



Toward Informative, Training-free, and Task-agnostic Inherently Interpretability based on Pre-trained Vision-Language Models

20, May 2026

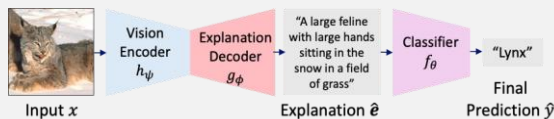
Interpretable Deep Learning Seminar

Shin'ya Yamaguchi, PhD

❖ Introducing recent our works on inherently interpretable models based on pre-trained vision-language models

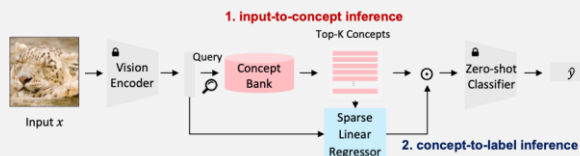
- ▷ Solving challenges of concept bottleneck models (CBMs) in *expressiveness*, *training-cost*, and *task-constraint*

XBM (AAAI'25)



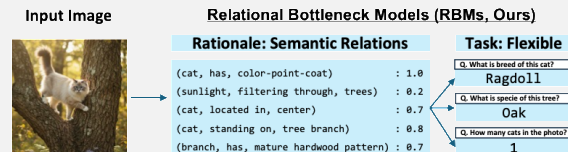
Can we provide natural language explanations beyond concepts?

Z-CBM (ICME'26)



Can we build CBMs w/o ANY target datasets and training?

RBM (On-going)



Can we build inherently interpretable models available for ANY tasks?

❖ **Background: Inherently Interpretable Models**

- ▷ Concept Bottleneck Models (CBMs)
- ▷ Vision-Language Models and CBMs
- ▷ Challenges and Research Questions

❖ **Toward Next-gen. Inherently Interpretable Models**

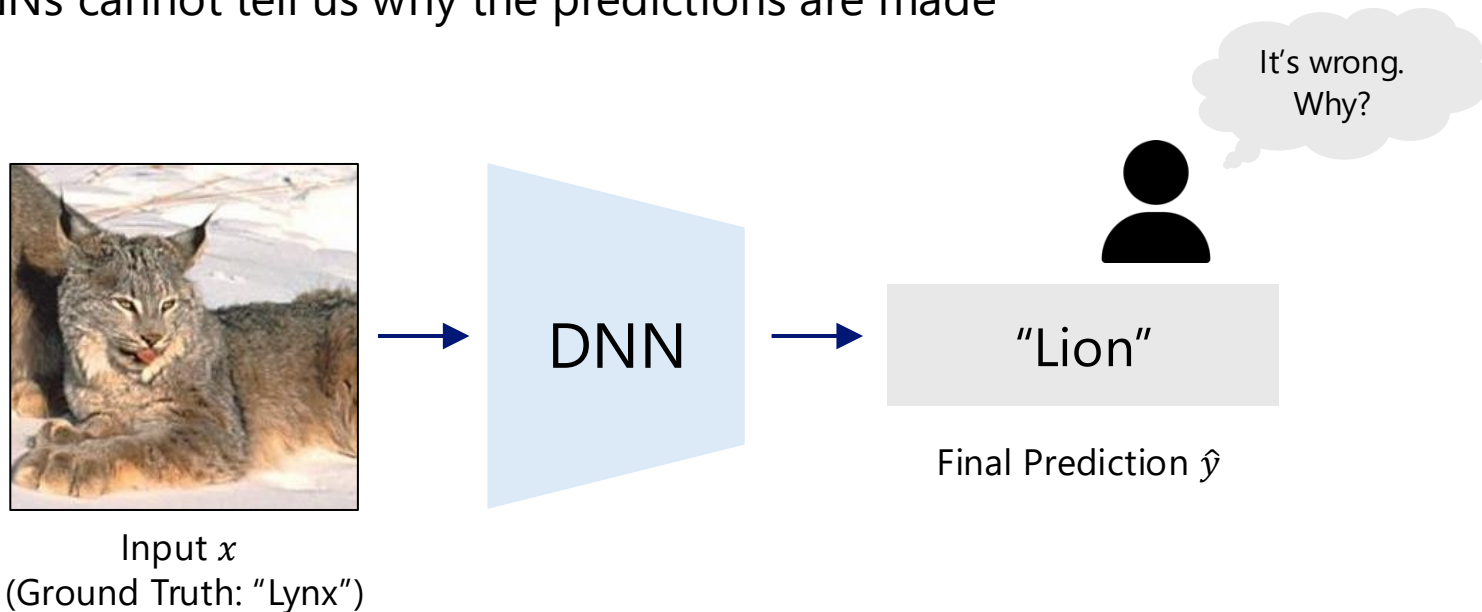
- ▷ Explanation Bottleneck Models (XBMs, AAI 2025)
- ▷ Zero-shot Concept Bottleneck Models (Z-CBMs, ICME 2026)
- ▷ Relational Bottleneck Models (RBMs, On-going)

❖ **Conclusion & Future Work**

Background: Inherently Interpretable Models

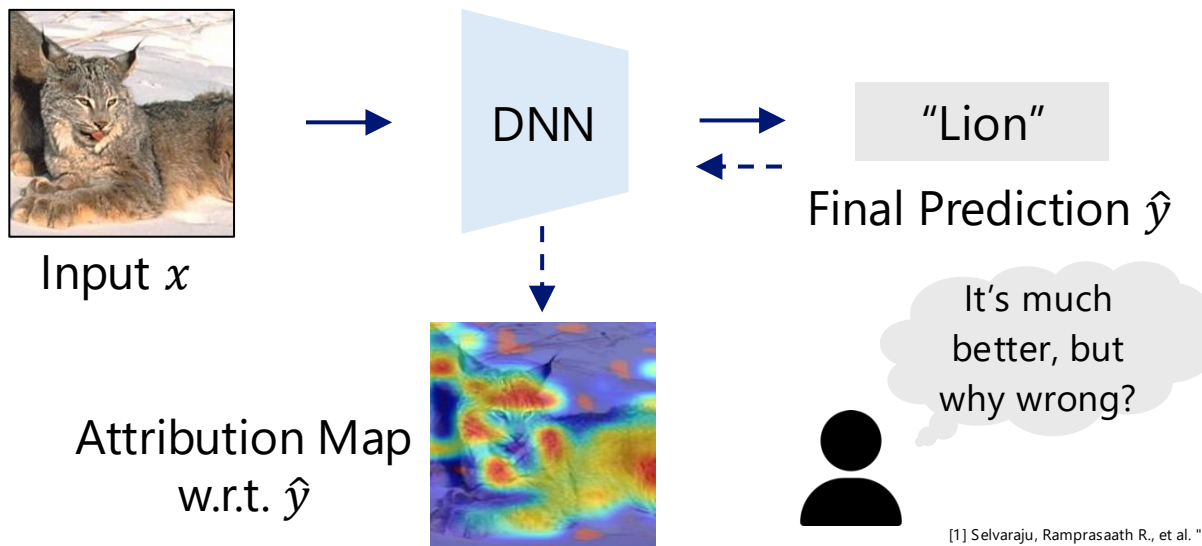
❖ High performance but less interpretable

