SEMANTIC INDUCTIVE BIAS FOR DISENTANGLEMENT THROUGH VISION-LANGUAGE MODELS AND ITERATIVE REFINEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Disentangled representation learning aims to recover meaningful factors of variation from observations. However, the observations alone do not uniquely determine a single factorization, making inductive bias necessary. The factors typically used in disentanglement, such as shape, color, pose, and position, are semantic concepts selected by humans. Recovering these factors benefits from inductive bias informed by semantic knowledge. Pretrained vision-language models (VLMs) acquire broad visual-semantic knowledge through large-scale vision-language pretraining and can provide this form of inductive bias. We therefore formulate factor discovery as a set-level comparative task, presenting small collections of unlabeled images as contact sheets from which a VLM identifies candidate semantic factors and provides visual evidence for each one. Rather than accepting these proposals directly, we iteratively validate and refine them according to sufficiency, minimality, and semantic independence using complementary image-based evidence. We evaluate the method on four disentanglement benchmark datasets using only 300 unlabeled target images and locally deployed open-weight models. The method consistently recovers the ground-truth factors on XYCS, dSprites, and 3D-Shapes, while MPI3D-C produces more variable semantic decompositions. The resulting representations are competitive with β-VAE, DAE, and InfoGAN trained using the same 300 images and remain competitive with baselines trained using all images in the target dataset.

1 INTRODUCTION

Disentangled representation learning aims to recover interpretable factors of variation from observations (Bengio et al., 2013; Higgins et al., 2018; Cha & Thiyagalingam, 2023). However, observations alone do not uniquely determine a single factorization, making additional inductive bias necessary (Locatello et al., 2019). Existing approaches introduce statistical objectives, regularization, or architectural constraints to favor particular decompositions of information, but these biases do not necessarily determine which factorization humans regard as meaningful (Chen et al., 2016; Higgins et al., 2017; Kim & Mnih, 2018; Chen et al., 2018; Cha & Thiyagalingam, 2023).

The factors typically used in disentanglement, such as shape, color, pose, scale, and position, are semantic concepts selected by humans (Bengio et al., 2013; Higgins et al., 2017; 2018; Locatello et al., 2019). This suggests that meaningful disentanglement should not be viewed only as statistical factorization, but as decomposing visual variation into semantically coherent factors. Recovering such factors requires inductive bias that favors semantically meaningful decompositions. Pretrained vision-language models (VLMs) are suitable for this role because large-scale image-text pretraining gives them broad knowledge of visual concepts and their linguistic descriptions (Radford et al., 2021; Patashnik et al., 2021; Shu et al., 2023). We focus on locally deployable models so that target images do not need to be sent to external model services, which is useful for scientific and proprietary data subject to access, security, or data-governance constraints.

We therefore formulate factor discovery as a set-level comparative task. Small collections of unlabeled images are presented jointly as contact sheets, from which a VLM identifies candidate semantic factors and provides visual evidence for each one. Rather than accepting these predictions

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

directly, we define the desired factorization through three conditions. Sufficiency means that systematic variation should not remain unexplained. Minimality requires redundant factors to be removed. Semantic independence requires distinct properties that can vary separately to be represented as separate factors. Candidate factors are iteratively validated using relative comparisons, image-level measurements, and factor-blind image descriptions, with an LLM assisting in reasoning about relations between factors. The resulting dictionary specifies the discovered factors together with their definitions, types, semantic domains, topology, and image-level measurements.

We evaluate the method on four disentanglement benchmark datasets. Using only 300 images from each dataset for factor discovery, the method consistently recovers the ground-truth factors on XYCS, dSprites, and 3DShapes, while MPI3D-C produces more variable semantic decompositions. We study sensitivity to the number of images, contact-sheet size, and model choice, together with ablations of the visual variation vocabulary and validation and refinement procedure. Finally, we evaluate the discovered factors as image-level disentangled representations and compare them with β-VAE (Higgins et al., 2017), DAE (Cha & Thiyagalingam, 2023), and InfoGAN (Chen et al., 2016). The baselines are trained using either the same 300 target images or all images in the target dataset. The discovered semantic representations are competitive with the 300 image baselines and remain competitive with models trained using substantially more target data.

