Concepts in Motion: Temporal Concept Bottleneck Model for Interpretable Video Classification

Patrick Knab [1 2] **Sascha Marton** [1] **Philipp J. Schubert** [2] **Drago Guggiana** [2] **Christian Bartelt** [1]

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 a class-conditioned VLM-based concept discovery module that extracts object- and action-centric textual concepts from training videos, yielding temporally expressive concept sets without manual concept annotation. Across multiple video benchmarks, this combination improves over global concept bottlenecks and remains competitive within the interpretable concept-bottleneck setting, while narrowing the gap to strong black-box video baselines that we report as contextual references.

1. Introduction

Modern deep learning models already achieve outstanding results in video understanding tasks such as video classification, action recognition, and event detection (Liu et al., 2021; Bertasius et al., 2021). 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 (Molnar et al., 2020; Knab et al., 2025b). Concept Bottleneck Models (CBMs) (Koh et al., 2020) 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 (Prasse et al., 2025; Yang et al., 2023; Sun et al., 2025; Schrodi et al., 2025), their extension to video remains largely unexplored (Jeyakumar et al., 2022). 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 (Lee et al., 2025b; Chen et al., 2025). Importantly, the challenge of variable video length is orthogonal to modeling long-range dependencies. While transformers (Vaswani et al., 2017) excel at capturing such dependencies (Bertasius et al., 2021), their dense temporal feature mixing obscures concept-level attributions, limiting their suitability for interpretable concept bottleneck modeling (Molnar et al., 2020; Hao et al., 2021). This highlights the need for architectures that handle variable-length inputs while preserving concept-level interpretability over time.

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 a *per-channel temporal self-attention* (diagonal attention) mechanism that isolates temporal reasoning for each concept to enable interpretation. In addition, we utilize a VLM-based concept discovery approach to generate object- and action-based, class-conditioned textual concepts from training videos. To illustrate how MoTIF extends beyond static images, Fig. 1 shows our framework: it processes a video, tracks its concepts through time, and explains which concepts drive the final prediction. MoTIF tracks concepts such as *bow*, *mount*, and *shoot* over time, whose combined activations support the final prediction. Temporal dependency maps indicate which windows provide contextual support for updating a concept channel, and are interpreted jointly with concept presence and contribution rather than as standalone evidence. This yields both *global concept contributions* over the full video and *local relevance* within specific windows.

[1]Technical University of Clausthal, Clausthal-Zellerfeld, Germany [2]Ramblr.ai Research. Correspondence to: Patrick Knab <patrick.knab@tu-clausthal.de>.

Accepted to the 2nd Workshop on Compositional Learning at ICML 2026, Seoul, South Korea.

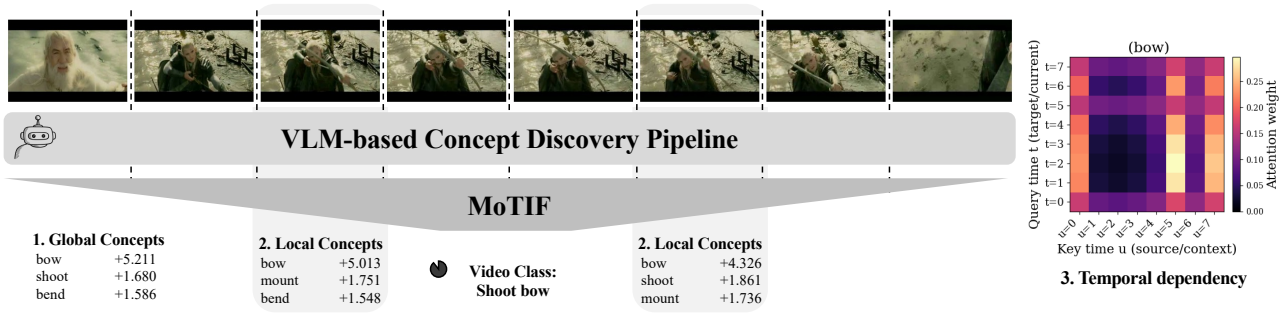


Figure 1. **MoTIF.** The method treats videos as windows and produces local concept explanations, global explanations for entire videos, and temporal dependency maps from the attention heads. Model represents MoTIF (ViT-L14) and sample frames are from HMDB51 (Kuehne et al., 2011), licensed under CC BY 4.0.

Our key contributions are:

- *MoTIF*, a Video CBM with per-channel temporal self-attention that **preserves concept independence** within Transformer blocks while modeling temporal dynamics at the level of individual concepts.
- *MoTIF* provides **three complementary explanation modes**: *(i)* global concept relevance via log-sum-exp (LSE) pooling, *(ii)* localized temporal explanations using windowed concept attributions, and *(iii)* attention-based temporal maps that visualize how a concept channel distributes its focus across time.
- A **VLM-based concept discovery pipeline** to extract textual object and action concepts directly from training video with weak supervision from class labels, improving concept coverage and downstream performance over global bottlenecks.

2. MoTIF

MoTIF is a transformer-based CBM for video classification that operates on sequences of window-level concept activations (Figure 1). A video is split into T temporal windows w_t, each embedded by a frozen vision backbone, and compared against a concept bank $\mathcal{C}$ of VLM-generated object and action concepts. The resulting cosine similarities form the activation matrix $X \in \mathbb{R}^{T\times C}$, where $C = |\mathcal{C}|$, which is processed by the temporal module before LSE pooling and classification.

The framework enforces strict *concept-wise attribution without cross-concept mixing*: temporal dependencies are modeled independently for each concept channel, so recurrent temporal patterns—*motifs*—remain attributable to individual concepts. We use the term *channel* to denote one dimension of the concept space. Each video n yields $\hat{y}^{(n)} \in \mathbb{R}^K$ for logits over K classes, $X^{(n)} \in \mathbb{R}^{T_n\times C}$ concept activations, where the second axis (C) indexes concepts; each slice $X^{(n)}_{:,c} \in \mathbb{R}^{T_n}$ corresponds to the temporal activation sequence of concept $\mathcal{C}_c$, with $c \in \{1,\dots,C\}$. Unlike CNN channels, these dimensions are semantically interpretable by design. After temporal processing, per-concept activations are refined via a nonnegative affine transformation, and a classifier produces per-time-step logits. Video-level predictions are obtained via log-sum-exp pooling, which also yields a time-importance profile for additional explanation. In contrast, a conventional transformer mixes channels during attention (Vaswani et al., 2017)[1], and algorithmic overviews are in Appendix D.

2.1. VLM-based concept discovery.

Concept discovery. The concept discovery stage converts raw visual inputs into structured, interpretable concept activations that serve as input to the MoTIF backbone. While formulated for video, images are naturally handled as the special case of a single temporal window. The stage only looks at a small number of videos from the train set. In our main setting, we sample videos per class and process each window w_t with a VLM A_{vlm} conditioned on the class label, which yields natural-language concept candidates

$$(\mathcal{C}^{\text{obj}}_{w_t}, \mathcal{C}^{\text{act}}_{w_t}) = A_{\text{vlm}}(w_t, y), \tag{1}$$

capturing objects and actions. This step is task-agnostic and does not rely on any predefined concept vocabulary. In addition this design allows the replacement of different models depending on the task.

Concept and video embeddings. The global concept bank $\mathcal{C}$ is constructed from the union of all discovered textual object and action concepts. To ensure semantic diversity and remove near-duplicates, concepts are embedded using CLIP (Radford et al., 2021) based models, on which MoTIF also builds on top of it, and filtered by cosine similarity: con-

[1]We also evaluate a variant with *full multi-head attention*, but emphasize that this removes explicit concept attribution. A detailed analysis of this and other design choices is provided in Section 3.2.1 and Appendix C

cepts with pairwise similarity greater than 0.9 are removed. For each window, depending on the backbone, visual embeddings are obtained either from a representative frame or a video-adapted CLIP variant (Appendix A.3). Concept activations are computed as cosine similarities between window embeddings and concept embeddings, yielding per-video activation matrices $X^{(n)} \in \mathbb{R}^{T_n \times C}$, where each channel corresponds to one interpretable concept and is processed independently. Appendix E provides evidence that concept quality matters: across discovery backbones, filtering thresholds, concept sources, and object/action splits, downstream performance changes systematically with the quality and diversity of the concept bank.

2.2. MoTIF's Transformer Bottleneck

2.2.1. Per-Channel Temporal Self-Attention

In standard transformers, query–key–value (QKV) projections are implemented as full linear layers ($W_Q \in \mathbb{R}^{C \times C}$), which mix channels and would obscure concept attribution. In MoTIF, we avoid mixing: before applying diagonal temporal attention, we add a fixed sinusoidal positional encoding to the concept activation sequence to encode temporal order. Each concept c then receives its own QKV projections via depthwise 1×1 convolutions,

$$\begin{aligned} Q, K, V \in \mathbb{R}^{T \times C}, Q_{:,c} = x_{:,c}\,\theta_Q^{(c)}, \\ K_{:,c} = x_{:,c}\,\theta_K^{(c)}, V_{:,c} = x_{:,c}\,\theta_V^{(c)}. \end{aligned} \tag{2}$$

Here $\theta_Q^{(c)}, \theta_K^{(c)}, \theta_V^{(c)} \in \mathbb{R}$ are channel-specific scalars applied uniformly across all T, not temporal filters. Equivalently, each projection uses a depthwise kernel $\Theta_Q, \Theta_K, \Theta_V \in \mathbb{R}^{C \times 1 \times 1}$, where $\theta_*^{(c)}$ is the c-th depthwise filter. Attention scores are computed *per concept* as $W_{c,t,u} = Q_{t,c}\,K_{u,c}$, where t denotes the query time step and u the key/value time step. A softmax over u yields attention weights $W \in \mathbb{R}^{C \times T \times T}$, so that each concept decides *which of its past or future activations to attend to*. The output is obtained as the weighted sum over V: $X_{t,c}^{(L)} = \sum_{u=1}^{T} W_{c,t,u}\,V_{u,c}$. This diagonal structure preserves concept attribution because evidence from one concept channel cannot flow into another.

$$\underbrace{\begin{bmatrix} (c_1 \to c_1) & \cdots & (c_1 \to c_C) \\ \vdots & \ddots & \vdots \\ (c_C \to c_1) & \cdots & (c_C \to c_C) \end{bmatrix}}_{\text{Full attention}} \quad \underbrace{\begin{bmatrix} (c_1 \to c_1) & & \\ & \ddots & \\ & & (c_C \to c_C) \end{bmatrix}}_{\text{Diagonal attention}} \tag{3}$$

The block concludes with per-channel normalization. In this work, we deliberately isolate temporal reasoning per concept to preserve attribution. Importantly, this is a design choice rather than a limitation: as shown in Section 3.2.1 enabling cross-concept interactions improves accuracy at the cost of interpretability, highlighting a controllable trade-off rather than a hard restriction.

Architectural extension. For most experiments in this paper, we report the results of MoTIF using diagonal attention within a standard transformer architecture. However, as shown by (Bertasius et al., 2021), separating spatial and temporal attention can further enhance performance. Therefore, we additionally evaluate MoTIF when 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.

Complexity. Diagonal attention avoids dense channel mixing, but computes a full $T \times T$ attention map for each concept channel, yielding attention-map compute and memory complexity of $\mathcal{O}(CT^2)$. By contrast, standard multi-head attention maintains one such map per head, resulting in attention-map memory $\mathcal{O}(HT^2)$ with $H \ll C$, while additionally incurring dense channel-mixing projection costs. Thus, diagonal attention trades efficiency for strict concept isolation when the number of concepts is large (see Appendix A.1).

Per-Concept Affine Transformation. Each refined activation $X_{t,c}^{(L)}$ can optionally be scaled and shifted by learnable concept-specific parameters, $\tilde{X}_{t,c} = \gamma_c X_{t,c}^{(L)} + \delta_c$, and then passed through a Softplus nonlinearity $Z_{t,c} = \mathrm{Softplus}(\tilde{X}_{t,c})$, which ensures nonnegative concept activations while avoiding dead units and maintaining differentiability everywhere. This transformation introduces a per-concept scale (γ_c) and bias (δ_c), allowing the model to adapt to differences in concept magnitude and activation thresholds.

2.2.2. Classification Head

From these activations, per-time-step logits are computed as $\ell_t = W_{\text{cls}} Z_{t,:} + b$ with $W_{\text{cls}} \in \mathbb{R}^{K \times C}$, 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 (Wang et al., 2018), which provides a smooth temporal aggregation and serves as a soft approximation to max-pooling, becoming sharper as $\tau \to 0$:

$$\hat{c} = \tau \log \sum_{t=1}^{T} m_t e^{c_t/\tau}, \qquad \hat{\ell} = \tau \log \sum_{t=1}^{T} m_t e^{\ell_t/\tau}, \tag{4}$$

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:

$$\mathcal{L} = \mathrm{CE}(\hat{\ell}, y) + \lambda_{\ell_1} \|W\|_1 + \lambda_{\text{sparse}} \frac{1}{(\sum_t m_t)C} \sum_{t,c} m_t |Z_{t,c}|. \quad (5)$$

2.2.3. Explanation Generation

To make predictions transparent, MoTIF decomposes them into time- and concept-resolved contributions. For a target class k, the contribution of each time step is $c_t^{(k)} = Z_{t,:} \odot W_{k,:}$, with score $s_t^{(k)} = \sum_{c=1}^{C} c_{t,c}^{(k)} + b_k$. Temporal importance weights are derived consistently with LSE pooling:

$$\pi_t^{(k)} = \frac{\exp(s_t^{(k)}/\tau)}{\sum_{u=1}^{T} \exp(s_u^{(k)}/\tau)}. \quad (6)$$

Aggregating over time yields global concept attributions, $\bar{c}^{(k)} = \sum_{t=1}^{T} \pi_t^{(k)} c_t^{(k)}$. MoTIF therefore provides three complementary views: **(1) Global concepts**, via $\bar{c}^{(k)}$; **(2) Local concepts**, active in high-weight windows (large $\pi_t^{(k)}$); **(3) Temporal dependencies**, revealed by per-concept attention maps ($W_{c,t,u}$) that show how occurrences of concepts relate across time (see Appendix B.3). Temporal attention peaks should not be interpreted in isolation as decisive visual evidence. Rather, the map indicates which temporal context is used when updating a given concept channel. We therefore interpret temporal maps jointly with (i) concept presence in the queried window, (ii) its contribution to the predicted class, and (iii) the temporal windows providing contextual support. Together, these expose both *which* concepts mattered and *when* they were decisive.

3. Experiments

3.1. Experimental Setup

Datasets. We evaluate on Breakfast Actions (Kuehne et al., 2014), HMDB51 (Kuehne et al., 2011), UCF101 (Soomro et al., 2012), and Something-Something V2 (SSv2) (Goyal et al., 2017; Materzynska et al., 2020), containing 10, 51, 101, and 174 classes, respectively. Together, they cover short and long videos, local and global temporal dependencies, and different levels of action granularity and viewpoint diversity.

Backbones. We use CLIP-based visual backbones with different architectures, scales, and temporal adaptation: CLIP RN/50, ViT-B/32, and ViT-L/14 (Radford et al., 2021), SigLIP ViT-L/14 (Zhai et al., 2023), and the video-adapted Perception Encoder with ViT-L/14 and ViT-G/14 variants (Bolya et al., 2025).

Concept Discovery. For VLM-based concept discovery (Section 2.1), we use Qwen-3 30B (Yang et al., 2025) to generate the concept sets used in the main experiments. As in prior CBMs, downstream performance depends on the quality and coverage of the resulting concept bank, especially for domain-specific and dynamic concepts. For ablations, we follow Yang et al. (2023) and construct smaller dataset-specific concept sets using GPT-5; Appendix E reports a targeted concept-bank analysis covering discovery backbones, filtering thresholds, concept types, concept-set variance, and a DCBM-style (Prasse et al., 2025) visual-concept robustness check based on SAM3 (Carion et al., 2026) segments and short visual snippets.

Baselines. Following (Prasse et al., 2025; Rao et al., 2024), we compare MoTIF against zero-shot baselines and supervised reference models. For fairness, zero-shot predictions are computed at the window level and aggregated by majority voting. As a supervised interpretable baseline, we train a Global CBM using the same concept vocabulary and backbone features as MoTIF, but collapse the temporal dimension by mean-pooling window-level representations before classification. This tests whether temporally localized concept modeling adds value beyond a global CBM, while retaining concept-based prediction but removing localized explanations and temporal concept dynamics. We additionally report strong black-box video models as contextual references rather than strict apples-to-apples baselines.

3.2. Experimental Results

Table 1. **Top-1 accuracy (%).** Mean ± standard deviation across train-test splits. Full results, including all backbones, are provided in Appendix C. The best result for each dataset is shown in **bold**, and the second-best result is underlined.

Method	Breakfast	HMDB51	UCF101	SSv2
Zero-shot PE-G/14	47.4 ± 5.4	60.7 ± 1.0	74.6 ± 0.9	2.2
Global CBM PE-G/14	75.8 ± 7.1	77.8 ± 0.8	97.5 ± 0.4	33.6
MoTIF ViT-L/14	71.0 ± 6.2	76.1 ± 0.5	94.8 ± 0.5	25.8
MoTIF PE-L/14	83.2 ± 6.2	81.8 ± 0.6	97.0 ± 0.3	37.3
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**
TSM (Lin et al., 2019)	59.1	73.5	95.9	61.7
NoFrame (Liu et al., 2021)	62.0	73.4	96.4	62.7
VideoMAE V2 (Wang et al., 2023)	–	**88.1**	**99.6**	76.8

Table 1 summarizes the main results; the complete backbone sweep and zero-shot baselines are reported in Appendix C. MoTIF improves over the Global CBM, showing that temporally localized concept activations provide more informative video representations than mean-pooled global concepts. Accuracy generally increases with backbone capacity, and PE-based backbones outperform CLIP variants at comparable scales. The space–time variant, MoTIF-ST (Bertasius et al., 2021), is particularly beneficial on SSv2, where classes depend on fine-grained temporal relations. SSv2 remains the most challenging dataset overall due to its abstract relational labels, such as "putting something onto something." On Breakfast, MoTIF improves over the Global

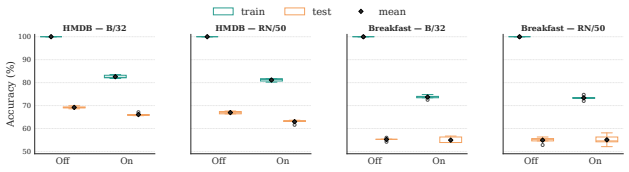


Figure 2. **Full vs. diagonal attention.** Training and test accuracy over five random seeds with and without enforcing diagonal attention. "Off" denotes full global attention, while "On" denotes the proposed diagonal attention mechanism.

CBM by 11–18 percentage points and surpasses two black-box baselines. For HMDB51, UCF101, and SSv2, gains are smaller but consistent, likely because these datasets contain shorter clips where mean pooling is already strong, especially with PE embeddings. Although a gap to some black-box models remains, especially on SSv2, MoTIF substantially narrows this gap while retaining an interpretable concept bottleneck.

Table 2. **Comparison with DANCE.** Top-1 accuracy on UCF101, HAA-100, and HAA-500. For fair comparison, we report only results available in the original publications.

Method	Backbone	UCF101	HAA-100	HAA-500
DANCE (Lee et al., 2025a)	Baseline w/o interp.	88.4	73.5	–
DANCE (Lee et al., 2025a)	DANCE	87.5	70.7	–
LF-CBM (Oikarinen et al., 2023)	Disentangled concepts	85.5	66.5	–
MoTIF	ViT-B/32	88.5 ± 0.6	61.3	55.3
MoTIF	PE-L/14	94.8 ± 0.4	87.8	80.9
MoTIF	PE-G/14	**98.0** ± 0.2	**89.9**	**84.1**

We further compare MoTIF against DANCE (Lee et al., 2025a), a neighboring baseline for explainable video action recognition. Unlike DANCE, MoTIF explicitly models temporally localized concept activations and their evolution across windows. To clarify this distinction, we report results on UCF101, HAA-100, and HAA-500 (Chung et al., 2021) in Table 2. For HAA-100, we follow the class merging protocol and use the reported results by Lee et al. (2025a). The MoTIF variants outperform DANCE on UCF101 and HAA-100, and additionally provides results on the larger HAA-500 benchmark. These results suggest that MoTIF scales well while preserving the interpretability benefits of CBMs.

3.2.1. Ablations

We conduct ablation studies to assess key design choices in MoTIF. Unless stated otherwise, experiments on Breakfast Actions and HMDB51 use CLIP ViT-B/32 and RN/50 backbones, with hyperparameters in Appendix A.1 and the GPT-5 based concepts listed in Appendix E.2. This setup spans both transformer-based (Dosovitskiy et al., 2020) and CNN-based (He et al., 2016) visual extractors, as well as datasets with different scale, variability, and action granularity. Additional ablations are provided in Appendix C.

Attention variant (full vs. diagonal). While MoTIF enforces concept isolation, we also evaluate a variant with full multi-head attention. Figure 2 compares train and test accuracy for both settings. This comparison explicitly evaluates the impact of cross-concept interactions on both predictive performance and interpretability. Diagonal attention exposes a clear accuracy–interpretability trade-off: enabling cross-concept attention recovers accuracy on demanding datasets, while diagonal attention preserves the cleanest single-concept attribution, allowing controlled analysis of how temporal interactions affect both prediction and explanation. On temporally demanding datasets such as SSv2, this trade-off becomes more pronounced, with full attention yielding up to 10.1% higher test accuracy. We further illustrate the effect on explanations in Section B.2: when channels are mixed, the most influential dimensions become harder to map back to stable individual concepts, substantially reducing interpretability.

Temporal sensitivity and dynamic concepts. A key challenge in video concept learning is preventing models from defaulting to temporally order-invariant reasoning (Bertasius et al., 2021). To assess whether MoTIF captures temporal structure, we evaluate temporal dependence in two complementary ways.

Table 3. **Temporal sensitivity analysis and bottleneck comparison (PE-L/14).** Shuffling indicates random permutation of windows at evaluation time; synthetic uses 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

First, we construct a controlled synthetic benchmark (1,989 sequences, matching the number of Breakfast training instances) in which class identity depends exclusively on frame ordering (ascending, descending, U-shaped, inverted-U, and periodic patterns), such that static appearance cues are uninformative by design (for details see Appendix B.4). MoTIF achieves 86.97% accuracy on ordered sequences but collapses to 21.06% under random frame shuffling (chance $\approx 20\%$), yielding a $4.1\times$ performance gap and confirming that, with fixed sinusoidal temporal positional encoding, MoTIF relies on temporal order when required. Moreover, on this benchmark the static image-level bottleneck baseline (Global CBM), achieves only 35.5% accuracy, underscoring that temporal modeling is essential when order defines the class. Second, in Table 3, we apply the same shuffling intervention to real-world datasets. While some classes admit appearance-based shortcuts, the real-data drops are modest

on Breakfast, HMDB51, and UCF101, and largest on SSv2. This pattern suggests that temporal order is most decisive on SSv2, whereas the other benchmarks remain partly solvable from appearance cues alone.

3.2.2. Explanations

A central goal of MoTIF is to provide interpretable insights into the decision-making process of video classification models. In these examples, the temporal maps are not intended as standalone semantic explanations. Instead, they indicate which temporal windows provide contextual support for updating the queried concept channel, and are interpreted jointly with the local and global concept evidence.

In Figure 3, the first example from Breakfast shows preparing a *sandwich*. Global and local concept evidence highlight *bread* and *bagel* as important for the prediction. For the concept *bagel*, the temporal map exhibits strong vertical stripes, indicating that many query frames use the same key frames as contextual support when updating that concept channel. This suggests that *bagel* is concentrated in a few characteristic windows, while its temporally propagated representation contributes across the sequence. The second example, from UCF101, depicts *kayaking*. Here, the most salient concepts are *paddle* and *kayak*, which remain stable over the short clip. The temporal map for *paddle* is more uniform across time, with slight emphasis at $u = 3$ and $u = 6$, indicating that contextual support for this concept is distributed across several windows rather than concentrated in a single temporal anchor. We provide further examples in the supplementary material (Appendix B.1 discusses misclassifications, and Appendix B.2 without diagonal attention), and all data will be released upon acceptance.

3.2.3. Concept Interventions

Bottleneck-sensitivity analysis. We evaluate MoTIF's bottleneck with two intervention types: *destructive interventions*, which test whether predictions depend on sparse concept activations, and *corrective interventions*, which test whether editing these activations can repair errors. For destructive analysis, Table 4 reports normalized *prediction overlap*, the fraction of post-intervention predictions matching the original MoTIF prediction, with the unperturbed model set to 1.0. We test global, local-slot, and window interventions. *Global Top-k* zeros the k most influential concept channels across all time steps, while *Global Rand.* removes k random channels as a non-targeted control. *Local Slot Top-k* zeros the k most influential individual concept-window entries, while *Window Top-k* zeros all concepts in the k most influential temporal windows and *Window Rand.* zeros random windows. For corrective interventions, we report *top-1 repair rate* on 30 misclassified instances whose ground-truth class appears in the top-5 logits. *Global Edit* manually changes up to k selected concepts across all windows, whereas *Local Edit* changes up to k selected concept-window entries.

Table 4 separates destructive sensitivity from corrective repair. The unperturbed model has value 1.0. Global top-k removal has the strongest destructive effect: at $k = 4$, overlap drops to 0.028 on Breakfast and 0.142 on HMDB51, while random global removal remains above 0.90 on both datasets. Local slot removal is weaker, which may reflect both temporal redundancy and the smaller perturbation size relative to global concept removal, whereas top-ranked window removal is consistently stronger than random window removal. For corrective interventions, global edits are substantially more effective than local edits: at $k = 4$, global edits repair 80% of Breakfast cases and 83% of HMDB51 cases, compared with 20% and 30% for local edits. Repair rates increase monotonically with k, suggesting that a small number of globally edited concepts is often sufficient to redirect the prediction to the correct class, while isolated concept-window edits provide more limited corrective control. Overall, these results show that MoTIF is both sensitive to meaningful concept perturbations and intervenable through targeted concept edits.

Qualitative intervention examples. The same mechanism also supports instance-level inspection. In Figure 1, zeroing the most influential global concept *bow* changes the prediction from the correct class to *run*; the original correct-class logit is 8.20, while the post-intervention *run* logit is 6.79. Removing windows 1–4, where the bow is handled, instead shifts the prediction to *talk* with logit 6.75. Conversely, corrective edits can repair errors. For UCF101 video `v_ApplyLipstick_g21_c01`, MoTIF predicts *Shaving-Beard* instead of *ApplyLipstick*; zeroing the misleading concept *barber chair*, which fires because the person is seated, changes the prediction to the correct class. For SSV2 instance `100255`, the model predicts *Pretending to scoop something up with something* instead of *Scooping something up with something*; setting the misleading local concept *pouring* to zero only in the first window repairs the prediction.

4. Related Work

Concept bottleneck models. CBMs (Koh et al., 2020) predict concepts as intermediate features before the final class prediction (Prasse et al., 2025; Yang et al., 2023; Havasi et al., 2022; Chauhan et al., 2023; Sawada & Nakamura, 2022). Recent research has focused on automatic concept discovery: DCLIP aligns data with CLIP's vision–language space (Menon & Vondrick, 2023), LaBo queries large language models for diverse candidate concepts (Yang et al., 2023), and DCBM leverages segmentation foundation models to extract object- and part-level concepts (Prasse et al.,

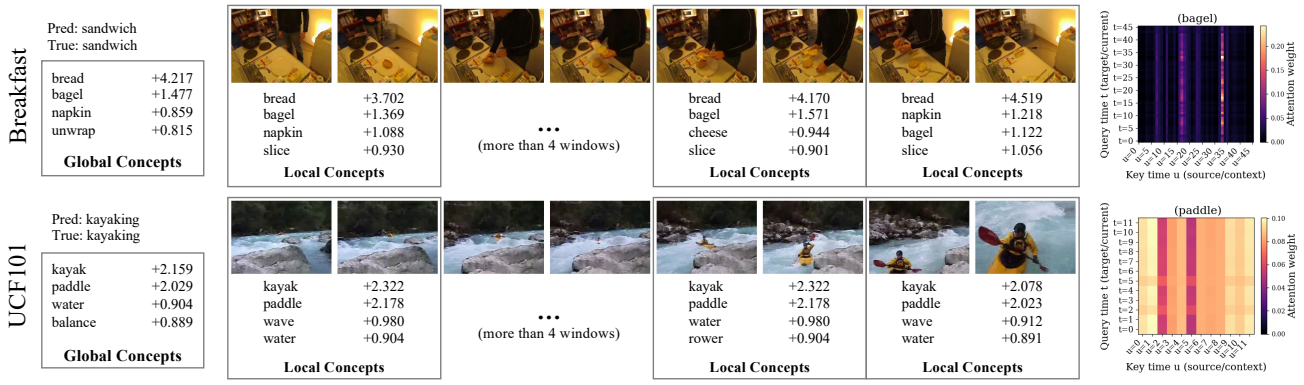


Figure 3. **MoTIF explanations.** Example videos from Breakfast and UCF101 with correct classifications, illustrating the three explanation modes supported by MoTIF (ViT-L14).

Table 4. **Concept interventions.** Destructive columns report normalized prediction overlap after intervention (lower is stronger). Corrective columns report top-1 repair rate on misclassified top-5 cases after manual oracle edits (higher is better).

Dataset	k	Destructive: normalized prediction overlap ↓					Corrective: top-1 repair rate ↑	
		Global Top-k	Global Rand.	Local Slot Top-k	Window Top-k	Window Rand.	Global Edit	Local Edit
Breakfast	0	1.000	1.000	1.000	1.000	1.000	–	–
	1	0.496	0.972	0.954	0.866	0.986	0.20	0.03
	2	0.229	0.950	0.933	0.775	0.977	0.47	0.10
	3	0.085	0.937	0.908	0.754	0.973	0.57	0.17
	4	0.028	0.909	0.891	0.732	0.960	0.80	0.20
HMDB51	0	1.000	1.000	1.000	1.000	1.000	–	–
	1	0.603	0.975	0.934	0.875	0.963	0.47	0.10
	2	0.374	0.959	0.892	0.801	0.947	0.60	0.23
	3	0.238	0.942	0.852	0.731	0.918	0.70	0.27
	4	0.142	0.926	0.809	0.674	0.886	0.83	0.30

2025). Other works have extended this idea to different modalities. For instance, Ismail et al. (2025) adapt the bottleneck principle to protein design, where interpretable biochemical features form the concept space. Wu et al. (2022) extend CBMs to medical time series data, demonstrating that clinically meaningful temporal features can act as concepts. More recently, Sun et al. (2025) apply the CBM framework to language models, where intermediate concepts correspond to interpretable linguistic or semantic units. More broadly, MoTIF's separation of information flow across concept channels is related to modular architectures such as Recurrent Independent Mechanisms (Goyal et al., 2021), which also structure computation through partially independent pathways.

CBMs have been widely studied for images, but remain underexplored for sequential data. Prior video extensions either use concepts extracted from video descriptions at a global level (Jeyakumar et al., 2022) or disentangle pose-based and textual concepts for comparison (Lee et al., 2025a). In contrast, MoTIF extends the CBM principle to enable both window-level and global concept attributions for more fine-grained explanations, while also improving performance and remaining flexible with respect to the use of different text–image aligned models and VLMs for concept discovery.

Video classification and action recognition. Video classification not only assigns a label to an entire sequence but, in the case of action recognition, also requires identifying the actions occurring within it (Pareek & Thakkar, 2021). Both tasks have progressed from CNN-based architectures (Liu et al., 2021; Lin et al., 2019; Tran et al., 2015) to transformer-based (Wang et al., 2024) models such as TimeSformer (Bertasius et al., 2021) and VideoMAE (Tong et al., 2022), which capture long-range dependencies and leverage large-scale self-supervised pre-training. MoTIF integrates these approaches by adapting a transformer block for temporal modeling alongside a per-channel scaling operation — distinct from temporal convolution — applied uniformly across all timesteps.

Temporal concept modeling. While concept-based explanations have been widely studied in the image domain (Knab et al., 2025a; Ghorbani et al., 2019), only a few works explore temporal dynamics (Gulshad et al., 2023; Kowal et al., 2024). Ji et al. (2023a) propose a spatio-temporal concept framework to analyze representations in 3D ConvNets. PCBEAR (Lee et al., 2025b) introduces static and dynamic pose concepts for action recognition, and Saha et al. (2024) extend TCAV to videos by computing concept importance scores across sequences. (Ji et al., 2023b) explain video models post hoc by automatically discovering spatial-temporal concepts from supervoxels and scor-

ing their importance for 3D ConvNet predictions. Related object-centric approaches process learned symbolic entities over time for downstream reasoning, including attention over learned object embeddings (Ding et al., 2021) and parallelized spatiotemporal slot binding for videos (Singh et al., 2024). In contrast, MoTIF integrates concept-specific temporal attention directly into the predictive model, enabling reasoning over evolving concepts rather than treating them as fixed inputs or external explanations.

5. Discussion

MoTIF introduces a transformer-based concept bottleneck architecture for temporal data. The core contribution of MoTIF is a temporal concept reasoning module that operates on sequences of interpretable concept activations. By combining per-concept temporal self-attention with transformer-style processing, MoTIF models *when* concepts recur and how their temporal patterns contribute to video-level predictions, while preserving explicit concept attribution. Across datasets, this design improves over zero-shot classification, global CBMs that average over windows, and prior video CBM baselines, showing that temporal reasoning can be integrated into concept bottleneck models without discarding concept-level structure.

Performance depends on concept quality as well as architecture. While MoTIF provides the reasoning structure, its performance depends, as in other CBMs (Rao et al., 2024; Oikarinen et al., 2023; Sawada & Nakamura, 2022), on the availability of concepts that reflect actions and interactions over time. VLM-based concept discovery is especially beneficial on temporally demanding datasets such as SSv2, and appendix experiments with SAM3-style visual concepts further suggest that MoTIF is not restricted to textual concept banks, although text remains stronger in our current setup. These findings suggest that architecture and concept discovery are complementary: MoTIF provides the temporal reasoning mechanism, while stronger concept banks improve its ability to handle challenging datasets.

MoTIF remains modular while preserving interpretable temporal concept analysis. We show that MoTIF can be paired with a range of embedding backbones, including CLIP, SigLIP (Zhai et al., 2023), and video-adapted Perception Encoders, with consistent gains across architectures. MoTIF-ST further extends the framework to space–time transformers, trading some of the strict per-concept separation of diagonal MoTIF for higher expressivity. At the same time, strict concept isolation is not free: its overhead is small on short clips but becomes substantially larger on long videos (Appendix A.1). Beyond predictive performance, MoTIF's three-level explanation interface—global, local, and temporal—exposes how concepts contribute across time without the entanglement introduced by standard attention mechanisms. This makes it possible to trace errors back to missing or ambiguous temporal concepts rather than opaque feature interactions, helping refine both the concept set and model behavior while maintaining competitive performance.

Challenges and future work. Selecting an appropriate temporal window size remains important, as actions span variable durations and temporal granularities. Adaptive or multi-scale windowing strategies could further improve robustness without sacrificing interpretability. More broadly, performance depends on concept-bank quality and coverage, especially for domain-specific and fine-grained dynamic concepts, suggesting that advances in concept quality may yield larger gains than architectural scaling alone. Finally, future work may explore selectively enabling concept interactions in a controlled and interpretable manner, reinforcing the value of architectures that make temporal reasoning explicit rather than implicit.

6. Conclusion

We introduced MoTIF, a transformer-based concept bottleneck framework for video that combines temporal concept reasoning with VLM-based concept discovery. MoTIF uses diagonal temporal attention to track concepts over time, providing a traceable account of *when* and *which* concepts support predictions. This exposes a controllable accuracy–interpretability trade-off: cross-concept attention can recover additional performance on demanding datasets, while diagonal attention preserves the cleanest concept-level decomposition. Overall, MoTIF provides a principled framework for learning, analyzing, intervening on, and explaining temporal concept representations, achieving state-of-the-art performance among CBMs while remaining competitive with black-box video classifiers.

Impact Statement

MoTIF aims to enhance the interpretability of video classification by explicitly exposing which semantic concepts are activated and when they occur within a temporal sequence. By providing global, local, and temporal explanations, MoTIF supports more transparent and trustworthy deployment of video models in high-stakes settings such as human–robot interaction, surveillance, and assistive technologies, where understanding model decisions over time is critical.

Acknowledgments

This research was supported in part by the German Federal Ministry for Economic Affairs and Climate Action of Germany (BMWK) and in part by the German Federal Ministry for Research, Technology, and Space (BMFTR), and was partially conducted during an internship at Ramblr.ai.

References

Bertasius, G., Wang, H., and Torresani, L. Is space-time attention all you need for video understanding? In *ICML*, volume 2, pp. 4, 2021.

Bolya, D., Huang, P.-Y., Sun, P., Cho, J. H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Rasheed, H., Wang, J., Monteiro, M., Xu, H., Dong, S., Ravi, N., Li, D., Dollár, P., and Feichtenhofer, C. Perception encoder: The best visual embeddings are not at the output of the network, 2025. URL https://arxiv.org/abs/2504.13181.

Carion, N., Gustafson, L., Hu, Y.-T., Debnath, S., Hu, R., Coll-Vinent, D. S., Ryali, C., Alwala, K. V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.-H., Zhou, Y., Momeni, L., HAZRA, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H. K., Dollar, P., Ravi, N., Saenko, K., Zhang, P., and Feichtenhofer, C. SAM 3: Segment anything with concepts. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=r35clVtGzw.

Chauhan, K., Tiwari, R., Freyberg, J., Shenoy, P., and Dvijotham, K. Interactive concept bottleneck models. In *Proceedings of the aaai conference on artificial intelligence*, volume 37, pp. 5948–5955, 2023.

Chen, D., Moutakanni, T., Chung, W., Bang, Y., Ji, Z., Bolourchi, A., and Fung, P. Planning with reasoning using vision language world model. *arXiv preprint arXiv:2509.02722*, 2025.

Chung, J., hsin Wuu, C., ru Yang, H., Tai, Y.-W., and Tang, C.-K. Haa500: Human-centric atomic action dataset with curated videos. In *ICCV 2021*, 2021.

Ding, D., Hill, F., Santoro, A., Reynolds, M., and Botvinick, M. Attention over learned object embeddings enables complex visual reasoning. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=lHmhW2zmVN.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Ghorbani, A., Wexler, J., Zou, J. Y., and Kim, B. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

Goyal, A., Lamb, A., Hoffmann, J., Sodhani, S., Levine, S., Bengio, Y., and Schölkopf, B. Recurrent independent mechanisms. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=mLcmdlEUxy-.

Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pp. 5842–5850, 2017.

Gulshad, S., Long, T., and van Noord, N. Hierarchical explanations for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3703–3708, 2023.

Hao, Y., Dong, L., Wei, F., and Xu, K. Self-attention attribution: Interpreting information interactions inside transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 12963–12971, 2021.

Havasi, M., Parbhoo, S., and Doshi-Velez, F. Addressing leakage in concept bottleneck models. *Advances in Neural Information Processing Systems*, 35:23386–23397, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Ismail, A. A., Oikarinen, T., Wang, A., Adebayo, J., Stanton, S. D., Bravo, H. C., Cho, K., and Frey, N. C. Concept bottleneck language models for protein design. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Yt9CFhOOFe.

Jeyakumar, J. V., Dickens, L., Cheng, Y.-H., Noor, J., Garcia, L. A., Echavarria, D. R., Russo, A., Kaplan, L. M., and Srivastava, M. Automatic concept extraction for concept bottleneck-based video classification, 2022. URL https://openreview.net/forum?id=66kgCIYQW3.

Ji, Y., Wang, Y., and Kato, J. Spatial-temporal concept based explanation of 3d convnets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15444–15453, 2023a.

Ji, Y., Wang, Y., and Kato, J. Spatial-temporal concept based explanation of 3d convnets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15444–15453, June 2023b.

Knab, P., Marton, S., and Bartelt, C. Beyond pixels: Enhancing LIME with hierarchical features and segmentation foundation models. In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025a. URL https://openreview.net/forum?id=JHs5p6nPbG.

Knab, P., Marton, S., Schlegel, U., and Bartelt, C. Which lime should i trust? concepts, challenges, and solutions, 2025b. URL https://arxiv.org/abs/2503.24365.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept bottleneck models. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Kowal, M., Dave, A., Ambrus, R., Gaidon, A., Derpanis, K. G., and Tokmakov, P. Understanding video transformers via universal concept discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10946–10956, 2024.

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

Kuehne, H., Arslan, A. B., and Serre, T. 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.

Lee, J., Lee, W., Park, G.-M., Kim, S. T., and Choi, J. Disentangled concepts speak louder than words: Explainable video action recognition. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=paRLw86ONU.

Lee, J., Lee, W., Park, G.-M., Kim, S. T., and Choi, J. Pcbear: Pose concept bottleneck for explainable action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2690–2699, June 2025b.

Lin, J., Gan, C., and Han, S. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7083–7093, 2019.

Liu, X., Pintea, S. L., Nejadasl, F. K., Booij, O., and van Gemert, J. C. No frame left behind: Full video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14892–14901, June 2021.

Materzynska, J., Xiao, T., Herzig, R., Xu, H., Wang, X., and Darrell, T. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *CVPR*, 2020.

Menon, S. and Vondrick, C. Visual classification via description from large language models. In *International Conference on Learning Representations*, 2023.

Molnar, C., Casalicchio, G., and Bischl, B. Interpretable machine learning – a brief history, state-of-the-art and challenges. In Koprinska, I., Kamp, M., Appice, A., Loglisci, C., Antonie, L., Zimmermann, A., Guidotti, R., Özgöbek, Ö., Ribeiro, R. P., