Concepts in Motion: Temporal Bottlenecks for Interpretable Video Classification

Patrick Knab
Technical University of Clausthal
patrick.knab@tu-clausthal.de

Sascha Marton
Technical University of Clausthal

Philipp J. Schubert
Ramblr.ai Research

Drago Guggiana
Ramblr.ai Research

Christian Bartelt
Technical University of Clausthal

Abstract

*Concept Bottleneck Models (CBMs) enable interpretable image classification by structuring predictions around human-understandable concepts, but extending this paradigm to video remains challenging due to the difficulty of extracting concepts and modeling them over time. In this paper, we introduce **MoTIF** (Moving Temporal Interpretable Framework), a transformer-based concept architecture that operates on sequences of temporally grounded concept activations, by employing per-concept temporal self-attention to model when individual concepts recur and how their temporal patterns contribute to predictions. Central to the framework is an agentic concept discovery module that automatically extracts object- and action-centric textual concepts from videos, and across multiple video benchmarks this combination narrows the performance gap between interpretable and black-box video models on temporally demanding datasets while maintaining faithful concept explanations.*

1. Introduction

Modern deep learning models already achieve outstanding results in video understanding tasks such as video classification, action recognition, and event detection [1, 14]. Despite their success, these models are commonly perceived as *black boxes* since their internal workings are not interpretable in a way that reveals their decision-making process [8, 16]. Concept Bottleneck Models (CBMs) [9] address this issue by enforcing an intermediate bottleneck layer of human-understandable concepts, which are then used by a linear classifier to generate the final prediction.

While CBMs have been extensively studied in the image domain [18, 21, 23, 28], their extension to video remains largely unexplored [7, 12]. Videos differ from images in that they contain a *temporal component*: concepts evolve over time, and many actions cannot be inferred from a single frame [3, 12]. While transformers [24] capture long-range dependencies [1], dense temporal feature mixing obscures concept-level attributions [6, 16], motivating architectures that preserve concept interpretability over time.

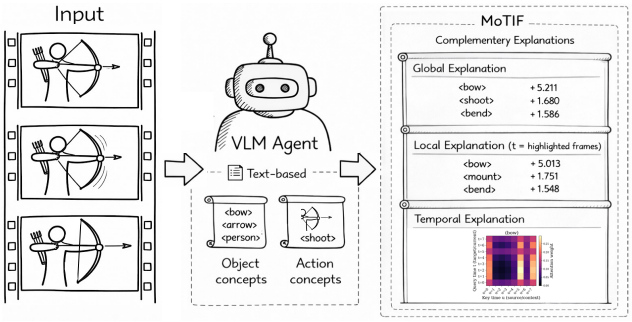

Figure 1. **MoTIF.** Overview: agentic concept discovery produces temporally grounded concept activations; diagonal temporal attention models per-concept dynamics and supports global, local, and temporal explanations.

In this work, we introduce **MoTIF** (**Mo**ving **T**emporal **I**nterpretable **F**ramework), a concept bottleneck model tailored for video classification. MoTIF builds on transformer-inspired blocks and introduces *per-channel temporal self-attention* (diagonal attention) to isolate temporal reasoning for each concept, paired with an agentic concept discovery pipeline that automatically extracts object- and action-centric textual concepts from video frames (Figs. 1 and 2).

Our key contributions are:

- *MoTIF*, a video CBM with **per-concept temporal self-attention** that preserves concept independence while modeling temporal dynamics.
- An **agentic, unsupervised concept discovery pipeline** that extracts object and action concepts directly from video windows.
- **Three complementary explanation modes**: *(i)* global

Figure 2. **MoTIF.** The framework takes videos as input and produces local concept explanations for local windows, global explanations for entire videos, and temporal dependency maps from the attention heads of the transformer module. Model represents MoTIF (ViT-L14) and sample frames are from HMDB51 [10], licensed under CC BY 4.0.

concept relevance via log-sum-exp pooling, *(ii)* localized temporal explanations from windowed concept attributions, and *(iii)* temporal dependency maps from the per-concept attention weights.

