CONCEPT-GUIDED DICTIONARY LEARNING FOR INTERPRETABLE CONCEPT EXTRACTION AND ATTRIBUTION IN LARGE VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive Large Vision-Language Models (LVLMs) generate text sequentially, conditioning each token on evolving multimodal states. This makes it difficult to assess whether predictions are grounded in **visual concepts** or instead reflect hallucination or bias. Existing concept-discovery approaches such as **TCAV**, **CRAFT**, and **CLIP-Dissect** are designed for encoder-only or contrastive models. At the same time, recent LVLM methods (CoX-LMM) depend on labeled concepts and simplified settings, limiting scalability.

We propose **Concept-Guided Dictionary Learning (CGDL)**, a weakly supervised and scalable framework for multimodal concept discovery in autoregressive LVLMs. CGDL first probes the model to surface textual concepts from a dataset. For each concept, it constructs positive and negative patch sets using concept-grounded crops using SAM and randomized backgrounds. A contrastive dictionary-learning stage then disentangles concept-aligned activations from residual noise, yielding sparse, monosemantic vectors that reveal **semantically aligned visual–textual interactions** and enable faithful attribution of predictions to visual evidence.

On **ImageNet-1k, MSCOCO**, CGDL outperforms recent interpretability methods with up to **4% higher sparsity**, **11% greater stability**, **17% lower overlap**, and strong attribution faithfulness, while scaling efficiently to large concept vocabularies. These results advance concept-based interpretability for LVLMs and provide a practical step toward transparent multimodal reasoning.

1 INTRODUCTION

Large Vision–Language Models (LVLMs) generate text *autoregressively*, conditioning each token on evolving multimodal states. This enables rich, context-sensitive prediction but raises unique interpretability challenges: predictions may stem from spurious correlations or hallucinations rather than *visual evidence*. Existing methods, such as saliency maps or training-based grounding, localize objects but do not reveal the higher-level *concepts* driving model decisions. Concept-based explanations offer more faithful insights, yet most discovery methods assume static feature spaces. Approaches like **TCAV** Kim et al. (2018), **CRAFT** Fel et al. (2023b), **CLIP-Dissect** Oikarinen & Weng (2023), and **Holistic** Fel et al. (2023a) cannot extend to autoregressive models and are restricted to single-object settings, while **CoX-LMM** Parekh et al. (2024) requires a labeled concept token and is restricted to single-object settings, limiting scalability (Table 1).

We propose **Concept-Guided Dictionary Learning (CGDL)**, a weakly supervised framework for scalable concept discovery in autoregressive LVLMs. CGDL first surfaces candidate concepts, constructs positive/negative patch sets, and then casts each as a one-vs-all representation learning problem with a two-basis decomposition. This formulation forces the dictionary to disentangle activations from residual noise, yielding sparse, non-overlapping vectors that faithfully align visual and textual modalities.

While most concept extraction methods rely on similar dictionary learning formulations, prior work has shown that the quality of discovered concepts depends more on how the data is presented to the dictionary than on the specific factorization algorithm itself Grobrügge et al. (2025); Sun et al. (2023). CGDL is not merely Semi-NMF with preprocessing; rather, it provides a general, model-agnostic

Table 1: Comparison of concept-based interpretability methods. Only **CGDL** supports weakly supervised, scalable discovery in autoregressive LVLMs; most prior methods target encoder-only (IE) models and single-object concepts, while they can not perform text grounding (TG).

Method	Type	TG	AutoReg	Key Limitation
TCAV Kim et al. (2018)	IE	✗	✗	Supervised
ACE Ghorbani et al. (2019)	IE	✗	✗	Single object
CRAFT Fel et al. (2023b)	IE	✗	✗	Single object
EAC Sun et al. (2023)	IE	✗	✗	Single object
CLIP-Dissect Oikarinen & Weng (2023)	IE	✓	✗	Single object
Holistic Fel et al. (2023a)	IE	✗	✗	Single object
CoX-LMM Parekh et al. (2024)	AutoReg	✓	✓	Single object
MCD Grobrügge et al. (2025)	IE	✗	✗	Single object
CGDL (Ours)	AutoReg	✓	✓	Unlimited-object

framework that reframes concept discovery as a contrastive one-vs-all decomposition. The two-basis design prevents atom collapse in large vocabularies, allowing CGDL to scale robustly to thousands of concepts Kim & Park (2008). By combining weakly supervised concept bags, weak localization, residual contrast, and two-basis factorization, CGDL achieves monosemantic, multimodal vectors at scale—something previous dictionary learning methods for LVLMs have not demonstrated. **Our contributions are:**

- We introduce **CGDL**, a weakly supervised framework for scalable concept discovery in autoregressive LVLMs, overcoming the single-object limitation of prior methods.
- We leverage the **Segment Anything Model (SAM)** Kirillov et al. (2023) for weak concept localization, providing high-quality positive/negative patch sets.
- We propose a contrastive residual extraction scheme that enforces monosemanticity by separating concept–vs.–No-concept activations.
- We show that CGDL yields faithful multimodal attribution, bridging visual and textual modalities using extensive experiments on **ImageNet-1k** and **MSCOCO** demonstrate superior sparsity, stability, and faithfulness over prior methods, scaling to 1k+ concepts with improved attribution quality (up to **+9% CLIPScore**, **+4% BERTScore**).

Together, these advances provide a practical step toward transparent multimodal reasoning, bridging the gap between autoregressive generation and human-understandable concept-based explanations.

2 Related Work

A central goal in interpretability is to uncover how internal representations encode semantically meaningful concepts. Early work introduced Concept Activation Vectors (CAVs) (Kim et al., 2018), which quantify model sensitivity to human-defined concepts but require curated examples, limiting scalability. Extensions sought to automate discovery (Ghorbani et al., 2019), yet dependence on segmentation quality hindered robustness.

Alternative approaches use saliency maps (Selvaraju et al., 2017; Strumbelj & Kononenko, 2014) or training-based grounding (Kang et al., 2024; Zhang et al., 2023; Ma et al., 2024) to highlight *where* evidence lies. While effective for localization, these methods do not articulate *what* abstract concepts the model internally represents, thus complementing but not replacing concept-based explanations.

To reduce supervision, some methods decomposed hidden states into interpretable factors. NMF (Liu et al., 2025), dictionary learning with prototypes (CRAFT) (Fel et al., 2023a), PCA, UMAP, and sparse autoencoders (Pach et al., 2025) have shown promise in vision-only settings, with unifying benchmarks emerging (Fel et al., 2023a). CLIP-based concept attribution (Oikarinen & Weng, 2023; Dreyer et al., 2025) further advanced automated discovery, but these remain restricted to vision encoders. While concept extraction overlaps with mechanistic interpretability Templeton (2024); Pach et al. (2025); Elhage et al. (2022), our focus is on *single-layer monosemantic concept discovery*

in autoregressive LVLMs rather than neuron- or circuit-level analysis. Alternative approaches such as attention maps (Jain & Wallace, 2019), causal tracing (Meng et al., 2023), and linear probes (Alain & Bengio, 2018) provide useful insights but suffer from weak causal grounding, poor scalability, or reliance on labeled supervision. CGDL complements these methods by enabling weakly supervised, scalable concept discovery with reusable concept vectors.

Autoregressive Large Vision–Language Models (LVLMs) present new challenges: activations evolve across time steps, residual streams encode multiple dependencies, and concepts rarely align with a single hidden state (Templeton, 2024). CoX-LMM (Parekh et al., 2024) adapted Semi-NMF Trigeorgis et al. (2014) to LVLM activations, but struggles with (i) reliance on tokenized object names, (ii) not being suitable for multi-token concepts, (iii) background noise from full-image extraction, and (iv) persistent polysemanticity in residual streams. One could extend the single-object concept extraction in CoX-LMM by extracting residual streams for sets of *(token, image)* pairs, but this remains inherently supervised and fails to overcome the aforementioned limitations.

While methods such as TCAV, CRAFT, and CLIP-Dissect pioneered concept-based interpretability, they target static encoders and cannot be applied to LVLMs. CoX-LMM incorporates many of these ideas into a dictionary-learning framework for autoregressive models, and thus serves as the most representative baseline. We therefore compare CGDL against CoX-LMM using its strongest dictionary learning variants (SNMF, SAE) to ensure fairness.

Unlike CoX-LMM and other concept extraction methods that rely on labeled tokens and often yield polysemantic vectors or assume single-object settings, we propose **Concept-Guided Dictionary Learning (CGDL)**. CGDL introduces contrastive residual extraction with spatially localized crops and candidate text, enforcing a clean separation between *concept* and *noise*. This produces faithful, monosemantic concept vectors and, to the best of our knowledge, is the first framework to scale weakly supervised concept discovery from single-object to multi-object settings.

3 PRELIMINARIES

3.1 DICTIONARY-LEARNING VIEW OF CONCEPT EXTRACTION

Recent work on concept extraction Ghorbani et al. (2019); Sun et al. (2023); Fel et al. (2023a); Parekh et al. (2024) shows that many methods can be framed as *dictionary learning*, where activations are approximated by a small set of interpretable bases.

Formally, given activations $S \in \mathbb{R}^{n \times d}$ with n samples and d-dimensional features, we posit K latent concepts and solve

$$\underset{U \in \mathbb{R}^{d \times K},\, V \in \mathbb{R}^{n \times K}}{\arg\min} \|S - VU^\top\|_F^2,$$

where $U = [u_1, \ldots, u_K]$ are the **concept bases** (CAVs) and $V = [v_1^\top; \ldots; v_n^\top]$ are the **activations**, with row v_i giving concept coordinates of sample i.

This factorization unifies prior approaches via constraints on (U, V):

$$\begin{cases} v_i \in \{e_1, \ldots, e_K\}, & \text{K-Means (ACE) Ghorbani et al. (2019),} \\ U^\top U = I, & \text{PCA Graziani et al. (2023),} \\ S \geq 0,\ U \geq 0,\ V \geq 0, & \text{NMF (CRAFT Fel et al. (2023a;b), ,} \\ U \text{ free},\ V \geq 0, & \text{Semi-NMF (SNMF) Trigeorgis et al. (2014); Parekh et al. (2024),} \\ V = \psi(S),\ \|v_i\|_0 \leq s, & \text{Sparse Autoencoder (SAE) Templeton (2024); Pach et al. (2025).} \end{cases}$$

Columns of U are concept vectors (CAVs), rows of V are per-sample activations. Special cases include PCA (orthogonal bases), NMF (nonnegative factors), SNMF (mixed-sign bases with nonnegative activations), K-Means (one-hot codes), and SAE (encoder-decoder with sparsity).

4 METHOD

We begin by formalizing LVLMs as black-box systems with accessible intermediate activations, taking multimodal inputs (text, image) and producing corresponding outputs (text/ set of tokens).

4.1 POSITIVE AND NEGATIVE CONCEPTS

We address the limitations of single-object dependence and polysemantic behavior in CoX-LMM by introducing *concept-example bags*—collections of image patches that serve as positive (concept-present) or negative (concept-absent) instances for each automatically discovered concept $c_k \in \{c_1, \ldots, c_K\}$.

Given an unlabeled image set $\mathcal{I} = \{I_k\}_{k=1}^{N}$, the LVLM f predicts candidate concepts for each image using a structured prompt (See Appendix B for details.):

$$C(I_k) = f(I_k, \texttt{prompt}) \subseteq \mathcal{V},$$

and the global vocabulary is

$$\mathcal{C} = \bigcup_{k=1}^{N} C(I_k).$$

For each $c \in \mathcal{C}$, we collect supporting images

$$\Phi(c) = \{I_k \mid c \in C(I_k)\}.$$

A patch operator $\mathcal{P}(I_k, c) \in \{\mathrm{crop}(I_k, c), \mathrm{sam}(I_k, c)\}$ localizes the region associated with c Kirillov et al. (2023). Because such patches may include background context, the resulting *concept-example bag* is a mixture of positives and negatives:

$$\mathcal{B}(c) = \{\, \mathcal{P}(I_k, c) : I_k \in \Phi(c) \,\} = \mathcal{B}^{+}(c) \cup \mathcal{B}^{-}(c).$$

Unlike prior approaches that rely on annotated single-object images, this formulation is *weakly supervised* and scales to multi-object datasets. For example, the concept "stripes" may emerge from zebras, tigers, or cats, without requiring manual concept labels.

