-Chun Zhu and Anima Anandkumar, International Conference on Learning Representations, 2022, https://openreview.net/forum?id=afoV8W3-lYp
- Emerging Properties in Self-Supervised Vision Transformers, Mathilde Caron and Hugo Touvron and Ishan Misra and Hervé Jégou and Julien Mairal and Piotr Bojanowski and Armand Joulin, 2021, 2104.14294, https://arxiv.org/abs/2104.14294

- Any Questions ?!

More Details: Datasets

HICO Dataset

- **Purpose**: Human-object interaction (HOI) recognition; predicts possible HOI categories for input images.
- **Content**: 600 HOI categories, 117 unique actions, 80 object classes.
- **Size**: 38,116 training images, 9,658 test images.
- **Splits**:
    - **Systematic-Easy Split**: Removes rare HOI classes.
    - **Systematic-Hard Split**: Removes non-rare HOI classes in addition to the rare ones.
- **Concepts**: Utilizes 600 HOI categories as default concepts; alternative concepts (actions, objects) also explored.

GQA Dataset

- **Purpose**: Visual question answering (VQA) with a focus on relational reasoning.
- **Content**: Includes semantics-labeled questions; provides both pretrained-CNN grid features and Faster R-CNN region features.
- **Splits**:
    - **Systematic Split**: Focuses on productivity in systematic generalization by excluding questions with more than 5 reasoning hops.
- **Concepts**: Derived from parsed question keywords, resulting in 1,615 concepts (verbs, nouns, adjectives with significant meanings).

More Details: Datasets

C.1 HICO

C.1.1 Original and systematic splits

Besides the official training/testing split, we adopt the splits for systematic generalization presented in (Hou et al., 2020). It offers two splits that follow different strategies to select held-out HOI categories. **Systematic-easy** only select *rare* HOI categories (with less than 10 training samples), while **Systematic-hard** select *non-rare* categories instead. Therefore, the training set of **Systematic-hard** will contain much fewer samples and become more challenging. Some basic statistics of these training/testing splits can be found in Table 5.

Splits	#Training samples	#Training HOIs	#Testing samples	#Testing HOIs
Original	38118	600	9658	600
Systematic-easy	37820	480	9658	600
Systematic-hard	9903	480	9658	600

Table 5: Statistics of the splits of HICO dataset.

C.1.2 Implementation of $\mathcal{L}_{\text{main}}$

In HICO, there might be multiple HOIs for a single image. We, therefore, formulate the HOI prediction task as a multi-class classification problem. Specifically, the model makes 600 binary classifications and $\mathcal{L}_{\text{main}}$ in equation 3 is a binary cross-entropy loss.

More Details: Sampling Techniques

Uniform Sampling

- **Equal Probability:** Every image feature in the queue is sampled with an equal probability (1/N), promoting feature diversity.
- **Benefit:** Encourages a wide variety of image features to be retrieved, potentially enhancing the overall performance.
- **Drawback:** Older features might become outdated if the model updates quickly, leading to unstable training.

“Most-Recent” Sampling

- **Freshness-Based:** Sampling probability is weighted based on the recency of the features, with newer features having a higher chance of being selected.
- **Benefit:** Prioritizes up-to-date features, stabilizing the training process.
- **Drawback:** Reduced diversity in sampled features due to the lower probability of selecting older features, which might negatively impact performance.

Backbones

E.1 RelViT with larger backbone models

As we mentioned in Section 3.1, the ViT backbone we use (PVTv2-b2) only has **25.4M** parameters, even less than the commonly-used ResNet-101 (**44.7M** parameters). Therefore, we testify RelViT with larger state-of-the-art ViT backbones: PVTv2-b3 (**45.2M** parameters) and Swin-base (**88M** parameters) (Liu et al., 2021) and provide the results on HICO and GQA below:

Table 10: Results with larger ViT models on HICO.

HICO mAP	Fang et al. (2018)	RelViT + EsViT (PVTv2-b2)	RelViT + EsViT (PVTv2-b3)	RelViT + EsViT (Swin-base)
Original	39.9	40.12	42.61	**43.98**
Systematic-easy	-	37.21	39.92	**42.04**
Systematic-hard	-	23.06	25.56	**28.36**

Table 11: Results with larger ViT models on GQA.

GQA overall accuracy	MCAN-Small (w/ bbox)	RelViT (PVTv2-b2)	RelViT (PVTv2-b3)	RelViT (Swin-base)
original	58.35	57.87	61.41	**65.54**
systematic	36.21	35.48	36.25	**37.51**

Preprocessing

1 INPUT PIPELINE

adopt the following data augmentation pipeline for the generating the additional views for o
auxiliary tasks

Randomly crop and resize the image into $(224, 224)$ with scale ratio $(0.2, 1.0)$;

Randomly jitter the color of the image on brightness, contrast saturation and hue with probabili
of $(0.4, 0.4, 0.4, 0.1)$, respectively;

Randomly turn the image into gray scale with probability 0.2;

Randomly apply Gaussian blur with kernel size 23 and sigma $(0.1, 2.0)$ and probability 0.5;

Randomly flip the image horizontally.

tably, we apply a random crop operation to ensure that all the input images for our auxiliary tasl
ntain the same number of patches.

Hyper Parameters

Table 3: Hyperparameters for RelViT.

Parameter	Value
Optimizer	AdamW with epsilon $1e^{-1}$ (HICO) / $1e$-5 (GQA)
Gradient clipping norm	No grad clipping (HICO) / 0.5 (GQA)
Base learning rate	$1.5e^{-4}$ (HICO) / $3e^{-5}$ (GQA)
Learning rate schedule	0.1 scale with milestones [15, 25] (HICO) / [8, 10] (GQA)
Batch size	16 (HICO) / 64 (GQA)
Total training epochs	30 (HICO) / 12 (GQA)
Temperature τ in DINO loss	0.04 for teacher and 0.1 for student, we don't use schedule.
Momentum m for teacher	0.999
Center m for center features	0.9
Sampling method	*"most-recent"* (HICO) / *uniform* (GQA)
Queue size $\|Q\|$	10