A General Concept-based Decomposition for Vision-Language Embeddings

Simone Alberto Peirone*
Politecnico di Torino
Torino, Italy
simone.peirone@polito.it

Ortal Senouf
EPFL
Lausanne, Switzerland
ortal.senouf@epfl.ch

Francesca Pistilli
Politecnico di Torino
Torino, Italy
francesca.pistilli@polito.it

Giuseppe Averta
Politecnico di Torino
Torino, Italy
giuseppe.averta@polito.it

Pascal Frossard
EPFL
Lausanne, Switzerland
pascal.frossard@epfl.ch

Abstract

Vision-Language Models are designed to capture the compositional structure of the world through multi-modal alignment; however, it remains unclear whether their representations truly align with the way a human would naturally describe the same visual input. Prior works attempt to decompose visual embeddings into their concept-level contributions, but their instance-level optimization is inherently local and does not capture the model's global semantic structure, which is essential for *compositional generalization* across different inputs, tasks, and domains. We introduce GRACE, a method that learns this concept-based structure at the feature space level, producing semantic and general concept decompositions. To achieve this, we propose a novel training objective that preserves the geometric relations of different samples in the concept space, enabling a better understanding of how concepts compose and relate to each other. We validate GRACE on several image and video benchmarks, showing that its concept decomposition is aligned with human perception and achieves faithful decompositions across different tasks and domains. With minimal overhead, GRACE makes existing Vision-Language Models more interpretable, while preserving most of their original performance. Project page with code at https://sapeirone.github.io/grace/.

1 Introduction

Humans have the remarkable ability to perceive and understand reality by decomposing it into atomic semantic abstractions [1]. In vision, few volumetric primitives are enough for our visual system to recognize familiar objects, even under occlusion or visual degradation [2]. This mechanism, known as *recognition-by-components*, forms the basis of human generalization. Natural language provides a powerful mechanism to express visual abstractions in a human-interpretable form and to enable their composition into rich descriptions of the world. Thus, we expect *interpretable* models

*Work carried out while visiting EPFL.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

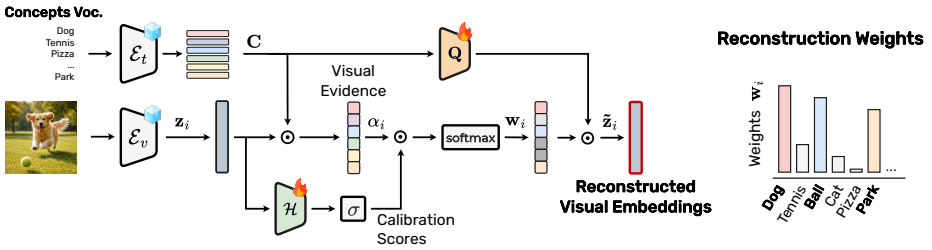


Figure 1: **The GRACE framework.** Given a visual sample, either an image or video, and a vocabulary $\mathcal{C}$ of human-interpretable concepts, a geometry-based reconstruction pipeline decomposes the visual embedding into semantically relevant concept contributions. These reconstructed embeddings can replace the original ones, improving interpretability and enabling model debugging.

to expose a reasoning structure that is well aligned with these abstractions, which is essential for *compositional generalization* [3], i.e., the ability to recognize novel compositions of known concepts. Vision–Language Models [4] align their visual representations with this language-based compositional structure, achieving strong generalization [5, 6]. However, visual data are inherently complex and often encode multiple, tightly entangled, semantic concepts. This challenge is particularly severe in videos, where objects, actions, and human motion co-occur and evolve over time [7–9]. As a result, it remains unclear whether their internal representations truly align with human-understandable concepts, which leads to a crucial question:

can we explicitly expose the interpretable concept-based structure of visual information in VLMs?

In this regard, the linear hypothesis suggests that high-level concepts are already represented linearly in feature space [10, 11], even across modalities [12]. Several works build on this intuition to decompose the visual embeddings of a VLM into text-based human-interpretable concepts. SpLiCE [13] learns a *per-sample* reconstruction of the embeddings from these concepts, while EXPLAIN-R [14] builds a concept-based surrogate representation that mimics the multi-modal alignment of the original model. Yet, many challenges remain unsolved. Instance-level decomposition may encourage shortcut learning and does not explicitly consider the different geometries of the visual and text modalities [15], while the surrogate representations in EXPLAIN-R are obtained by collecting concept evidence from external captioning models [16], rather than leveraging the intrinsic concept-based structure of the underlying VLM. These models encode concept-level interpretability but lack a *general* mechanism to expose it. Indeed, concept-based interpretability should capture not only the concept-level contributions within individual inputs, but also how concepts combine and relate across inputs, tasks, and domains, which is an essential property for generalization [5].

We address these challenges by introducing GRACE, a novel concept-based framework that learns an interpretable decomposition of visual representations at the feature space level. We leverage cross-modal matching over a large concept vocabulary to capture concept evidence directly from visual inputs, and propose a geometry-based reconstruction pipeline to ensure that the matched concepts are semantically relevant. Rather than requiring exact reconstruction of the original visual embeddings, our method is designed to preserve the pairwise geometric relations between the samples in the concept space, which encourages the model to better understand how different concepts compose and relate to each other, thus better capturing the global semantic structure of the concept space. Sparsity and consistency priors promote the selective use of relevant concepts and ensure consistent concept usage across semantically similar inputs.

We evaluate GRACE across a range of benchmarks to show that its concept decompositions are well aligned with human perception. Its reconstructed embeddings preserve or even outperform the original embeddings, depending on the task, by exposing semantically relevant information and demonstrating strong *zero-shot* generalization. We summarize our contributions as follows: (i) we explore the connection between interpretability and generalization, addressing key limitations in existing approaches for concept-based decomposition; (ii) we propose GRACE to expose the concept-based structure of VLMs, modeling how concepts compose within and across samples; (iii) we validate our method on several image and video benchmarks, showing its decomposition extracts sparse semantic concepts with strong generalization; (iv) GRACE improves alignment with human perception by +10.6% on ConVisBench, while producing more faithful reconstructions (+7.50% on EgoMCQ). Our approach overcomes many limitations of similar previous approaches, and introduces

an effective way to decompose visual embeddings into their concept-level contribution, improving interpretability while providing a practical tool to debug and better understand the model behavior.

2 Background

We consider different strategies to build models that expose human-understandable concepts. In Concept Bottleneck Models, the raw input is explicitly mapped to high-level concepts, which are then used to solve the task, typically leveraging concept-level supervision. Another line of research investigates how high-level concepts emerge as linearly separable directions in the representation space of VLMs, and how their contributions can be disentangled. Our method brings together both views: it relies on vision-language models to match visual inputs with text-based human-understandable concepts, and learns a general concept-based decomposition of the embedding space.

Concept Bottleneck Models. Concept Bottleneck Models (CBMs) [17] decompose the internal representations of the model into a set of human-understandable concepts, and then learn a mapping function from the concept space to the output space of a specific target task, e.g., an image classification task. The bottleneck is trained with concept-level supervision obtained either from human annotations [17] or by using LLMs [18, 19] to generate plausible concept vocabularies. Recent works also apply CBMs to video tasks [7, 9], learning to model the temporal evolution of concepts. While the bottleneck is commonly trained jointly with the downstream task, Post-hoc CBMs [20] extend concept-based interpretability to pretrained models by projecting their embeddings onto a set of interpretable activation vectors. CBMs learn task-specific mappings to the concept space, limiting the transferability of the learned representations across different tasks.

Sparse Autoencoders. The *linear representation hypothesis* [21–24] suggests that high-level concepts are represented linearly in feature space. Under this principle, Sparse Autoencoders (SAE) [25] learn a linear decomposition of a model's internal activations into sparse, monosemantic latent features, following an unsupervised activation reconstruction process. SAEs have emerged as a relevant tool for concept-based interpretability in language [25–27] and multimodal models [28–33]. The number of latent features is set to greatly exceed the size of the input activations to encourage disentanglement. Latents are then assigned a human-interpretable label by inspecting their activation patterns across different samples, possibly using an LLM to generate such explanations [34]. However, this approach shifts interpretability to the explanation phase, which may generate oversimplified or broad labels [35]. Some hybrid approaches combine SAE-based decomposition and CBM interpretability in a single framework. DN-CBM [29] first decomposes CLIP visual embeddings into sparse representations and then assigns each feature the closest label from a given vocabulary. ConceptScope [32] exploits SAE-based decomposition to distinguish between target and bias concepts in the input. Recent works have questioned whether SAEs detect truly monosemantic features, such that each latent corresponds to a single, clearly defined concept [36]. Others question whether SAE features generalize across data scarcity, class imbalance, label noise, and covariate shift, or merely match the performance of simple baselines [37]. G-SAE [38] introduces a conditioning objective to encourage monosemanticity. Yet, even this approach requires ground-truth concept labels.

Concept-based interpretability in VLMs. Recent works [13, 14, 39] explore an interpretability-first approach by exploiting the linear representation hypothesis across modalities in CLIP-like models [11, 12]. These works first define a vocabulary of semantic concepts through language, enabling direct human-aligned interpretability, and then leverage a bottleneck-like approach to project the embeddings into the concept space and reconstruct them back. These works place strong emphasis on the faithfulness of the reconstructed representations, i.e., whether the reconstruction preserves the capabilities of the original embeddings on the downstream tasks. Even though they demonstrate the linearity of VLMs in concept space across modalities, they suffer from key limitations. The resulting mappings are either local, relying on instance-level decomposition [13, 39], which may result in shortcut learning, or leverage captioning models for concept detection [14].

GRACE learns a task-agnostic decomposition of visual embeddings that does not require concept-level supervision, leveraging the internal cross-modal structure of VLMs. Human-aligned interpretability of the decomposition is built into the model by design, as we force the model to reconstruct its features from a sparse set of predefined concepts, avoiding ambiguous post-hoc explanations. Our method integrates principles from SAE and CBM approaches to learn a general concept-based decomposition of the embedding space matching text-based and human-understandable concepts with visual inputs.

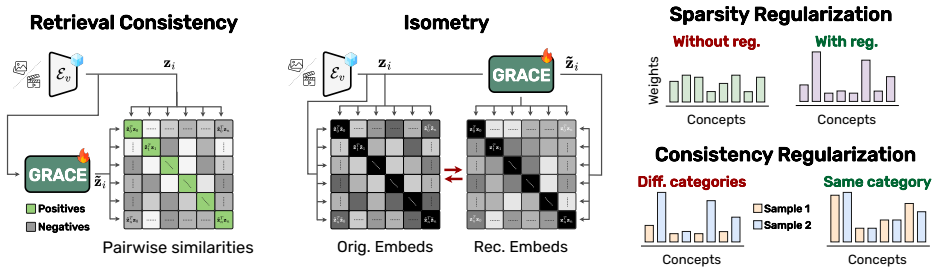


Figure 2: **Geometry-based reconstruction and regularization objectives in GRACE.** *Retrieval consistency* ensures that the reconstructed visual embeddings are close to their original counterparts, *Isometry* preserves the same pairwise similarities between the samples in the batch before and after the reconstruction, while *Sparsity* and *Consistency* regularization favor sparse reconstruction weights and consistency across different sample categories.

