MAX PLANCK INSTITUTE FOR INFORMATICS
SIC Saarland Informatics Campus

CFM: Language-aligned Concept Foundation Model for Vision

Kai Wittenmayer, Sukrut Rao, Amin Parchami-Araghi, Bernt Schiele, Jonas Fischer

Paper Code

ECCV 2026 | MALMÖ | SEPT 8–12

Motivation

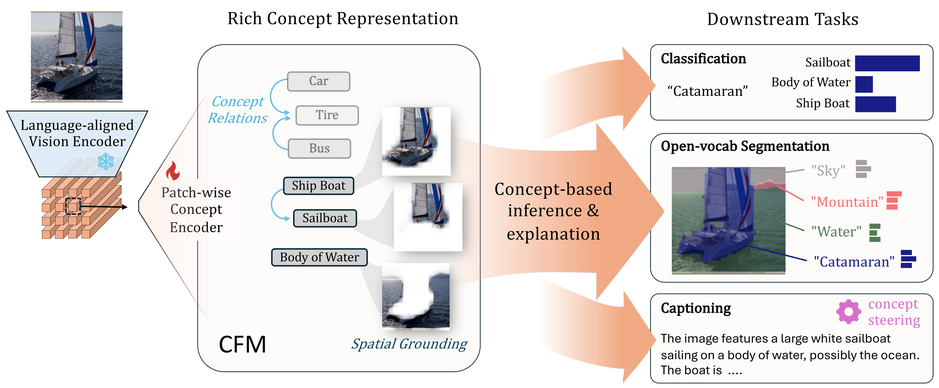
Foundation models
❌ opaque
✅ applicable to various downstream tasks

→

CFM
✅ interpretable
✅ applicable to various downstream tasks

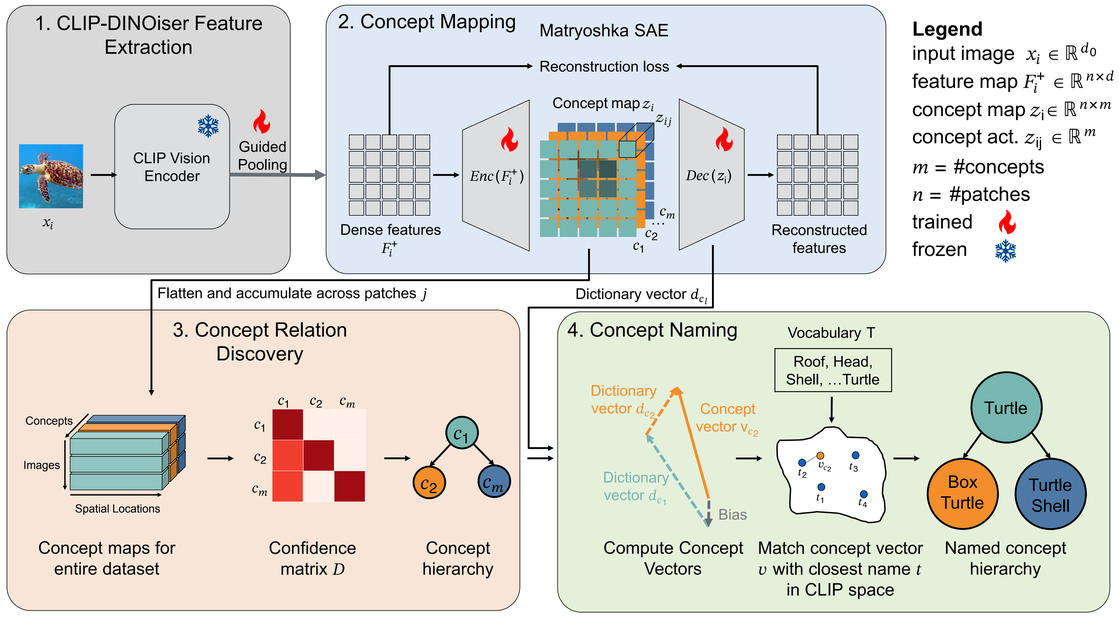
←

Concept bottleneck models
✅ interpretable
❌ mostly restricted to image classification



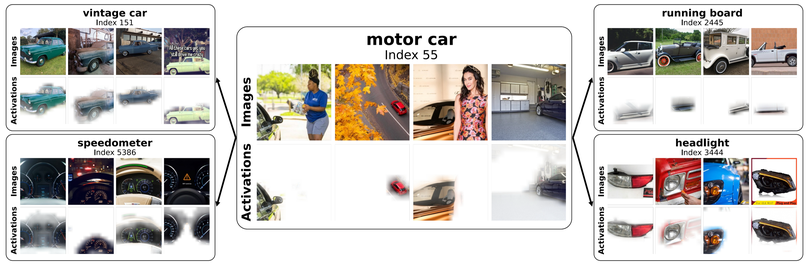
Concepts: ✅ spatially grounded ✅ fine-grained ✅ explicit hierarchies ✅ language-aligned

