# Beyond Accuracy: Ensuring Correct Predictions With Correct Rationales

**Tang Li** **Mengmeng Ma** **Xi Peng**

DeepREAL Lab: https://deep-real.github.io
Department of Computer & Information Science, University of Delaware
{tangli, mengma, xipeng}@udel.edu

## Abstract

Large pretrained foundation models demonstrate exceptional performance and, in some high-stakes applications, even surpass human experts. However, most of these models are currently evaluated primarily on prediction accuracy, overlooking the validity of the rationales behind their accurate predictions. For the safe deployment of foundation models, there is a pressing need to ensure *double-correct predictions*, *i.e.*, correct prediction backed by correct rationales. To achieve this, we propose a two-phase scheme: First, we curate a new dataset that offers structured rationales for visual recognition tasks. Second, we propose a rationale-informed optimization method to guide the model in disentangling and localizing visual evidence for each rationale, without requiring manual annotations. Extensive experiments and ablation studies demonstrate that our model outperforms state-of-the-art models by up to 10.1% in prediction accuracy across a wide range of tasks. Furthermore, our method significantly improves the model's rationale correctness, improving localization by 7.5% and disentanglement by 36.5%. Our dataset, source code, and pretrained weights: https://github.com/deep-real/DCP

## 1 Introduction

Large foundation models, such as CLIP [1] and GPT-4V [2], exhibit exceptional performance or even surpass human experts in some high-stakes applications, such as medical diagnosis [3] and autonomous driving [4, 5]. However, most of these models are currently evaluated primarily on prediction accuracy, overlooking a critical aspect for ensuring safety, *i.e.*, the validity of the reasons behind their accurate predictions. Understanding the *rationales* - the "how" and "why" behind model predictions - is crucial for developing safe predictions. Fig. 1 shows typical examples of unsafe predictions: CLIP might predict accurately yet based on wrong rationales, whereas GPT-4V might make wrong predictions based on rationales that are plausible to humans. To build trust in real-world deployment, a natural question arises: *Can models make double-correct predictions, i.e., correct predictions backed by correct rationales?*

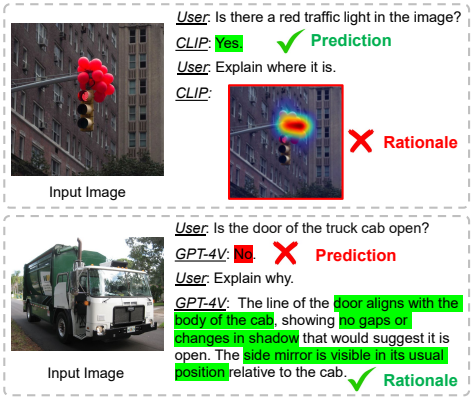


Figure 1: Unsafe prediction examples. **Correct prediction, incorrect rationale**: CLIP identifies a red light, but wrongly based on red balloons. **Incorrect prediction, correct rationale**: GPT-4V incorrectly predicts a closed door, yet based on plausible visual evidence.

Correct rationales generally align with how humans would reason about the same decision and are based

38th Conference on Neural Information Processing Systems (NeurIPS 2024).


https://doi.org/10.52202/079017-1367

on valid *visual evidence* [6, 7, 8]. There are existing attempts to provide rationales for machine learning models' predictions. They either explicitly force the models to make decisions based on human-understandable concepts by introducing bottleneck layers [9, 10], or implicitly inject commonsense knowledge into models by contrastive learning between similar yet distinct textual concepts [11, 12]. However, none of them ensures *double-correct predictions*. Observations from our previous research [13] and recent studies in the field [14, 15] reveal that these models might provide *incorrect rationales*, as they fail to base the rationales on valid *visual evidence*.

To this end, we develop *double-correct predictions* by focusing on two foundational aspects:

**i) "What" are the correct rationales? Structured rationale acquisition.** Existing vision datasets typically provide ground truth labels of predictions, whereas missing the rationales behind these decisions [16, 17]. To fill this gap, we curate a new dataset that offers over 4,000 unique *textual rationales* designed for predicting the 1,000 categories in *ImageNet* [18], structured in a tree format. This design differs from existing knowledge graphs [19, 20, 21], which either provide irrelevant knowledge for the vision task or are too coarse-grained, providing insufficient information. Our rationale dataset is tailored to capture the detailed reasoning processes for visual recognition.

**ii) "Where" are the correct rationales? Rationale-informed optimization.** The other challenge in developing double-correct predictions is the absence of pixel-wise annotations for rationales' *visual evidence*. Although some datasets provide segmentation masks of object parts [17, 22], they lack sufficient rationale coverage and are limited to small-scale use cases [23]. To address this issue, we propose a rationale-informed optimization method to guide the model in disentangling and localizing the visual evidence of rationales, without requiring manual annotations. Our method can be integrated into the existing model training process without architectural changes and extra parameters.

We evaluate the proposed method on a wide range of benchmark datasets and tasks. For prediction correctness, our model outperforms state-of-the-art models in zero-shot, linear probe, and fine-tuning settings by 2.6%, 2.0%, and 10.1%. For rationale correctness, the empirical results exhibit that our model significantly improves ground truth rationale localization and rationale disentanglability by 7.5% and 36.5%. Furthermore, the extensive qualitative results and ablation studies demonstrate the effectiveness of the proposed method.

Our contribution includes: **1)** We curate a new structured rationale dataset. **2)** A faithful explanation method tailored for explaining CLIP-ViT predictions. **3)** A principled optimization method that seamlessly integrates structured rationale information to develop *double-correct predictions*. **4)** Empirical results in a wide range of benchmark datasets and tasks including image classification and retrieval demonstrate the superior prediction and rationale correctness of our model.

## 2 Problem Formulation

In this section, we first formally define *rationales*, then provide the mathematical formulation of the *double-correct prediction* problem.

**Definition 1** *(Rationales) Given a category $y$, rationales are a set of $K$ underlying abstract notions $\{r_k^y\}_{k=1}^K$ and relations that capture the reasoning process leading to the recognition of $y$.*

In the real world, rationales can be represented through textual descriptions [24, 25]. For example, when recognizing a specific breed of dog in an image, the rationales could be a set of concepts such as the shape of the ears, the color of the fur, and the size of the dog. Mathematically, given a textual rationale $r$, we assume the existence of a ground truth labeling function $V(x, r)$ that can provide the pixel-wise annotations of *visual evidence* corresponding to $r$ on an input $x$.

**Definition 2** *(Double-Correct Predictions) A correct prediction is double-correct when it is backed by correct rationales that are based on valid visual evidence.*

Denote $(x, y) \sim P(X, Y)$ as a data point sampled from the training distribution $P(X, Y)$, $g(\cdot)$ as an explanation method that attributes the prediction of text $r$ to a group of pixels in input $x$ depending on model $f$, $\ell(\cdot)$ as the task-specific loss function, and $\mathcal{F}$ as a function class that is model-agnostic for the prediction task. To ensure the model $f$ makes *double-correct prediction*, we propose to solve the following constrained optimization problem:

$$\min_{f \in \mathcal{F}} \mathcal{R}(f) := \mathbb{E}_{(x,y)\sim P(X,Y)}[\ell(f(x), y)] \quad \text{s.t. } g(x, r; f) = V(x, r), \ \forall r \in \{r_k^y\}_{k=1}^K. \tag{1}$$

The problem in Eq. 1 is challenging to solve, since we neither have access to the rationales $\{r_k^y\}_{k=1}^K$, nor to the ground truth labeling functions $V(\cdot)$. There are existing attempts that employ domain experts to manually collect textual descriptions of rationales [22, 26], or pixel-wise annotations of object parts on the image [17]. However, these approaches are often limited to small-scale datasets, and impractical in large-scale settings due to the high cost of fine-grained annotations [27, 23].