3 Methodology

We introduce GRACE, a concept-based framework that decomposes the visual embedding space of any vision-language model into a combination of human-interpretable textual concepts, while maintaining multi-modal alignment and preserving the geometry of the samples in the concept space, i.e., the pairwise relations between the visual samples should remain the same when mapping through the concept space. We address key limitations in prior work by extracting concept evidence directly from the visual inputs, thereby exposing the inherent concept-based structure of the input embeddings, and learning a general and task-agnostic decomposition that reasons about concept composition within and across different samples. Our method learns a concept decomposition that reflects strong semantic relationships rather than task- or sample-specific cues, enabling effective transfer of the learned concept structure across tasks, datasets, and domains. We illustrate GRACE in Figure 1.

Problem definition. We denote by $\mathcal{E}_v$ and $\mathcal{E}_t$ the vision and text encoders of a pretrained vision-language model, mapping the two modalities in a shared $\mathbb{R}^d$ space. We treat $\mathcal{E}_v$ as the *target model*, i.e., the model we aim to make interpretable, and indicate as $\mathbf{z}_i = \mathcal{E}_v(x_i)$ the visual embedding of any sample x_i. $\mathcal{C}$ is a large vocabulary of K textual concepts, representing individual semantic elements, e.g., simple actions or objects, that may appear in the visual input. We look for a mapping $h\colon \mathbb{R}^d \to \mathbb{R}^K$, such that $h \circ \mathcal{E}_v$ yields a decomposition encoding the relative contribution of the concepts in any given visual input. To ensure that the matched concepts represent relevant semantic information for the input, we introduce a mapping $r\colon \mathbb{R}^K \to \mathbb{R}^d$ that maps back to the original embeddings and minimize a reconstruction objective $\ell(\tilde{\mathbf{z}}_i, \mathbf{z}_i)$ between the original visual embeddings $\mathbf{z}_i$ and their reconstruction $\tilde{\mathbf{z}}_i = (r \circ h)(\mathbf{z}_i)$. The choice of ℓ is critical, as it determines *how* and *how much* of the original embeddings is preserved through the concept-based decomposition.

Intuition. For any given visual sample, we detect initial concept evidence by matching its visual embedding against the text embeddings of a predefined concept vocabulary through cross-modal cosine similarity. We then look for a mapping back to the original visual embeddings by learning the optimal reconstruction weights from the visual evidence. Prior works enforce strict ℓ_2 reconstruction [13], which is fundamentally limited by the modality gap and the different geometries of the two modalities [15, 40]. Instead, we opt for a relaxed reconstruction objective that preserves the geometry of the reconstructed visual embeddings in concept space, i.e., the pairwise similarity of features before and after reconstruction. This approach goes beyond instance-level reconstruction and encourages the model to better understand how concepts are composed within and across samples, improving compositional reasoning. We push the reconstruction weights to capture rich and meaningful semantic information from the input, while promoting (i) *sparsity*, i.e., samples should be reconstructed by few concepts, and (ii) *consistency*, i.e., reconstruction weights should be consistent across semantic categories.

3.1 Geometry-based reconstruction

Collecting visual concept evidence. We exploit the cross-modal *bag-of-words* nature of vision-language models [12, 13, 41] to detect individual concepts from the vocabulary in a given visual sample by measuring the cosine similarity between the visual embeddings $\mathbf{z}_i$ and the text embedding

of the concept $\mathbf{c}_j$. We refer to these similarities as *visual evidence* α_i, computed as $\alpha_{i,j} = \cos(\mathbf{z}_i, \mathbf{c}_j)$. Concept vocabularies can be obtained from multiple sources, including human supervision [17], LLMs [18, 42], or by mining frequent terms from captions. In this work, we adopt the mining strategy for simplicity and scalability, even though GRACE is robust to the selection of the concepts vocabulary. However, synonyms in the vocabulary can skew the visual evidence toward uniformly positive values, resulting in many detected concepts. To mitigate this phenomenon, we introduce a gating mechanism similar to GLU [43] that calibrates the relevance of each concept for the given input. We define the *calibrated evidence* $\tilde{\alpha}_i$ as:

$$\tilde{\alpha}_i = \alpha_i \odot \mathcal{H}(\mathbf{z}_i), \tag{1}$$

where $\mathcal{H}$ is an MLP with a sigmoid activation function on the output. Desirably, the calibrated evidence should preserve only a few concepts per sample, with a long tail of concepts with near-zero weight, directly improving interpretability and the semantic relevance of the matched concepts.

Visual embeddings reconstruction. We ensure that the calibrated evidence reflects meaningful concepts by using them to reconstruct the original embeddings. Prior works directly minimize the ℓ_2 distance between the reconstructed and original embeddings subject to a sparsity penalty on the reconstruction coefficients [13], e.g., $\min_{\mathbf{w}_i} ||\mathbf{z}_i - \mathbf{w}_i^T \mathbf{C}||_2^2 + ||\mathbf{w}_i||_1$, or learn a *surrogate* concept-based representation $\tilde{\mathbf{z}}_i$ that preserves the multi-modal alignment of the original model [14], e.g., $\cos(\tilde{\mathbf{z}}_i, \mathbf{t}_j) \approx \cos(\mathbf{z}_i, \mathbf{t}_j)$ for any possible combination (i, j) of visual and text samples in the batch. However, these approaches overlook the fact that the visual and text representations are not interchangeable and have different geometries [15, 40], which prevents a full reconstruction of the visual embeddings as a linear combination of text-encoded concepts. Instead, we relax such a strict reconstruction process using a simpler geometry reconstruction objective, which preserves the pairwise similarities between the visual samples in the batch when mapped to the concept space beyond instance-level reconstruction. Formally, we compute our *reconstructed visual embeddings* as:

$$\tilde{\mathbf{z}}_i = \sigma(\tilde{\alpha}_i^T / \tau)^T \mathbf{C}\mathbf{Q} = \mathbf{w}_i^T \mathbf{C}\mathbf{Q}, \tag{2}$$

where σ is the softmax function over the evidence, τ is a temperature parameter and $\mathbf{Q} \in \mathbb{R}^{d \times d}$ is a learnable orthogonal transformation, i.e., $\mathbf{Q}\mathbf{Q}^T = \mathbb{I}$, of the concept embeddings. The orthogonal transformation maps the concepts in a subspace compatible with the visual embeddings without further entangling them, i.e., the dot product before and after the projection is preserved $(\mathbf{Q}^T \mathbf{c}_i)^T (\mathbf{Q}^T \mathbf{c}_j) = \mathbf{c}_i^T \mathbf{c}_j$. We discuss the importance of this property for interpretability in Sec. F.

Geometry reconstruction. We define a dual *geometry reconstruction* objective that ensures: (i) *retrieval consistency*, such that the reconstructed embeddings are ranked correctly against the original embeddings, and (ii) *isometry*, enforcing that the pairwise cosine similarities of the reconstructed embeddings match the original representations. We illustrate these objectives in Figure 2. For retrieval consistency, we use a contrastive loss in which reconstructed embeddings are pushed closer to the original counterparts, maximizing their mutual information, and further apart from other samples in the batch. For isometry, we minimize the distance between the cross-similarity matrices of the original and reconstructed embeddings:

$$\mathcal{L}_{rec} = \underbrace{\sum_{\mathbf{z}_i \in \mathcal{B}} \left[-\log \frac{e^{\tilde{\mathbf{z}}_i^T \mathbf{z}_i / \tau}}{\sum_k e^{\tilde{\mathbf{z}}_i^T \mathbf{z}_k / \tau}} \right]}_{\text{Retrieval consistency}} + \underbrace{||\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^T - \mathbf{Z}\mathbf{Z}^T||_F^2}_{\text{Isometry}}, \tag{3}$$

where $\mathbf{Z}, \tilde{\mathbf{Z}} \in \mathbb{R}^{|\mathcal{B}| \times d}$ are the ℓ_2-normalized original and reconstructed visual embeddings in the batch. The *retrieval consistency* term enforces that, for any reconstructed embedding $\tilde{\mathbf{z}}_i$, its closest match among the original embeddings is $\mathbf{z}_i$, i.e., $\operatorname{argmax}_j \cos(\tilde{\mathbf{z}}_i, \mathbf{z}_j) = i$. The *isometry term* ensures that the cosine similarity between any two reconstructed samples is close to that of their original embeddings, i.e., $\cos(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_j) \approx \cos(\mathbf{z}_i, \mathbf{z}_j)$. Together, these two objectives preserve the unimodal pairwise relations of the original and reconstructed embeddings, capturing the global semantic structure of the model, without imposing strict instance-level reconstruction.

3.2 Sparsity and consistency properties

Our reconstruction objective does not impose any constraint on the reconstruction weights $\mathbf{w}_i$. Yet, we argue that a *good* concept-based decomposition should satisfy two key properties: *sparsity* and

consistency. Visual samples should be compressed into a few significant concepts, thereby improving interpretability, while promoting consistency in the reconstruction weights across different categories, e.g., food-related concepts are more likely to appear in images or videos set in a *kitchen*, but less likely in *sport* videos. We discuss in Sec. 4.1.1 that this property emerges naturally from the data.

Sparsity. We parametrize the reconstruction weights to follow a categorical distribution $\text{Cat}(\mathbf{w}_i)$. To explicitly control the sparsity of the reconstruction process, we assume $\mathbf{w}_i$ from Eq. 2 to be drawn from a symmetric Dirichlet distribution $\mathbf{w}_i \sim \text{Dir}(\omega)$, whose density is proportional to $\prod_j \mathbf{w}_{i,j}^{\omega-1}$ and ω is a concentration parameter in $(0, 1]$. For lower values of ω, sparse reconstruction weights are more likely, while higher values favor more uniform weights. The concentration parameter ω acts as a tunable hyperparameter that controls the sparsity of the reconstruction. The model is trained to minimize the negative log-likelihood of the reconstruction weights under the Dirichlet prior:

$$\mathcal{L}_{reg} = -\sum_{i=1}^{|\mathcal{B}|} \left[(\omega - 1) \sum_{k=1}^{K} \log \left(\mathbf{w}_{i,k} + \epsilon \right) \right], \tag{4}$$

where ϵ prevents instabilities as $\mathbf{w}_{i,k} \to 0$.

Consistency. We encourage the model to predict similar concepts for semantically similar samples, e.g., videos from the same location. This approach builds on prior work in disentangled representations which seek to isolate shared and unique latent factors across samples and modalities [44–46]. We encourage the distribution of the reconstruction weights $(\mathbf{w}_i, \mathbf{w}_j)$ to be similar for samples (i, j) with overlapping categories, e.g., videos captured in the same location, and dissimilar otherwise:

$$\mathcal{L}_{con} = \sum_{i=1}^{|\mathcal{B}|} \sum_{j=1}^{|\mathcal{B}|} \delta_{ij} \text{D}(\mathbf{w}_i || \mathbf{w}_j), \tag{5}$$

where δ_{ij} is $+1$ if samples i and j are from the same category and -1 otherwise. D is the Jensen-Shannon Divergence (JSD) [47] between the reconstruction weights of the two samples and is defined as $\text{D}(p||q) = \frac{1}{2}\left[\text{KL}(p||m) + \text{KL}(q||m)\right]$, with $m = \frac{1}{2}(p + q)$.