Our main contributions are:

- We formulate meaningful disentanglement through sufficiency, minimality, and semantic independence, and use pretrained VLM knowledge as semantic inductive bias for factor discovery.
- We introduce a set-level comparative procedure that discovers semantic factors from small unlabeled image collections presented as contact sheets and associates each candidate with observable visual evidence.
- We develop an evidence-based validation procedure that tests candidate factors against the images rather than accepting model outputs directly, allowing redundant, unsupported, or semantically bundled factors to be identified and refined.
- We show semantic factor discovery using only 300 images per dataset with locally deployed models, study sensitivity to sample size, contact-sheet layout, model choice, and prompt vocabulary, and compare the resulting representations with established disentanglement methods trained using either the same 300 target images or all images in the target dataset.

2 Related Work

2.1 Disentangled Representation Learning

Disentangled representation learning aims to separate observations into interpretable factors of variation. Early approaches introduced different inductive biases to encourage such separation. InfoGAN maximizes mutual information between selected latent variables and generated observations (Chen et al., 2016), while β-VAE restricts latent capacity through a weighted variational objective (Higgins et al., 2017). FactorVAE and β-TCVAE further encourage factorization by penalizing dependencies between latent variables through total correlation (Kim & Mnih, 2018; Chen et al., 2018). These approaches differ in how they constrain the latent representation, but the resulting factorization is still determined by the imposed statistical or architectural assumptions.

Disentanglement has been approached through a range of statistical, architectural, and objective-based constraints. However, these different formulations do not remove the fundamental ambiguity of purely unsupervised disentanglement. Locatello et al. (2019) showed that, without inductive biases on both the model and the data, the underlying disentangled factors cannot in general be identified from observations alone. Their large-scale empirical study further showed that the learned representations are highly sensitive to modeling choices and random seeds. This makes the choice of inductive bias central to determining which factorization is recovered. Our work uses semantic knowledge as this bias to identify human-interpretable factors of variation from unlabeled images.

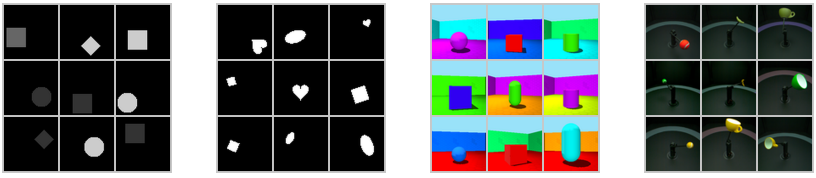

Figure 1: Example 3×3 contact sheets presented to the Semantic Perceiver. Images are randomly sampled from the benchmark datasets and presented jointly, allowing factors of variation to be identified through comparison across the image set. The examples shown are from XYCS, dSprites, 3D-Shapes, and MPI3D-C, from left to right.

2.2 Vision-Language Models for Semantic Concepts

Large-scale vision-language pretraining provides representations in which visual content can be associated with natural-language concepts. CLIP demonstrated that image-text supervision can support zero-shot recognition through natural-language descriptions of visual concepts (Radford et al., 2021). More recent work has used this semantic structure for model interpretation and concept-based representations. CLIP-Dissect uses CLIP to assign semantic descriptions to internal features of pretrained vision models (Oikarinen & Weng, 2022), while label-free concept bottleneck methods use pretrained semantic models to construct interpretable concept representations without requiring concept annotations (Oikarinen et al., 2023; Yu et al., 2025). Language-guided bottleneck approaches additionally use language models to generate candidate concepts and vision-language models to associate them with images (Oikarinen et al., 2023; Yang et al., 2023).

Most closely related to our motivation, Zang et al. (2024) show that pretrained vision-language models contain generic visual concepts that can be extracted using natural-language descriptions. Together with related concept-based methods, this supports the view that pretrained multimodal models encode semantic visual knowledge that can be accessed without requiring concept-level annotations. Our objective differs from concept-based classification or network interpretation. We do not begin with a predefined dataset-specific concept set or seek concepts that explain a classifier. Instead, we identify which semantic factors vary across an unlabeled image collection. We formulate this as a comparative task over image sets and then test the proposed factorization for sufficiency, minimality, and semantic independence using image-based evidence.

3 Method

3.1 Semantic factorization

Given an unlabeled image dataset $\mathcal{X}$, our goal is to discover a set of semantic factors of variation

$$\mathcal{F} = \{f_1, \ldots, f_m\},$$

without knowing either the factors or their number in advance. Each factor f_i is represented by a semantic name and description, a type, a semantic domain, and a topology. The type is either continuous or categorical. For continuous factors, the domain is defined by semantic low and high endpoints, while categorical factors are defined by a set of semantic categories. The topology specifies whether the factor is linear, circular, or unordered. After discovery, an image-level value can be assigned to each factor.