4.2 CONCEPT-GUIDED DICTIONARY LEARNING

According to Fel et al. (2023a), most prior concept-expansion methods rely on dictionary learning for concept extraction; CoX-LMM is no exception. The key difference lies in the data passed to the dictionary-learning algorithm, which critically affects concept quality Grobrügge et al. (2025); Sun et al. (2023).

Relying on long, ambiguous token sequences from open-ended captioning typically restricts analysis to the residual stream of a single token, introducing overlapping concepts Templeton (2024) and preventing dictionary elements from representing clean, disentangled semantics. Multi-object images further exacerbate this, as activations mix features from different objects.

We address these issues with *Concept-Guidance*, a contrastive residual extraction scheme. The model is prompted to output either the target concept c_k or $\texttt{No-}c_k$, ensuring that activations are aligned with a single concept and enforcing monosemanticity:

> $\texttt{prompt}_{cg}$ = "Does the image contain c_k? If yes, output c_k; otherwise, output No-c_k."

This binary design yields cleaner, concept-focused residuals.

For each concept bag $\mathcal{B}(c_k)$, we query the model with the above prompt and collect residual activations. Following Koh et al. (2020); Alam et al. (2025), we decompose the LVLM into an **embedding function** g (vision encoder, bridging, decoder attention) and an **output function** h (projection and softmax). Based on Fel et al. (2023a); Parekh et al. (2024), we analyze the penultimate residual layer (language_model.norm). For an image crop $x^{c_k} \in \mathcal{B}(c_k)$ and cached prefix $\hat{y}_{<t}$, the embedding function produces

$$a_t^{(l)} = g^{(l)}(x^{c_k}, \texttt{prompt}_{cg}, \hat{y}_{<t}) \in \mathbb{R}^p, \tag{1}$$

where p is the residual dimension Geva et al. (2020). The output function then predicts

$$\hat{y}_t = h^{(l)}(a_t^{(l)}). \tag{2}$$

According to Geva et al. (2020), we can average residuals across tokens in each response to obtain a concept-level embedding without losing semantic meaning.

$$s_m = \tfrac{1}{T_m} \sum_{t=1}^{T_m} a_t^{(l)} \in \mathbb{R}^p, \tag{3}$$

where T_m is the number of generated tokens for sample m. Collecting M such samples yields

$$S = [s_1, \ldots, s_M] \in \mathbb{R}^{M \times p}, \tag{4}$$

which serves as input for concept extraction. Unlike approaches that rely solely on token embeddings, this formulation captures information from multi-token concepts (e.g., *hot dog*) rather than splitting them into isolated tokens (*hot*, *dog*).

Dictionary learning then decomposes S into concept and negation bases:

$$S \approx VU^T, \quad V \in \mathbb{R}^{M \times 2},\ U \in \mathbb{R}^{p \times 2}.$$

Here U contains basis vectors for c_k and No-c_k, while V captures sample activations. When restricting each bag to two bases (concept vs. negation), the per-bag cost reduces to $\mathcal{O}(M_c p + M_c^2)$, where M_c is the number of samples in $\mathcal{B}(c_k)$. Over K concepts, the total complexity scales as $\mathcal{O}(K(M_c p + M_c^2))$, which remains efficient for large concept sets, consistent with the per-iteration complexity of Semi-NMF Kim & Park (2008); Ding et al. (2010). This contrasts with full NMF, where larger dictionary sizes ($K \gg 2$) lead to cubic dependence on K, and with Sparse Autoencoders (SAE), where training requires backpropagation through millions of parameters. By reducing each bag to two bases, our approach avoids interference among dictionary atoms and remains scalable for LVLM interpretability.

This formulation contrasts each concept against all others (akin to one-vs-all classification), capturing inter-concept relations without requiring oversized dictionaries. Unlike prior methods that entangle concepts across long sequences, Concept-Guidance yields disentangled, monosemantic residuals suitable for large-scale LVLM interpretability.

4.3 Positive concept vector identification

Following visual grounding in Parekh et al. (2024), for each feature $u \in U$ from the dictionary decomposition, we extract the *maximum activated crops (MAC)*—the top-α_{MAC} image patches that most strongly activate it, i.e., the highest values in the k-th column of the activation matrix V:

$$X_{k,\text{MAC}} = \{i \mid v_i^{(k)} \text{ is among the top-}\alpha_{\text{MAC}} \text{ values of } v^{(k)}\}.$$

Since each concept bag contains both c_k (positive) and $\texttt{No}{-}c_k$ (negative) samples, we select the basis

$$k^* = \arg\max_{j \in \{0,1\}} \big|\{i \in X_{k,\text{MAC}} \mid \hat{y}_i = c_k\}\big|,$$

i.e., the one aligned with the majority of positive outputs. We then define $u_{k^*} \in \mathbb{R}^p$ as the concept vector and $X_{k^*,\text{MAC}}$ as its supporting crops. Repeating this across all concept bags yields a dictionary $U = [u_1, \ldots, u_K] \in \mathbb{R}^{p \times K}$ with visual groundings $X_{\text{MAC}} = \{X_{1,\text{MAC}}, \ldots, X_{K,\text{MAC}}\}$.

For textual grounding, we use the LVLM's output function directly. Recall that for a residual activation $a_t^{(l)}$, the model predicts the next token as

$$y_t = h^{(l)}(a_t^{(l)}). \tag{2}$$

Analogously, for each concept feature u_k, we compute its token distribution by applying the same output function:

$$q_k = h^{(l)}(u_k) \in \mathbb{R}^{|V|}$$

where $|V|$ denotes the size of the model's output vocabulary (i.e., the number of distinct tokens in the LVLM). We then select the top-τ tokens (e.g., 50) with the highest scores, remove stopwords and noise, and define the resulting set as $X_{k,\text{MAT}}$ for concept u_k. Collecting across all concepts yields $X_{\text{MAT}} = \{X_{1,\text{MAT}}, \ldots, X_{K,\text{MAT}}\}$.

4.4 Concept Attribution and Multi-modal Alignment

Following Kim et al. (2018), we project residual activations at layer l onto the learned concept subspace $\widehat{U} = [u_1, \ldots, u_K] \in \mathbb{R}^{p \times K}$. At the token level, concept scores are

$$\alpha_j^{(t)} = \cos\_sim(u_j, a_t^{(l)}),$$

while at the phrase level we use the mean-pooled activation $s_m = \frac{1}{T_m} \sum_{t=1}^{T_m} a_t^{(l)}$ to compute

$$\alpha_{m,j} = \cos\_sim(u_j, s_m).$$

Thus $a_t^{(l)}$ provides fine-grained token attribution, whereas s_m captures phrase/sentence-level semantics. The most activated concept is

$$k^* = \arg\max_j \alpha_{m,j},$$

and its groundings $X_{\text{MAC}}[k^*]$ (visual) and $X_{\text{MAT}}[k^*]$ (textual) provide multimodal explanations.

In the *text-only mode*, this procedure further allows us to assess whether purely textual prompts activate the same feature directions as multi-modal inputs, thereby quantifying *multi-modal alignment*. This shared projection space links visual and textual semantics, supporting faithful cross-modal interpretability.

5 Experiments

5.1 Model and Data

We evaluate three recent instruction tuned LVLMs—**Qwen2-VL-7B** Wang et al. (2024), **Qwen2.5-VL-7B** Team (2025b), and **Gemma-3n-E4B** Team (2025a)—keeping all models *frozen* to ensure post hoc interpretability. Results for the two Qwen models are reported in the Appendix E.

For concept learning, we collect **300 examples per class** from ImageNet and MSCOCO, extracting features from the *penultimate norm* layer, shown to yield high-quality embeddings (Parekh et al., 2024; Kim et al., 2018). Evaluation uses a **disjoint set of 50 images per class** from the validation splits of ImageNet (1,000 classes) and MSCOCO (10 randomly selected objects).

Unlike CoX-LMM, which requires **ImageNet** labels and **MSCOCO** caption tokens during extraction, CGDL is weakly supervised: we collect a large pool of concept examples and retain the top 1,000 most frequent concepts for ImageNet and the top 10 for MSCOCO, with each concept bag capped at 1,600 cropped patches. This setup allows us to study scalability across very different concept set sizes, while keeping comparisons fair against CoX-LMM, which requires ground-truth labels for multiple concepts. Images are resized to 500 pixels in width and cropped into 200×200 windows with 0.2 overlap, yielding on average 5–6 crops per image. This makes the effective number of samples per bag comparable to the 300 full images used in CoX-LMM.

We compare **CGDL** against CoX-LMM Parekh et al. (2024) using two dictionary-learning methods: Sparse Autoencoders (SAE) Pach et al. (2025) and Semi-Nonnegative Matrix Factorization (SNMF) Trigeorgis et al. (2014). We also include a simple baseline, **SIMPLE**, which tests whether residuals with the highest l_2-norm align with concepts.

Ground-truth concepts are defined by image class labels, with correctness measured by top-1 alignment. For faithfulness testing, we evaluate on the same 10 MSCOCO objects.

5.2 Evaluation Metrics

Concept discovery can be framed as a special case of dictionary learning. Following the evaluation protocol of Fel et al. (2023a), we first assess the quality of discovered features using three standard metrics: **Sparsity** (↑), **Stability** (↓), and **Overlap** (↓). Based on these results, SNMF emerges as the most suitable method for downstream evaluation.

Scalability is evaluated with both 10 and 1,000 concepts, demonstrating that purity and uniqueness are maintained as the number of concepts grows. Attribution quality Parekh et al. (2024) is measured

using **CLIPScore** and **BERTScore**, which remain standard for text–image and text–text alignment. CLIP Radford et al. (2021) provides cross-modal contrastive alignment, while BERT Devlin et al. (2019) captures contextual semantics through bidirectional encoding.

Faithfulness is assessed using **concept insertion** and **concept deletion** curves Fel et al. (2023a); Kadir et al. (2023), which quantify performance shifts as important concepts are progressively inserted or removed. We report results across the top-1, top-2, and top-3 concepts ranked by their influence on model output.

Finally, qualitative analyses highlight representative extracted concepts and their textual groundings, and illustrate their application to binary classification and text-to-concept alignment.

5.3 RESULTS

Table 2 reports quantitative results for **concept vectors** extracted from **Gemma-3n-E4B** on ImageNet and MSCOCO. We compare CGDL with CoX-LMM across three dictionary-learning settings (SNMF, SAE, SIMPLE). CGDL–SNMF consistently achieves the best performance, with the highest sparsity, lowest stability, and lowest overlap. SAE also benefits from concept guidance, while SIMPLE remains weak with low sparsity and high overlap. These results highlight that concept guidance substantially improves the quality of learned concept vectors, particularly under SNMF Fel et al. (2023a); Parekh et al. (2024).

Method	**Dictionary Learning**	**ImageNet**			**MSCOCO**		
		Spars. ↑	Stab. ↓	Overlap ↓	Spars. ↑	Stab. ↓	Overlap ↓
CoX-LMM	SNMF	0.96	0.13	0.25	**1.00**	0.02	0.16
	SAE	0.84	0.21	0.27	0.97	0.17	0.28
	SIMPLE	0.07	0.79	0.68	0.63	0.91	0.84
CGDL	SNMF	**1.00**	**0.02**	**0.08**	**1.00**	**0.00**	**0.06**
	SAE	0.90	0.16	0.10	**1.00**	0.05	0.16
	SIMPLE	0.52	0.64	0.67	0.80	0.49	0.54

Table 2: Concept vector evaluation on ImageNet and MSCOCO. Higher sparsity is better; lower stability and overlap are better. CGDL–SNMF yields the best results across both datasets.

Table 3 presents attribution results for **Gemma-3n-E4B** on ImageNet and MSCOCO. We evaluate two aspects of alignment: (i) **BERTScore**, which measures semantic correspondence between top-activated concept groundings and ground-truth labels, and (ii) **CLIPScore**, which assesses multimodal consistency between concept groundings and input images. We compare our method (CGDL) against CoX-LMM under three settings: **Text-only**, where activations are probed via class names; **Image-only**, where short visual descriptions are used; and **Combined**, where the model predicts concept presence (c_k vs. `UNK`). Random baselines correspond to assigning concepts uniformly at random, which yields nearly constant values due to dataset distribution. Across datasets and metrics, **CGDL with SNMF** consistently achieves stronger alignment and outperforms CoX-LMM in all modalities, despite requiring substantially less supervision. This advantage holds across both small-scale (10 concepts, MSCOCO) and large-scale (1,000 concepts, ImageNet) evaluations.

We evaluate the *concept attribution ranking* using established *faithfulness metrics* Fel et al. (2023a), namely **concept deletion (C-Deletion)** and **concept insertion (C-Insertion)**, on the MSCOCO validation set. For each image, the model is prompted to classify the input, and the top-3 activated concepts are identified from the residual embeddings (Sec. 4.4). Attribution faithfulness is then measured as follows:

1. **C-Deletion:** progressively set to zero the coordinates corresponding to the most influential concept directions, ranked by gradient magnitude with respect to the highest-probability token, and record the drop in output probability.
2. **C-Insertion:** start from a zero vector and gradually add concept coordinates in the same order, recording the corresponding probability increase.

Table 3: Comparison of CGDL and CoX-LMM on CLIPScore and BERTScore across datasets and concept types. Higher metric values indicate better alignment.

Method	Dataset	Concept	Metric	Random	Text-only	Image-only	Combined
CGDL	ImageNet	1,000	CLIPScore	0.52 ± 0.04	–	**0.62 ± 0.08**	**0.67 ± 0.09**
			BERTScore	0.71 ± 0.06	0.78 ± 0.07	0.84 ± 0.08	0.86 ± 0.10
	MSCOCO	10	CLIPScore	0.48 ± 0.04	–	0.60 ± 0.06	0.64 ± 0.08
			BERTScore	0.71 ± 0.01	**0.89 ± 0.06**	**0.91 ± 0.05**	**0.93 ± 0.07**
CoX-LMM	ImageNet	1,000	CLIPScore	0.49 ± 0.04	–	0.57 ± 0.03	0.58 ± 0.05
			BERTScore	0.71 ± 0.00	0.74 ± 0.08	0.75 ± 0.06	0.82 ± 0.09
	MSCOCO	10	CLIPScore	0.51 ± 0.04	–	0.57 ± 0.10	0.55 ± 0.05
			BERTScore	0.71 ± 0.01	0.83 ± 0.01	0.79 ± 0.09	0.73 ± 0.11

Scores are averaged across tokens for each image, then aggregated over the validation set. Results are summarized in Figure 1.

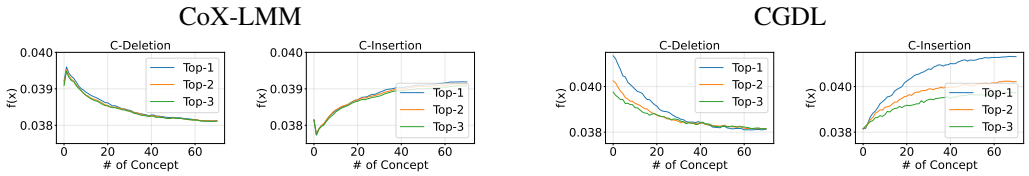

Figure 1: Faithfulness comparison of CoX-LMM and CGDL using concept deletion and insertion. CoX-LMM yields relatively flatter curves with weak separation across ranks, whereas CGDL preserves a clear order (Top-1 >Top-2 >Top-3), with sharper degradation under deletion and stronger recovery under insertion. This shows that CGDL produces more faithful and discriminative concept rankings.

Figure 2: Qualitative examples of concept representations extracted from ImageNet. CoX-LMM (left) concepts often ground to multiple unrelated or overlapping tokens (e.g., “canine” linked to “hot dog”), reflecting polysemantic vectors. In several cases, concepts mix distinct animals: for example, *tiger* grounds across multiple concepts, while *lion* is misrepresented as *tiger*. Such cases illustrate non-monosemantic behavior, where concept vectors capture blended rather than distinct semantics. In contrast, CGDL (right) discovers fine-grained and monosemantic concepts (e.g., fur, dog, stripes).

Figure 2 compares ImageNet concepts extracted by CoX-LMM (left) and CGDL (right). CGDL yields fine-grained, monosemantic representations (e.g., fur, stripes), whereas CoX-LMM produces

entangled groundings that mix semantics (e.g., tiger conflated with lion or multiple animals). Figure 3 shows attribution in three settings: (i) binary classification, (ii) open-ended classification, and (iii) text–image alignment. In each case, attribution is explained by retrieving the nearest concept examples to the residual activation, demonstrating robust multimodal alignment.

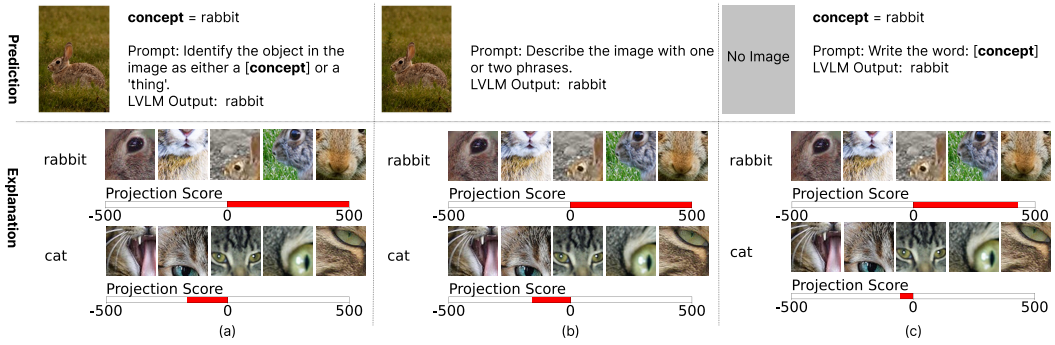

Figure 3: Attribution-based explanation using concepts (a) Aligned image-text concepts yield strong feature attribution. (b) An image-only input without concept information still triggers a relevant feature. (c) Text alone activates semantically meaningful features, demonstrating robust multimodal alignment.

6 Ablation Study

We test multiple variants of the contrastive prompt template (See AppendixB for details). Table 4 shows only minor fluctuations in CLIPScore, indicating robustness to phrasing. Qwen-2 exhibits slightly higher variance than Qwen2.5 and Gemma-3n.

Table 4: Mean ± std. of CLIPScore across binary prompts on MSCOCO-10. Abbreviations: Q-2 = Qwen-2, Q-2.5 = Qwen2.5, G-3n = Gemma-3n.

Prompt	Q-2	Q-2.5	G-3n	Prompt	Q-2	Q-2.5	G-3n
P1	0.57±0.10	0.65±0.13	0.62±0.08	P3	0.57±0.14	0.62±0.07	0.62±0.06
P2	0.58±0.07	0.63±0.11	0.64±0.08	P4	0.59±0.11	0.63±0.08	0.64±0.09

Conclusion

We introduced *Concept-Guided Dictionary Learning* (CGDL), a weakly supervised framework that enforces monosemanticity and grounds concepts directly within LVLMs, yielding faithful multimodal alignment. CGDL is flexible, efficient, and improves concept quality across dictionary learning methods. Limitations include the lack of hierarchical organization and the requirement that models understand basic language instructions—though this holds for most modern LVLMs. Future work will extend CGDL to capture hierarchical concepts and evaluate beyond LVLMs.

Reproducibility Statement We release a reproducible pipeline requiring only a Hugging Face model card, access token, and dataset directory with `train/val` splits. CGDL and CoX-LMM can be run via `scripts/run_full_pipeline.sh` and `scripts/run_full_pipeline_dl.sh`, respectively. Code and configs are shared anonymously at `https://anonymous.4open.science/r/xl-vlms-30C1`, with installation and usage detailed in the `README`. All experiments used a single NVIDIA RTX 3090 (24GB) GPU with fixed random seeds.

Ethics Statement. This work poses no direct societal risks beyond those inherent to model interpretability. Our method may surface biased or harmful concepts present in LVLMs, which should be interpreted responsibly and not used to reinforce stereotypes.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2018. URL https://arxiv.org/abs/1610.01644.

Hasan Md Tusfiqur Alam, Devansh Srivastav, Md Abdul Kadir, and Daniel Sonntag. Towards interpretable radiology report generation via concept bottlenecks using a multi-agentic rag. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part III*, pp. 201–209, Berlin, Heidelberg, 2025. Springer-Verlag. ISBN 978-3-031-88713-0. doi: 10.1007/978-3-031-88714-7_18. URL https://doi.org/10.1007/978-3-031-88714-7_18.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019.

Chris H.Q. Ding, Tao Li, and Michael I. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):45–55, 2010.

Maximilian Dreyer, Jim Berend, Tobias Labarta, Johanna Vielhaben, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. Mechanistic understanding and validation of large ai models with semanticlens, 2025. URL https://arxiv.org/abs/2501.05398.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition, 2022. URL https://arxiv.org/abs/2209.10652.

Thomas Fel, Victor Boutin, Mazda Moayeri, Rémi Cadène, Louis Bethune, Léo andéol, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation, 2023a. URL https://arxiv.org/abs/2306.07304.

Thomas Fel, Agustin Picard, Louis Béthune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2711–2721, June 2023b.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. *CoRR*, abs/2012.14913, 2020. URL https://arxiv.org/abs/2012.14913.

Amirata Ghorbani, James Wexler, James Zou, and Been Kim. *Towards automatic concept-based explanations*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Mara Graziani, An phi Nguyen, Laura O'Mahony, Henning Müller, and Vincent Andrearczyk. Concept discovery and dataset exploration with singular value decomposition. In *ICLR 2023 Workshop on Pitfalls of limited data and computation for Trustworthy ML*, 2023. URL https://openreview.net/forum?id=iOlYmD1PtC8.

Arne Grobrügge, Niklas Kühl, Gerhard Satzger, and Philipp Spitzer. Towards human-understandable multi-dimensional concept discovery, 2025. URL https://arxiv.org/abs/2503.18629.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Md Abdul Kadir, Amir Mosavi, and Daniel Sonntag. Evaluation metrics for xai: A review, taxonomy, and practical applications. In *2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES)*, pp. 000111–000124, 2023. doi: 10.1109/INES59282.2023.10297629.

Weitai Kang, Gaowen Liu, Mubarak Shah, and Yan Yan. Segvg: Transferring object bounding box to segmentation for visual grounding, 2024. URL https://arxiv.org/abs/2407.03200.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), 2018. URL https://arxiv.org/abs/1711.11279.

Hyunsoo Kim and Haesun Park. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM Journal on Matrix Analysis and Applications*, 30(2):713–730, 2008.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. URL https://arxiv.org/abs/2304.02643.

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*, pp. 5338–5348. PMLR, 2020.

Jingjing Liu, Nian Wu, Xianchao Xiu, and Jianhua Zhang. Robust orthogonal nmf with label propagation for image clustering. *arXiv preprint arXiv:2504.21472*, 2025.

Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models, 2024. URL https://arxiv.org/abs/2404.13013.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2023. URL https://arxiv.org/abs/2202.05262.

Tuomas Oikarinen and Tsui-Wei Weng. Clip-dissect: Automatic description of neuron representations in deep vision networks, 2023. URL https://arxiv.org/abs/2204.10965.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. *arXiv preprint arXiv:2504.02821*, 2025.

Jayneel Parekh, Pegah Khayatan, Mustafa Shukor, Alasdair Newson, and Matthieu Cord. A concept-based explainability framework for large multimodal models. *arXiv preprint arXiv:2406.08074*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

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

Erik Strumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41:647–665, 2014. URL https://api.semanticscholar.org/CorpusID:2449098.

Ao Sun, Pingchuan Ma, Yuanyuan Yuan, and Shuai Wang. Explain any concept: Segment anything meets concept-based explanation, 2023. URL https://arxiv.org/abs/2305.10289.

Gemma Team. Gemma 3n. 2025a. URL https://ai.google.dev/gemma/docs/gemma-3n.

Qwen Team. Qwen2.5-vl, January 2025b. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

Adly Templeton. *Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet*. Anthropic, 2024.

George Trigeorgis, Konstantinos Bousmalis, Stefanos Zafeiriou, and Björn W. Schuller. A deep semi-nmf model for learning hidden representations. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pp. II–1692–II–1700. JMLR.org, 2014.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Lei Zhang, Chunyuan Li, and Jianwei Yang. Llava-grounding: Grounded visual chat with large multimodal models, 2023. URL https://arxiv.org/abs/2312.02949.

A Monosemantic vs. Polysemantic Representations

A central challenge in interpreting large vision–language models (LVLMs) lies in *superposition* and *feature entanglement* in high-dimensional residual streams (Elhage et al., 2022). Here, *features* can be understood as vector directions in activation space that encode candidate concepts. Ideally, such vectors should be *monosemantic*—each aligned with a single interpretable concept. In practice, however, LVLMs often learn *polysemantic* vectors, where a single direction is activated by multiple, semantically unrelated concepts.

For instance, in Fig. 4(a), a feature vector f_1 responds to both "cat" and "chair." Such overlap can arise when these concepts frequently co-occur in training data, leading the model to conflate them. When f_1 activates, it is therefore ambiguous whether the cause was the presence of a cat, a chair, or both. This ambiguity breaks the one-to-one mapping between features and concepts, making attribution unreliable.

In contrast, monosemantic features (Fig. 4b) provide clean alignment: e.g.

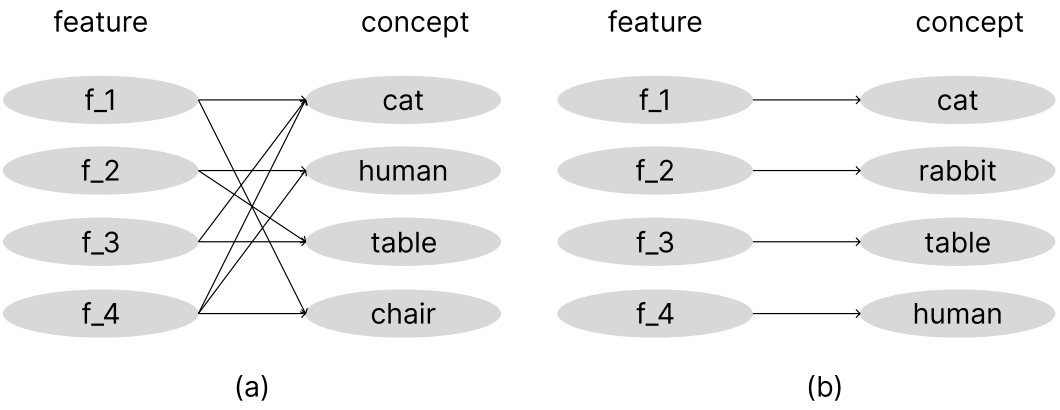

Figure 4: Comparison between prior concept decomposition methods and our proposed approach. (a) Previous methods (e.g., Parekh et al. (2024)) often produce polysemantic features (e.g., f_3 activates for "chair" and "cat"). (b) Our method encourages monosemantic features (e.g., f_1 for "cat," f_3 for "table").

Apart from a limited number of studies Templeton (2024); Pach et al. (2025), most existing models do not offer disentangled representations, and tools for analyzing and extracting monosemantic features remain scarce. This paper addresses this gap through **CGDL**.

B PROMPTS

Concept generation prompt. We use the following instruction to extract candidate concept text from images: *"Identify every visible object, item, concept, and pattern in the image. Output only a single-word, comma-separated list. No explanations or sentences."* This prompt generates concept tokens directly from the dataset without requiring manual labels, thereby creating a concept-to-image mapping.

Concept-Guidance prompt

- P1: Detect whether the image contains c_k. If yes, return c_k; otherwise, return `UNK`.
- P2: Does the image contain c_k? If yes, output c_k; otherwise, output No-c_k.
- P3: Is there a clear instance of c_k in this image? Reply with c_k or `thing`, nothing else.
- P4: Recognize whether the concept c_k is present in the picture. Use only c_k or `UNK` as your answer.

C MODELS

C.1 GEMMA-3N E4B-IT TEAM (2025A)

Gemma-3n E4B-IT (4-billion-parameter model) is trained using a nested subnetwork approach based on the Matryoshka Transformer (MatFormer) architecture. Each Transformer layer supports multiple capacity levels, implemented as top-left submatrices of full-size weight tensors. The model is instruction-tuned on a mixture of multilingual and multimodal data, including text, images, audio, and video inputs. It is trained autoregressively to predict the next token, with a maximum context length of 32k tokens.

C.2 QWEN2.5-VL-7B-INSTRUCTTEAM (2025B) AND QWEN2-VL-7B-INSTRUCT WANG ET AL. (2024)

Qwen2.5-VL-7B-Instruct is a 7-billion parameter multimodal instruction-tuned language model, designed for vision-language tasks. It accepts both text and image inputs and generates text outputs. The model supports a maximum context length of 8192 tokens, enabling it to handle long conversational and reasoning scenarios. Training is performed via supervised instruction tuning on paired text-image datasets. The model is optimized with an autoregressive next-token prediction objective using cross-entropy loss, conditioning on both textual and visual contexts. Large-scale distributed training with mixed precision improves efficiency. This approach enhances the model's instruction-following capabilities and generalization to diverse vision-language tasks.

D DATASET DESCRIPTIONS

We evaluate our models on four datasets: **ImageNet**, **MSCOCO (10 classes)**, **CIFAR100**, and **DTD (Describable Textures Dataset)**.

- **ImageNet**: A large-scale visual dataset with 1000 object categories and high-resolution images, commonly used for image classification tasks.
- **MSCOCO (10 classes)**: A subset of the MSCOCO dataset with 10 object categories, featuring complex scenes and multiple annotated objects per image.

E RESULTS

E.1 QWEN2

Table 6 reports the performance of **CGDL** and **CoX-LMM** on four benchmark datasets. We evaluate alignment using **CLIPScore (CS)** and **BERTScore (BS)**, where higher is better.

Dataset	Total Classes	Train Samples/Class	Val Samples/Class
ImageNet	1000	300	50
MSCOCO (10)	10	300	50
CIFAR100	100	300	100
DTD	47	95	24

Table 5: Overview of datasets used for training and validation, including number of classes and samples per class.

Table 6: Comparison of **CGDL** and **CoX-LMM** across datasets. Metrics: CS = CLIPScore, BS = BERTScore. Higher is better. Best non-random results are **bolded**.

Method	Dataset	#C	Metric	Rand	Text	Img	Comb
CGDL	ImageNet	1k	CS	0.53 ± 0.02	–	**0.62 ± 0.07**	**0.63 ± 0.08**
			BS	0.80 ± 0.05	**0.86 ± 0.06**	**0.88 ± 0.08**	**0.86 ± 0.09**
	CIFAR100	100	CS	0.51 ± 0.02	–	**0.63 ± 0.04**	**0.63 ± 0.05**
			BS	0.83 ± 0.08	**0.86 ± 0.07**	**0.87 ± 0.08**	**0.91 ± 0.08**
	DTD	47	CS	0.53 ± 0.06	–	**0.63 ± 0.05**	**0.62 ± 0.05**
			BS	0.74 ± 0.04	**0.84 ± 0.06**	**0.83 ± 0.07**	**0.87 ± 0.06**
	MSCOCO	10	CS	0.52 ± 0.03	–	**0.64 ± 0.06**	**0.62 ± 0.06**
			BS	0.82 ± 0.02	**0.88 ± 0.06**	**0.89 ± 0.05**	**0.94 ± 0.08**
CoX-LMM	ImageNet	1k	CS	0.53 ± 0.03	–	0.54 ± 0.03	0.53 ± 0.05
			BS	0.82 ± 0.04	0.82 ± 0.03	0.84 ± 0.04	0.86 ± 0.09
	CIFAR100	100	CS	0.53 ± 0.04	–	0.53 ± 0.06	0.53 ± 0.05
			BS	0.78 ± 0.06	0.84 ± 0.06	0.80 ± 0.03	0.73 ± 0.07
	DTD	47	CS	0.51 ± 0.05	–	0.54 ± 0.05	0.52 ± 0.05
			BS	0.80 ± 0.07	0.84 ± 0.06	0.83 ± 0.07	0.77 ± 0.07
	MSCOCO	10	CS	0.53 ± 0.03	–	0.57 ± 0.04	0.58 ± 0.06
			BS	0.82 ± 0.04	0.83 ± 0.01	0.83 ± 0.05	0.82 ± 0.03

Across all datasets, CGDL consistently outperforms CoX-LMM. Notably, on **MSCOCO**, CGDL achieves the strongest gains: BS improves from 0.82 (CoX-LMM) to **0.94**, and CS improves from 0.58 to **0.64**. Similarly, on **CIFAR100**, the BS of CoX-LMM drops to 0.73 in the combined setting, while CGDL achieves a significantly higher **0.91**. These results highlight the robustness of CGDL in both low- and high-concept regimes.

E.2 QWEN2.5

Table 7 shows results for the Qwen2.5 backbone. Again, CGDL achieves the strongest improvements across all datasets. In particular, on **MSCOCO**, CGDL improves BS from 0.90 to **0.95** in the combined setting. On **CIFAR100**, CGDL reaches **0.92**, compared to 0.87 with CoX-LMM. These consistent gains indicate that CGDL scales effectively from small (10 concepts) to large-scale (1k concepts) benchmarks.

Method	Dataset	#C	Metric	Rand	Text	Img	Comb
CGDL	ImageNet	1k	CS	0.51 ± 0.06	–	**0.63 ± 0.06**	**0.64 ± 0.05**
			BS	0.82 ± 0.04	**0.87 ± 0.01**	**0.88 ± 0.07**	**0.87 ± 0.07**
	MSCOCO	10	CS	0.53 ± 0.05	–	**0.64 ± 0.07**	**0.64 ± 0.08**
			BS	0.83 ± 0.06	**0.89 ± 0.07**	**0.90 ± 0.06**	**0.95 ± 0.08**
	CIFAR100	100	CS	0.50 ± 0.07	–	**0.62 ± 0.05**	**0.64 ± 0.06**
			BS	0.80 ± 0.08	**0.88 ± 0.06**	**0.88 ± 0.09**	**0.92 ± 0.07**
	DTD	47	CS	0.51 ± 0.06	–	**0.63 ± 0.08**	**0.63 ± 0.07**
			BS	0.79 ± 0.07	**0.87 ± 0.07**	**0.85 ± 0.09**	**0.89 ± 0.08**
CoX-LMM	ImageNet	1k	CS	0.51 ± 0.05	–	0.56 ± 0.04	0.55 ± 0.06
			BS	0.81 ± 0.06	0.83 ± 0.04	0.85 ± 0.06	0.86 ± 0.03
	CIFAR100	100	CS	0.53 ± 0.07	–	0.58 ± 0.06	0.57 ± 0.04
			BS	0.78 ± 0.07	0.84 ± 0.06	0.85 ± 0.05	0.87 ± 0.06
	MSCOCO	10	CS	0.52 ± 0.03	–	0.60 ± 0.03	0.60 ± 0.02
			BS	0.81 ± 0.04	0.88 ± 0.04	0.90 ± 0.04	0.90 ± 0.07
	DTD	47	CS	0.53 ± 0.03	–	0.56 ± 0.03	0.56 ± 0.06
			BS	0.81 ± 0.04	0.83 ± 0.05	0.82 ± 0.06	0.88 ± 0.05

Table 7: Comparison of CGDL and CoX-LMM across datasets using CLIPScore (CS) and BERTScore (BS). Best non-random results are **bolded**.