3.3 Training and inference

We train GRACE with a combination of these objectives, $\mathcal{L} = \mathcal{L}_{rec} + \lambda_r \mathcal{L}_{reg} + \lambda_c \mathcal{L}_{con}$, where λ_r and λ_c control the weight of the regularization terms, and are chosen to minimize the reconstruction objective on the training data. We first train GRACE on large-scale image and video datasets to learn a general task-agnostic concept decomposition, which can then be directly applied to unseen tasks and domains. At inference time, the reconstructed embeddings $\tilde{\mathbf{z}}_i$ can be used in place of the original embeddings, while the reconstruction weights $\mathbf{w}_i$ provide an interpretable bottleneck encoding the relative contribution of each concept.

4 Experiments

We consider a diverse set of image and video benchmarks to evaluate GRACE. With *Concept-based Similarity Estimation*, we measure how the reconstruction weights align with human perception, capturing semantic concepts from the input. With *classification* and *retrieval* tasks, we demonstrate that the interpretable decomposition from GRACE generalizes well across different tasks and domains. Together, these benchmarks show the connection between interpretability and generalization in VLMs.

4.1 Video Benchmarks

We train GRACE on EgoClip [57], a collection of 3.8M video-text pairs derived from Ego4D [58]. Then, we evaluate our method on text-to-video retrieval on **EgoMCQ** [57], action recognition on **Charades-Ego** [59] and video similarity estimation on **ConVisBench** [60], a task where the goal is to determine the semantic similarity of two given videos. Together, we use these benchmarks to demonstrate that our reconstructed embeddings preserve vision-language alignment and generalize across different tasks and domains, while capturing rich semantic information. We use the Perception-Encoder [56] backbone and mine the concept vocabulary from the EgoClip captions. For the consistency regularization term, we use the scenario labels from Ego4D.

Table 1: **Concept-based Similarity Estimation on ConVisBench.** Given a video pair, the model has to predict their semantic similarity across different axes, e.g., *Main Objects* or *Location*. Performance is measured by Spearman's ρ and Kendall's τ correlations against human judgment.

Model	Method	Main Action ρ / τ	Main Subjects ρ / τ	Main Objects ρ / τ	Location ρ / τ	Order of Actions ρ / τ	Average ρ / τ
Text-to-Text							
mPLUG-Owl3 [48]	SBERT	52.1 / 36.2	45.5 / 31.1	55.1 / 38.9	28.4 / 19.6	49.9 / 34.9	46.2 / 32.1
InternVL2.5 [49]	SBERT	39.4 / 27.4	31.2 / 21.4	45.0 / 31.1	29.9 / 20.5	42.0 / 29.5	37.5 / 26.0
Qwen2.5-VL (7B) [50]	SBERT	42.3 / 29.5	45.4 / 31.7	45.4 / 31.8	38.5 / 26.7	40.9 / 28.4	42.5 / 29.6
Qwen3.5 (9B) [51]	SBERT	36.5 / 25.3	37.1 / 25.8	42.9 / 30.0	34.8 / 23.9	37.0 / 25.6	37.6 / 26.1
Video-to-video							
VideoMAE [52]	Cosine Sim.	13.0 / 8.7	23.1 / 15.6	13.2 / 8.8	37.8 / 26.4	15.1 / 10.2	20.4 / 14.0
DINOv2 [53]	Cosine Sim.	33.3 / 22.6	40.9 / 28.6	37.4 / 25.8	57.4 / 41.7	34.6 / 24.0	40.7 / 28.5
Cross-modal							
CLIP [4]	CLIPScore	35.7 / 24.6	39.0 / 27.0	39.0 / 27.1	36.0 / 24.9	30.9 / 21.1	36.1 / 24.9
InternVideo [54]	CLIPScore	43.1 / 30.1	54.8 / 38.7	54.5 / 38.5	40.9 / 28.4	40.2 / 28.1	46.7 / 32.8
LLaVA-OV [55]	VQAScore	51.1 / 36.1	55.8 / 39.6	58.3 / 41.4	46.5 / 32.3	48.1 / 33.9	52.0 / 36.6
Perception-Encoder [56]	CLIPScore	51.1 / 36.0	57.9 / 41.3	57.5 / 41.2	44.4 / 31.3	47.6 / 33.5	51.7 / 36.7
Concept-based methods (using Perception-Encoder backbone)							
SpLiCE [13]	CLIPScore	50.6 / 35.8	41.0 / 28.0	53.0 / 37.4	31.9 / 21.7	47.0 / 32.8	44.7 / 31.1
EXPLAIN-R [14]	CLIPScore	41.1 / 28.8	29.9 / 20.2	44.9 / 31.3	29.9 / 20.5	40.0 / 27.7	37.2 / 25.7
GRACE	CLIPScore	43.7 / 30.6	47.0 / 32.9	44.3 / 31.2	32.0 / 22.1	40.3 / 27.9	41.5 / 28.9
SpLiCE [13]	Weights (L1)	51.3 / 35.4	47.7 / 32.8	**62.1 / 44.1**	43.7 / 29.9	49.5 / 34.5	50.9 / 35.4
GRACE	Weights (JSD)	**54.0 / 38.2**	**60.9 / 43.5**	58.6 / 42.2	**56.3 / 40.1**	**51.6 / 36.6**	**56.3 / 40.1**

Table 2: **Accuracy and Faithfulness on video retrieval (EgoMCQ) and action recognition (Charades) tasks.** GRACE and EXPLAIN-R are trained on EgoClip and evaluated in *zero-shot*.

Method	Zero-Shot	EgoMCQ			Charades	
		Intra Acc.	Inter Acc.	Faith.	Acc.	Faith.
Perception Encoder [56]	✓	37.64	84.88	N/A	31.30	N/A
SpLiCE [13]	✓	34.60	79.04	70.36	**26.94**	29.70
EXPLAIN-R [14]	✓	30.10	76.17	63.70	21.73	25.92
GRACE	✓	35.43	81.35	77.86	25.99	35.15
GRACE (+*distill.*)	✓	**35.50**	**82.94**	**80.48**	25.97	**39.53**

Concept-based Similarity Estimation. In ConVisBench, the goal is to predict whether two videos are similar with respect to a predefined semantic axis, e.g., the *location* in which the video was recorded or the *main subject*. Performance is measured as Spearman's and Kendall's correlation between the predicted similarity score and the human similarity annotation for each pair. Notably, all the previous methods evaluated on this benchmark are not explicitly concept-based. *Text-to-Text* methods measure the similarity of the video captions with SBERT [61], while *Video-to-Video* methods use the cosine similarity between global video embeddings. *Cross-modal* methods evaluate similarity using CLIPScore [62]. For GRACE, we compute similarity using either CLIPScore on the reconstructed embeddings or the distance between the two samples' reconstruction weights. Notably, the latter approach achieves the highest correlation with human judgment, demonstrating that the reconstruction weights capture semantically meaningful concepts in the videos, as shown in Table 1. GRACE also outperforms the non-interpretable baseline, showing that the concept bottleneck isolates semantically relevant information while filtering out irrelevant noise.

Video Retrieval and Action Recognition. EgoMCQ consists of 39k multiple-choice *text-to-video* questions where, given a textual query, the model has to select the correct video clip among five candidates. We report Inter-Video accuracy, where clips are from different videos, and Intra-Video accuracy, where they are from the same video. We use cosine similarity between the reconstructed visual embeddings and the text embedding of the query to find the best match. **Charades-Ego** is a dataset of 7.8K videos with annotated actions from a taxonomy of 157 classes. Performance is evaluated using mAP. We define a zero-shot classifier using the text embeddings of the action labels. For both benchmarks, we also evaluate faithfulness, i.e., measuring how well the reconstructed embeddings preserve the predictions of the original embeddings, as defined by [14]. As shown in Table 2, GRACE consistently improves on faithfulness, despite not relying on instance-level optimization or caption-based evidence as in SpLiCE and EXPLAIN-R, respectively. For a fair comparison, we also integrate the multi-modal distillation objective from [14]. This addition provides marginal gains, showing that our uni-modal reconstruction is sufficient to preserve multi-modal alignment.

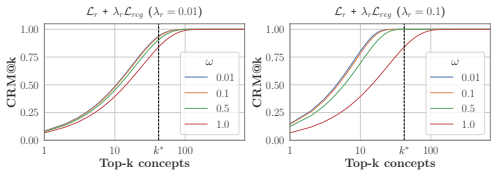


Figure 3: **Cumulative Reconstruction Mass (CRM@k) for different sparsity regimes.** k^* is the average number of concepts predicted by SpLiCE.

Method	mAP (↑)	Disent. (↑)	Complet. (↑)	L_0 Norm (↓)
SpLiCE [13]	86.1	8.9	18.3	633.2
GRACE ($\mathcal{L}_{rec}$)	72.5	53.4	93.8	28.5
GRACE (+$\mathcal{L}_{con}$)	73.3	53.9	93.8	27.8

Table 3: **Evaluation of the concepts disentanglement with DCI metrics [63].** We use linear probing to predict the video scenario, e.g., *cooking* or *sports*, from just the reconstruction weights.

Table 4: **Analysis of the concept vocabulary.** Few well-selected concepts are sufficient to retain most of the reconstruction quality of the more redundant larger vocabulary (left). The GPT-based vocabulary achieves good reconstruction despite minimal human effort in concepts definition (right).

Vocabulary	Size	ConVisBench (ρ / τ)	CRM@50
Full Voc.	2139	56.3 / 40.1	0.73
Pruned ($\tau = 0.5$)	79	42.1 / 29.5	1.00
Pruned ($\tau = 0.9$)	343	51.4 / 36.4	0.99

Vocabulary	Size	EgoMCQ	ConVisBench (ρ / τ)
EgoClip	700	35.43 / 81.35	51.0 / 36.2
Full Voc.	2139	36.06 / 82.08	56.3 / 40.1
GPT-5.4	414	34.76 / 80.28	52.3 / 37.1

4.1.1 Analyses of the concept decomposition

Sparsity of the decomposition. We define the *cumulative reconstruction mass at k (CRM@k)* as the sum of the top-k largest reconstruction weights, and report its trend in Figure 3 as the sparsity parameter ω varies. Low values of ω reduce the number of non-zero reconstruction weights. GRACE achieves near-perfect reconstruction with few concepts, thanks to its general reconstruction objective.

Disentanglement of the reconstruction weights. We adapt the DCI metrics [63] to our setting, by training a set of sparse logistic regression classifiers to predict the video scenario, i.e., a coarse label describing the main activity in the video, from just the reconstruction weights, and evaluate their behavior in terms of (i) *sparsity*, i.e., how many concepts are required for the prediction (L_0 Norm); (ii) *disentanglement*, measuring the separation of concepts across scenarios; (iii) *completeness*, reflecting whether each scenario is explained by few concepts; and (iv) *informativeness*, corresponding to prediction accuracy (mAP). As shown in Table 3, SpLiCE achieves good prediction accuracy but poor disentanglement and sparsity, indicating substantial overlap in the concepts matched across different scenarios. In contrast, GRACE is slightly less accurate but exhibits stronger disentanglement, indicating that scenario-level disentanglement already emerges from the data. More details in Sec. B.

Concept redundancy and sufficiency. We consider the Singular Value Decomposition of the concept matrix to estimate redundancy in the vocabulary. Let s_i denote the singular values of $\mathbf{C}$ in decreasing order. We define k as the smallest integer such that $\sum_{i=1}^{k} s_i^2 / \sum_{i=1}^{K} s_i^2 \geq \tau$, where τ is a threshold hyperparameter that controls the amount of retained concept variance, and thus the completeness of the selected concepts. Then, we perform a QR factorization with column pivoting of the concept matrix, iteratively selecting the concept with the largest residual contribution. The retained concept embeddings form a set of *sufficient* concepts, based on the choice of τ. The concept matrix can be heavily compressed into few concepts, without significant performance degradation (Table 4-left).