The discovered factors are refined according to three criteria. Sufficiency requires the factor set to contain enough factors to describe the consistent visual variation in the images. Minimality requires the factor set to contain no redundant factors, so that each factor contributes information not captured by the others. Semantic independence requires distinct semantic properties to be represented by separate factors, so that changing one factor corresponds to a change in one semantic property without necessarily changing the others. This does not require statistical independence. We apply these criteria through the validation and refinement procedure in Section 3.3.

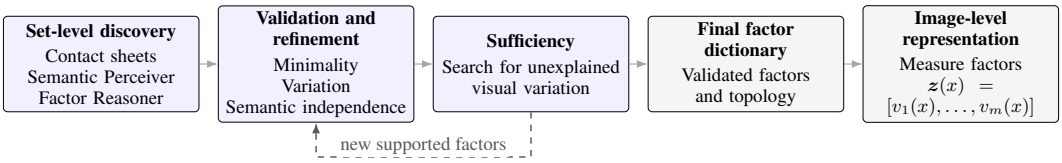

Figure 2: Overview of the proposed method. Factors discovered from image sets are refined, tested for sufficiency, and assembled into a final factor dictionary, which is then used to obtain image-level semantic representations.

3.2 Set-level Semantic Factor Discovery

We formulate factor discovery as a comparative task over image sets. Rather than processing images independently, we randomly sample small collections of images and present them jointly as 3×3 contact sheets, as illustrated in Figure 1. The VLM serves as the *Semantic Perceiver*, comparing images within each sheet and converting the visual differences into candidate semantic factors.

To guide this comparison, we provide the Perceiver with a fixed, dataset-independent visual variation vocabulary. The vocabulary includes common geometric and appearance changes such as translation, rotation, scale, shear, flip, brightness, contrast, saturation, hue, and sharpness, which are widely used in data augmentation and image processing (Van der Walt et al., 2014; Cubuk et al., 2019; 2020). It also covers broader sources of variation, including object identity, object count and arrangement, and viewpoint and illumination. Because the vocabulary describes general forms of visual variation rather than dataset-specific factors, we use it unchanged across all datasets and do not provide the dataset identity or ground-truth factors to the models. The same vocabulary is included in the Semantic Perceiver prompts used for set-level discovery and the sufficiency search. Full prompts for both the Semantic Perceiver and Factor Reasoner are provided in Section C of the supplementary material.

Candidate factors are collected from independently sampled contact sheets. The LLM serves as the *Factor Reasoner*, comparing proposals across sheets and combining descriptions that refer to the same semantic property into an initial candidate dictionary. Differences in naming or granularity are resolved at this stage so that repeated observations of the same variation are represented consistently.

A candidate enters the validation procedure only when it is independently identified across multiple contact sheets. This reduces candidates arising from a single model response and provides an initial factor set supported by repeated visual observations.

We explicitly evaluate the role of the visual variation vocabulary by comparing the full vocabulary with transformation-only, semantic-class-only, and free-form discovery in Section 4.4.

3.3 Factor Validation and Refinement

The candidate factors are iteratively evaluated using the criteria introduced in Section 3.1. The Perceiver first provides image-level measurements for the current factor set. These measurements, together with visual and semantic evidence, are then used to identify redundant factors, verify that each factor corresponds to observable variation, and determine whether a factor combines semantic properties that should be represented separately. Newly introduced factors pass through the same procedure. The overall discovery, validation, and measurement procedure is summarized in Figure 2.

Once the current factor set stabilizes, sufficiency is evaluated by searching for systematic visual variation not represented by the existing factors. Any newly supported factors are added to the candidate set and evaluated through the same refinement procedure. When no additional supported variation is found, a final conditional minimality test determines whether every remaining factor is still needed.

Minimality. Minimality is evaluated at two stages of the procedure. During refinement, candidates with strongly associated measurements or similar visual evidence are examined to determine whether they represent the same underlying property. Equivalent candidates are represented as a single factor, while semantically distinct properties remain separate. After the sufficiency search, each

remaining factor is tested against other factors whose measurements are strongly associated with it. We compare images that differ in the candidate factor but are otherwise similar with respect to these associated factors. If the candidate no longer shows a consistent visual difference under these comparisons, the variation attributed to the candidate can be explained by the existing factors and it is removed. Details of how strongly associated factors are identified in both stages are provided in Section A.5 of the supplementary material.