## 3 Double-Correct Predictions

To bridge the gaps, in Sec. 3.1 we present how to acquire rationales $\{r_k^y\}_{k=1}^K$, in Sec. 3.2 we propose a new explanation method $g(\cdot)$, and in Sec. 3.3 we develop *double-correct predictions* without $V(\cdot)$.

### 3.1 Structured Rationale Dataset

In this section, we curate a new rationale dataset to offer $\{r_k^y\}_{k=1}^K$ in Eq. 1. According to Def. 1, rationales are structured human knowledge. Therefore, *ontologies* that encapsulate complex, interconnected information while maintaining semantic relationships between entities [28, 29], present a proper tool to represent rationales. The benefits are bi-directional: **i)** in the human-to-machine direction, it offers a standardized, machine-readable format; **ii)** in the machine-to-human direction, ontology structure mirroring how humans organize and retrieve information to explain the model's decision-making process.

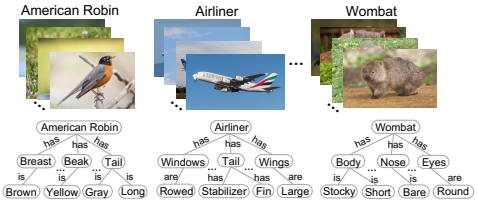


Figure 2: Our structured rationales capture the major attributes and their sub-attributes that lead to the recognition of objects. Our dataset offers over 4,000 unique rationales covering all 1,000 categories from *ImageNet* [18].

**Acquire structured rationales:** Different from existing works that are limited to small-scale manual annotation, we generate our rationale dataset in a scalable manner. Specifically, we utilize Large Language Models (LLMs) like GPT-4 [2] to extract the structured rationales. Existing studies prove that GPT-4 has expert-level expertise in commonsense [30] and domain knowledge [31]. However, we find that directly querying LLMs would yield inconsistent tree structures that can hardly be used by machine learning models. To address this issue, we provide a series of exemplary structured rationales before the query, employing in-context learning [32] to extract *standardized* rationales in a `.JSON` format. See Appendix A for our full prompt and rationale examples.

**Rationale dataset statistics:** Our dataset covers all 1,000 categories in the *ImageNet* [18]. For each category, we generate an ontology tree with a maximum height of two. As illustrated in Fig. 2, the root node is the category, the children of the root are the attributes, and the leaves are the sub-attributes. The edges represent the relationships between nodes. Combining attributes and sub-attributes, our dataset contains over 4,000 unique rationales. Our rationale ontology trees capture the reasoning processes leading to the recognition of the corresponding root categories.

**Can we trust the rationales extracted from GPT-4?** Although there are plenty of works showing GPT-4's remarkable capabilities [30, 31], it still could suffer from *hallucinations* [33, 34]. However, evaluations on the generation quality are largely missing from existing works that generate data from LLMs [9, 35, 36]. To fill this gap and ensure the quality of our rationale data, we conduct comprehensive human and machine evaluations. As detailed in Sec. 4.1, on a 5-point Likert scale across three metrics, 964 out of 1,000 categories are scored as having high-quality rationales ($\geq$4.0).

In contrast to existing Knowledge Graphs [19, 20, 21] that either offer knowledge *unrelated* to the visual prediction task, or are too coarse-grained that provide *insufficient* information, our structured rationales are tailored for visual recognition tasks in a fine-grained attribute level. Furthermore, our dataset can expand to accommodate new rationales, providing flexibility to dealing with evolving datasets where more data becomes available. For example, our rationale ontologies can be seamlessly integrated following the *ImageNet* [18] category ontology derived from WordNet [20].

### 3.2 Faithful Explanation Method

In this section, we develop a new explanation method to implement $g(\cdot)$ in Eq. 1. To incorporate both image and text inputs, we instantiate the model $f$ using the CLIP-ViT architectures [37] because of


https://doi.org/10.52202/079017-1367

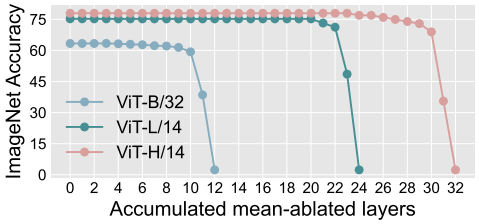


Figure 3: Multi-head Self Attention (MSA) accumulated mean-ablation study. Based on Eq. 2, we replace the direct effects of MSAs up to a specific layer with their mean values calculated across the *ImageNet* [18] validation set. Most of the performance gains can be attributed to the final layers of the ViT.

Table 1: Weakly-supervised segmentation accuracy on *ImageNet-Seg* [45]. We threshold explanation heatmaps from CLIP-ViT-L-14 as segmentation masks. Our method outperforms existing explanation methods in segmentation accuracy, demonstrating the high faithfulness of our explanations.

| Exp. Methods | Pixel Acc. ↑ | mIoU ↑ | mAP ↑ |
|---|---|---|---|
| LRP [46] | 52.81 | 33.57 | 54.37 |
| rollout [39] | 60.63 | 40.64 | 74.47 |
| row attention | 65.67 | 43.83 | 76.05 |
| GradCAM [41] | 70.27 | 44.50 | 70.30 |
| Chefer et al. [40] | 69.21 | 47.47 | 78.29 |
| TextSpan [44] | 75.21 | 54.50 | 81.61 |
| Ours | **76.27** | **58.04** | **82.17** |

their proven capability [1, 38]. Existing methods for explaining the ViT model either directly use the attention maps as explanations [39], or weigh them using gradients [40, 41]. However, these methods might be *unfaithful* to the ViT predictions. This is because the computation of each ViT prediction involves queries, keys, and values, whereas the attention maps only capture the inner products of queries and keys, ignoring information in values that also affect predictions [42, 40]. Therefore, explanations based on attention maps might not fully reflect the reasons behind ViT predictions.

**Decompose ViT outputs:** Recent works [43, 44] prove that, for ViT models, the image embeddings can be decomposed into the contributions of each token within each attention head. Let $\phi$ and $\theta$ parameterize the image- and text-encoder of the CLIP-ViT model, $P$ is the projection matrix, $L$, $M$, $N$ are the numbers of layers, heads, and image tokens, $a_i^{l,m}$ is the output of the $m$-th attention head in layer $l$ for the $i$-th image token, then the embedding of image $I$ can be decomposed as:

$$e_I = f_\phi(I) = P\text{ViT}(I) = \sum_{l=1}^{L}\sum_{m=1}^{M}\sum_{i=0}^{N} P a_i^{l,m}. \tag{2}$$

By contracting along layers and heads, [44] calculates the contribution of the $i$-th image token to the final image embedding using $\sum_{l=1}^{L}\sum_{m=1}^{M} P a_i^{l,m}$.

**Faithful explanations weighted by mean-ablation results:** As indicated by our mean-ablation results in Fig. 3, the final layers contribute the most to the predictions, whereas the earlier layers have minimal impact. Thus, noise from early layers could obscure key information by a naive summation across all layers as in [44]. To address this issue, we weigh each layer's contribution based on its importance, measured by the corresponding performance drop in the mean-ablation study. Denote the performance drop of layer $l$ as $\Delta_l$, we calculate the contribution of the $i$-th image token by:

$$e_i = \sum_{l=1}^{L} w^l \sum_{m=1}^{M} P a_i^{l,m}, \quad \text{where } w^l = \frac{\Delta_l}{\sum_{j=1}^{L}\Delta_j}. \tag{3}$$

Note that $e_i$ is projected onto the image-text embedding space by $P$. Thus, we can use $g(I, r) = \{\langle e_i, f_\theta(r)\rangle\}_{i\in I}$ to calculate the explanations of rationale $r$ on an image $I$, *i.e.*, *visual evidence*.