Construction of the concept vocabulary. We show in Table 4-right the impact on performance of different concept vocabularies, including one obtained by prompting GPT-5.4 for an exhaustive list of plausible text-based concepts that could appear in videos of human activities. This approach assumes little knowledge about the videos and minimal human effort, yet it achieves good performance.

4.2 Image benchmarks

We evaluate our method on several tasks, including **COCO** [64] and **Flickr30k** [65], **ImageNet** [66], **Flower102** [67] and **Food101** [68], measuring task performance and faithfulness. We train GRACE on ImageNet and evaluate it on these tasks without fine-tuning, using the CLIP ViT-B/32 backbone. We adapt EXPLAIN-R to the same training and validation protocol, and report its original performance when dataset-specific training is available. We use a vocabulary mined from ImageNet captions, and compare different vocabularies in Sec. G. As for video tasks, we also consider a variant of GRACE that leverages captions during training to improve multi-modal alignment.

Table 5: **Performance and Faithfulness on Image-to-Text (I→T) and Text-to-Image (T→I) retrieval tasks (Flickr30k and COCO)**. We report retrieval metrics and faithfulness.

Method	Concepts Vocab.	Zero Shot	Flickr30k I→T			Flickr30k T→I			COCO I→T			COCO T→I		
			R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Task Metrics														
CLIP (ViT-B) [4]	N/A	✓	67.17	88.90	93.80	63.60	86.78	92.26	51.90	81.26	90.24	47.48	77.10	87.08
SpLiCE [13]	LAION	✓	36.43	65.03	75.36	36.83	63.50	74.20	32.26	62.94	76.44	29.92	59.22	73.00
EXPLAIN-R [14]	ImageNet	✓	45.67	73.97	83.30	**49.30**	**78.45**	**87.10**	**44.22**	**75.28**	**87.30**	**47.42**	**78.04**	**88.12**
GRACE	ImageNet	✓	48.48	76.17	84.50	42.84	69.97	79.26	41.80	71.98	83.66	33.56	63.34	75.92
GRACE (+*distill.*)	ImageNet	✓	**52.30**	**79.30**	**86.97**	46.87	74.43	83.41	42.78	72.84	84.46	36.34	65.82	78.66
EXPLAIN-R [14]	Target	✗	74.53	92.58	96.24	73.60	92.67	96.30	58.03	85.68	93.19	56.30	84.98	93.09
Faithfulness														
SpLiCE [13]	LAION	✓	40.25	70.41	80.65	40.50	69.56	80.04	38.12	72.40	86.00	38.26	72.74	85.66
EXPLAIN-R [14]	ImageNet	✓	44.48	72.62	81.42	42.52	70.02	79.44	46.72	77.82	88.32	39.68	70.10	83.52
GRACE	ImageNet	✓	55.24	84.27	91.39	51.77	81.56	89.64	54.78	86.42	94.38	47.36	81.60	91.42
GRACE (+*distill.*)	ImageNet	✓	**62.72**	**90.19**	**95.40**	**55.94**	**85.65**	**92.85**	**58.16**	**89.80**	**96.22**	**49.76**	**84.24**	**93.66**
EXPLAIN-R [14]	Target	✗	66.37	86.32	91.26	60.31	82.79	89.42	61.99	89.31	95.51	55.03	86.64	94.46

Table 6: **Accuracy and faithfulness on image classification.** For zero-shot evaluations, we train on ImageNet and evaluate the quality of the reconstructed embeddings on Flowers102 and Food101.

Method	Concepts Vocab.	Zero Shot	ImageNet		Flowers102		Food101	
			Acc.	Faith.	Acc.	Faith.	Acc.	Faith.
CLIP (ViT-B) [4]	N/A	✓	61.92	N/A	63.39	N/A	80.88	N/A
SpLiCE [13]	LAION	✓	42.91	51.33	-	26.12	-	51.80
EXPLAIN-R [14]	ImageNet	✓	**50.77**	50.19	15.51	16.59	45.86	45.81
GRACE	ImageNet	✓	48.04	60.15	**32.14**	**41.78**	62.67	66.83
GRACE (+*distill.*)	ImageNet	✓	47.84	**60.68**	28.82	41.18	**63.24**	**67.49**
EXPLAIN-R [14]	Target-specific	✗	52.37	53.38	-	37.73	-	63.62

Image-Text Retrieval. GRACE produces significantly higher-quality reconstructed embeddings, while being competitive with previous methods on the task metrics (Table 5). Our reconstructed embeddings are much closer to the original embeddings in terms of multi-modal alignment (*higher faithfulness*), even though the model is slightly less confident in the relative ranking of the image-text pairs, especially at lower recall thresholds. We attribute this gap to the use of caption-derived concept evidence in EXPLAIN-R, which simplifies vision–language alignment.

Image Classification. Our approach consistently outperforms SpLiCE and EXPLAIN-R in both accuracy and faithfulness, despite not having access to caption-based evidence and not using instance-level optimization (Table 6). EXPLAIN-R suffers substantial performance degradation in the zero-shot setting, whereas the geometry-preserving reconstruction process of GRACE yields a more transferable concept decomposition that generalizes across domains. This method leverages an external captioning model to extract concept evidence, which limits its applicability. In contrast, our method avoids this assumption by learning the most relevant concepts directly from visual features.

4.2.1 Robustness to severe concept shift and bias analysis

We define five different folds over the ImageNet classes, such that the overlap between the train and test classes is minimized. Completely disjoint concept sets are not possible, as some common concepts are ubiquitous and occur frequently in many classes. We train GRACE on the train set of each fold and evaluate zero-shot on the corresponding test set, comparing with the model trained on all the classes. Even though the reconstructed embeddings are tested on zero-shot classes, GRACE maintains performance very close to the model trained on all classes, as shown in Figure 4. These experiments demonstrate the robustness of GRACE to severe shifts in the concept distribution across training and evaluation splits. A GRACE model trained on a set of classes with a given class-concept distribution can still produce reasonable decompositions when tested on classes with a very significant shift in their class-concept distributions.

Bias analysis. We analyze the average top-5 concept activations for a subset of the ImageNet classes, showing some interesting patterns in Table 7. Locations, e.g., "*chile*" and "*indonesia*", often appear in the volcano class, showing a strong entity-location correlation. Similarly, "*tokyo*" is the strongest

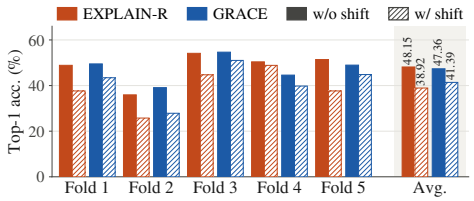


Figure 4: **Transferability of the decomposition.** We train GRACE on custom ImageNet folds with minimal concept overlap, demonstrating robust transfer under severe concept-distribution shifts.

Label	Most activated Concepts
volcano	volcano, **chile**, view towards, **indonesia**, ...
rickshaw	**tokyo**, flower delivery, bike, fast delivery, ...
cassette player	**casio g**, radio station, cassette, ...
digital clock	clock, **casio g**, temperature, heartbeat, ...
digital watch	**casio g**, chronograph, clock, time, ...
oscilloscope	ultrasound, heartbeat, synthesis, **casio g**, ...

Table 7: **Bias analysis of the decomposition.** We report the most activated concepts for some representative ImageNet classes, showing interesting patterns, e.g., strong *entity-location* correlation.

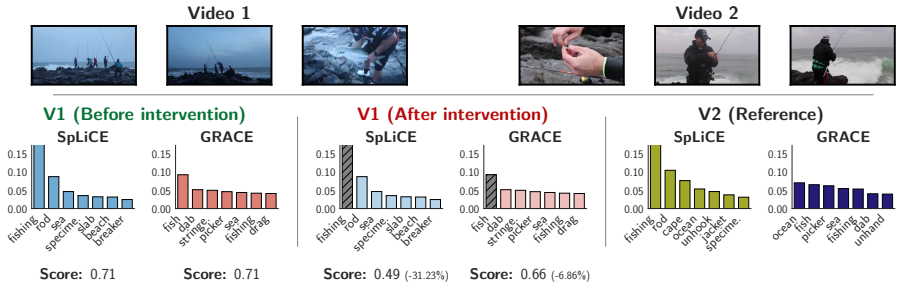


Figure 5: **Intervention on the reconstruction weights on ConVisBench [60].** We artificially zero out a concept from the predicted reconstruction weights and estimate the similarity *before* and *after* the intervention. SpLiCE focuses on a few concepts: when one of these is artificially removed, the similarity score drops significantly, even though the two videos remain similar in terms of subject, location, and objects. GRACE captures the relevant concepts more uniformly.

activated concept in the rickshaw class. "*casio g*" is often activated for devices featuring digital displays, even though these objects are not necessarily watches.

4.3 Qualitative results

We compare the concepts activated by SpLiCE and GRACE on ConVisBench videos. Both methods recover meaningful concepts for each video, but SpLiCE tends to concentrate most of the weights in the very first few concepts, which is a direct consequence of its instance-level optimization. Figure 5 shows that this behavior makes SpLiCE sensitive to concept interventions, i.e., when some matched concepts are artificially removed to simulate prediction errors. In this setting, we consider a pair of videos from ConVisBench and compute their reconstruction weights. We zero out the weight of a predicted concept, rescale the remaining ones, and measure similarity before and after the intervention. GRACE is only slightly affected by the intervention, as the similarity score between the two videos remains almost the same, while SpLiCE exhibits a more significant drop, even though they remain similar in terms of location, main subject, and objects. More details in Sec. I.

5 Conclusions

We explore how to make existing VLMs more interpretable. We argue that these models already encode a latent concept-based structure, though it is not explicitly exposed. Therefore, we introduce GRACE, a novel framework that captures this structure to learn a general decomposition of visual embeddings that can generalize across tasks and domains. Our approach overcomes many limitations of similar previous approaches, as it does not rely on per-sample optimization or external concept evidence. Instead, we exploit a reconstruction process that preserves the pairwise geometric relations among samples and encourages a selective and efficient use of the concept vocabulary. We show that GRACE decomposes visual embeddings into highly semantic concepts, which are well aligned with human judgement. The decomposition transfers well to unseen domains, is robust to severe shift in the concept distribution, and can be used to debug dataset biases.

Acknowledgments

This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them. We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support. We also acknowledge the support of the European Laboratory for Learning and Intelligent Systems (ELLIS). Simone's work was carried out during a research visit supported by the "ELSA – European Lighthouse on Secure and Safe AI" Mobility Grant.