2. Method

MoTIF is an interpretable concept bottleneck model for video classification. Each video is represented as a sequence of concept activations $X \in \mathbb{R}^{T\times C}$, where T denotes temporal windows and C semantic concepts. The framework preserves attribution by modeling temporal dependencies independently per concept using diagonal self-attention. Refined concept activations are classified per time step and aggregated with log-sum-exp (LSE) pooling to obtain the video-level prediction and temporal importance profile.

2.1. Agentic Concept Discovery

We convert the raw video into structured, semantically interpretable concept activations. Given a video, we partition it into T temporal windows $\{w_k\}_{k=1}^{T}$, enabling localized concept detection and explicit temporal structure.

Concept proposal and bank construction. Each window is processed by a vision–language agent that generates textual concept candidates describing objects and actions present in the scene. All discovered concepts are merged into a global concept bank $\mathcal{C}$. To ensure semantic diversity, concepts are embedded using the same vision–language model and filtered by cosine similarity to remove near-duplicates.

Concept grounding. For each window, a visual embedding is extracted by a Clip-based method [4, 19], and matched against all concept embeddings using cosine similarity. This yields a concept activation matrix

$$X^{(n)} \in \mathbb{R}^{T_n\times C},$$

where each entry measures the presence of concept c in window t. Each channel therefore corresponds to one interpretable concept with a temporal activation profile, forming the input to the MoTIF bottleneck.

2.2. Transformer Bottleneck Model

MoTIF models temporal dependencies directly in concept space while preserving attribution.

Diagonal temporal self-attention. Standard transformers mix information across channels, which would obscure concept attribution. Instead, MoTIF applies self-attention independently to each concept channel, using depthwise 1×1 convolutions to compute the query, key, and value projections. For each concept $c \in \mathcal{C}$, attention is computed only across time:

$$X_{t,c}^{(L)} = \sum_{u=1}^{T} W_{c,t,u} V_{u,c},$$

where weights $W_{c,t,u}$ determine how strongly concept activations at different time steps influence each other. This allows the model to capture temporal *motifs* such as recurring or causally related concept occurrences, while preventing cross-concept mixing.

Equation 1 shows the difference between full and diagonal attention: each channel learns to construct its own temporal filter by weighting past activations differently at each step (diagonal). Unlike fixed convolutional kernels, the attention weights adapt to the input sequence, yet the restriction to depthwise (per-concept) projections guarantees that evidence for one concept does not leak into another.

$$\underbrace{\begin{bmatrix} (c_1\to c_1) & \cdots & (c_1\to c_C) \\ \vdots & \ddots & \vdots \\ (c_C\to c_1) & \cdots & (c_C\to c_C) \end{bmatrix}}_{\text{Full attention}} \quad \underbrace{\begin{bmatrix} (c_1\to c_1) & & \\ & \ddots & \\ & & (c_C\to c_C) \end{bmatrix}}_{\text{Diagonal attention}} \tag{1}$$

The block concludes with per-channel normalization and a lightweight feed-forward network (two depthwise 1×1 convolutions with GELU and dropout). Diagonal attention reduces the channel-mixing cost from $\mathcal{O}(C^2T)$ to $\mathcal{O}(CT)$, but it requires computing a full $T\times T$ attention map for every channel, yielding $\mathcal{O}(CT^2)$. The MoTIF transformer bottleneck can also be extended to a space-time transformer architecture (MoTIF-ST), as detailed in Appendix A.4, demonstrating that MoTIF is not restricted to a single transformer design.

Per-Concept Affine Transformation. Each refined activation $X_{t,c}^{(L)}$ is optionally scaled and shifted with learnable concept-specific parameters, $\tilde{X}_{t,c} = \gamma_c X_{t,c}^{(L)} + \delta_c$, and passed through Softplus, $Z_{t,c} = \mathrm{Softplus}(\tilde{X}_{t,c})$, to obtain nonnegative, fully differentiable concept activations without dead units. The per-concept scale γ_c and bias δ_c adapt to differing activation magnitudes and thresholds.