Our method significantly improves the explanation accuracy, as shown in Tab. 1. In contrast to attention-based explanations [39], our method fully utilizes the information from queries, keys, and values that are used for ViT predictions. Compared to gradient-weighted attention maps [40, 41], our method cuts down the computational complexity from $O(n^2)$ to $O(n)$ over $n$ image tokens.

### 3.3 Rationale-informed Optimization

In this section, we develop *double-correct predictions* by disentangling and localizing rationales without pixel-wise human annotations $V(\cdot)$ in Eq. 1.

**Disentanglement via reconstruction:** Drawing insights from our previous research [47, 13], we propose to contrast between explanation heatmaps of rationales to guide the model training in a *self-supervised* manner. Specifically, we enforce the following two constraints: **i)** the image embeddings for different rationales within the same category are disentangled, and **ii)** the aggregated image embedding of all rationales within the same category aligns with the text embedding of the category.

https://doi.org/10.52202/079017-1367

Mathematically, the backbone objective is to learn a mapping function $f \in \mathcal{F}$ such that for each image-text pair $(I, T) \sim P(\mathbf{I}, \mathbf{T})$, the embeddings $f_\phi(I)$ and $f_\theta(T)$ are aligned in a shared space if they are a correct match, where $T$ is a text description of category $y$. Let $\ell(\cdot)$ be the InfoNCE loss [48]. $h(g(I, r)) = \sum_i e_i \cdot \mathbb{1}(g(I, r)_i > \tau)$ extracts the image embedding of a given rationale. $\mathcal{D}(\cdot, \cdot)$ is a distance metric such as L2 distance. $\tau$, $\epsilon$, and $\delta$ are thresholding hyperparameters. For all $r, r' \in \{r_k^y\}_{k=1}^K$, we propose to develop *double-correct predictions* by optimizing:

$$\begin{aligned} &\min_{f \in \mathcal{F}} \ \mathcal{R}(f) := \mathbb{E}_{(I,T)\sim P(\mathbf{I},\mathbf{T})}[\ell(f_\phi(I), f_\theta(T))] && \triangleleft \textbf{ Correct Predictions} \\ &\text{s.t.} \ \underbrace{\mathcal{D}(h(g(I, r)), h(g(I, r'))) \geq \epsilon,}_{\text{Disentanglement}} \ \underbrace{\mathcal{D}(\textstyle\sum_r h(g(I, r)), f_\theta(y)) \leq \delta.}_{\text{Reconstruction}} && \triangleleft \textbf{ Correct Rationales} \end{aligned} \tag{4}$$

Intuitively, the reconstruction term prevents the disentanglement from collapsing into trivial solutions, thereby ensuring localization. Solving Eq. 4 often leads to a non-convex problem, wherein methods such as stochastic gradient descent (SGD) cannot guarantee constraint satisfaction [49, 50]. To address this issue, we leverage Karush–Kuhn–Tucker (KKT) conditions [51, 52] and introduce Lagrange multipliers $\lambda$ and $\gamma$ to convert the constrained problem into its unconstrained counterpart:

$$\begin{aligned} \min_{f \in \mathcal{F}} \{\mathcal{R}(f) := \mathbb{E}_{(I,T)\sim P(\mathbf{I},\mathbf{T})}[\ell(f(I, T))] \\ + \lambda \mathcal{D}(h(g(I, r)), h(g(I, r'))) \ + \ \gamma \mathcal{D}(\textstyle\sum_r h(g(I, r)), f_\theta(y))\}. \end{aligned} \tag{5}$$

Our method has the following merits: **i)** In contrast to existing works that rely on expensive pixel-wise annotations to localize objects [53, 54], the proposed rationale-informed optimization achieves a more fine-grained, attribute-level localization without manual annotations. **ii)** Our method can be integrated into vision-language model training without architectural changes and extra parameters.

# 4 Experiments

In this section, we first evaluate the quality of our curated rationale dataset in Sec. 4.1. To best validate *double-correct predictions*, we then conduct a series of experiments to compare the proposed method with existing methods in Secs 4.2 - 4.7. The experimental results prove that our model achieves superior prediction and rationale correctness on a wide range of benchmark datasets and tasks.

## 4.1 Evaluation of Rationale Quality

**Metrics:** We focus on three essential aspects of the rationale quality. (1) *Factual Consistency*: whether the rationales are consistent with facts. (2) *Comprehensiveness*: whether the rationales provide sufficient information necessary to predict the category. (3) *Visual Disentanglement*: whether the rationales are visually disentanglable or non-overlap. We rate them on a 5-point Likert scale scoring system, where higher scores indicate better performance. For example, in Factual Consistency, score 5 means 100% of the generated rationales are consistent with facts, score 4 means 75%, score 3 means 50%, score 2 means 25%, and score 1 means completely wrong.

**Evaluators:** (1) *Human Evaluators*: We recruited four human evaluators, who are mostly graduate students. They are asked to conduct assessments based on commonsense knowledge and perform Internet searches for validation. On average, it takes them around one minute per sample. (2) *Machine Evaluators*: The latest GPT-4o and GPT-4v models (date accessed: Aug. 6th, 2024). For each evaluation, we perform three independent runs and calculate the average scores. Note that expanding human evaluations to the entire dataset is not scalable. To this end, we first prove the reliability of machine evaluations, then use it to automatically evaluation the entire dataset.

**Human evaluations:** We sample three independent groups of data from our rationale dataset, each consisting of 50 categories and their corresponding rationales. Specifically, categories were randomly selected from their superclasses: Animals (20), Objects & Artifacts (15), Natural Scenes (5), Plants (5), and Human Activities (5). This ensures that not only each superclass is represented but also that our results are robust [55]. As shown in Tab. 2, The dataset consistently achieves scores of 4.61 or higher on the average of evaluators for each metric, indicating that over 90.3% of the rationales for each category are highly factual, comprehensive, and visually disentanglable.

**Machine evaluations:** Note that the scores of all three metrics are almost identical between machines and humans. The Pearson Correlation coefficient of 0.82 reveals the strong positive correlation


https://doi.org/10.52202/079017-1367

Table 2: Evaluation results of rationale quality. Both machine and human evaluators receive the same instructions about the metrics. The scores for all three metrics are nearly identical between machine and human evaluators, indicating that over 90.3% of our rationales are of high quality.

| Evaluators | Factual Consistency | Comprehensiveness | Visual Disentanglement |
|---|---|---|---|
| GPT-4o | 4.89±0.05 | 4.55±0.06 | 4.66±0.06 |
| GPT-4v | 4.92±0.03 | 4.67±0.05 | 4.70±0.02 |
| Machine Avg. | 4.91 | 4.61 | 4.68 |
| Human-A | 4.85±0.11 | 4.64±0.19 | 4.42±0.15 |
| Human-B | 4.97±0.02 | 4.77±0.02 | 4.20±0.11 |
| Human-C | 4.78±0.04 | 4.60±0.11 | 4.78±0.10 |
| Human-D | 4.81±0.08 | 4.65±0.05 | 4.77±0.07 |
| Human Avg. | 4.85 | 4.66 | 4.54 |

between machine and human evaluators. Based on this observation, we further conduct machine evaluations on the entire dataset efficiently. Our results indicate that 964 out of 1,000 categories have high-quality rationales ($\geq$4.0). See detailed results for the entire dataset in Appendix. B.

### 4.2 Benchmark Datasets and Implementation Details

**Backbone model:** Due to the computational cost of training large vision-language models (VLMs) from scratch, we focus on fine-tuning experiments. Specifically, we fine-tune the ViT-B/32 variant of CLIP on the *ImageNet* [18] dataset combined with our curated rationale dataset. To maintain simple and interpretable rationales, the ontology graph for each category is limited to a maximum depth of two, allowing for the extraction of five to six independent concepts on average.