- ▷ Naïve DNNs cannot tell us why the predictions are made



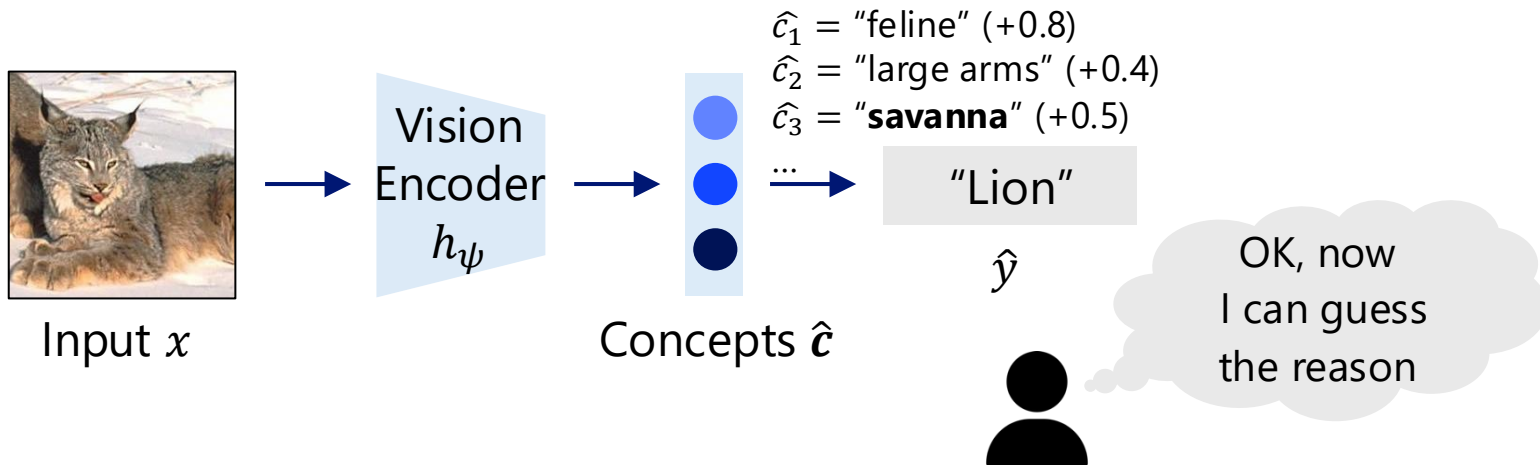
Improving Interpretability of DNNs

- ❖ Explanation via Visual Attribution Map (e.g., GradCAM [1])
 - ▷ Visualizing contributions to final prediction in input spaces
 - ▷ It hardly provides **semantic meanings** of the regions



Concept Bottleneck Models (CBMs)

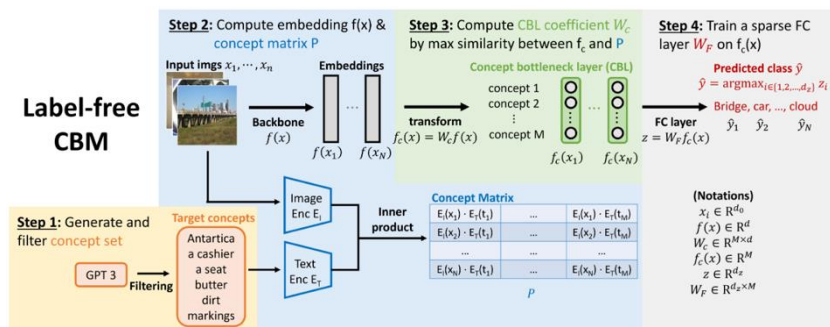
- ❖ Inherently interpretable models via intermediate concept prediction [2]
 - ▷ Concept := Ingredient of data (provided by annotations)
 - ▷ Ensuring interpretability by *input-to-concept* & *concept-to-output* predictions
 - ▷ It provides more semantic explanation by *pre-defined* concepts (noun phrases)



CBMs and Vision-Language Models

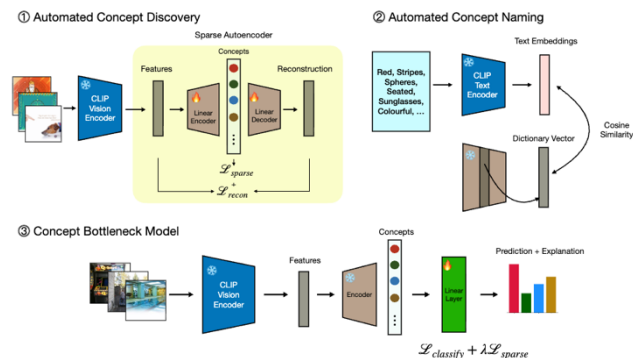
- ❖ Recent CBMs are built upon foundation VLMs like CLIP encoders
 - ▷ Enabling learning *input-to-concept* mapping w/o manual concept annotations by embedding LLM-generated concepts on the multi-modal text encoder

Label-free CBMs (ICLR'23, [3])



[3] Oikarinen, Tuomas, et al. "Label-free concept bottleneck models." ICLR 2023.

DN-CBMs (ECCV'24, [4])



[4] Rao, Sukrut, et al. "Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery." ECCV 2024.

Challenges

① Limited Expressiveness of Explanation

CBMs rely on pre-defined to represent the explanations for output → limited vocabulary



② Additional Training Cost

CBMs require training with labeled datasets → data collection and GPU computation cost



③ Task Dependency

CBMs' applicability is restricted to their learned tasks → limited transferability to unseen tasks



Research Question

① Toward Informative Explanation

Can we provide natural language explanations beyond concepts?

② Toward Training-free CBMs

Can we build CBMs w/o ANY target datasets and training?

③ Toward Task-agnostic Interpretability

Can we build inherently interpretable models available for ANY tasks?

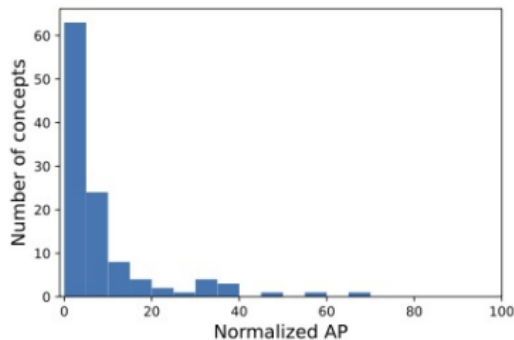
Explanation Bottleneck Models (XBMs, AAAI 2025)

SY, and Kosuke Nishida. "Explanation bottleneck models."
Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 39. No. 20. 2025.