Classification Head. From these activations, per-time-step logits are computed as $\ell_t = W_k Z_{t,:} + b$ with $W_k \in \mathbb{R}^{K \times C}$, where K denotes the number of target classes and C the number of concepts. Since videos vary in length, we apply *log-sum-exp (LSE) pooling* across time [26], which smoothly interpolates between mean-pooling($\tau \to 0$) and max-pooling ($\tau \to \infty$):

$$\hat{c} = \tfrac{1}{\tau} \log \sum_{t=1}^{T} m_t e^{\tau c_t}, \qquad \hat{\ell} = \tfrac{1}{\tau} \log \sum_{t=1}^{T} m_t e^{\tau \ell_t}, \tag{2}$$

where $m_t \in 0, 1$ are masks for padded windows. We denote the pooled concept vector by $\hat{c}$ and the pooled logits by $\hat{\ell}$. The pooled logits $\hat{\ell}$ form the video-level prediction.

Training objective. The model is trained with class-weighted cross-entropy on $\hat{\ell}$, complemented with two regularizers: an ℓ_1 penalty on W to encourage sparsity, and an activation sparsity penalty on Z:

$$\mathcal{L} = \mathrm{CE}(\hat{\ell}, y) + \lambda_{\ell_1} \|W_k\|_1 + \lambda_{\text{sparse}} \frac{1}{(\sum_t m_t) C} \sum_{t,c} m_t |Z_{t,c}|. \tag{3}$$

2.3. Explanation

MoTIF decomposes predictions into concept- and time-resolved contributions. For class k, per-time-step contributions are $c_t^{(k)} = Z_{t,:} W_{k,:}$, with score $s_t^{(k)} = \sum_{c=1}^{C} c_{t,c}^{(k)} + b_k$, and temporal importance weights follow naturally from LSE pooling:

$$\pi_t^{(k)} = \frac{\exp(s_t^{(k)}/\tau)}{\sum_{u=1}^{T} \exp(s_u^{(k)}/\tau)}.$$

This yields three complementary explanation views: global concept importance, decisive local concepts, and temporal dependencies via per-concept attention, exposing both which concepts mattered and when.

3. Experiments

3.1. Experimental Setup

Datasets. We evaluate on Breakfast Actions [11], HMDB51 [10], UCF101 [22], and SSv2 [5, 15]. They cover short vs. long clips, local vs. global temporal dependencies, and varying abstraction and viewpoint diversity.

Table 1. **Performance comparison (% Top-1 accuracy).** Mean ± standard deviation across train-test splits on Breakfast Actions, HMDB51, UCF101, and SSv2 with agentic concept discovery.

	Method	Breakfast	HMDB51	UCF101	SSv2
Zero-Shot	CLIP-ViT-L/14 [19]	31.1 ± 4.7	45.7 ± 0.1	70.6 ± 0.5	0.9
	PE-L/14 [2]	41.4 ± 7.0	56.7 ± 0.6	74.6 ± 0.9	2.2
	PE-G/14 [2]	47.4 ± 5.4	60.7 ± 1.0	74.6 ± 0.9	2.2
Global-CBM	CLIP-ViT-L/14 [19]	57.0 ± 8.0	71.0 ± 1.1	93.4 ± 0.7	22.0
	PE-L/14 [2]	72.4 ± 8.3	76.4 ± 0.8	96.3 ± 0.1	31.3
	PE-G/14 [2]	75.8 ± 7.1	77.8 ± 0.8	97.5 ± 0.4	33.6
MoTIF (ours)	MoTIF (ViT-L/14)	71.0 ± 6.2	76.1 ± 0.5	94.8 ± 0.5	25.8
	MoTIF-ST (ViT-L/14)	72.6 ± 6.5	75.8 ± 0.6	94.8 ± 0.4	27.7
	MoTIF (PE-L/14)	83.2 ± 6.2	81.8 ± 0.6	97.0 ± 0.3	37.3
	MoTIF-ST (PE-L/14)	85.4 ± 6.3	80.8 ± 1.0	97.2 ± 0.2	39.6
	MoTIF (PE-G/14)	**87.5** ± 4.9	83.0 ± 0.6	98.0 ± 0.2	40.4
	MoTIF-ST (PE-G/14)	87.3 ± 7.1	82.1 ± 1.0	98.4 ± 0.3	41.9
Black-Box	TSM [13]	59.1	73.5	95.9	61.7
	No frame left behind [14]	62.0	73.4	96.4	62.7
	VideoMAE V2 [25]	–	**88.1**	**99.6**	**76.8**