**Baseline models:** We compare our model with state-of-the-art VLMs that use ViT-B/32 as their vision encoders, including large-scale pretrained models (CLIP [1], DeCLIP [56]), knowledge-augmented model (NegCLIP [11]), and fine-grained alignment models (FILIP [57], PyramidCLIP [53]). For fair comparisons, we also compare our model with *ImageNet* [18] fine-tuned models using the same CLIP initialization and augmented text descriptions as our model, including full model fine-tuning (-ft) and vision-encoder-only fine-tuning (-ft-vision).

**Evaluation datasets:** We validate the prediction correctness of the models on image classification and image-text retrieval tasks. For image classification (zero-shot, linear probe), experiments are carried out on nine benchmark datasets, including *CUB* [17], *Caltech101* [58], *OxfordPets* [59], *Food101* [60], *SUN397* [61], *StanfordCars* [62], *DTD* [63], *CIFAR-10* [64], and *CIFAR-100* [64]. For retrieval, we conduct experiments on *Flickr30K* [65] and *MSCOCO* [66]. To evaluate the correctness of rationales, we evaluate the models' rationale localizability on *CUB-Part* [67] and *PartImageNet* [68] that provide ground truth segmentation masks of object parts, *e.g.*, "head" and "body". Furthermore, we evaluate the rationale disentanglability on the aforementioned nine benchmark datasets. More details can be found in Appendix D.

**Implementation details:** We follow the same architecture design as CLIP [1] for ViT-B/32. The input resolution of image encoder is 224×224 and the maximum context length of text encoder is 77. We train our model using an AdamW [69] optimizer and the cosine learning rate scheduler with a linear warmup. Specifically, the learning rate linearly increases from 0 to the peak value within 10% of the total steps, and then decreases with a cosine anneal strategy. Our learning rate is set to 5e-7 and train the model for eight epochs. More details can be found in Appendix D.

### 4.3 Evaluation Metrics

**Prediction correctness:** We use standard category *prediction accuracy* to evaluate the prediction correctness for zero-shot, linear probe, and fine-tuned settings.

**Rationale correctness:** We define two new metrics to measure rationale correctness.

i) *Rationale localizability*. We evaluate the correctness of rationales using ground truth segmentation masks of object parts [67, 68]. Following the standard evaluation protocol [70], we threshold the rationale explanation heatmaps to segmentation masks and calculate a mean Intersection over Union

Table 3: Comparison of prediction accuracy (%) on nine benchmark datasets. Our results are on the average of three trials of experiments using different random seeds. We highlight the **best results** and the second best results. Surprisingly, different from most interpretability methods that compromise benchmark performance, our method also enhances prediction accuracy.

| Metrics | Models | C10 | C100 | CUB | CAL | PETS | F101 | SUN | CARS | DTD | AVG |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Zero-shot Accuracy (%) | CLIP | **91.3** | 65.1 | 51.5 | 87.9 | 87.0 | **84.4** | 63.2 | 59.4 | 44.5 | 70.5 |
| | DeCLIP | 91.2 | 66.4 | 51.2 | **89.5** | 79.5 | 74.6 | 63.4 | 50.6 | 42.7 | 67.7 |
| | NegCLIP | 85.7 | 60.9 | 37.4 | 81.0 | 79.7 | 71.1 | 57.0 | 45.4 | 37.5 | 61.7 |
| | FILIP | 86.9 | 65.5 | 37.5 | 91.9 | 88.1 | 82.8 | 69.1 | 55.4 | 49.3 | 69.6 |
| | PyramidCLIP | 81.5 | 53.7 | 52.7 | 81.7 | 83.7 | 67.8 | 65.8 | **65.0** | 47.2 | 66.6 |
| | CLIP-ft | 83.6 | 59.5 | 46.3 | 83.6 | 81.6 | 78.7 | 54.2 | 45.3 | 33.9 | 63.0 |
| | CLIP-ft-vision | 86.1 | 56.0 | 42.2 | 81.0 | 79.8 | 65.1 | 56.7 | 42.2 | 38.7 | 60.9 |
| | Ours | 90.8 | **68.1** | **56.0** | 89.3 | **88.5** | 84.3 | **70.6** | 62.3 | **47.7** | **73.1** |
| Linear Probe Accuracy (%) | CLIP | 95.1 | 80.5 | 71.4 | 93.0 | 90.0 | **88.8** | 76.6 | 81.1 | 76.5 | 83.7 |
| | DeCLIP | **96.5** | **84.7** | 65.0 | 94.8 | 89.2 | 85.0 | 75.0 | 81.6 | 78.5 | 83.4 |
| | NegCLIP | 94.3 | 79.3 | 71.8 | 98.7 | 89.5 | 85.6 | 78.6 | 75.0 | 81.3 | 83.8 |
| | FILIP | 95.1 | 82.4 | 77.0 | 99.1 | 88.3 | 83.4 | 78.7 | 76.8 | 88.3 | 85.5 |
| | PyramidCLIP | 96.0 | 82.5 | 72.3 | 96.4 | 87.8 | 83.3 | 77.5 | 82.6 | 77.3 | 84.0 |
| | CLIP-ft | 93.1 | 76.5 | 70.7 | 98.1 | 88.1 | 81.7 | 75.8 | 58.6 | 76.3 | 79.9 |
| | CLIP-ft-vision | 93.7 | 77.9 | 71.7 | 98.3 | 88.6 | 84.3 | 76.4 | 73.9 | 75.4 | 82.2 |
| | Ours | 95.6 | 82.7 | **77.2** | **99.3** | **92.9** | 88.1 | **79.8** | **83.0** | **88.9** | **87.5** |

(mIoU ↑) score with the ground truth masks across different object parts. Specifically, the dynamic threshold $\tau = \mu + \sigma$, where $\mu$ and $\sigma$ are the mean and standard deviation of importance values of all pixels in a heatmap. The pixel with an importance value larger than $\tau$ is set to 1, otherwise 0.

ii) *Rationale disentanglability*. As shown in Fig. 4, for the CLIP model [1], the visual evidence of different rationales is entangled. Specifically, we treat the disentanglement between the visual evidence of different rationales as an important metric to evaluate whether the model can distinguish rationales. Specifically, we treat rationale explanation heatmaps $\mathbf{m}$ and $\mathbf{m}'$ as vectors and calculate $1 - |\langle \mathbf{m}, \mathbf{m}' \rangle|$ as an intuitive measure of disentanglability, the higher metric value the better.

### 4.4 Evaluation on Prediction Correctness

**Zero-shot image classification:** We compare our model against other state-of-the-art and fine-tuned VLMs on zero-shot image classification tasks. The results are shown in Tab. 3. On the average of nine datasets, our model outperforms the second-best result by 2.6%. The results indicate the strong transferability of our model to other vision datasets.

**Linear probe:** Following the common practice [1, 71], we conduct linear probe experiments on the nine image classification datasets. As shown in Tab. 3, our model outperforms the second-best result by 2.0%. These results demonstrate the superior vision representations learned by our model.

**Fair comparison with fine-tuned models:** As shown in Tab. 3, our model outperforms the best fine-tuned model by 10.1% and 5.3% on zero-shot and linear probe results. This suggests that the proposed Rationale-informed Optimization is essential in improving the model's performance.

### 4.5 Evaluation on Rationale Correctness

**Rationale localizability:** We compare our model with state-of-the-art and fine-tuned VLMs. As shown in Tab. 4, our model significantly improves the localization accuracy of rationales by 7.5% and 6.0% on *CUB-Part* [67] and *PartImageNet* [68]. This suggests that even without using explicit region annotations, our method significantly enhances the model's localizability of rationales.

**Rationale disentanglability:** We compare the rationale disentanglement performance of our model with state-of-the-art and fine-tuned models. As shown in Tab. 5, on the average of nine image classification datasets, our model outperforms the second-best result by 36.5%. This significant improvement reveals that our model can distinguish between different rationales.