Variation. A candidate must correspond to an observable change in the dataset. The Perceiver first identifies images in which the candidate property appears to differ on independently sampled contact sheets. Because this comparison explicitly specifies the candidate, the response may be influenced by the proposed factor itself. We therefore perform a second check in which the same images are described without mentioning the candidate factor. The Factor Reasoner then evaluates whether the differences expressed in these general descriptions are consistent with the proposed variation.

Semantic independence. A candidate may combine multiple semantic properties that can vary separately. The Factor Reasoner examines the candidate description, semantic domain, and its relation to the other factors to identify possible decompositions into simpler components. Since this decomposition is initially proposed from semantic information, we verify that the proposed components are also supported by the images. The components are therefore measured on the images, and the candidate is separated only when these measurements support the proposed decomposition. Newly introduced components return to the refinement procedure, where their variation and redundancy are evaluated in the same way as the original candidates.

Sufficiency. Sufficiency tests whether the current dictionary misses any systematic visual variation. We search for missing factors using contact sheets selected from the current image-level measurements. First, for each discovered factor, we group images with similar values of that factor, so that the Perceiver can examine which other properties still vary while that factor remains approximately constant. We also select a group of images that are similar across all discovered factors. If these images still show consistent visual differences, this suggests that the current dictionary is missing a factor that explains those differences. The Perceiver compares these sheets while being provided with the current dictionary and proposes additional varying properties. Newly proposed factors are required to receive support across multiple contact sheets before being added to the candidate set and evaluated through the same validation and refinement procedure.

3.4 Semantic factor dictionary and measurement

After validation is complete, the topology of each remaining factor is determined. This produces the final semantic factor dictionary, which contains the number of discovered factors and, for each factor, its name, description, type, semantic domain, topology, and the visual evidence supporting it. Once the dictionary is fixed, the Semantic Perceiver can measure each discovered factor for an individual image. For a continuous factor, it estimates a value in $[0, 1]$ using the semantic low and high endpoints as references. For a categorical factor, it assigns one of the discovered categories. For an image x, the measurements of all discovered factors form its image-level semantic representation,

$$\boldsymbol{z}(x) = [v_1(x), \ldots, v_m(x)],$$

where $v_i(x)$ denotes the measurement of factor f_i.

Factor discovery and factor measurement are treated as separate problems. Discovery determines which semantic factors are supported by variation across the dataset, whereas measurement estimates the value of an already discovered factor for an individual image. The former relies primarily on relative comparisons across images, while the latter requires an absolute reading within the semantic domain defined for each factor. Consequently, a factor may be discovered reliably even when its image-level measurements are imperfect. Exact prompts, association statistics, thresholds, stopping criteria, and implementation details are provided in the supplementary material.

4 Experiments

We evaluate semantic factor discovery, robustness to observation and model settings, the contribution of individual method components, and the quality of the resulting image-level representations.

Table 1: Semantic factor discovery over five seeds.

DATASET	FACTOR RECOVERY	COMPLETE RUNS	DISCOVERED FACTORS
XYCS	$4/4$	$5/5$	5.6 ± 0.55
dSprites	$5/5$	$5/5$	5.0 ± 0.00
3D-shapes	5.8 ± 0.45 / 6	$4/5$	9.6 ± 1.82
MPI3D-C	–	–	18.2 ± 3.03

4.1 EXPERIMENTAL SETUP

We use Qwen3-VL-32B-Instruct (Bai et al., 2025a) as the Semantic Perceiver and Qwen3-14B (Yang et al., 2025) as the Factor Reasoner. For each dataset, 300 images are randomly sampled and presented as 3×3 contact sheets. All experiments are run on an NVIDIA A100 GPU with 40 GB of memory.

Datasets. We evaluate the method on four disentanglement benchmarks with different levels of visual complexity, namely XYCS (Cha & Thiyagalingam, 2023), dSprites (Higgins et al., 2017), 3D-Shapes (Kim & Mnih, 2018), and MPI3D-Complex (real world shapes), referred to as MPI3D-C (Gondal et al., 2019). XYCS and dSprites contain controlled two-dimensional variation in geometric and appearance properties. 3D-shapes consists of rendered 3D scenes with controlled variation in object and scene properties, while MPI3D-C contains substantially more complex real-world images captured in a robotic environment. Ground-truth generative factors are used only for evaluation and are not provided during factor discovery.