Backbones. We use CLIP-based encoders with different capacity and temporal adaptation: CLIP (RN/50, ViT-B/32, ViT-L/14) [19], SigLIP [29], and Perception Encoder (PE) [2]. SigLIP replaces CLIP's contrastive softmax with a pairwise sigmoid loss, improving scalability. PE is trained on video–text pairs and aggregates frame-level information. We use SigLIP ViT-L/14 and PE ViT-L/14 and ViT-G/14.

Concept discovery. For the main results (Table 1), we use Qwen-3 30B [27] to generate dataset-specific concept banks, prompting with up to five training videos per class and extracting concepts on the training split only to avoid leakage. For ablations, we follow Yang et al. [28] and construct smaller concept sets using an LLM (GPT-5) [17]. Concept vocabularies are reported in Appendix E.

Baselines. Following [18, 20], we compare to zero-shot and linear-probe baselines. Zero-shot is computed at the window level with majority voting to match MoTIF's temporal granularity. As a supervised baseline (Global CBM), we train a linear classifier on mean-pooled window embeddings, i.e., a global bottleneck without temporal localization.

3.2. Experimental Results

3.2.1. Performance Evaluation

Table 1 reports Top-1 accuracy for MoTIF across backbones and datasets (hyperparameters in Appendix A.1; splits in Appendix C.1). Global CBM consistently improves over zero-shot, and MoTIF improves over both, with gains increasing with backbone capacity—suggesting that a temporal concept bottleneck can retain (and sometimes improve) accuracy while adding interpretability.

Both transformer variants perform well; the space–time variant (MoTIF-ST) [1] is consistently stronger on SSv2. PE backbones outperform CLIP variants at comparable scale. SSv2 remains the hardest benchmark due to abstract, relational classes. On Breakfast, MoTIF improves over Global CBM; on HMDB51, UCF101, and SSv2 gains are smaller, likely because many clips are short and mean pooling is

often sufficient—especially with PE embeddings. A gap to black-box baselines (TSM [13], NoFrameLeftBehind [14], VideoMAE V2 [25]) remains, particularly on SSv2; nevertheless, MoTIF surpasses two reported baselines on Breakfast. The largest SSv2 improvements stem from agentic concept discovery (Table 9).

3.2.2. Ablations

We ablate key design choices in MoTIF. Unless noted, Breakfast and HMDB51 use CLIP ViT-B/32 and RN/50 (Appendix A.1), covering transformer- and CNN-based encoders and complementary dataset characteristics (scale, variability, granularity). Additional ablations are in Appendix C.

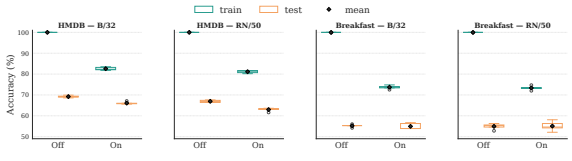

Figure 3. **Full vs. diagonal attention.** Train/test accuracy with and without diagonal attention over five seeds.