References

[1] Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988. 1

[2] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987. 1

[3] Thaddäus Wiedemer, Prasanna Mayilvahanan, Matthias Bethge, and Wieland Brendel. Compositional generalization from first principles. In *Advances in Neural Information Processing Systems*, 2023. 2

[4] 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*, 2021. 2, 7, 9, 17

[5] Elias Kempf, Simon Schrodi, Max Argus, and Thomas Brox. When and how does CLIP enable domain and compositional generalization? In *International Conference on Machine Learning*, 2025. 2

[6] Prasanna Mayilvahanan, Thaddäus Wiedemer, Evgenia Rusak, Matthias Bethge, and Wieland Brendel. Does CLIP's generalization performance mainly stem from high train-test similarity? In *International Conference on Learning Representations*, 2024. 2

[7] Jongseo Lee, Wooil Lee, Gyeong-Moon Park, Seong Tae Kim, and Jinwoo Choi. Disentangled concepts speak louder than words: Explainable video action recognition. In *Advances in Neural Information Processing Systems*, 2025. 2, 3

[8] Qihang Fang, Chengcheng Tang, Bugra Tekin, Shugao Ma, and Yanchao Yang. HuMoCon: Concept discovery for human motion understanding. In *Computer Vision and Pattern Recognition*, 2025.

[9] Patrick Knab, Sascha Marton, Philipp J Schubert, Drago Guggiana, and Christian Bartelt. Concepts in motion: Temporal concept bottleneck model for interpretable video classification. In *2nd Workshop on Compositional Learning: Safety, Interpretability, and Agents*, 2026. 2, 3

[10] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. 2

[11] Isabel Papadimitriou, Huangyuan Su, Thomas Fel, Sham M. Kakade, and Stephanie Gil. Interpreting the linear structure of vision-language model embedding spaces. In *Second Conference on Language Modeling*, 2025. 2, 3

[12] Darina Koishigarina, Arnas Uselis, and Seong Joon Oh. CLIP behaves like a bag-of-words model cross-modally but not uni-modally. In *International Conference on Learning Representations*, 2026. 2, 3, 4

[13] Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio Calmon, and Himabindu Lakkaraju. Interpreting CLIP with sparse linear concept embeddings (SpLiCE). In *Advances in Neural Information Processing Systems*, 2024. 2, 3, 4, 5, 7, 8, 9, 17

[14] Nhat Hoang-Xuan, Xiyuan Wei, Wanli Xing, Tianbao Yang, and My T. Thai. Advancing interpretability of CLIP representations with concept surrogate model. In *Advances in Neural Information Processing Systems*, 2025. 2, 3, 5, 7, 9, 17, 19, 21

[15] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. CyCLIP: Cyclic contrastive language-image pretraining. In *Advances in Neural Information Processing Systems*, 2022. 2, 4, 5

[16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023. 2, 17, 19

[17] 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*, 2020. 3, 5

[18] 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 *Computer Vision and Pattern Recognition*, 2023. 3, 5

[19] Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *International Conference on Learning Representations*, 2023. 3

[20] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *International Conference on Learning Representations*, 2023. 3

[21] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning*, 2024. 3

[22] Matthew Trager, Pramuditha Perera, Luca Zancato, Alessandro Achille, Parminder Bhatia, and Stefano Soatto. Linear spaces of meanings: compositional structures in vision-language models. In *International Conference on Computer Vision*, 2023.

[23] Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Language models implement simple word2vec-style vector arithmetic. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2024.

[24] Lukas Muttenthaler, Lorenz Linhardt, Jonas Dippel, Robert A Vandermeulen, Katherine Hermann, Andrew Lampinen, and Simon Kornblith. Improving neural network representations using human similarity judgments. In *Advances in Neural Information Processing Systems*, 2023. 3

[25] Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024. 3

[26] Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. In *International Conference on Machine Learning*, 2025.

[27] Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, 2025. 3

[28] Dahye Kim, Xavier Thomas, and Deepti Ghadiyaram. Revelio: Interpreting and leveraging semantic information in diffusion models. In *International Conference on Computer Vision*, 2025. 3

[29] Sukrut Rao, Sweta Mahajan, Moritz Böhle, and Bernt Schiele. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In *European Conference on Computer Vision*, 2024. 3

[30] Harrish Thasarathan, Julian Forsyth, Thomas Fel, Matthew Kowal, and Konstantinos G. Derpanis. Universal sparse autoencoders: Interpretable cross-model concept alignment. In *International Conference on Machine Learning*, 2025.

[31] Thomas Fel, Ekdeep Singh Lubana, Jacob S. Prince, Matthew Kowal, Victor Boutin, Isabel Papadimitriou, Binxu Wang, Martin Wattenberg, Demba E. Ba, and Talia Konkle. Archetypal SAE: Adaptive and stable dictionary learning for concept extraction in large vision models. In *International Conference on Machine Learning*, 2025.

[32] Jinho Choi, Hyesu Lim, Steffen Schneider, and Jaegul Choo. ConceptScope: Characterizing dataset bias via disentangled visual concepts. In *Advances in Neural Information Processing Systems*, 2025. 3, 19

[33] Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. In *Advances in Neural Information Processing Systems*, 2025. 3

[34] Gonçalo Santos Paulo, Alex Troy Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models. In *International Conference on Machine Learning*, 2025. 3

[35] George Ma, Samuel Pfrommer, and Somayeh Sojoudi. Revising and falsifying sparse autoencoder feature explanations. In *Advances in Neural Information Processing Systems*, 2025. 3

[36] Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. In *International Conference on Learning Representations*, 2026. 3

[37] Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. Are sparse autoencoders useful? a case study in sparse probing. In *International Conference on Machine Learning*, 2025. 3

[38] Ruben Härle, Felix Friedrich, Manuel Brack, Björn Deiseroth, Stephan Waeldchen, Patrick Schramowski, and Kristian Kersting. Measuring and guiding monosemanticity. In *Advances in Neural Information Processing Systems*, 2025. 3

[39] Shin'ya Yamaguchi, Kosuke Nishida, Daiki Chijiwa, and Yasutoshi Ida. Zero-shot concept bottleneck models. In *IEEE International Conference on Multimedia and Expo*, 2026. 3

[40] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*, 2022. 4, 5

[41] 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 *International Conference on Learning Representations*, 2023. 4

[42] Andong Tan, Fengtao Zhou, and Hao Chen. Explain via any concept: Concept bottleneck model with open vocabulary concepts. In *European Conference on Computer Vision*, 2024. 5

[43] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International Conference on Machine Learning*, 2017. 5

[44] Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, 2021. 6

[45] Marco Fumero, Florian Wenzel, Luca Zancato, Alessandro Achille, Emanuele Rodolà, Stefano Soatto, Bernhard Schölkopf, and Francesco Locatello. Leveraging sparse and shared feature activations for disentangled representation learning. In *Advances in Neural Information Processing Systems*, 2023.

[46] Paul Pu Liang, Zihao Deng, Martin Q Ma, James Y Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. In *Advances in Neural Information Processing Systems*, 2023. 6

[47] J Lin. Divergence measures based on the Shannon entropy. *IEEE transactions on information theory*, 37 (1):145–151, 1991. 6

[48] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mPLUG-Owl3: Towards long image-sequence understanding in multi-modal large language models. In *International Conference on Learning Representations*, 2025. 7

[49] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 7

[50] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. 7

[51] Qwen Team. Qwen3.5: Towards native multimodal agents, 2026. URL `https://qwen.ai/blog?id=qwen3.5`. 7

[52] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022. 7

[53] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 7

[54] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. InternVideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 7

[55] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025. 7

[56] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Abdul Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Dollar, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In *Advances in Neural Information Processing Systems*, 2025. 6, 7

[57] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. In *Advances in Neural Information Processing Systems*, 2022. 6, 16, 17, 20, 21

[58] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *Computer Vision and Pattern Recognition*, 2022. 6, 16

[59] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-Ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018. 6, 17

[60] Benedetta Liberatori, Alessandro Conti, Lorenzo Vaquero, Yiming Wang, Elisa Ricci, and Paolo Rota. ConViS-Bench: Estimating video similarity through semantic concepts. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. 6, 10, 17, 20, 23, 24

[61] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. 7

[62] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021. 7, 17

[63] Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018. 8, 18

[64] 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 *European Conference on Computer Vision*, 2014. 8, 18

[65] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *International Conference on Computer Vision*, 2015. 8, 18

[66] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, 2009. 8, 16, 17, 21

[67] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian conference on computer vision, graphics & image processing*. IEEE, 2008. 8, 17

[68] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 8, 17

[69] Cian Eastwood, Andrei Liviu Nicolicioiu, Julius Von Kügelgen, Armin Kekić, Frederik Träuble, Andrea Dittadi, and Bernhard Schölkopf. DCI-ES: An extended disentanglement framework with connections to identifiability. In *International Conference on Learning Representations*, 2023. 18

[70] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. LLaVA-NeXT-Interleave: Tackling multi-image, video, and 3D in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 19

[71] Bingyi Cao, Koert Chen, Kevis-Kokitsi Maninis, Kaifeng Chen, Arjun Karpur, Ye Xia, Sahil Dua, Tanmaya Dabral, Guangxing Han, Bohyung Han, Joshua Ainslie, Alex Bewley, Mithun Jacob, René Wagner, Washington Ramos, Krzysztof Choromanski, Mojtaba Seyedhosseini, Howard Zhou, and Andre Araujo. TIPSv2: Advancing vision-language pretraining with enhanced patch-text alignment. In *Computer Vision and Pattern Recognition*, 2026. 19

[72] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Computer Vision and Pattern Recognition*, 2017. 19

[73] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, 2022. 21

[74] Chenming Shang, Shiji Zhou, Hengyuan Zhang, Xinzhe Ni, Yujiu Yang, and Yuwang Wang. Incremental residual concept bottleneck models. In *Computer Vision and Pattern Recognition*, 2024. 23

Appendix

This appendix is organized as follows:

- **Implementation details**: In Sec. A, we provide additional implementation details about GRACE and the adaptation of previous works to our setting.
- **Disentanglement in reconstruction**: In Sec. B, we present a more complete description of the metrics used to evaluate concepts disentanglement in GRACE.
- **Concepts redundancy**: In Sec. C, we evaluate the redundancy in the concept decompositions produced by our approach.
- **Visual grounding**: In Sec. D, we replace the image backbone with TIPSv2 end evaluate spatial grounding of the concept decomposition.
- **Captions integration in GRACE**: In Sec. E, we describe the integration of a multi-modal distillation loss that can improve the multi-modal alignment of our method when captions are available during training.
- **Concepts projection in the reconstruction**: In Sec. F, we analyze how GRACE encourages semantic alignment of the predicted reconstruction weights with human perception.
- **Robustness to the choice of the concept vocabulary**: In Sec. G, we compare the use of different concept vocabularies in image and video benchmarks, showing the robustness of our method to different concepts sources.
- **Visual evidence and captions-based evidence**: In Sec. H, we compare two strategies to identify concepts in a visual input, i.e., *visual evidence*, based on multi-modal comparison between the input and the embeddings of a given concept vocabulary, as in GRACE, and *captions-based evidence*, which leverages external captioning models to identify concepts.
- **Additional qualitative results**: In Sec. I, we present more qualitative results showing the reconstruction weights predicted by our method and its robustness to concepts interventions.
- **Limitations and future works**: In Sec. J, we describe the main limitations of this work and propose some possible future research directions.

A Implementation details

A.1 Training and evaluation protocol

For video benchmarks, we train GRACE for 5 epochs on EgoClip [57], using a batch size of 128 and a learning rate of 0.0001. For image benchmarks, we train GRACE for 25 epochs on ImageNet [66], using batch size 512 and learning rate 0.0001. During training, we select the model with the lowest reconstruction loss ($\mathcal{L}_{rec}$) on the training set. For experiments that incorporate captions, we select the best model according to the faithfulness of the vision-language alignment on the training set. We use $\tau = 0.05$ for the temperature parameter in the reconstruction equation (Eq. 2). We train GRACE on NVIDIA RTX 4090 GPUs.