Semantic evaluation. Semantic evaluation compares the discovered factors with the known generative factors of each benchmark. For each discovered factor, we compute a type-dependent correspondence score with every ground-truth factor using their image-level measurements. The discovered factor is matched to the ground-truth factor with the highest score, and the match is counted when the score is at least 0.4. A ground-truth factor is considered recovered when at least one discovered factor is matched to it. The correspondence measure depends on the factor types, with full definitions provided in Section A.6 of the supplementary material.

This recovery measure assumes a direct correspondence between discovered and ground-truth factors. However, the same visual variation can also be represented using a different set of semantic factors. For example, two-dimensional position can be represented using Cartesian coordinates (x, y) or polar coordinates (r, θ). We therefore distinguish ground-truth factor recovery from ground-truth variation coverage.

For XYCS, dSprites, and 3D-Shapes, the ground-truth factors correspond closely to semantic visual properties, so ground-truth factor recovery is directly reported. For MPI3D-C, we instead consider whether the discovered semantic factor set covers the variation associated with all seven ground-truth factors.

Representation evaluation. We evaluate whether the discovered factors provide numerical representations of individual images. The discovered factors are applied to 1,000 images, and the resulting image-level representations are evaluated using standard disentanglement metrics. Factor discovery uses 300 images from each target dataset, so we train β-VAE, DAE, and InfoGAN using the same 300 images. Because the Semantic Perceiver also uses visual-semantic knowledge acquired during pretraining, this comparison does not use the same amount of training data for all methods. We therefore additionally train the baselines using all images in the target dataset.

4.2 SEMANTIC FACTOR DISCOVERY

Across five random seeds, we evaluate ground-truth factor recovery, the number of runs with complete recovery, and the total number of discovered factors in Table 1. A complete run recovers all ground-truth factors under the evaluation defined in Section 4.1.

Table 2: Robustness and model dependence of semantic factor discovery. For robustness, $a/b\,(N)$ denotes recovery of a out of b ground-truth factors with N discovered factors. MPI3D-C reports only N. Model comparisons report factor-set F1.

	SETTING	XYCS	dSprites	3D-Shapes	MPI3D-C
Robustness					
Target images	150	$3/4\,(4)$	$5/5\,(5)$	$6/6\,(9)$	– (21)
	300	$4/4\,(5)$	$5/5\,(5)$	$6/6\,(12)$	– (20)
	600	$4/4\,(6)$	$5/5\,(5)$	$6/6\,(12)$	– (13)
	1200	$4/4\,(5)$	$5/5\,(5)$	$6/6\,(12)$	– (14)
Sheet size	2×2	$4/4\,(5)$	$5/5\,(5)$	$6/6\,(10)$	– (15)
	3×3	$4/4\,(5)$	$5/5\,(5)$	$6/6\,(12)$	– (20)
	4×4	$4/4\,(6)$	$5/5\,(5)$	$5/6\,(12)$	– (20)
	5×5	$4/4\,(6)$	$5/5\,(5)$	$5/6\,(13)$	– (37)
Model dependence		**Factor-set F1**			
Semantic Perceiver	Qwen3-VL-4B	.60	.33	.29	.19
	Qwen3-VL-8B	.67	.57	.40	.25
	Qwen3-VL-32B	.89	.60	.56	.15
Factor Reasoner	Qwen3-4B	.80	.73	.67	.38
	Qwen3-8B	.89	.60	.67	.38
	Qwen3-14B	.89	.60	.56	.15

XYCS and dSprites recover all ground-truth factors in every run. dSprites returns exactly five factors in every run, while XYCS discovers 5.6 ± 0.55 factors. 3D-shapes recovers 5.8 ± 0.45 of six ground-truth factors, with object orientation missed in one of the five runs.

For MPI3D-C, factor recovery is not reported because its ground-truth generative factors do not necessarily correspond one-to-one with semantic visual factors. The method discovers 18.2 ± 3.03 semantic factors on average, and the resulting image-level representation is evaluated against the ground-truth factors in Section 4.5.

4.3 Robustness and model dependence

Sensitivity to the number of images, contact-sheet size, and model size is evaluated using ground-truth factor recovery and the total number of discovered factors in Table 2.

Factor recovery remains stable across a broad range of observation settings. dSprites recovers all five factors under every tested sample size and sheet size, while XYCS recovers all four except with 150 images, where vertical position is not recovered separately. 3D-shapes recovers all six factors for every sample size and for 2×2 and 3×3 sheets, but misses object orientation with 4×4 and 5×5 sheets. Increasing the number of images or images per sheet therefore does not consistently improve recovery.

MPI3D-C is more sensitive to these settings. The number of discovered factors decreases from 21 and 20 with 150 and 300 images to 13 and 14 with 600 and 1200 images, while increasing the sheet size from 4×4 to 5×5 increases the number from 20 to 37. Since larger settings do not consistently improve recovery and can substantially change the granularity of the discovered factorization, we use 300 images and 3×3 contact sheets as the default settings.

For model comparisons, we vary Qwen3 model size and additionally evaluate Qwen2.5-VL (Bai et al., 2025b), Gemma3 (Team et al., 2025), and Mistral Small (Mistral AI Team, 2025) as Semantic Perceivers, together with Qwen2.5 (Yang et al., 2024), Gemma3 (Team et al., 2025), Llama3.1 (Grattafiori et al., 2024), and Mistral Small (Mistral AI Team, 2025) as Factor Reasoners. We use factor-set F1, which uses the same factor correspondences while additionally penalizing missing, duplicate, and additional factors. Details of the model configurations, correspondence measure, and score calculation are provided in the supplementary material.

Within the Qwen3 family, Semantic Perceiver size has a more consistent effect than Factor Reasoner size on XYCS, dSprites, and 3D-shapes. Cross-family results further show that factor-set F1 depends on both model choice and dataset, with complete results provided in Section B.4.

Table 3: Ablation of the visual variation vocabulary. Each entry reports exact recovery of the ground-truth factors, followed in parentheses by the total number of discovered factors.

VOCABULARY	XYCS	dSprites	3D-shapes
Full	$4/4\,(5)$	$5/5\,(5)$	$6/6\,(9)$
Transformations only	$4/4\,(6)$	$5/5\,(5)$	$6/6\,(11)$
Semantic classes only	$4/4\,(6)$	$5/5\,(9)$	$6/6\,(9)$
Free-form	$4/4\,(5)$	$5/5\,(6)$	$4/6\,(11)$

4.4 Ablation studies

All ablations use the default model configuration, 300 images, and 3×3 contact sheets.

Visual variation vocabulary. We compare the full visual variation vocabulary with transformation-only, semantic-class-only, and free-form discovery. The same guidance is used during set-level discovery and the sufficiency search, with exact prompts provided in Section C of the supplementary material.

Table 3 shows that all settings recover the complete ground-truth factor set on XYCS and dSprites. On 3D-shapes, free-form discovery recovers four of six factors, while the other settings recover all six. The full vocabulary gives the most consistently compact factorization while maintaining complete recovery across the three datasets.

Validation and refinement. We next ablate the multiple-sheet support requirement and the validation and refinement components, with complete component-wise results provided in Section B.5.

Removing all checks jointly preserves complete recovery but increases the number of additional factors from four to twelve, including an increase from nine to sixteen factors on 3D-shapes. Removing semantic independence instead leaves wall and floor color combined as a single background-color factor and reduces factor recovery on 3D-shapes to three of six factors. The procedure therefore affects both compactness and the separation of distinct semantic properties.

4.5 Disentangled representation quality

We evaluate the image-level semantic representations on 1,000 images using metrics covering the three families of disentanglement measures identified by Carbonneau et al. (2022). These include z-diff and z-min-var as intervention-based metrics, DCI disentanglement (DCI-D) and SAP as predictor-based metrics, and MIG and JEMMIG as information-based metrics. Table 4 compares the image-level semantic representations obtained from the discovered factors with β-VAE, DAE, and InfoGAN trained using either the same 300 images used for factor discovery or all images in the target dataset. Baseline training and representation evaluation details are provided in Section D of the supplementary material.

When the baselines are trained using the same 300 images, the discovered semantic representations have the highest mean on all six metrics for XYCS and dSprites and on five of six metrics for 3D-shapes and MPI3D-C. The exception on both datasets is z-min-var, with the difference on 3D-shapes small relative to the standard deviations.

With baselines trained using all images, the discovered semantic representations retain the highest mean on all six metrics for XYCS, on JEMMIG for dSprites, on z-diff, JEMMIG, and SAP for 3D-shapes, and on z-diff, DCI-D, JEMMIG, and MIG for MPI3D-C. Baselines using all images therefore reduce or reverse some of the differences, particularly on dSprites and 3D-shapes.

Across the five random seeds, the discovered semantic representations generally have small standard deviations and show lower sensitivity to random seed than the 300 image baselines for many metrics. The main exception is z-min-var on 3D-shapes. These results show that the discovered factors provide competitive disentangled representations from 300 images, while the all image comparison accounts for the substantially larger amount of target data available to the conventional baselines.

Table 4: Disentanglement quality over five runs. Entries are mean ± standard deviation. Ours uses 300 images for factor discovery. Baselines are trained using either 300 images or all images, corresponding to 42,135 for XYCS, 737,280 for dSprites, 480,000 for 3D-shapes, and 460,800 for MPI3D-C. Higher values indicate better performance for all reported metrics.

DATASET	METHOD	z-diff ↑	z-min-var ↑	DCI-D ↑	JEMMIG ↑	MIG ↑	SAP ↑
XYCS	Ours (300)	**1.000** ± .001	**.949** ± .040	**.520** ± .014	**.697** ± .023	**.511** ± .081	**.670** ± .066
	β-VAE (300)	.609 ± .066	.324 ± .030	.014 ± .005	.314 ± .019	.030 ± .013	.154 ± .082
	DAE (300)	.818 ± .112	.586 ± .140	.138 ± .073	.392 ± .094	.125 ± .083	.329 ± .176
	InfoGAN (300)	.570 ± .107	.343 ± .033	.011 ± .002	.283 ± .010	.016 ± .003	.103 ± .050
	β-VAE (all)	.719 ± .168	.482 ± .190	.065 ± .059	.368 ± .070	.112 ± .098	.269 ± .175
	DAE (all)	.882 ± .056	.714 ± .036	.447 ± .177	.577 ± .102	.330 ± .024	.519 ± .023
	InfoGAN (all)	.823 ± .113	.605 ± .092	.199 ± .091	.335 ± .035	.210 ± .089	.191 ± .115
dSprites	Ours (300)	.781 ± .082	.645 ± .053	.171 ± .036	**.525** ± .015	.229 ± .035	.392 ± .064
	β-VAE (300)	.505 ± .089	.305 ± .066	.012 ± .005	.304 ± .021	.031 ± .015	.134 ± .069
	DAE (300)	.683 ± .121	.486 ± .134	.079 ± .047	.362 ± .069	.123 ± .073	.299 ± .114
	InfoGAN (300)	.502 ± .121	.345 ± .094	.015 ± .009	.308 ± .012	.028 ± .017	.092 ± .067
	β-VAE (all)	.758 ± .063	.598 ± .090	**.281** ± .119	.491 ± .061	**.273** ± .110	.396 ± .088
	DAE (all)	**.807** ± .027	**.661** ± .022	.208 ± .047	.401 ± .039	.219 ± .053	**.503** ± .079
	InfoGAN (all)	.689 ± .100	.555 ± .122	.084 ± .032	.331 ± .021	.110 ± .044	.285 ± .111
3D-shapes	Ours (300)	**.990** ± .008	.693 ± .190	.546 ± .016	**.789** ± .015	.546 ± .038	**.368** ± .058
	β-VAE (300)	.687 ± .093	.577 ± .094	.270 ± .071	.502 ± .066	.302 ± .106	.091 ± .034
	DAE (300)	.648 ± .071	.542 ± .086	.198 ± .053	.410 ± .050	.208 ± .069	.050 ± .022
	InfoGAN (300)	.845 ± .094	.705 ± .132	.184 ± .029	.393 ± .041	.178 ± .080	.270 ± .124
	β-VAE (all)	.840 ± .035	.732 ± .057	.568 ± .064	.690 ± .041	.573 ± .059	.208 ± .077
	DAE (all)	.846 ± .020	**.778** ± .016	**.671** ± .085	.712 ± .046	**.602** ± .046	.286 ± .064
	InfoGAN (all)	.391 ± .054	.284 ± .071	.025 ± .017	.334 ± .026	.025 ± .009	.035 ± .016
MPI3D-C	Ours (300)	**.675** ± .052	.319 ± .044	**.318** ± .019	**.540** ± .034	**.171** ± .032	.135 ± .048
	β-VAE (300)	.276 ± .014	.272 ± .026	.086 ± .006	.441 ± .038	.027 ± .023	.060 ± .042
	DAE (300)	.406 ± .042	.314 ± .025	.091 ± .011	.307 ± .013	.059 ± .028	.075 ± .040
	InfoGAN (300)	.477 ± .084	.373 ± .068	.098 ± .035	.328 ± .017	.056 ± .029	.075 ± .034
	β-VAE (all)	.280 ± .019	.239 ± .029	.072 ± .009	.477 ± .030	.024 ± .020	.063 ± .026
	DAE (all)	.502 ± .122	**.405** ± .094	.167 ± .081	.325 ± .043	.160 ± .092	**.153** ± .065
	InfoGAN (all)	.262 ± .017	.212 ± .017	.034 ± .010	.348 ± .013	.015 ± .006	.029 ± .011

5 CONCLUSION

We introduced a framework for discovering semantically meaningful factors of variation from unlabeled images using pretrained vision-language models as semantic inductive bias. Candidate factors are discovered from image sets and refined according to sufficiency, minimality, and semantic independence to form image-level semantic representations.

Across XYCS, dSprites, and 3D-Shapes, the method recovers most or all ground-truth factors using only 300 images and remains robust across a range of sample sizes and contact-sheet settings. All Semantic Perceiver and Factor Reasoner models are deployed locally, so target images do not need to be sent to external model services. The discovered factors also provide competitive disentangled representations compared with β-VAE, DAE, and InfoGAN, including baselines trained using all images in the target dataset.

The results also identify several limitations. Factor discovery depends on the choice and size of the Semantic Perceiver and Factor Reasoner, indicating sensitivity to differences in pretrained visual-semantic knowledge and reasoning ability. Performance is also less stable on MPI3D-C, where the number of discovered factors varies substantially across observation settings and model configurations. This suggests that semantic factor discovery becomes more difficult as visual variation becomes more complex and realistic. Future work should investigate more model-independent discovery procedures, stronger validation of candidate factors, and methods that scale more reliably to complex real-world image collections.

AI USE STATEMENT

Generative AI tools were used to aid and polish the writing and to draft parts of the manuscript. All AI-assisted text was reviewed, edited, and verified by the authors before inclusion in the paper. The authors take responsibility for the final content of the manuscript, including all text, claims, and interpretations.

Ethics statement

This work evaluates machine learning methods using established disentanglement benchmark datasets and does not involve human participants or personal data. We are not aware of additional ethical concerns specific to the experiments presented in this work

Reproducibility statement

The main paper describes the semantic factor discovery, validation, refinement, and evaluation procedures. The supplementary material provides the complete model configuration, contact-sheet construction, stopping criteria, association statistics and thresholds, baseline training settings, representation evaluation protocol, and additional experimental results. Exact prompt templates used by the Semantic Perceiver and Factor Reasoner are also provided in the supplementary material. The code will be made publicly available upon acceptance.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Ke-Qin Chen, Junyang Lin, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Marc-André Carbonneau, Julian Zaidi, Jonathan Boilard, and Ghyslain Gagnon. Measuring disentanglement: A review of metrics. *IEEE transactions on neural networks and learning systems*, 35 (7):8747–8761, 2022.

Jaehoon Cha and Jeyan Thiyagalingam. Orthogonality-enforced latent space in autoencoders: An approach to learning disentangled representations. In *International Conference on Machine Learning*, pp. 3913–3948. PMLR, 2023.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 29, pp. 2172–2180, 2016.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 113–123. IEEE, 2019.

Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. Randaugment: Practical automated data augmentation with a reduced search space. *Advances in neural information processing systems*, 33:18613–18624, 2020.

Muhammad Walid Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Pitz-Paal, Peter Sefidgar, Bernhard Schölkopf, Stefan Bauer, and Francesco Locatello. On the transfer of inductive bias from simulation to the real world: A study on real-world disentanglement. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International conference on machine learning*, pp. 2649–2658. PMLR, 2018.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pp. 4114–4124. PMLR, 2019.

Mistral AI Team. Mistral small 3 technical report. *arXiv preprint arXiv:2501.XXXXX*, 2025.

Tuomas Oikarinen and Tsui-Wei Weng. Clip-dissect: Automatic description of neuron representations in deep vision networks. *arXiv preprint arXiv:2204.10965*, 2022.

Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*, 2023.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 2065–2074. IEEE, 2021.

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

Yang Shu, Xingzhuo Guo, Jialong Wu, Ximei Wang, Jianmin Wang, and Mingsheng Long. Clipood: Generalizing clip to out-of-distributions. In *International conference on machine learning*, pp. 31716–31731. PMLR, 2023.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19187–19197. IEEE, 2023.

Lu Yu, Haoyu Han, Zhe Tao, Hantao Yao, and Changsheng Xu. Language guided concept bottleneck models for interpretable continual learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14976–14986. IEEE, 2025.

Yuan Zang, Tian Yun, Hao Tan, Trung Bui, and Chen Sun. Pre-trained vision-language models learn discoverable visual concepts. *arXiv preprint arXiv:2404.12652*, 2024.