Attention variant (full vs. diagonal). We compare MoTIF's diagonal attention to full multi-head attention (Figure 3), isolating the effect of cross-concept interactions. Full attention can recover accuracy on demanding datasets, whereas diagonal attention preserves faithful concept attribution; on SSv2, full attention yields up to 10.1% higher test accuracy, but with degradation in explanation quality as discussed in Section B.2.

Temporal sensitivity and dynamic concepts. To test whether MoTIF captures temporal order rather than defaulting to order-invariant reasoning [1], we construct a synthetic benchmark (1,989 sequences, matching Breakfast training size) where labels depend solely on frame order (Appendix B.4). MoTIF achieves 86.97% accuracy, dropping to 21.06% under random shuffling (chance ≈20%), a 4.1× gap, while the Global CBM reaches only 35.5%, confirming that temporal modeling is essential when order defines the class.

Table 2. **Temporal sensitivity and bottleneck comparison (PE-L/14).** Shuffling randomly permutes windows at test time; synthetic has five temporal classes.

Setting	Basic	Shuffled
Synthetic (MoTIF)	86.97	21.06
Synthetic (Global CBM)	35.5	35.5
Breakfast (MoTIF)	87.3	85.2–86.6
HMDB51 (MoTIF)	79.9	77.1–78.5
UCF101 (MoTIF)	94.7	94.5–94.6
SSv2 (MoTIF)	30.0	26.9–27.4

Second, applying the same shuffling intervention to real datasets (Table 2) shows the largest degradation on SSv2, which also benefits most from agent-generated temporal concepts, indicating that temporal reasoning is decisive when relevant dynamics are captured.

3.2.3. Concept Interventions

To illustrate concept-level control, we revisit Figure 2. Ablating the most influential concept *bow* (zeroing its channel) flips the prediction from the correct class to *run* (logit 8.20→6.79), while removing windows 1–4, where the bow is handled, shifts the output to *talk* (logit 6.75). We further evaluate (i) top-k concept removal, (ii) random removal, and (iii) random noise insertion ($\epsilon \sim \mathcal{N}(0, \sigma^2)$, σ=0.5). As shown in Table 3, accuracy drops sharply only when removing the most influential concepts, whereas random removal or noise has minor effects, indicating that MoTIF relies on a small, semantically meaningful concept subset and remains robust to perturbations[1].

Table 3. **Concept interventions.** Accuracy after removing or perturbing k concept channels (RN/50).

Dataset	k	Top-k Removal	Random Removal	Random Insertion
Breakfast	0	1.000	1.000	1.000
	1	0.496	0.989	0.979
	2	0.229	0.972	0.947
	3	0.085	0.951	0.951
	4	0.028	0.923	0.908
HMDB51	0	1.000	1.000	1.000
	1	0.603	0.974	0.978
	2	0.374	0.963	0.961
	3	0.238	0.949	0.945
	4	0.142	0.935	0.921

4. Discussion and Conclusion

MoTIF introduces a transformer-based concept bottleneck architecture for video that performs temporal reasoning directly on sequences of interpretable concept activations. Its key innovation is diagonal temporal attention, which models *when* concepts occur while preventing cross-concept mixing and preserving faithful attribution. Across datasets, MoTIF consistently outperforms zero-shot and global bottleneck baselines, demonstrating that temporal transformer modeling and interpretability can be achieved jointly.

Performance depends strongly on concept quality. Agentic concept discovery substantially improves accuracy—especially on temporally demanding datasets such as SSv2—by providing concepts that capture temporal structure. Architecture and concept discovery are thus complementary: MoTIF enables temporal reasoning, while high-quality concepts unlock its full potential.

The framework is modular and compatible with modern embedding backbones (CLIP, SigLIP, PE) and space–time extensions (MoTIF-ST). It further provides global, local, and temporal explanation views, exposing both *which* concepts matter and *when*, and enabling targeted interventions.

[1] k=0 corresponds to perfect accuracy since values are normalized to MoTIF's baseline predictions.