A.2 Video benchmarks

We train GRACE on the large-scale Ego4D dataset [58], covering a large range of different human daily activities over 3.6k recording hours. Specifically, we use the EgoClip [57] subset, which consists of 3.8M short video clips paired with human-annotated captions. Videos are annotated with a *scenario* label that identifies the main activities appearing in the video, such as *cooking* or *playing sports*. For the consistency loss $\mathcal{L}_{con}$, we use the scenario label to group together samples we expect to activate similar concepts. Then, we evaluate GRACE on different benchmarks in zero-shot to test the multi-modal alignment of the reconstructed visual embeddings and the quality and interpretability of the reconstruction weights.

For video benchmarks, we use two concept vocabularies: a *small* vocabulary, built from verbs and nouns in EgoClip captions grouped via the official fixed taxonomy provided by the authors (700 concepts), and a *large* vocabulary, consisting of the top 25% most frequent verbs and nouns across all

captions (approx. 2.1K concepts). In all experiments, we use the same vocabulary at both training and inference time.

EgoMCQ [57] is a multiple-choice text-to-video retrieval benchmark, designed to evaluate video-text alignment. It consists of 39K questions, each comprising a textual query and five candidate video clips, only one of which is the correct match. The benchmark includes two evaluation splits: (i) an *Inter* split, where candidate clips are drawn from different videos, and (ii) an *Intra* split, where all candidates are sampled from the same video. For retrieval, we compute cosine similarity between the text embedding of the query and the reconstructed visual embeddings from GRACE. Performance is evaluated using accuracy.

Charades-Ego [59] is a large-scale dataset of first and third person recordings, covering around 68 hours of daily human activities. We evaluate our method on the test split of the Action Recognition benchmark, covering 846 short videos annotated with one or more activity labels from a taxonomy of 157 classes. For this task, we define a zero-shot classifier by extracting the text embeddings of the class labels and computing their similarity with respect to the reconstructed visual embeddings. Performance is measured with mean Average Precision (mAP).

ConVisBench [60] is a dataset consisting of 610 video pairs sampled from 543 different videos. Each pair is annotated according to human-judged similarity across five different macro categories: *Main Action*, *Main Subjects*, *Main Objects*, *Location* and *Order of Actions*. Videos are also annotated with short captions. Similarity is collected by showing the video pairs to different annotators and asking them to evaluate the similarity of the two videos on a 1-to-5 scale. We evaluate GRACE on the *Concept-based Video Similarity Estimation* task, where the goal is to predict a similarity score based on the semantic similarity between the two videos, which is compared with human judgment with the Spearman's rank correlation ρ and Kendall's τ. These metrics evaluate how the predictions correlate with the similarity scores collected from human annotators. We compute the similarity score for a video pair using the CLIPScore [62], i.e., we compute the cosine similarity between the reconstructed visual embedding of the first video and the text embedding of the second video's caption, and vice versa, and then average the two scores. Moreover, we also compute the distance between the reconstruction weights of the visual embeddings of the two videos, as predicted by GRACE. This approach allows us to directly evaluate the expressivity of the reconstruction weights, and how they align with human interpretable semantic similarities or differences between the two videos.

A.2.1 Adaptation of SpLiCE and EXPLAIN-R to video benchmarks

SpLiCE [13] and EXPLAIN-R [14] were originally proposed for image benchmarks, and we adapt them to work on video benchmarks with some changes. For **SpLiCE**, we tune the ℓ_1 penalty parameter λ to achieve a reasonable average sparsity in the reconstruction weights. Specifically, we choose $\lambda = 0.1$ when using the *small* concepts vocabulary on EgoMCQ and Charades, which results in approximately 42.2 and 37.9 non-zero weights on the two datasets, respectively. On ConVisBench, we set $\lambda = 0.15$. For **EXPLAIN-R**, we uniformly sample 8 frames from each video clip and extract frame-level captions using BLIP2 [16]. This process takes approximately 3k GPU hours on EgoClip. Following the original protocol, we then compute the concept encoding by extracting lemmas that appear in at least three captions and filtering out those that occur in more than 10% of the total captions, obtaining a vocabulary of 10990 concepts.

A.3 Image benchmarks

We train GRACE on ImageNet [66] with a CLIP ViT-B/32 backbone [4]. Following the protocol introduced in EXPLAIN-R [14], we generate 10 captions per image using BLIP2 [16] and build the concept vocabulary by filtering out both rare and overly frequent concepts extracted from the captions. For $\mathcal{L}_{con}$, we use the image labels as category label, encouraging the model to learn similar reconstruction weights for visual samples belonging to the same class. The visual and text encoders are kept frozen, and we pre-extract the visual embeddings before training to significantly speed up the training process.

We evaluate GRACE on different Image Classification and Image-Text Retrieval tasks without additional fine-tuning, following the evaluation splits defined by EXPLAIN-R. For Image Classification, we consider **ImageNet** [66] (validation split), **Food101** [68] (test split) and **Flowers102** [67] (test split). For these tasks, we define a zero-shot classifier using the ground truth labels of each dataset,

Table 8: **Analysis of concepts redundancy.** Representative macro-clusters obtained with K-Means on the text embeddings of the LAION concept vocabulary (left). Median number of distinct macro-clusters among the top-25 concepts per class, with and without the calibration module (right).

Macro cluster	Size	Concepts
Kitchen	83	*kitchen*, *kitchen design*, *small kitchen*, *kitchen pantry*, ...
Staring	101	*staring*, *gazing*, *focused*, *observing*, *admiring*, ...
Glass window	74	*glass window*, *glass door*, *window*, *old window*, ...

Model	# Clusters
GRACE	**16.0**
GRACE w/o calibration	9.0

and compare their text embeddings with the reconstructed visual embeddings using cosine similarity. For Image-Text Retrieval, we include **Flickr30k** [65] (entire dataset) and **COCO** [64] (validation set).

B Disentanglement in the reconstruction

In this section, we evaluate the impact of the consistency regularization described in Sec. 3.2. The regularization term enforces consistency in the reconstruction weights for samples sharing the same category, and penalizes similarity when samples belong to different categories. We evaluate the effectiveness of this approach in disentangling the concepts that contribute the most in each category by following the protocol defined in [63, 69], which we summarize in the following.

DCI Metrics. In this framework, a set of sparse regressors f is trained on the target representations $\mathbf{c} \in \mathbb{R}^D$ to predict the latent factors $\mathbf{z} \in \mathbb{R}^K$, and evaluate disentanglement with respect to these predicted factors. $\mathbf{P} \in \mathbb{R}^{D \times K}$ represents the relative importance of each code $\mathbf{c}_i$ in predicting a given latent factor $\mathbf{z}_j$. The classifiers are then evaluated in terms of *informativeness*, *disentanglement* and *completeness*. **Informativeness** is the accuracy of the regressor in predicting the ground truth latent factors on a held-out test set. **Disentanglement** measures how close each code $\mathbf{c}_i$ is to capturing one single latent factor, and is defined as $D_i = (1 - H_K(P_{i\cdot}))$, where $H_K(P_{i\cdot})$ is the entropy of the relative importance of code i over all the latent factors. In the ideal case, each code contributes to predicting a single latent factor, hence resulting in $D_i = 1$. **Completeness** measures whether each latent factor is captured by one or more codes, and is defined as $C_j = (1 - H_D(P_{\cdot j}))$. When a latent factor is captured by a single code, the corresponding completeness is $C_j = 1$.

We adapt these metrics to our setting by evaluating the quality of the reconstruction weights from GRACE (*codes*) with respect to some ground truth category label (*latent factors*). Specifically, we consider the *scenario* annotations in Ego4D, a set of discrete labels that describe the general context of a video in the dataset, as category labels for computing the metrics. Since these labels are binary, rather than continuous latent factors, we use logistic regression classifiers with ℓ_1 penalties to predict the category labels from the reconstruction weights and define $\mathbf{P}_{ij}$ as the weight assigned by classifier j to the reconstruction weight i, normalized with respect to all the classifiers. For *informativeness*, we compute the mean average precision on the category predictions. From EgoMCQ, we derive a subset of videos from the 25 most common scenarios, and divide them into 27453 and 9206 training and validation samples.

C Concepts redundancy

To analyze redundancy within the concept-based decomposition, we cluster the LAION vocabulary concepts using K-Means applied to their text embeddings and select the optimal number of clusters via the silhouette score. This produces 120 distinct clusters with a median size of 110.5 samples. We present representative examples of the resulting macro-clusters in Table 8-left, and assign each macro-cluster the label closest to the centroid. Each cluster contains synonyms and related words. Following clustering, we take the top-25 most activated concepts for each class and map them to their corresponding macro-clusters. This results in a median of 16.0 distinct clusters among the top-25 concepts per class, indicating relatively low redundancy. The calibration module controls the contribution of each concept in the reconstruction and is crucial for reducing redundancies. Indeed, when this module is removed the median number of distinct macro concepts in Top-25 drops to 9.0, showing significantly higher redundancy (Table 8-right).

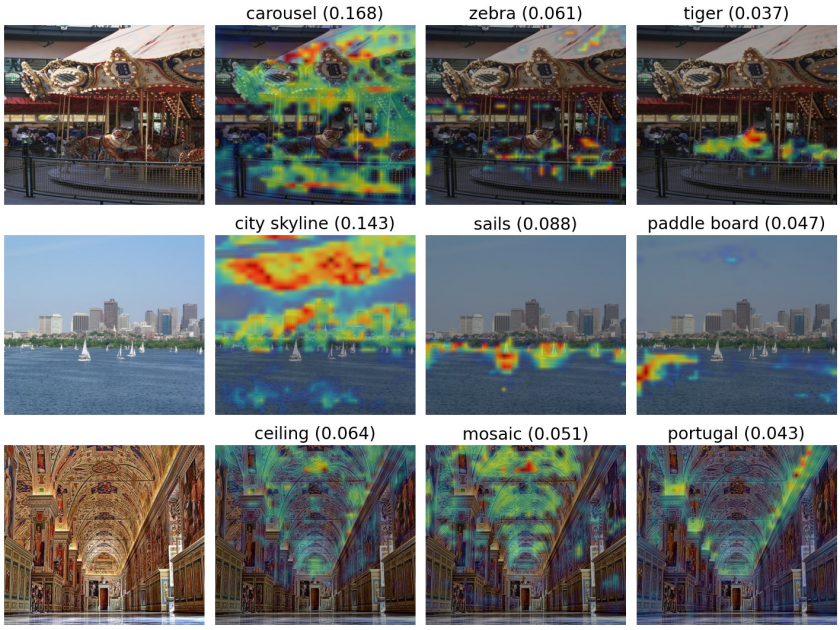


Figure 6: **Visual grounding of the concept decomposition on ADE20k.** Using a GRACE variant trained on TIPSv2 [71], we compute the concept decomposition of the CLS token, select the top three highest-scoring concepts, and show their contribution in the visual patches.

D Visual grounding

GRACE performs concept-based decomposition of the embedding corresponding to the CLS token, leveraging its alignment with text. In CLIP and PerceptionEncoder, the embeddings corresponding to spatial visual patches are not text-aligned, so this decomposition cannot be applied to them directly.

Table 9: **Concept grounding on ADE20k.** We report AUPRC.

Method	AUPRC
BLIP-2 [16]	0.098
LLaVA-NeXT [70]	0.302
ConceptScope [32]	0.399
GRACE (TIPSv2)	**0.482**

To address this, we extend GRACE to a different VLM backbone, TIPSv2 [71], which outputs text-aligned visual patch embeddings that can be decomposed with GRACE. We train GRACE with TIPSv2 on ImageNet, using the original reconstruction objective on the CLS token. At inference time, we apply the same concept-based decomposition to the spatial patches and evaluate spatial grounding on ADE20k [72], selecting the best class-concept mapping on the training set by following the evaluation protocol from ConceptScope [32].

As shown in Table 9, GRACE substantially improves concept grounding compared to ConceptScope (0.482 vs 0.399), showing that concepts can be reliably localized in the input image. These experiments further show that GRACE can be easily implemented on top of different backbones, including CLIP, PerceptionEncoder, and TIPSv2, for both global and dense concept-based decomposition tasks. In Fig. 6, we select a few samples from ADE20k, compute the concept decomposition of the CLS token, select the top three highest-scoring concepts, and show their contribution in the visual patches.

E Captions integration in GRACE

The reconstructed visual embeddings from GRACE achieve high faithfulness with respect to the original embeddings and strong downstream task performance, despite not relying on instance-level optimization or captions during training or inference. For a fair comparison with EXPLAIN-R, which uses captions to identify concepts in the visual input and to produce surrogate representations that preserve the vision-language alignment of the target model, we also consider a variant of GRACE that leverages captions at training time to further improve the reconstruction process. Specifically, we incorporate the distillation loss from [14] to push the reconstructed embeddings to match the

Table 10: **Concept projection in the reconstruction.** We consider different strategies to project the concept embeddings before reconstruction, and evaluate retrieval accuracy on EgoMCQ [57] and semantic alignment of the reconstruction weights on ConVisBench [60].

Projection	**EgoMCQ**			**ConVisBench**
	Intra Acc.	Inter Acc.	Faith.	ρ / τ
Identity	32.59	75.93	67.08	**53.9 / 38.4**
Linear	<u>36.08</u>	<u>81.33</u>	<u>79.94</u>	48.3 / 34.2
MLP	**36.15**	81.12	**80.37**	47.9 / 33.9
Linear (Orthogonal) - GRACE	35.43	**81.35**	77.86	<u>51.0</u> / <u>36.2</u>

original embeddings in terms of cosine similarity with the text embeddings of the captions in the batch. For any visual $\mathbf{z}_i$ and text $\mathbf{t}_j$ embeddings in the batch $\mathcal{B}$, we indicate the distribution of the pairwise image-to-text cosine similarities as:

$$P_i = \frac{\exp(\mathbf{z}_i^T \mathbf{t}_i / \tau)}{\sum_k \exp(\mathbf{z}_i^T \mathbf{t}_k / \tau)}, \tag{6}$$

where τ is a temperature parameter, and the same distribution for the reconstructed embeddings $\tilde{\mathbf{z}}_i$ as:

$$\tilde{P}_i = \frac{\exp(\tilde{\mathbf{z}}_i^T \mathbf{t}_i / \tau)}{\sum_k \exp(\tilde{\mathbf{z}}_i^T \mathbf{t}_k / \tau)}. \tag{7}$$

Similarly, we indicate with Q_i and $\tilde{Q}_i$ the distributions of the text-to-image similarities. The distillation loss $\mathcal{L}_d$ is defined as the KL divergence between the distributions obtained from the original and reconstructed embeddings:

$$\mathcal{L}_d = \frac{1}{2|\mathcal{B}|} \left[\sum_{i=1}^{|\mathcal{B}|} \mathrm{KL}(P_i || \tilde{P}_i) + \sum_{j=1}^{|\mathcal{B}|} \mathrm{KL}(Q_j || \tilde{Q}_j) \right]. \tag{8}$$

In our experiments, we observe that the distillation provides a quite marginal contribution on both image and video benchmarks, indicating that uni-modal reconstruction with our geometry-aware objectives alone is sufficient to produce faithful representations with respect to the original embeddings.

F Concepts projection in the reconstruction

Visual embeddings in GRACE are reconstructed from a linear combination of the text embeddings of the concepts in the vocabulary. In the reconstruction process, concepts are projected to facilitate the reconstruction (Eq. 2). Ideally, this process should limit concept entanglement, which would limit the interpretability of the reconstruction weights, thus preserving the pairwise relations between the concepts in the projected space. In Table 10, we compare different strategies in terms of constraints and capacity, starting from a simple identity mapping and evaluating the quality of the reconstruction in terms of multi-modal alignment on EgoMCQ [57] and semantic alignment of the reconstruction weights on ConVisBench [60].

The identity mapping achieves the best performance on ConVisBench: freezing the concept embeddings encourages the model to predict highly semantic weights to meet the reconstruction objectives. However, the performance and faithfulness on EgoMCQ are quite poor, as the identity is not sufficient to reconstruct the geometric relations of the original samples. Replacing the identity with a linear projection or an MLP significantly improves multi-modal alignment and faithfulness of the reconstructed visual embeddings, at the cost of a lower interpretability of the reconstruction weights, as they become less correlated with human-judged similarity on ConVisBench. The orthogonal linear projection in GRACE strikes a good balance between the two aspects, as it achieves close to optimal multi-modal alignment and consistently better correlation on ConVisBench. During training, we force $\mathbf{Q}$ to be orthogonal through a Cayley transform.

G Robustness to the choice of the concept vocabulary

We present strong empirical evidence that GRACE is robust to the choice of the concept vocabulary.

Table 11: **Comparison of different concept vocabularies on Image Classification tasks.** We compare different versions of GRACE using the ImageNet [66] and LAION [73] vocabularies, showing quite consistent performance across different vocabularies.

Method	Captions	Concepts Vocab.	ImageNet		Flowers102		Food101	
			Acc.	Faith.	Acc.	Faith.	Acc.	Faith.
GRACE	✗	ImageNet	48.04	60.15	**32.14**	**41.78**	62.67	66.83
GRACE	✗	LAION	47.36	59.87	25.11	31.92	61.60	64.84
GRACE	✓	ImageNet	47.84	60.68	28.82	41.18	**63.24**	**67.49**
GRACE	✓	LAION	**48.27**	**60.96**	28.57	36.79	60.80	64.65

Table 12: **Comparison of different concept vocabularies on Image-Text Retrieval tasks.** We compare GRACE using the ImageNet [66] and LAION [73] vocabularies, showing the robustness of our method to different vocabularies.

Method	Captions	Concepts Vocab.	Flickr30k I→T			Flickr30k T→I			COCO I→T			COCO T→I		
			R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Task Metrics														
GRACE	✗	ImageNet	48.48	76.17	84.50	42.84	69.97	79.26	41.80	71.98	83.66	33.56	63.34	75.92
GRACE	✗	LAION	48.24	75.85	84.23	42.49	69.55	78.88	41.32	71.88	83.70	32.36	61.54	74.18
GRACE	✓	ImageNet	52.30	**79.30**	86.97	46.87	74.43	83.41	42.78	72.84	84.46	**36.34**	65.82	78.66
GRACE	✓	LAION	**52.56**	79.29	**87.18**	**47.24**	**74.62**	**83.66**	**43.48**	**73.88**	**84.80**	35.94	**66.60**	**79.18**
Faithfulness Metrics														
GRACE	✗	ImageNet	55.24	84.27	91.39	51.77	81.56	89.64	54.78	86.42	94.38	47.36	81.60	91.42
GRACE	✗	LAION	55.19	83.99	91.13	51.89	81.82	89.91	54.38	87.08	94.84	46.82	81.44	91.14
GRACE	✓	ImageNet	62.72	90.19	95.40	55.94	85.65	92.85	58.16	89.80	96.22	49.76	84.24	93.66
GRACE	✓	LAION	**63.10**	**90.45**	**95.79**	**56.69**	**86.37**	**93.35**	**58.56**	**90.58**	**96.88**	**50.26**	**85.44**	**94.50**

ImageNet vs. LAION vocabularies. For image benchmarks, we compare vocabularies derived from ImageNet and LAION captions. The ImageNet vocabulary was introduced in the EXPLAIN-R paper [14], mined from a large corpus of ImageNet captions extracted with BLIPv2, while the LAION vocabulary is built from bigrams extracted from LAION captions. The ImageNet and LAION vocabularies include 18490 and 15000 concepts, respectively. We compare the impact of using different concepts vocabularies in GRACE on Image Retrieval (Table 12) and Image Classification tasks (Table 11). Task metrics and faithfulness are quite similar across all benchmarks, showing the robustness of GRACE to different concept vocabularies.

Video benchmarks. For video benchmarks, we draw concepts from human-annotated captions at two levels of granularity: "EgoClip" and "Full Voc.". "EgoClip" uses the official Ego4D taxonomy of verbs and nouns [57], with related terms grouped into broader macro-concepts (e.g., "brush", "clean", "dust", "mop", "sanitize", and "scrub" are merged into a single "clean" concept). "Full Voc." instead keeps all verbs and nouns from the captions ungrouped. These vocabularies contain 700 and 2139 concepts, respectively. We also test a vocabulary generated by prompting GPT5.4 to produce an exhaustive list of plausible text-based concepts likely to appear in human activity videos. These comparisons are reported in Table 4-right. We find that ConVisBench benefits considerably from richer, more exhaustive vocabularies, since they allow for finer-grained concept decomposition, whereas retrieval performance (EgoMCQ) is largely less affected by this choice.

Table 13: **Comparison of visual evidence and captions-based evidence.** *Visual evidence* identifies concepts in the input through the similarity of the visual embeddings and the text embeddings of the concepts in the vocabulary. *Captions-based evidence* leverages an external captioning model to identify concepts in the input, which is the default approach for EXPLAIN-R.

Method	Evidence	EgoMCQ		Charades		ImageNet		Flowers102		Food101		Flickr30k		COCO	
		Acc.	Faith.	Acc.	Faith.	Acc.	Faith.	Acc.	Faith.	Acc.	Faith.	R@1	Faith.	R@1	Faith.
EXPLAIN-R	Captions	53.14	63.70	21.73	25.92	50.77	50.19	15.51	16.59	45.86	45.81	45.67	44.48	**44.22**	46.72
EXPLAIN-R	Visual	57.15	73.94	24.88	**44.50**	**52.51**	**67.72**	26.67	35.00	55.65	61.03	37.78	47.14	35.16	51.34
GRACE	Visual	**58.39**	**77.86**	**25.99**	35.15	48.04	60.15	**32.14**	**41.78**	**62.67**	**66.83**	**48.48**	**55.24**	41.80	**54.78**

Concept filtering. Using Singular Value Decomposition (SVD) on the concept matrix, we estimate the redundancy present in the concept vocabulary and show that it can be substantially compressed into fewer concepts without significantly hurting the downstream performance (Table 4-left).

Together, these experiments validate GRACE's robustness along three axes: the source of the concept vocabulary (ImageNet vs. LAION, human-annotated captions vs. GPT), its granularity ("EgoClip" vs. "Full Voc." for video benchmarks), and the impact of aggressive filtering (concept pruning).

H Visual evidence and captions-based evidence

GRACE identifies concepts in the input by computing the cosine similarity between the visual embedding $\mathbf{z}_i$ of the input and the text embedding $\mathbf{c}_j$ of the concepts in the vocabulary, i.e., $\alpha_{i,j} = \cos(\mathbf{z}_i, \mathbf{c}_j)$. We refer to this signal as *visual evidence*. In contrast, EXPLAIN-R relies on an external captioning model to generate n captions for each input and constructs a concept encoding vector $\boldsymbol{\alpha}_i$, where $\alpha_{i,j}$ corresponds to the relative frequency of concept j in the generated captions for that sample. We refer to this signal as *caption-based evidence*.

Visual evidence does not depend on external models and introduces negligible overhead at inference time. More importantly, visual evidence provides a direct means to probe how semantic concepts are represented within the target model's embeddings. To further demonstrate the benefit of visual evidence, even beyond our method, we replace concept-based evidence in EXPLAIN-R with visual evidence, while keeping the rest of the method unchanged. Specifically, we define a concept encoding vector $\boldsymbol{\nu}_i$, such that $\nu_{i,j}$ is 1 if concept j is among the top-k concepts matched by cosine similarity with the visual embedding, and 0 otherwise:

$$\nu_{i,j} = \begin{cases} 1 & \text{if j} \in \text{TopK}_l\left[\cos(\mathbf{z}_i, \mathbf{c}_l)\right] \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

We compare visual and captions-based evidence on video and image benchmarks in Table 13. EXPLAIN-R with visual evidence outperforms the original captions-based version on several benchmarks, in both task metrics and faithfulness. These results demonstrate that visual evidence is sufficient to produce surrogate embeddings that preserve the performance of the original model. GRACE is designed to exploit and refine visual evidence, which results in consistent improvements across the benchmarks.

I Additional qualitative results

We present additional qualitative results, showing the concepts matched in the reconstruction process by SpLiCE and GRACE (Figure 7) and the robustness of our method to concept interventions (Figure 8).

Both SpLiCE and GRACE recover semantically meaningful concepts from the input video, even though the reconstruction weights from SpLiCE tend to concentrate in the first few concepts. This behavior indicates that SpLiCE is over-focusing on these concepts, compared to GRACE that distributes the reconstruction weights more evenly over the most relevant concepts. With the intervention, we evaluate how much the model is relying on these concepts to predict the similarity between the two videos. Even when removing the most significant concept, the two videos remain similar on multiple semantic axes, e.g., the location or the main subject. In this scenario, the similarity score predicted by SpLiCE drops significantly, while GRACE is more robust to these interventions by distributing the reconstruction weights more evenly across the concepts present in the input.

J Limitations and future works

Our approach is a significant effort towards making Vision-Language Models more interpretable, decomposing visual embeddings into human-understandable concepts. However, the notion of a "*concept*" is itself not uniquely defined and may be considered quite subjective. Empirically, in our work, we define a concept as a semantic unit that a human would use to describe the same input, while acknowledging that this definition is necessarily approximate and not very rigorous. Because of this definition, and due to the potential incompleteness of the vocabulary source, some concepts may

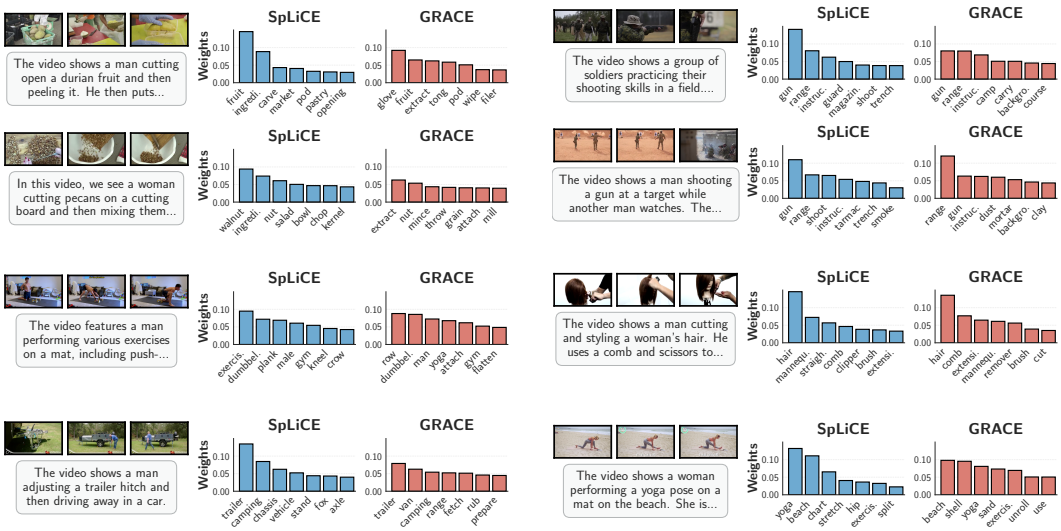
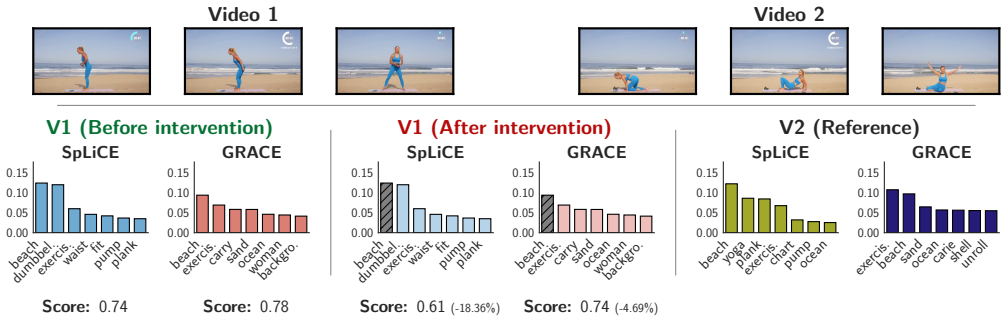
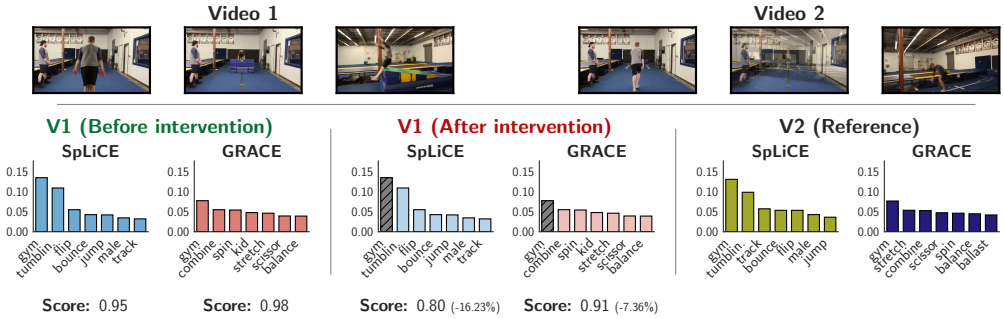


Figure 7: **Reconstruction weights on ConVisBench [60].** For each video, we show a few uniformly sampled frames, the ground truth caption and the Top-k reconstruction weights from the two methods. Both SpLiCE and GRACE recover semantically meaningful concepts from the input video, even though SpLiCE tends to allocate more reconstruction mass to the first few concepts.

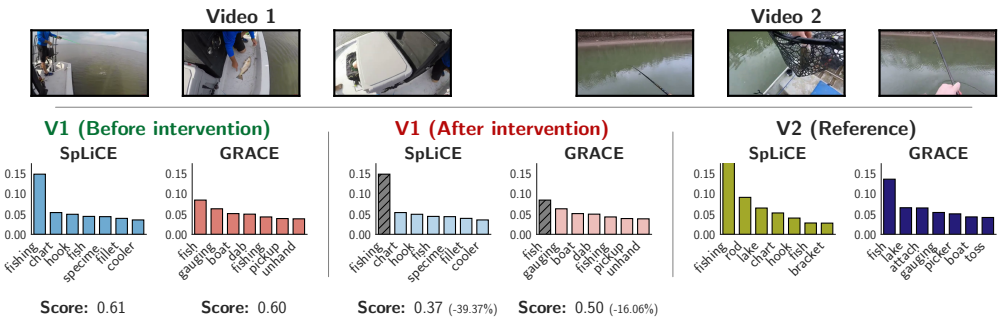
be absent from the concept vocabulary or may not be readily describable through text [74], limiting the coverage and, consequently, the interpretability of the resulting decomposition. Modeling the residual information that cannot be expressed through concepts is a very relevant future research direction. Regarding the interpretability and alignment of the concept decompositions with human perception, we observe a higher correlation with human judgments on ConVisBench compared to existing approaches. This suggests that the concepts extracted by our model are better aligned with the factors considered by human annotators when assessing the similarity between different videos. Even though this is a strong indication that the model embeds a concept-based structure, we do not establish a strict causal relation between the concept decomposition produced by GRACE and the model behavior on downstream tasks. Notably, none of the prior methods reported on ConVisBench establish a strict causal implication between the concept decomposition and the model's behavior; they instead rely on correlation evidence, i.e., alignment with human judgment. Although we do not claim a formal causal relation between the concept decomposition produced by GRACE and the model's behavior, our approach improves the understanding of the concept-based structure of VLMs like CLIP. Similarly, our approach does not model how concepts evolve over time in a video, as we produce clip-level explanations. We provide preliminary evidence about the spatial grounding of the concepts obtained from the decomposition. However, our approach does not model temporal grounding of the decomposed concepts. We leave these as future work.



(a) Comparison between videos `1514_005` and `1514_008`.



(b) Comparison between videos `1525_004` and `1525_005`.



(c) Comparison between videos `3637_006` and `3941_003`.

Figure 8: **Concept intervention on ConVisBench [60].** We artificially zero out the concept corresponding to the highest reconstruction weight to test the ability of these models in capturing all the relevant concepts in the input. After the intervention, the two videos remain similar according to multiple semantic axes. However, the similarity score predicted by SpLiCE drops significantly after the intervention, while GRACE shows a more robust behavior.

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 claims made in the abstract and introduction reflect the contributions of the paper in terms of method and experimental results.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: We discuss the limitations of the approach in the conclusions.

   Guidelines:

   - The answer [N/A] 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.
   - 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: [N/A]

Justification: We do not have theoretical results.

Guidelines:

- The answer [N/A] 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: Implementation details of the proposed model are reported in the paper and appendix. We will also release the code upon acceptance.

   Guidelines:

   - The answer [N/A] 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: [No]

Justification: We are not currently providing the code as part of the submission. We will release the code upon acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: We reported implementation details on the training and evaluation processes in the paper and appendix.

   Guidelines:

   - The answer [N/A] 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: [No]

   Justification: Error bars are not reported due to computational complexity.

   Guidelines:

   - The answer [N/A] 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: We include a section in the appendix about computational resources.

   Guidelines:

   - The answer [N/A] 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: We have read the NeurIPS Code of Ethics and we confirm we have conducted our research in accordance to it.

   Guidelines:

   - The answer [N/A] 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: This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: These safeguards do not apply to our work.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - 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: We use publicly available datasets and models with proper citation.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: We do not release new assets as part of this submission.

    Guidelines:

    - The answer [N/A] 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: [No]

    Justification: The paper does not involve crowdsourcing.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not involve research with human subjects

    Guidelines:

    - The answer [N/A] 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.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.