**Fair comparison with fine-tuned models:** To evaluate whether our model's performance gain can be obtained by solely introducing information from our rationale dataset, we conduct fair comparison


https://doi.org/10.52202/079017-1367

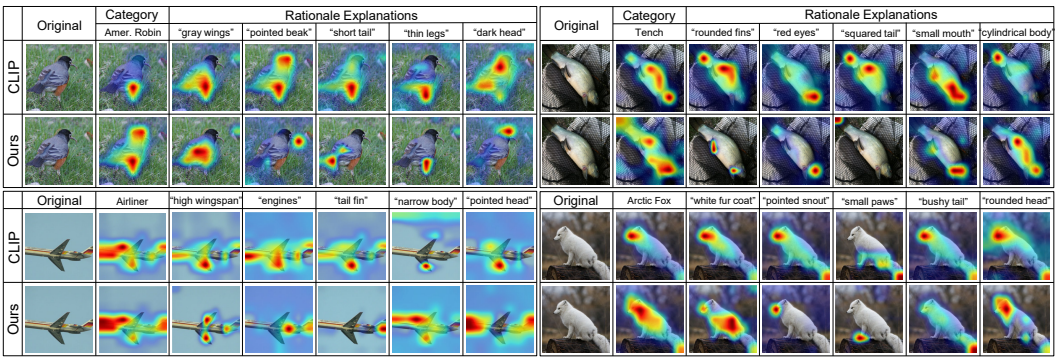


Figure 4: Qualitative results of rationale disentanglement and localization. The rationales' visual evidence of the CLIP model [1] typically highlights the entire object, lacking precise localization. In contrast, our model can correctly localize rationales, thereby enhancing trust in its predictions.

Table 4: Comparison of rationale localizability on *CUB-Part* [67] and *PartImageNet* [68]. As detailed in Sec. 4.3, we threshold rationales' explanation heatmaps as segmentation masks and calculate their mIoU (↑) with ground truth masks of corresponding object parts. Our model significantly improves the localization accuracy of fine-grained object parts. Full table in Appendix C.

| Models | Training Size | CUB-Part | | | | | | | PartImageNet |
|---|---|---|---|---|---|---|---|---|---|
| | | Head | Beak | Tail | Wings | Eyes | Torso | Avg. | |
| CLIP | 400M | 16.6 | 3.1 | 9.9 | 25.5 | 3.3 | 28.0 | 14.4 | 5.2 |
| DeCLIP | 88M | 6.9 | 2.0 | 5.1 | 16.2 | 1.5 | 18.4 | 8.3 | 3.7 |
| NegCLIP | 400M+COCO | 15.1 | 3.0 | 7.5 | 26.1 | 2.5 | 29.4 | 13.9 | 5.2 |
| FILIP | 340M | 10.3 | 2.4 | 7.3 | 20.5 | 3.4 | 23.7 | 11.2 | 4.0 |
| PyramidCLIP | 143M | 10.7 | 2.9 | 6.0 | 17.0 | 1.8 | 20.5 | 9.8 | 3.9 |
| CLIP-ft | 400M+IN | 13.5 | 3.3 | 5.8 | 22.9 | 2.1 | 25.5 | 12.1 | 4.5 |
| CLIP-ft-vision | 400M+IN | 7.4 | 2.5 | 7.9 | 26.4 | 1.6 | 22.0 | 11.3 | 4.4 |
| Ours | 400M+IN | **25.3** | **10.1** | **12.7** | **32.6** | **15.7** | **35.2** | **21.9** | **11.2** |

experiments with fine-tuned CLIP models. We compare our model with baseline (CLIP-zs), full model fine-tuning (CLIP-ft), and vision-only fine-tuning (CLIP-ft-vision). All fine-tuned models use the same CLIP initialization and receive the same language supervision as our model. As shown in Tabs. 4& 3, our model outperforms the best fine-tuned model by 9.8% and 41.1% on rationale localizability and disentanglability. This indicates that naive fine-tuning using augmented information without constraints would deteriorate the rationale correctness of the model.

**Qualitative results:** In Fig. 4, we show the visualizations of our visual evidence of different rationales. As shown, the rationales' visual evidence of the CLIP model [1] are entangled and mislocalized. In contrast, the rationales' visual evidence of our model are visually distinct and correctly localized.

### 4.6 Ablation Study

**Ablation on rationale disentanglement:** The "w/o disen." refers to a variant of our method without rationale disentanglement constraint. As shown in Tab. 7, the rationale localizability decreased by 10.4%, indicating the model might not learn to distinguish between rationales without constraints.

**Ablation on reconstruction:** The "w/o recon." refers to a variant of our method without reconstruction constraint. As shown in Tab. 7, the rationale localizability and prediction accuracy drastically decreased by 13.3% and 30.2%. This reveals that recklessly optimizing the disentanglement between rationale can easily fall into trivial solutions.

**Generalize to different rationale sets:** According to DCLIP [9], using the text embeddings of concepts as a bottleneck layer to force the CLIP model [1] to predict based on them can improve prediction accuracy and interpretability. Specifically, the final prediction will be made by the average embedding similarity between the image and all concepts, namely $\hat{y} = \text{argmax}_y \frac{1}{K} \sum_{k=1}^{K} \langle f_\phi(I), f_\theta(c_k^y) \rangle$. We

https://doi.org/10.52202/079017-1367

Table 5: Comparison of rationale disentanglement. We conduct experiments on nine image classification datasets. Our results are on the average of three trials using different random seeds.

| Models | C10 | C100 | CUB | CAL | PETS | F101 | SUN | CARS | DTD | AVG |
|---|---|---|---|---|---|---|---|---|---|---|
| CLIP | 0.249 | 0.445 | 0.475 | 0.442 | 0.540 | 0.481 | 0.519 | 0.353 | 0.287 | 0.420 |
| DeCLIP | 0.303 | 0.297 | 0.359 | 0.395 | 0.388 | 0.392 | 0.360 | 0.332 | 0.258 | 0.343 |
| NegCLIP | 0.386 | 0.319 | 0.456 | 0.401 | 0.440 | 0.491 | 0.495 | 0.389 | 0.261 | 0.404 |
| FILIP | 0.367 | 0.359 | 0.267 | 0.260 | 0.305 | 0.384 | 0.427 | 0.371 | 0.378 | 0.346 |
| PyramidCLIP | 0.299 | 0.300 | 0.428 | 0.418 | 0.391 | 0.318 | 0.397 | 0.359 | 0.283 | 0.355 |
| CLIP-ft | 0.378 | 0.346 | 0.469 | 0.389 | 0.434 | 0.374 | 0.383 | 0.339 | 0.251 | 0.374 |
| CLIP-ft-vision | 0.327 | 0.393 | 0.433 | 0.431 | 0.491 | 0.475 | 0.423 | 0.357 | 0.274 | 0.400 |
| Ours | **0.697** | **0.714** | **0.831** | **0.821** | **0.920** | **0.823** | **0.749** | **0.776** | **0.734** | **0.785** |

Table 6: Comparison of zero-shot image-text retrieval accuracy (%). Double-correct prediction enhances the model's visual understanding. (Note that NegCLIP is trained on *MSCOCO* [66])

| Models | MSCOCO | | Flickr30K | |
|---|---|---|---|---|
| | I2T | T2I | I2T | T2I |
| CLIP | 32.5 | 28.6 | 64.0 | 60.9 |
| DeCLIP | 32.6 | 22.1 | 59.8 | 46.2 |
| NegCLIP | - | - | 69.3 | 68.1 |
| FILIP | 33.6 | 36.4 | 52.9 | 53.3 |
| PyramidCLIP | 37.1 | **37.6** | 69.0 | **69.6** |
| CLIP-ft | 24.3 | 25.1 | 42.5 | 41.6 |
| CLIP-ft-vision | 25.9 | 27.2 | 49.1 | 55.5 |
| Ours | **38.4** | 37.1 | **69.5** | 68.9 |

Table 7: Ablation study on proposed constraints using *CUB-Part* [67] and *CUB* [17].

| Models | mIoU (↑) | Acc. (%) |
|---|---|---|
| Ours w/o disen. | 11.5 ± 1.3 | 43.3 ± 0.8 |
| Ours w/o recon. | 8.6 ± 0.8 | 25.8 ± 0.7 |
| Ours (full) | **21.9** ± 1.6 | **56.0** ± 0.5 |

Table 8: Comparison of rationale-based prediction accuracy (%) on *ImageNet* [18].

| Model | CLIP | CLIP-ft | Ours |
|---|---|---|---|
| + concepts | 63.1 | 67.5 | **70.5** |
| + rand. str. | 63.3 | 68.6 | 68.3 |
| Δ | +0.2 | +1.1 | -2.2 |

use the concept set provided in [9] rather than our training rationale dataset. As shown in Tab. 8, our model can generalize to an unseen concept set with improved prediction accuracy.