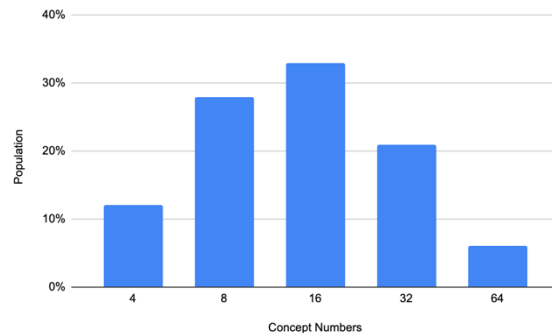
❖ Dependency on Pre-defined Concepts

- ▷ Vocabulary for rationale explanation is limited
- ▷ The number of concepts is hard to increase
 - Increasing pre-defined concepts makes training difficult [5]
 - Providing too many concepts is less interpretable [5]

Concepts are hard to be learned



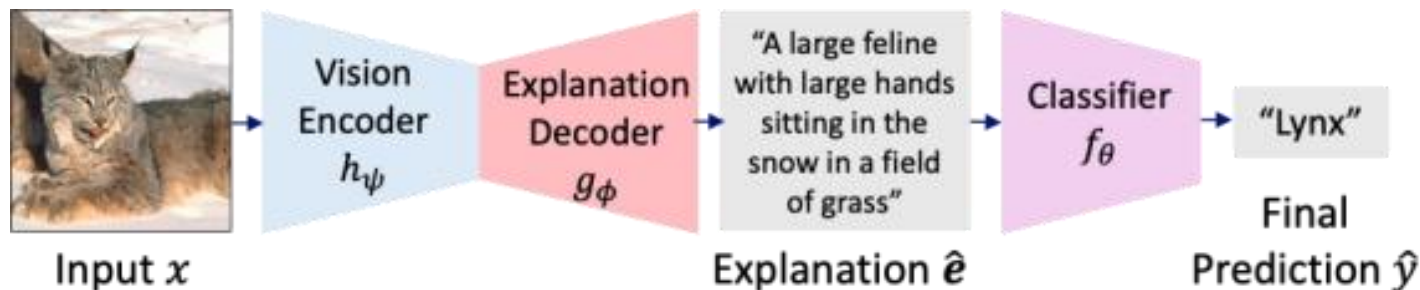
Human prefers concepts ≤ 16



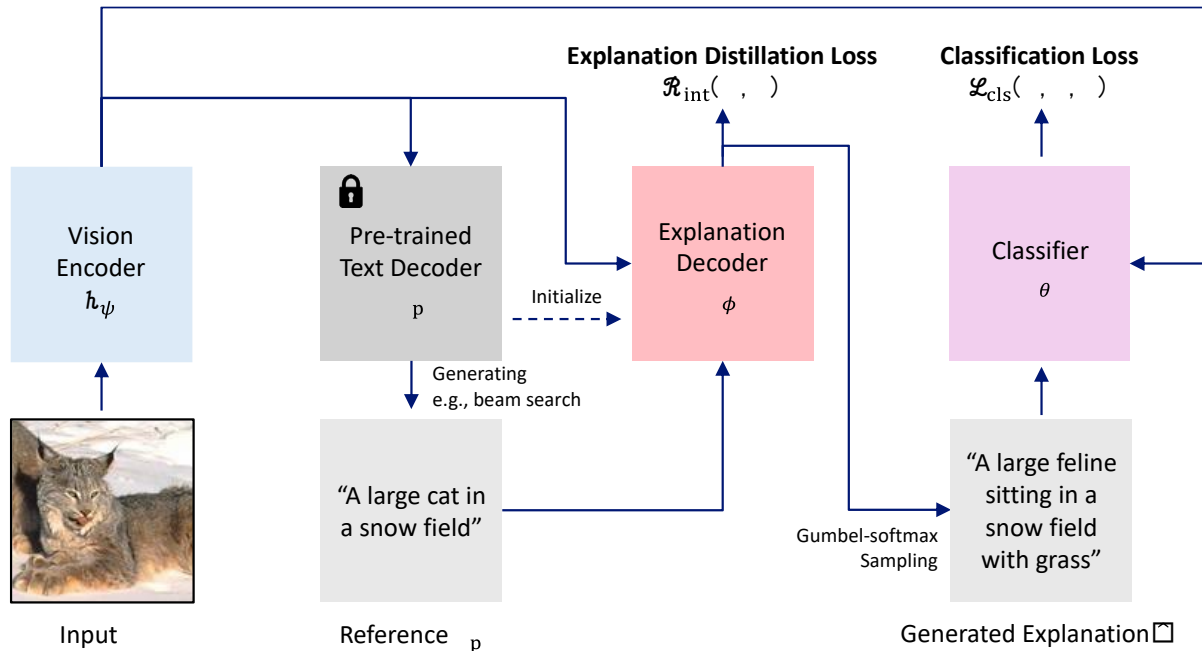
❖ Explanation Bottleneck Models (XBMs)

▷ Key Idea: Providing explanations in **natural language**

- Generating text explanations from input → predicting final labels
- Leveraging strong pre-trained vision-language encoder-decoder like BLIP [6]



- ❖ Optimizing encoder h_ψ , decoder g_ϕ , classifier f_θ via target task (classification) and **explanation distillation**



Objective Function

$$\min_{\theta, \phi, \psi} \mathcal{L}_{cls}(\theta, \phi, \psi) + \lambda \mathcal{R}_{int}(\phi, \psi)$$

$$\mathcal{L}_{cls}(\theta, \phi, \psi) = \mathbb{E}_{(x,y) \in \mathcal{D}} \ell_{CE}(f_\theta(g_\phi(h_\psi(x))), y)$$

$\mathcal{R}_{int}(\phi, \psi)$: For Maintaining text quality

- ❖ Challenge of naïve optimization only with \mathcal{L}_{cls}
 - ▷ Text decoder g_ϕ ignores structures of natural language
 - Output texts become broken and meaningless
- ❖ Our Idea: Penalizing model distribution $p_\phi(\mathbf{e})$ by pre-trained model's one $q(\mathbf{e})$

$$\mathcal{R}_{\text{int}}(\phi, \psi) = D_{\text{KL}}(q \parallel p_\phi) = \sum_{\mathbf{e} \in \mathcal{E}} q(\mathbf{e}|x) \log \left(\frac{q(\mathbf{e}|x)}{p_\phi(\mathbf{e}|x)} \right) = \mathbb{E}_{\mathbf{e} \sim q(\mathbf{e}|x)} \log \left(\frac{q(\mathbf{e}|x)}{p_\phi(\mathbf{e}|x)} \right)$$

- ▷ To avoid back prop via all sampling process, approximate it by sequence-level knowledge distillation [7]

$$\mathcal{R}_{\text{int}}(\phi, \psi) \approx - \sum_{\mathbf{e} \in \mathcal{E}} \mathbb{I}_{\mathbf{e}=\mathbf{e}_p} \log p_\phi(\mathbf{e}|x) = -\log p_\phi(\mathbf{e} = \mathbf{e}_p|x)$$

XBM's Explanation Styles

❖ XBMs provide three-styles of explanations

- ▷ Explanation Text \hat{e} \rightarrow Output of decoder g_ϕ
- ▷ Concept Phrase c \rightarrow Generated from \hat{e} by an NLP parser
- ▷ Cross-Attn. Map \rightarrow Visualizing cross-attn. in classifier f_θ



\hat{e} = "A white large truck parked in a parking lot with a building in the background"

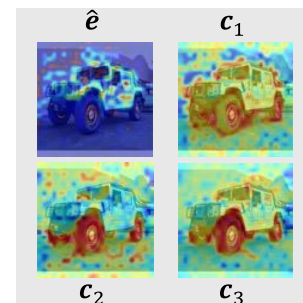
(i) Text Explanation

c_1 = "A white large truck" (0.48)

c_2 = "a parking lot" (0.23)

c_3 = "a building in the background" (0.20)

(ii) Concept Phrase w/ Score

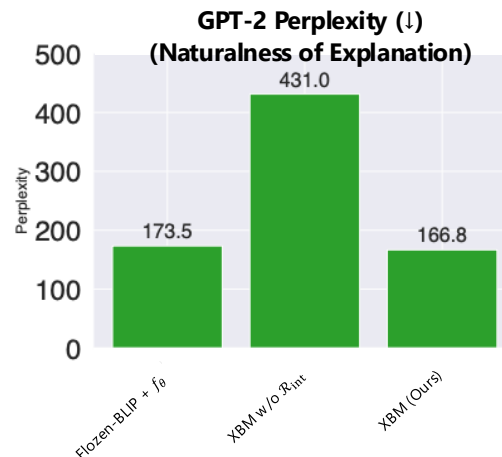
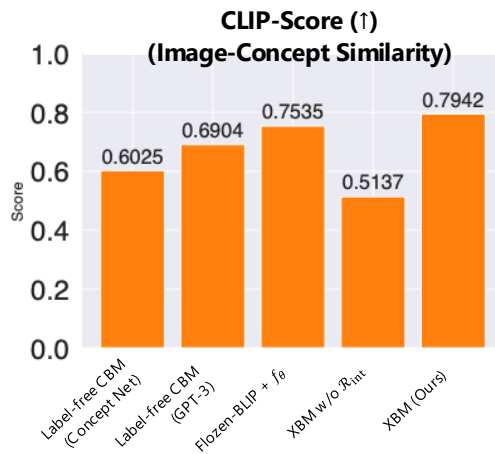
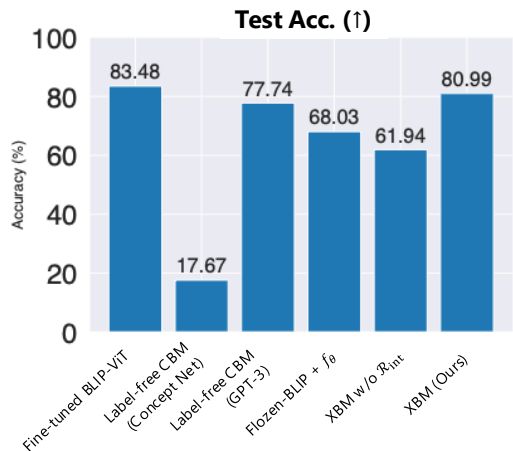


(iii) Cross-Attention Heatmap



Quantitative Evaluation

- ❖ XBMs outperformed Label-free CBMs in Acc. and Interpretability
 - ▷ It also improves Frozen-BLIP → Optimization is valuable
 - ▷ Ablation Study → \mathcal{R}_{int} helps to avoid explanation collapse

Performance Evaluation on Target Dataset: Bird (CUB)





- ❖ XBMs focus on more detailed information on classifying labels
 - ▷ Penalizing by \mathcal{R}_{int} plays crucial role to ensure the sanity of output

	Bird (Yellow Bellied Flycatcher)	ImageNet (Lynx)
		
Pre-trained BLIP (Caption)	A bird perched on a wire fence with leaves on the ground and a blurry background.	Cat walking through the grass in the woods at night with it's eyes open.
Label-free CBMs (Top-3 Concept)	olive-colored sides (0.77) green head (0.55) a small, green body (0.52)	feline (0.98) long, sharp claws (0.53) mau (0.17)
XBMs w/ o R_{int} (Text Explanation)	2222222222222 2222222222	when when when when when when when when when when
XBMs (Text Explanation)	A small green and yellow bird perched on a wire fence with leaves on the side.	Furry feline walking in the woods at night with its eyes open and one paw on the ground.

Output Transition during Training

❖ XBMs gradually focus on detailed features according to optimization

<u>Input</u> (California Gull)	<u>Epoch 0</u>	<u>Epoch 20</u>	<u>Epoch 40</u>
	"Someone standing on a rock in front of the water"	"A seagull standing on a rock by the water's edge"	"A seagull with a yellow beak standing on beach"
	"Animals that are standing on the sand near the water"	"A seagull standing on a beach next to a bunch of sea lions"	"A seagull with a beak on a beach next to a group of sea lions"
Top-10 Word Occurrence	seagull, standing, water, beach, looking, back, body, sand, rock, sky	seagull, standing, water, beach, back, body, sky, sand, rock, grass	seagull, standing, water, beak , beach, back, body, yellow , sky, mouth

❖ Explanation Bottleneck Models (XBMs)

- ▷ Generating text explanation without pre-defined concepts
- ▷ Penalizing text decoders by explanation distillation
- ▷ The optimization may enhance multi-modal understanding
- ▷ Experimental analysis confirms that XBMs can provide accurate and high-quality explanation qualitatively and quantitatively



Zero-shot Concept Bottleneck Models (Z-CBMs, ICME 2026)

SY, Kosuke Nishida, Daiki Chijiwa, and Yasutoshi Ida. "Zero-shot Concept Bottleneck Models." *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2026

Challenge | Additional Training Cost



❖ CBMs require target task training with concept annotations

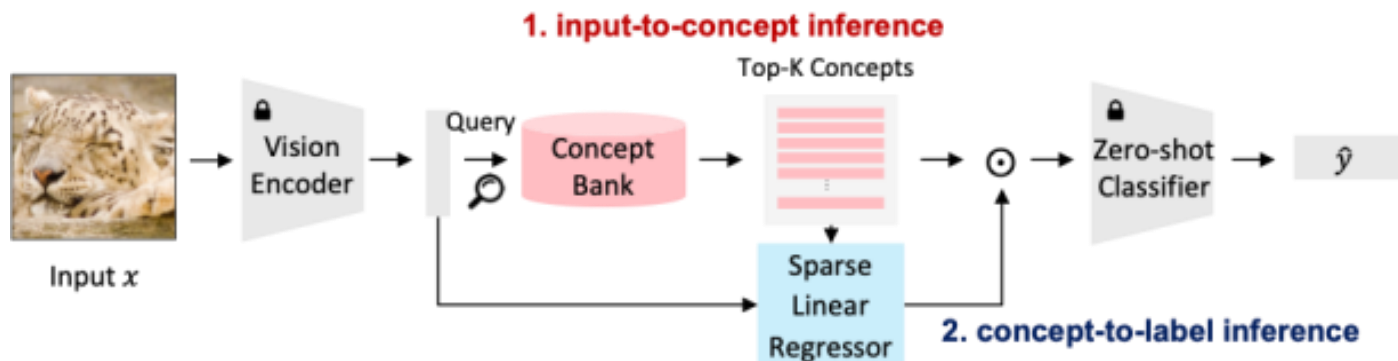
- ▷ Learning mappings of *input-to-concept* and *concept-to-output* with neural network parameters
- ▷ We can avoid the manual concept annotations with LLMs/VLMs, but **it still requires labeled target datasets**

Can we build CBMs w/o ANY target datasets and training?

Proposed Method | Z-CBMs

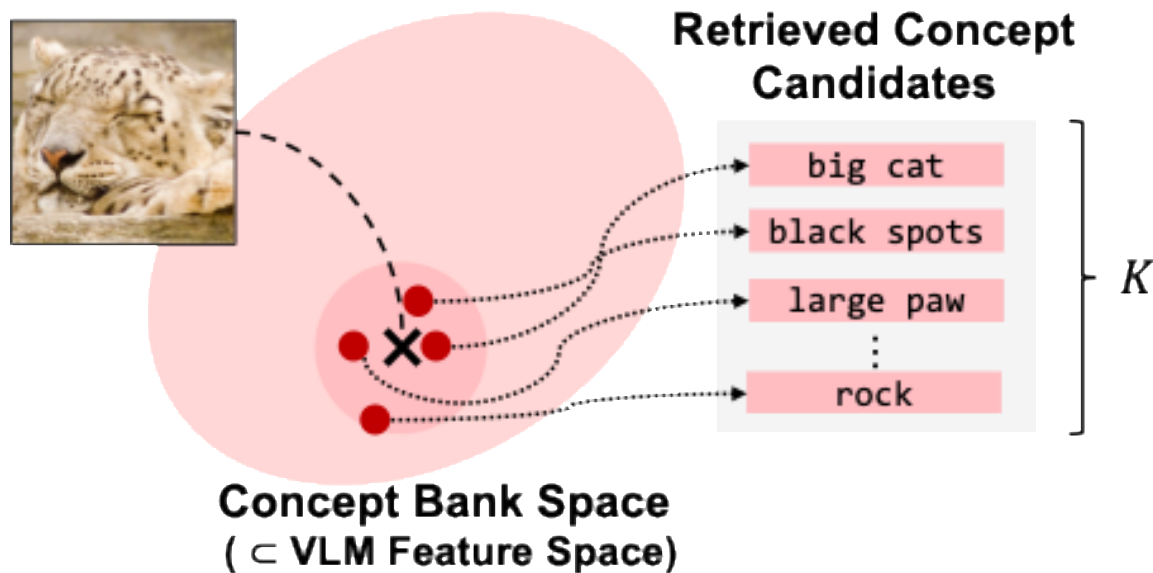
❖ Zero-shot Concept Bottleneck Models (Z-CBMs)

- ▷ Idea: Combining VLMs and large concept bank for zero-shot inference
- ▷ Providing input-to-concept and concept-to-label predictions via
 - **Concept Retrieval**: Searching candidates from containing millions concepts
 - **Concept Regression**: Estimating concept contributions via sparse regression



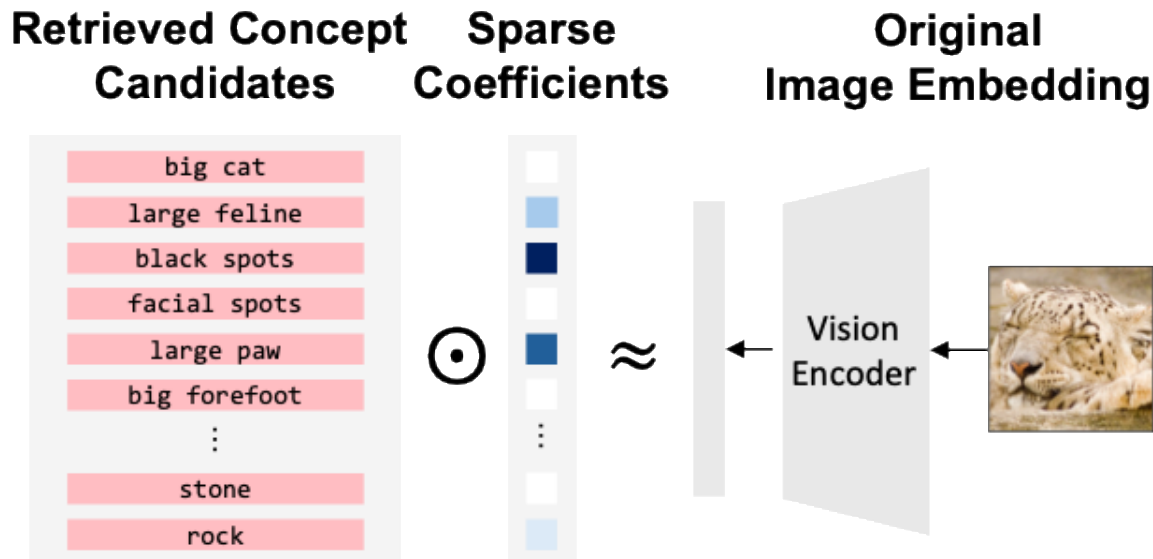
Z-CBM | Concept Retrieval

- ❖ Retrieving top- K concept candidates from input image
 - ▷ Supported by efficient vector search engine (e.g., faiss)



Z-CBM | Concept Regression

- ❖ Selecting essential concepts and estimating their contributions by sparse regression (e.g., lasso) as the coefficients
 - Reducing the redundancy in predicted concepts



- ❖ Final prediction is done by cosine similarity between label text vectors and the weighted sum of concept vectors

Algorithm 1 Zero-shot Inference of Z-CBMs

Require: Input x , concept bank C , image encoder f_V , text encoder f_T

Ensure: Predicted label \hat{y} , concepts C_x , importance weight W_{C_x}

- 1: # Retrieving top-K concepts from input
 - 2: $C_x \leftarrow \underset{c \in C}{\text{Ret}_K}(f_V(x), f_T(c))$
 - 3: $F_{C_x} \leftarrow [f_T(c_1), \dots, f_T(c_K)]$
 - 4: # Predicting importance weights by sparse linear regression
 - 5: $W_{C_x} \leftarrow \arg \min_{W \in \mathbb{R}^K} \|f_V(x) - F_{C_x} W\|_2^2 + \lambda \|W\|_1$
 - 6: # Predicting label by importance weighted sum concept vectors
 - 7: $\hat{y} \leftarrow \arg \max_{y \in \mathcal{Y}} \text{Sim}(F_{C_x} W_{C_x}, f_T(t_y))$
-

- ❖ Z-CBM can predict valid concepts that related to input images
 - ▷ High SigLIP-Score: Z-CBM can predict more related concepts to input image than training-based CBMs
 - ▷ High Concept Recall: Sparse regression is key for providing effective concepts
 - ▷ Sensitivity to Intervention: Z-CBM are rigorously based on content of concepts

Table 2. SigLIP-Score (Image-Concept Similarity)

Method	Avg. of 12 datasets
Label-free CBM	0.5485
LaBo	0.5419
CDM	0.5714
Z-CBM (ALL)	0.6309

Table 3. Concept Recall (%) (Coverage of GT Concepts)

Method	Avg. of 12 datasets
Z-CBM (Cosine Similarity)	58.51
Z-CBM (Linear Regression)	76.87
Z-CBM (Lasso)	85.27

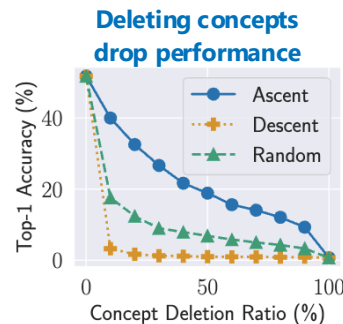


Figure 3. Concept Deletion

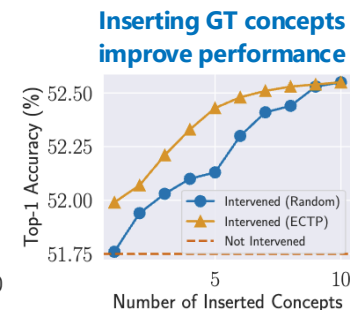
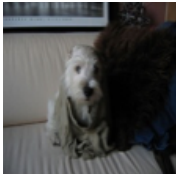




Figure 4. Concept Insertion

❖ Z-CBM focuses on various concepts in input images

▷ Sparse regression contributes to the redundancy (e.g., concepts about “coat”)

	Label-free CBM	Z-CBM (Linear Reg.)	Z-CBM (Lasso)
GT: West Highland White Terrier 	Predicted: Pajamas bath robe (3.62) matching trouser (2.01) apparel (0.69) baby product (0.16) sewing pattern (0.12)	Predicted: Sealyham Terrier wheaten terrier mix (2.650) white wheaten dog (2.474) dog costume (2.174) dog coat (2.027) animal coat (2.024)	Predicted: West Highland White Terrier NOT maltese dog terrier (0.433) beige blanket coat (0.412) white wheaten dog (0.389) modern sofa (0.269) cosy doggy jumper (0.247)
GT: Toy Store 	Predicted: Toy Store a comic store (4.38) a variety of toys (3.74) toys (0.04) retailer (0.04) soft toys (0.01)	Predicted: Toy Store game collections (1.869) game room (1.683) comic book place (1.569) comic room (1.490) NOT banner store (1.435)	Predicted: Toy Store merchandise displays (0.538) hobby store (0.531) comic store (0.511) displayed toys (0.396) store view (0.363)
GT: Chain Link Fence 	Predicted: Hook locking shackle (1.35) a rod (1.33) hangs from a wire (0.97) a handle for leverage (0.41) a loop for hanging (0.38)	Predicted: Turnstile NOT macro rope (2.408) NOT rope (2.117) macro rust (2.040) barded wire tree (1.933) NOT sambucas stem (1.885)	Predicted: Chain pruned branch (0.441) rust steel (0.367) iron railing (0.343) NOT macro rope (0.332) curly branch (0.327)

- ❖ Z-CBMs achieve competitive performance to training-based CBMs
 - ▷ It stably performs on any backbone VLMs
 - ▷ Choice of concept bank is crucial to achieve better performance

Table 4. Top-1 accuracy on 12 classification datasets with CLIP ViT-B/32. Complete results appear in Appendix D-A.

Setting	Method	Avg. of 12 datasets
Zero-Shot	Zero-shot CLIP	53.73
	ConSe	10.82
	Z-CBM (Flickr30K)	52.62
	Z-CBM (CC3M)	52.98
	Z-CBM (CC12M)	53.97
	Z-CBM (YFCC15M)	53.94
	Z-CBM (ALL)	54.28
Training Head	Linear Probe CLIP	78.98
	Label-free CBM	74.87
	LaBo	74.04
	CDM	76.39
	LP-Z-CBM (ALL)	78.31

Table 5. Performance of Z-CBMs varying backbone VLMs on ImageNet.

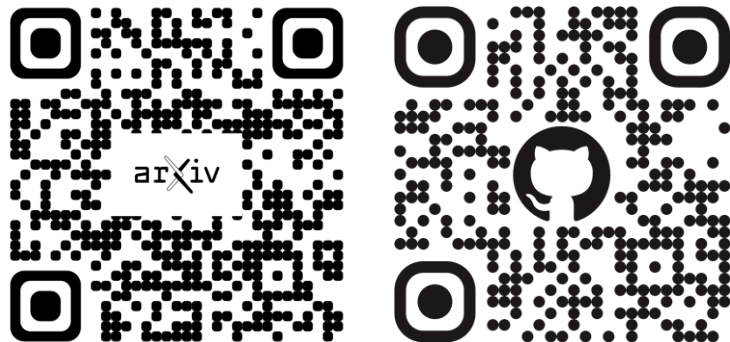
Backbone VLM	Top-1 Acc. (Black Box)	Top-1 Acc. (Z-CBM)	SigLIP-Score (Z-CBM)
CLIP ViT-B/32	61.88	62.70	0.6498
CLIP ViT-L/14	72.87	73.19	0.6608
OpenCLIP ViT-H/14	77.20	77.81	0.6790
OpenCLIP ViT-G/14	79.03	78.27	0.6810
DFN ViT-H/14	83.85	83.40	0.7038

Table 6. Performance of Z-CBMs varying concept banks on ImageNet with CLIP ViT-B/32.

Concept Bank	Vocab. Size	Top-1 Acc.	SigLIP-Score
Zero-shot CLIP	N/A	61.88	N/A
Label-free CBM w/ GPT-3 (ImageNet Class)	4K	58.00	0.5896
CDM w/ GPT-3 (ImageNet Class)	4K	62.52	0.6193
GPT-3 (ImageNet Class)	4K	59.18	0.5407
Noun Phrase (Flickr30K)	45K	61.52	0.5539
Noun Phrase (CC3M)	186K	62.38	0.5904
Noun Phrase (CC12M)	2.58M	62.42	0.6242
Noun Phrase (YFCC15M)	2.20M	62.45	0.6375
Noun Phrase (ALL)	5.12M	62.70	0.6498

❖ Zero-shot Concept Bottleneck Models (Z-CBMs)

- ▶ Inferring input-to-concept and concept-to-output mapping through **concept retrieval** and **concept regression**
- ▶ Pre-trained VLMs, concept bank, sparse regression are keys
- ▶ This work opens a new problem setting and Z-CBMs can be a good baseline in this setting



Relational Bottleneck Models (RBMs, On-going)

SY, Kosuke Nishida. "Relational Bottleneck Models: Inherently Interpretable Decoding for Large Vision Language Models." TBA. 2026

Remaining Challenges of CBMs

❖ Expressiveness

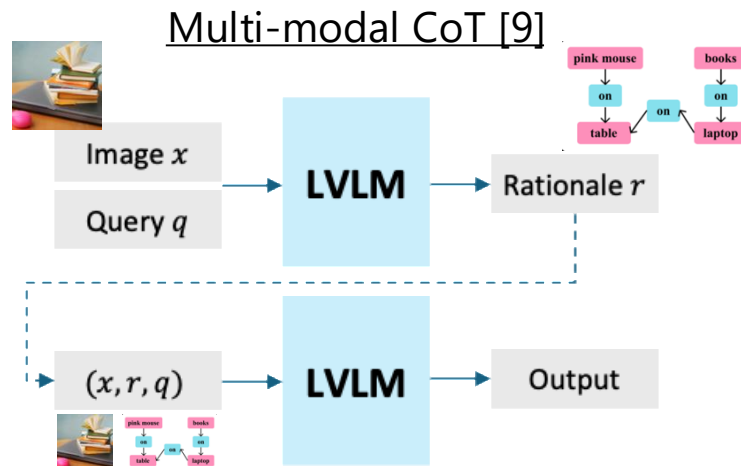
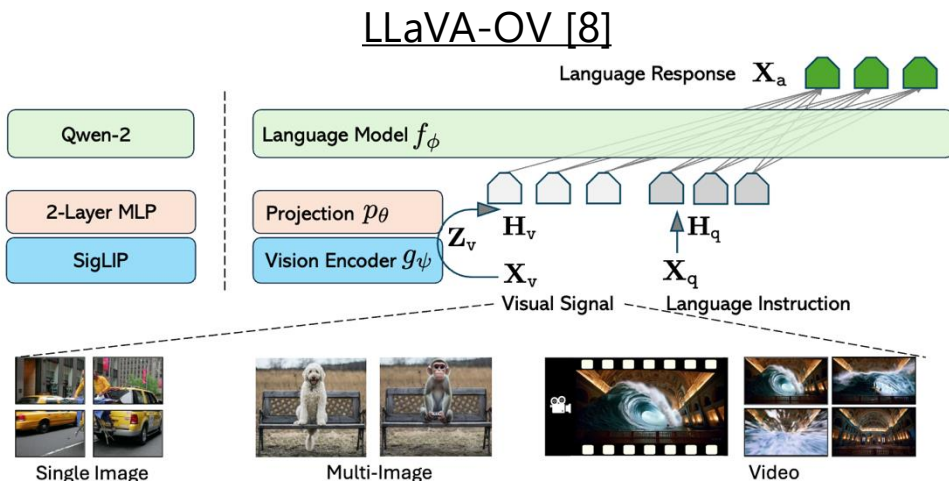
- ▷ CBMs use individual concepts and hardly capture complex **semantic relations**
- ▷ XBMs represent the rationales by free text, but it's not **structured**

❖ Generalizability

- ▷ Applicable tasks are restricted to **specific tasks** (e.g., classification)
- ▷ Z-CBMs can alleviate this limitation, but it is still limited to **discriminative** tasks

Can we build inherently interpretable models available for ANY tasks with semantic structured rationales?

- ❖ LLMs extended to accept input image
 - ▷ Post-trained on diverse vision-language datasets
 - **Emerging zero-shot reasoning capability for unseen visual tasks**
 - ▷ Reasoning capability can be enhanced by Chain-of-Thought (CoT)

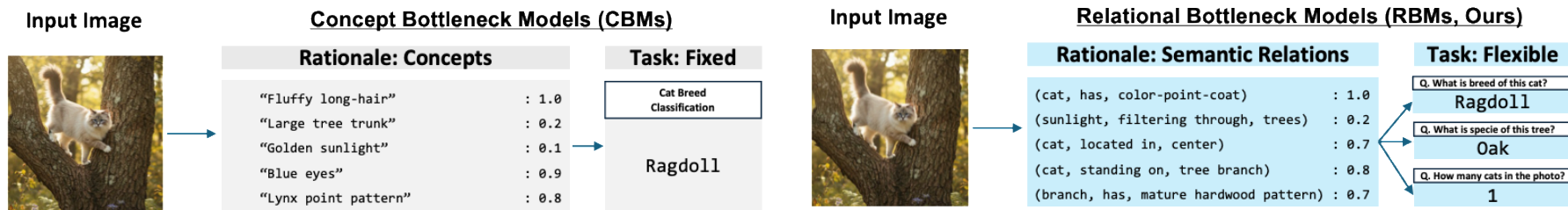


[8] Li, Bo, et al. "Llava-ovision: Easy visual task transfer." *arXiv preprint arXiv:2408.03326* (2024)

[9] Mitra, Chanchrik, et al. "Compositional chain-of-thought prompting for large multimodal models." *CVPR* 2024

❖ Relational Bottleneck Models

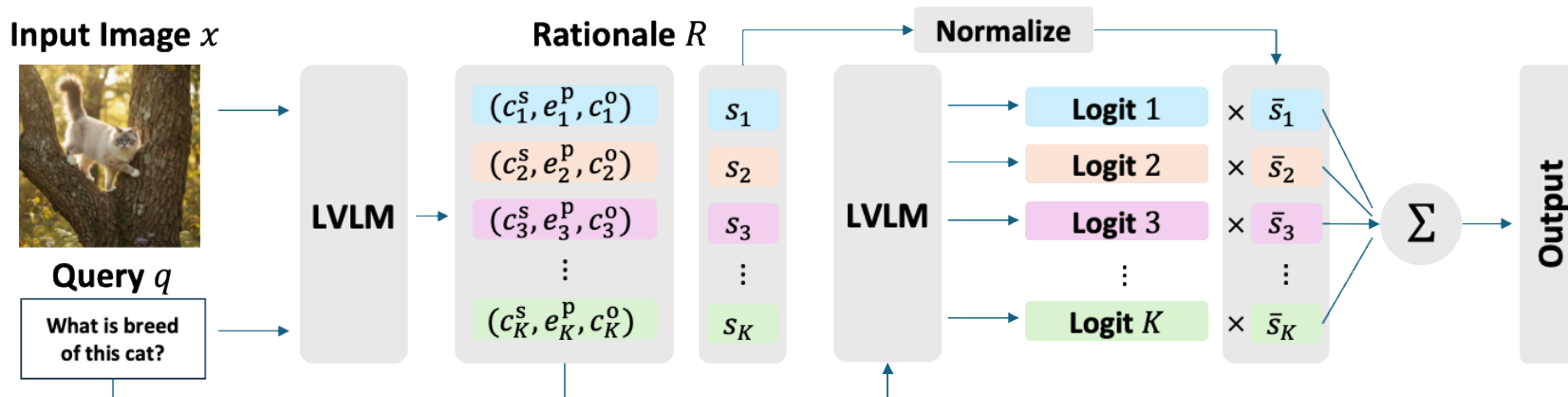
- ▷ Idea: Leveraging LVLMs' capability to build inherently interpretable models
- ▷ Key Features:
 - **Relations as Rationale**: directly representing semantic relations of concepts
 - **Zero-shot Interpretability**: rationale-dependent inference for ANY tasks
 - **Plug-and-Play**: converting ANY LVLMs to be interpretable w/o training



❖ Two-stage Inference Strategy

- ▷ **SRG**: Generating rationales with subject-predicate-object triplets
- ▷ **RGD**: Forming next-token distribution by ensemble of triplet-conditional dists.

1. Structured Rationale Generation (SRG) 2. Rationale-Grounded Decoding (RGD)



❖ Set of Relational Rationale R

$$R := \{(r_k, s_k) \mid r_k = (c_k^s, e_k^p, c_k^o), s_k \in [0, 1]\}_{k=1}^K$$

**Relational Triplet
between Concepts** **Self-Verbalized
Importance**

- ▷ r_k : Instead of isolated concepts (e.g., *Cat*), RBMs represent rationales by the semantic relation between concepts (e.g., (*Cat*, *Stands on*, *Branch*))
- ▷ s_k : Relative importance for LVLMs (will be used in decoding)

❖ We use a specialized instruction prompt to generate R

- ❖ Deriving a next-token distribution \hat{p}_θ based **only** on K triplet rationales
 - ▷ \hat{p}_θ is formalized by **Power-of-Experts (PoE)** of rationale-conditional dists.
 - \overline{s}_k : Normalized self-verbalized importance s_k

$$\widehat{p}_\theta(y_i | R, q) \propto \prod_{k=1}^K p_\theta(y_i | \mathbf{y}_{<i}, r_k, q)^{\overline{s}_k}$$

- ▷ Discarding image inputs from the context ensures the inherently interpretability
- ▷ We compute K logits for rationale-conditional distributions by batch parallel

❖ RBMs achieve strong causal dependency on the rationale

▷ Better/Worse quality rationales make better/worse performance

→ RBMs is actually based on rationales

- Random intervened rationales do not necessarily cause 0 accuracy because there are “yes/no” questions in GQA, which can be correctly answered w/o rationales randomly

	GQA		
	Self (↑)	Gemini3 (↑)	Random (↓)
	Qwen3-VL-8B		
CoT	51.96	55.73	26.10
LVLm-CBM	46.73	48.16	24.06
RBM	54.98	59.34	12.39
	Gemma3-12B		
CoT	24.64	38.78	4.50
LVLm-CBM	12.61	20.51	0.80
RBM	17.42	21.29	0.58
	InternVL3.5-8B		
CoT	54.56	56.81	32.64
LVLm-CBM	45.76	46.95	21.55
RBM	48.03	48.82	11.80

Experiments | Qualitative Analysis

❖ RBMs can describe accurate logical and semantic relations to output

GQA



Q. What type of food is on top of the plate near the fork?

LVLM-CBM

pancake with butter : 0.7
scramble eggs : 0.6
bacon strips : 0.6

pancake with butter ✓

RBM (Ours)

(pancake, is located near, fork) : 0.95
(pancake, has on top, butter) : 0.85
(pancake, is on, plate) : 0.9

pancake ✓

MMMU-Pro



Q. Who is the artist of the work?

LVLM-CBM

sleeping man at desk : 0.4
cat watching scene : 0.2
text in Spanish : 0.5

Pablo Picasso ✗

RBM (Ours)

(the image, depicts, a sleeping man) : 0.7
(the man, surrounded by, bats and owls) : 0.7
(the image, contains, Spanish text) : 0.6

Francisco Goya ✓

❖ RBMs achieve higher accuracy than LVLM-CBMs on diverse tasks

▷ However, they inferior to Direct/CoT that use images in decoding

→ Limitation of Bottleneck

	GQA	TextVQA	MME	COCO	MMMU	MMMU-Pro	Avg.
Qwen3-VL-8B							
Direct	60.73	78.34	89.30	21.02	42.00	23.29	52.45
CoT	51.96	60.80	67.02	3.08	33.11	15.66	38.61
LVLM-CBM	46.73	63.28	69.47	14.43	35.33	14.28	40.59
RBM	54.89	68.76	69.00	20.83	37.78	15.14	44.40
Gemma3-12B							
Direct	44.17	76.08	84.16	33.14	38.33	17.86	48.96
CoT	24.64	71.82	85.80	19.49	45.11	20.58	44.57
LVLM-CBM	12.61	56.70	62.64	9.10	36.56	18.55	32.69
RBM	17.42	60.40	64.03	18.94	44.56	20.35	37.61
InternVL3.5-8B							
Direct	61.41	78.34	87.19	52.93	47.56	22.83	58.38
CoT	54.56	74.14	92.92	51.70	34.11	14.68	54.08
LVLM-CBM	45.76	54.04	65.63	10.32	39.78	16.88	38.73
RBM	48.03	69.44	71.32	18.96	38.56	15.90	43.70

❖ Relational Bottleneck Models (RBMs)

- ▶ Providing task-agnostic interpretability based on LVLMs
- ▶ Generating semantic triplets as rationales, and predict answer from the triplets
- ▶ Improved interpretability from LVLM-CBMs

❖ Limitations

- ▶ Accuracy drop from the LVLM baselines using images for final answer
- ▶ Imperfect causal-dependency on rationales due to non-linear transform

Conclusion & Future Work

- ❖ **We presented three works going beyond CBMs based on VLMs**
 - ▷ XBMs: Providing flexible rationales beyond concepts by natural language
 - Fine-tuning encoder-decoder VLMs w/ Explanation Distillation
 - ▷ Z-CBMs: Performing interpretable classification w/o additional training
 - Concept retrieval + sparse regression
 - ▷ RBMs: Achieving task-agnostic interpretability with semantic triplet rationales
 - Converting LVLMs to be inherently interpretable by plug-and-play decoding

❖ Performance Drop from Black-box Models

- ▷ Can we achieve better accuracy-interpretability trade-off?
- ▷ Can we design totally new interpretable models for high-performance?

❖ Other Modalities and More

- ▷ Can we extend existing frameworks to minor or rare modality (e.g., time series)?
- ▷ Can we build interpretable models accepting ANY modality input?

❖ Different Perspectives: LLMs and Agents

- ▷ Is LLM's reasoning (e.g., CoT) *interpretable*?
- ▷ Can LLM-based agents make a contract with humans?