Method



Legend
input image $x_i \in \mathbb{R}^{d_0}$
feature map $F_i^+ \in \mathbb{R}^{n \times d}$
concept map $z_i \in \mathbb{R}^{n \times m}$
concept act. $z_{ij} \in \mathbb{R}^m$
m = #concepts
n = #patches
trained 🔥
frozen ❄

Interpretable Concepts



Concept Evaluation

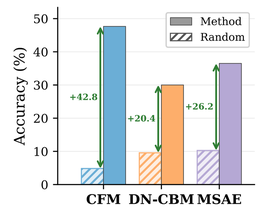
Spatial Grounding

Methods	Locality ↑ Part	Locality ↑ Stuff	Consistency ↑ Part	Consistency ↑ Stuff	Impurity ↓ Part	Impurity ↓ Stuff	C^2-Score ↑ IMN
SALF-CBM	16.4	19.5	9.9	6.7	0.841	1.143	0.298
PatchSAE	32.2	34.6	13.1	12.7	0.353	0.663	0.295
CFM	**44.3**	**44.5**	**15.3**	**13.1**	**0.333**	**0.539**	**0.465**

Provides **consistent** and **well-grounded concepts**

SALF-CBM: Benou et al. 2025; PatchSAE: Lim et al. 2025

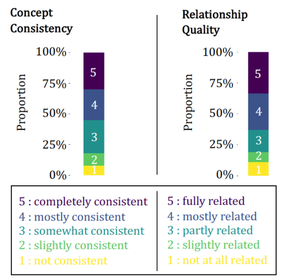
Concept Naming



Improves **embedding-based naming**

DN-CBM: Rao et al. 2024; MSAE: Zaigrajew et al. 2025

User Study

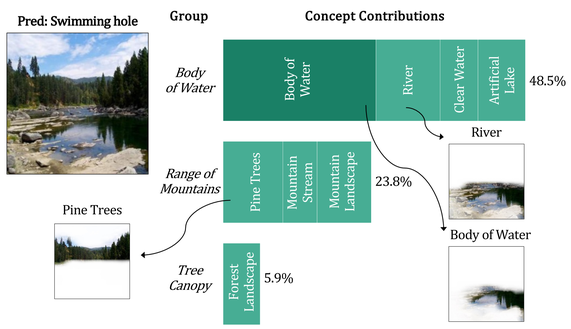


Discovers consistent concepts and meaningful **unsupervised hierarchies**

Amazon MTurk 1000 concepts, covered with 5 users per concept

Performance Evaluation

Classification

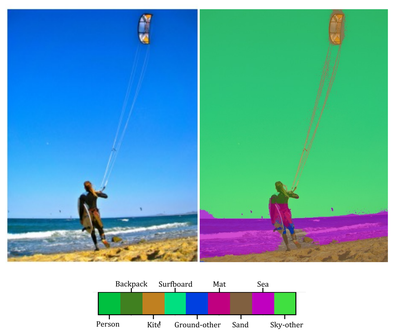


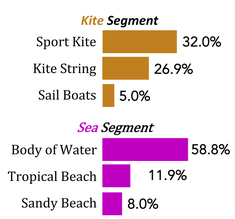
Model	Spatial Grounding	IMN	Places
DN-CBM	✗	**79.5**	55.1
SALF-CBM	✓	76.3	49.4
CFM	✓	78.9	**55.4**

Competitive accuracy whilst providing **spatially grounded and hierarchically ordered explanations**

DN-CBM: Rao et al. 2024; SALF-CBM: Benou et al. 2025

Open Vocabulary Segmentation



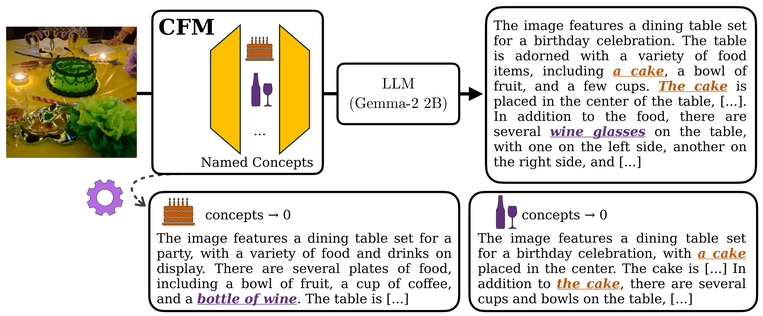


Methods	OVS avg.
CLIP-DINOiser	41.4
CLIP-DINOiser +(AnyUp)	42.2
CFM	41.3
CFM +(AnyUp)	41.9

Performs on par with opaque baselines, while being the only method providing **patch- or pixel-level explanations**

CLIP-DINOiser: Wysoczańska et al. 2024; AnyUp Wimmer et al. 2026

Steerable Captioning



Metrics ↑	CLIP-DINOiser	CFM
CLIP-Sc	0.72	0.72
RefCLIP-Sc	0.73	0.72
Capture	0.35	0.35

Maintains performance similar to opaque backbone, while additionally allowing for **steering**

CLIP-Sc: Hessel et al. 2021; Capture: Dong et al. 2024