**Ablation using random string:** WaffleCLIP [72] shows that random concept strings as bottlenecks can achieve similar performance gains in DCLIP [9]. We conduct an ablation study using the random strings provided by [72]. As shown in Tab. 8, since our model can distinguish between different rationales, the random strings deteriorate the prediction accuracy of our model.

### 4.7 Evaluation on Retrieval Tasks

**Zero-shot image-text retrieval:** We evaluate our model on zero-shot image-text retrieval tasks. As shown in Tab. 6, the improved rationale correctness also benefits retrieval tasks.

**Rationale-based text-to-image retrieval:** To better evaluate the rationale correctness of our model, we conduct a novel retrieval task: rationale-based text-to-image retrieval. The model should retrieve the image with a specified rationale presented. As shown in Fig. 5, in contrast to the CLIP model [1] that entangles rationales with specific categories, our model precisely understands the semantic meaning of rationales independent to categories.

## 5 Related Works

**Vision model explainability.** A widely adopted branch of explainability methods *post hoc* generates heatmaps to identify the image regions most crucial to the model's predictions, *e.g.*, GradCAM [41], LIME [7], and SHAP [73]. Although useful for revealing the correlations between inputs and outputs, such explanations might be ambiguous, and fail to correspond to high-level concepts that humans easily understand [8]. Methods like TCAV [74] curate attribute datasets to explain vision models using concepts familiar to humans. However, such methods can fail when the models do not learn these concepts [75]. Another branch of methods attempts to design specific architecture to *intrinsically interpret* model predictions, *e.g.*, CBM [76] and ProtoPNet [77]. However, they cannot guarantee the model learns the semantic meanings of the concepts correctly [14] and yield compromised prediction


https://doi.org/10.52202/079017-1367

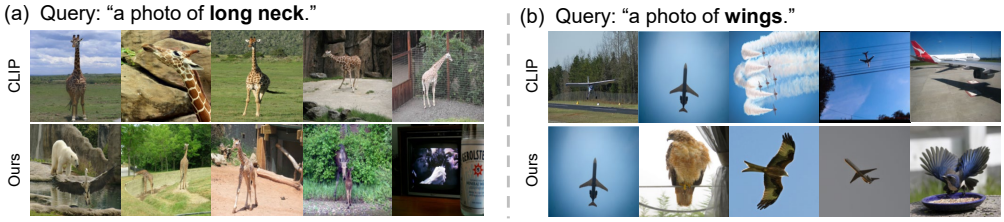


Figure 5: Qualitative results of zero-shot text-to-image retrieval on *MSCOCO* [66]. The task is to retrieve the top-5 images with a given rationale presented. The CLIP results reveal a significant entangle of rationales with a specific category, such as “long neck” with giraffes and “wings” with airliners. In contrast, our model treats rationales independently from categories, thus offering diverse retrieval results. For example, the “long neck” found in birds, giraffes, dears, and bottles.

accuracy [76]. Different from existing works, our method incorporates explanations to guide the model training, achieving accurate predictions backed by correct rationales.

**Knowledge augmentation for vision-language models.** Visual models often learn spurious correlations that stem from data biases unrelated to the causal explanation of interest [78, 79], whereas external knowledge allows models to learn the right features [77, 80]. Existing attempts for injecting knowledge into the models are often from the language modality. K-LITE [81] enrich the image caption using knowledge from WordNet [20] and Wikitionary [82]. NegCLIP [11] and DANCE [12] improve the commonsense understanding of CLIP by generating hard negative captions, the latter uses knowledge from ConceptNet [19]. StructureCLIP [83] leverages scene-graphs [21] to incorporate knowledge into text embeddings. However, our results (Tabs. 4& 7) reveal that solely augmenting information in the language cannot guarantee the model learning correct features. In contrast to these works, our method offers supervision signals from both modalities to ensure double-correctness.

**Contrastive vision-language alignment.** Different from conventional multimodal learning that fuses different modalities [84, 85], large-scale vison-language pretrained models, such as CLIP [1] and ALIGN [71], exhibit promising zero-shot transferability to downstream tasks. However, their global alignment objective is coarse, which only learns the existence of objects like bag-of-word while ignoring their localizations [11]. Recent attempts like PyramidCLIP [53] and X-VLM [54] leverage object region annotations to align word phrases with image regions. DeCLIP [56] and FILIP [57] align text with image regions through self-supervised learning. However, their supervision is limited to a coarse, object-level granularity. Different from these works, our method offers fine-grained, concept-level supervisory signals of rationales without expensive manual annotations.

## 6 Limitation

While our study advances the double-correctness of predictions, it is not without limitations. First, the absence of explicit ground truth for rationale localization in large-scale datasets remains a significant challenge. We mitigated this by leveraging a self-supervised rationale disentanglement and localization method, but this approach depends heavily on the quality of the structured rationale ontologies. Second, our methods, though effective, are computationally intensive, which may limit their applicability in resource-constrained scenarios.

## 7 Conclusion

We introduce a new concept of *double-correct predictions* aimed at training vision-language foundation models to make accurate predictions backed by correct rationales, thereby enhancing their safety for real-world deployment. To support this, we establish a solid foundation for the development of double-correct predictions. Specifically, we develop a unique dataset with structured rationales that clearly outline the reasoning processes necessary for visual recognition tasks. Furthermore, we propose a principled rationale-informed optimization method tailored for double-correct prediction. Our comprehensive empirical evaluations demonstrate that our method significantly enhances the double correctness of vision-language model predictions.

https://doi.org/10.52202/079017-1367

## Acknowledgments

This work is supported by the NSF CAREER Award No. 2340074, the NSF SAFE Award No. 2416937, the NSF III CORE Award No. 2412675, and the DoD DEPSCoR Award AFOSR FA9550-23-1-0494. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the supporting entities.

## References

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

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

[4] Xiwen Liang, Yangxin Wu, Jianhua Han, Hang Xu, Chunjing Xu, and Xiaodan Liang. Effective adaptation in multi-task co-training for unified autonomous driving. In *Advances in Neural Information Processing Systems*, 2022.

[5] Xiwen Liang, Minzhe Niu, Jianhua Han, Hang Xu, Chunjing Xu, and Xiaodan Liang. Visual exemplar driven task-prompting for unified perception in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9611–9621, 2023.

[6] Randy L Teach and Edward H Shortliffe. An analysis of physician attitudes regarding computer-based clinical consultation systems. *Computers and Biomedical Research*, 14(6):542–558, 1981.

[7] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[8] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[9] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *The Eleventh International Conference on Learning Representations*, 2022.

[10] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2022.

[11] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022.

[12] Shuquan Ye, Yujia Xie, Dongdong Chen, Yichong Xu, Lu Yuan, Chenguang Zhu, and Jing Liao. Improving commonsense in vision-language models via knowledge graph riddles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2634–2645, 2023.

[13] Tang Li, Mengmeng Ma, and Xi Peng. Deal: Disentangle and localize concept-level explanations for vlms. In *European Conference on Computer Vision*, pages 383–401. Springer, 2025.

[14] Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do concept bottleneck models learn as intended? *ICLR Workshop*, 2021.

 https://doi.org/10.52202/079017-1367

[15] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024.

[16] Li Fei-Fei, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

[17] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[18] 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*, pages 248–255. Ieee, 2009.

[19] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[20] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[21] 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*, 123:32–73, 2017.

[22] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

[23] Ribana Roscher, Bastian Bohn, Marco F Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. *Ieee Access*, 8:42200–42216, 2020.

[24] Lawrence W Barsalou. Grounded cognition. *Annu. Rev. Psychol.*, 59:617–645, 2008.

[25] Jeffrey Mark Siskind. Grounding language in perception. *Artificial Intelligence Review*, 8:371–391, 1994.

[26] Roxana Daneshjou, Mert Yuksekgonul, Zhuo Ran Cai, Roberto Novoa, and James Y Zou. Skincon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis. *Advances in Neural Information Processing Systems*, 35:18157–18167, 2022.

[27] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020.

[28] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. Knowledge graph embedding: A survey of approaches and applications. *IEEE transactions on knowledge and data engineering*, 29(12):2724–2743, 2017.

[29] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2):494–514, 2021.

[30] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. arxiv. *arXiv preprint arXiv:2303.12712*, 2023.

[31] Zhengliang Liu, Hanqi Jiang, Tianyang Zhong, Zihao Wu, Chong Ma, Yiwei Li, Xiaowei Yu, Yutong Zhang, Yi Pan, Peng Shu, et al. Holistic evaluation of gpt-4v for biomedical imaging. *arXiv preprint arXiv:2312.05256*, 2023.

[32] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[33] Timothy R McIntosh, Tong Liu, Teo Susnjak, Paul Watters, Alex Ng, and Malka N Halgamuge. A culturally sensitive test to evaluate nuanced gpt hallucination. *IEEE Transactions on Artificial Intelligence*, 2023.

[34] Mikaël Chelli, Jules Descamps, Vincent Lavoué, Christophe Trojani, Michel Azar, Marcel Deckert, Jean-Luc Raynier, Gilles Clowez, Pascal Boileau, and Caroline Ruetsch-Chelli. Hallucination rates and reference accuracy of chatgpt and bard for systematic reviews: Comparative analysis. *Journal of Medical Internet Research*, 26:e53164, 2024.

[35] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19187–19197, 2023.

[36] Ziyuan Qin, Hua Hui Yi, Qicheng Lao, and Kang Li. Medical image understanding with pretrained vision language models: A comprehensive study. In *The Eleventh International Conference on Learning Representations*, 2022.

[37] 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*, 2020.

[38] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022.

[39] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

[40] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021.

[41] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[42] Yibing Liu, Haoliang Li, Yangyang Guo, Chenqi Kong, Jing Li, and Shiqi Wang. Rethinking attention-model explainability through faithfulness violation test. In *International Conference on Machine Learning*, pages 13807–13824. PMLR, 2022.

[43] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

[44] Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. In *The Twelfth International Conference on Learning Representations*, 2023.

[45] Matthieu Guillaumin, Daniel Küttel, and Vittorio Ferrari. Imagenet auto-annotation with segmentation propagation. *International Journal of Computer Vision*, 110:328–348, 2014.

[46] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In *Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25*, pages 63–71. Springer, 2016.

 https://doi.org/10.52202/079017-1367

[47] Tang Li, Fengchun Qiao, Mengmeng Ma, and Xi Peng. Are data-driven explanations robust against out-of-distribution data? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3821–3831, 2023.

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

[49] Alexander Robey, George J Pappas, and Hamed Hassani. Model-based domain generalization. *Advances in Neural Information Processing Systems*, 34:20210–20229, 2021.

[50] Fengchun Qiao and Xi Peng. Topology-aware robust optimization for out-of-distribution generalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[51] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[52] Mengmeng Ma, Tang Li, and Xi Peng. Beyond the federation: Topology-aware federated learning for generalization to unseen clients. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

[53] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, Rongrong Ji, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *Advances in neural information processing systems*, 35:35959–35970, 2022.

[54] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *International Conference on Machine Learning*, pages 25994–26009. PMLR, 2022.

[55] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[56] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations*, 2022.

[57] Y. L. et al. Filip: Fine-grained interactive language-image pre-training. *ICLR*, 2022.

[58] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.

[59] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

[60] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 446–461. Springer, 2014.

[61] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

[62] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

[63] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[64] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[65] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[66] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[67] Oindrila Saha, Zezhou Cheng, and Subhransu Maji. Improving few-shot part segmentation using coarse supervision. In *European Conference on Computer Vision*, pages 283–299. Springer, 2022.

[68] 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*, pages 128–145. Springer, 2022.

[69] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[70] R. S. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. *ICCV*, 2017.

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

[72] Karsten Roth, Jae Myung Kim, A Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. Waffling around for performance: Visual classification with random words and broad concepts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15746–15757, 2023.

[73] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[74] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[75] Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. From" where" to" what": Towards human-understandable explanations through concept relevance propagation. *arXiv preprint arXiv:2206.03208*, 2022.

[76] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

[77] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

[78] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[79] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12556–12565, 2020.

[80] Tang Li, Jing Gao, and Xi Peng. Deep learning for spatiotemporal modeling of urbanization. *Advances in Neural Information Processing Systems Workshops (Best Paper Award)*, 2021.

[81] S. S. et al. K-lite: Learning transferable visual models with external knowledge. *NeurIPS*, 2022.

 https://doi.org/10.52202/079017-1367

[82] Christian M Meyer and Iryna Gurevych. *Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography*. na, 2012.

[83] Yufeng Huang, Jiji Tang, Zhuo Chen, Rongsheng Zhang, Xinfeng Zhang, Weijie Chen, Zeng Zhao, Zhou Zhao, Tangjie Lv, Zhipeng Hu, et al. Structure-clip: Towards scene graph knowledge to enhance multi-modal structured representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2417–2425, 2024.

[84] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2302–2310, 2021.

[85] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18177–18186, 2022.

# Appendix

## A Full Prompts

Our Prompt to Obtain Structured Rationales

```
American Robin = {
  "nodes": [
    {"id": "American Robin", "label": "American Robin"},
    {"id": "Breast", "label": "Breast"},
    {"id": "Tail", "label": "Tail"},
    {"id": "Beak", "label": "Beak"},
    {"id": "Eyes", "label": "Eyes"},
    {"id": "Red", "label": "Red"},
    {"id": "Gray", "label": "Gray"},
    {"id": "Yellow", "label": "Yellow"},
    {"id": "Round", "label": "Round"},
    {"id": "Long", "label": "Long"}
  ],
  "edges": [
    {"source": "American Robin", "target": "Breast", "relation": "has"},
    {"source": "American Robin", "target": "Tail", "relation": "has"},
    {"source": "American Robin", "target": "Beak", "relation": "has"},
    {"source": "American Robin", "target": "Eyes", "relation": "has"},
    {"source": "Breast", "target": "Red", "relation": "is"},
    {"source": "Tail", "target": "Gray", "relation": "is"},
    {"source": "Beak", "target": "Yellow", "relation": "is"},
    {"source": "Eyes", "target": "Round", "relation": "are"},
    {"source": "Tail", "target": "Long", "relation": "is"}
  ]
}

Airliner = {
  "nodes": [
    {"id": "Airliner", "label": "Airliner"},
    {"id": "Wings", "label": "Wings"},
    {"id": "Tail", "label": "Tail"},
    {"id": "Fuselage", "label": "Fuselage"},
    {"id": "Engines", "label": "Engines"},
    {"id": "Windows", "label": "Windows"},
    {"id": "Logo", "label": "Logo"},
    {"id": "Large", "label": "Large"},
    {"id": "Horizontal stabilizer", "label": "Horizontal stabilizer"},
    {"id": "Cylindrical", "label": "Cylindrical"},
    {"id": "Under wings", "label": "Under wings"},
    {"id": "Rowed", "label": "Rowed"},
    {"id": "Tail fin", "label": "Tail fin"}
  ],
  "edges": [
    {"source": "Airliner", "target": "Wings", "relation": "has"},
    {"source": "Airliner", "target": "Tail", "relation": "has"},
    {"source": "Airliner", "target": "Fuselage", "relation": "has"},
    {"source": "Airliner", "target": "Engines", "relation": "has"},
    {"source": "Airliner", "target": "Windows", "relation": "has"},
    {"source": "Airliner", "target": "Logo", "relation": "has"},
    {"source": "Wings", "target": "Large", "relation": "are"},
    {"source": "Tail", "target": "Horiz. stabilizer", "relation": "has"},
```


https://doi.org/10.52202/079017-1367

```
    {"source": "Fuselage", "target": "Cylindrical", "relation": "is"},
    {"source": "Engines", "target": "Under wings", "relation": "are"},
    {"source": "Windows", "target": "Rowed", "relation": "are"},
    {"source": "Tail", "target": "Tail fin", "relation": "has"}
  ]
}

What are useful visual concepts for distinguishing a {category_name}
in a photo? These features should be visually distinctable and have
limited overlap with each other. These features should include
attributes and their relations. For each item, you should be concise
and precise, and use no more than five words. No ambiguous answers.
Show your answer using a tree structure in JSON format strictly
following the examples shown above. Only contains two depths of
nodes (depth 1: attributes, depth 2: subattributes). No connections
between node with the same depth. Do not contain a node without an
edge connected to it. No other explanations, only provide the graph.
```

## B Machine Evaluation on Full Dataset

Table 9: The machine evaluation results on the quality of the full rationale dataset.

| Evaluators | Factual Consistency | Comprehensiveness | Visual Disentanglement |
|---|---|---|---|
| GPT-4v | 4.74 | 4.39 | 4.52 |
| GPT-4o | 4.89 | 4.59 | 4.61 |

## C Full Table

Table 10: Evaluation on *PartImageNet* [68] with ground truth region of parts using ViT-B/32 vision encoder. We summarize the annotated parts for different categories into 13 common parts. We apply thresholds to the explanation heatmaps and calculate their mIoU with ground truth masks. Our model improves the localization accuracy of each part, even though they appear significantly different across categories, such as "wings" for birds and airliners.

| Model | Head | Body | Foots | Tail | Hands | Fin | Wings | Tiers | Mirror | Seat | Seal | Engine | Mouth | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CLIP | 8.4 | 9.6 | 3.9 | 2.5 | 4.7 | 3.5 | 5.5 | 4.2 | 0.9 | 2.5 | 11.1 | 3.8 | 7.6 | 5.2 |
| DeCLIP | 5.9 | 6.5 | 3.2 | 2.4 | 3.7 | 1.8 | 5.2 | 3.0 | 0.5 | 1.7 | 5.4 | 2.7 | 5.9 | 3.7 |
| NegCLIP | 8.3 | 8.1 | 4.7 | 2.1 | 5.2 | 3.7 | 5.7 | 4.1 | 0.7 | 2.0 | 12 | 3.7 | 6.9 | 5.2 |
| FILIP | 5.4 | 6.9 | 3.2 | 2.6 | 3.9 | 3.0 | 5.8 | 3.4 | 0.5 | 1.8 | 6.9 | 3.1 | 4.9 | 4.0 |
| PyramidCLIP | 4.9 | 7.4 | 3.3 | 1.7 | 5.7 | 2.8 | 5.2 | 3.8 | 0.8 | 1.6 | 7.7 | 1.8 | 4.0 | 3.9 |
| CLIP-ft | 6.9 | 7.3 | 3.5 | 1.9 | 4.9 | 2.6 | 5.5 | 4.2 | 0.5 | 1.9 | 9.7 | 2.6 | 6.5 | 4.5 |
| CLIP-ft-vision | 6.4 | 6.9 | 3.4 | 2.1 | 4.4 | 2.5 | 5.4 | 4.1 | 0.7 | 2.1 | 9.3 | 2.9 | 6.4 | 4.4 |
| Ours | **16.7** | **21.1** | **8.9** | **7.2** | **10.4** | **7.9** | **11.5** | **10.1** | **3.2** | **5.5** | **26.3** | **4.7** | **12.3** | **11.2** |

## D Implementation Details

Table 11: Datasets for classification task.

| Dataset | Abbreviation | Classes | Train Size | Test Size |
|---|---|---|---|---|
| CIFAR-10 | C10 | 10 | 50,000 | 10,000 |
| CIFAR-100 | C100 | 100 | 50,000 | 10,000 |
| Describable Textures | DTD | 47 | 3,760 | 1,880 |
| Stanford Cars | CARS | 196 | 8,144 | 8,041 |
| Food-101 | F101 | 101 | 75,750 | 25,250 |
| Oxford-IIIT Pets | PETS | 37 | 3,680 | 3,669 |
| SUN397 | SUN | 397 | 19,850 | 19,850 |
| Caltech-101 | CAL | 102 | 3,060 | 6,085 |
| CUB-200-2011 | CUB | 200 | 5,994 | 5,794 |

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly outline the main claims of the paper, delineating both the theoretical and experimental contributions, which are supported by the results presented.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper comprehensively discusses the limitations of the proposed methods, including robustness against violations of underlying assumptions and scalability concerns.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.


https://doi.org/10.52202/079017-1367

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper provides a detailed presentation of full assumptions and definitions, as presented in Secs. 2& 3. Each equation and its definitions are clearly numbered and cross-referenced.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper fully discloses all necessary details for reproducing the main experimental results, including comprehensive descriptions of the methodologies, experimental setups, and parameter settings. In addition, the code and specific datasets are provided as well.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provides an Anonymous GitHub link with open access to both the data and code used in the experiments, complete with detailed instructions in the supplemental material that enable faithful reproduction of the main experimental results. This includes exact commands, necessary environment details, and scripts for preprocessing data, ensuring that other researchers can replicate the study's findings without ambiguity.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper details all aspects of the experimental settings, including data splits, hyperparameter selection processes, and the types of optimizers used.

   Guidelines:


https://doi.org/10.52202/079017-1367

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper provides statistical measures such as error bars and confidence intervals for all major experimental results, such as Tab. 7. These measures are correctly defined, and the paper details the variability factors they capture, including train/test splits and initialization randomness.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper adequately details the computational resources required for each experiment, including the types of compute workers (CPU or GPU), memory specifications, and execution times.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research presented in the paper adheres fully to the NeurIPS Code of Ethics, ensuring ethical considerations are addressed and complied with throughout the study.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper effectively discusses both the potential positive and negative societal impacts of the research conducted. It acknowledges the benefits of the proposed technology in enhancing data-driven decision-making processes while also addressing possible negative implications, such as unsafe predictions.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: The paper outlines comprehensive safeguards for the responsible release of data and models, particularly those with potential for high misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.


https://doi.org/10.52202/079017-1367

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly credits the creators and original owners of all assets used, including datasets, models, and code. Each asset is clearly cited, with references to the original sources and explicit mention of licenses and terms of use.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new assets introduced in the paper, including datasets and models, are well documented with comprehensive details provided in structured templates.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve any experiments or research activities that include crowdsourcing or direct interactions with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve any experiments or research activities that include crowdsourcing or direct interactions with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.


https://doi.org/10.52202/079017-1367