A clear accuracy–interpretability trade-off emerges: cross-concept attention improves accuracy, whereas diagonal attention preserves attribution fidelity. Overall, MoTIF establishes a strong and scalable foundation for interpretable temporal reasoning in video, combining competitive performance with transparent, concept-level explanations.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Icml*, page 4, 2021. 1, 3, 4

[2] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network, 2025. 3, 1, 7

[3] Delong Chen, Theo Moutakanni, Willy Chung, Yejin Bang, Ziwei Ji, Allen Bolourchi, and Pascale Fung. Planning with reasoning using vision language world model. *arXiv preprint arXiv:2509.02722*, 2025. 1

[4] Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, Muhammad Maaz, Yale Song, Tengyu Ma, Shuming Hu, Suyog Jain, et al. Perceptionlm: Open-access data and models for detailed visual understanding. *arXiv preprint arXiv:2504.13180*, 2025. 2

[5] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 3

[6] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Self-attention attribution: Interpreting information interactions inside transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12963–12971, 2021. 1

[7] Jeya Vikranth Jeyakumar, Luke Dickens, Yu-Hsi Cheng, Joseph Noor, Luis Antonio Garcia, Diego Ramirez Echavarria, Alessandra Russo, Lance M. Kaplan, and Mani Srivastava. Automatic concept extraction for concept bottleneck-based video classification, 2022. 1

[8] Patrick Knab, Sascha Marton, Udo Schlegel, and Christian Bartelt. Which lime should i trust? concepts, challenges, and solutions, 2025. 1

[9] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020. 1

[10] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011. 2, 3

[11] H. Kuehne, A. B. Arslan, and T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of Computer Vision and Pattern Recognition Conference (CVPR)*, 2014. 3

[12] Jongseo Lee, Wooil Lee, Gyeong-Moon Park, Seong Tae Kim, and Jinwoo Choi. Pcbear: Pose concept bottleneck for explainable action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2690–2699, 2025. 1

[13] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7083–7093, 2019. 3, 4, 7

[14] Xin Liu, Silvia L. Pintea, Fatemeh Karimi Nejadasl, Olaf Booij, and Jan C. van Gemert. No frame left behind: Full video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14892–14901, 2021. 1, 3, 4, 6, 7

[15] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. pages 1049–1059, 2020. 3, 8

[16] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. Interpretable machine learning – a brief history, state-of-the-art and challenges. In *ECML PKDD 2020 Workshops*, pages 417–431, Cham, 2020. Springer International Publishing. 1

[17] OpenAI, Josh Achiam, Steven Adler, and Sandhini Agarwal et al. Gpt-4 technical report, 2024. 3

[18] Katharina Prasse, Patrick Knab, Sascha Marton, Christian Bartelt, and Margret Keuper. DCBM: Data-efficient visual concept bottleneck models. In *Forty-second International Conference on Machine Learning*, 2025. 1, 3, 10

[19] 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*, pages 8748–8763. PmLR, 2021. 2, 3, 1, 7

[20] Sukrut Rao, Sweta Mahajan, Moritz Böhle, and Bernt Schiele. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXVII*, page 444–461, Berlin, Heidelberg, 2024. Springer-Verlag. 3, 10

[21] S. Schrodi, J. Schur, M. Argus, and T. Brox. Selective concept bottleneck models without predefined concepts. *Transactions on Machine Learning Research (TMLR)*, 2025. 1, 10

[22] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision*, 2(11):1–7, 2012. 3

[23] Chung-En Sun, Tuomas Oikarinen, Berk Ustun, and Tsui-Wei Weng. Concept bottleneck large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[25] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14549–14560, 2023. 3, 4, 7

[26] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern recognition*, 74:15–24, 2018. 3

[27] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. 3

[28] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19187–19197, 2023. 1, 3

[29] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986, 2023. 3, 1, 7

[30] Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A Ehinger, and Benjamin IP Rubinstein. Invertible concept-based explanations for cnn models with non-negative concept activation vectors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11682–11690, 2021. 4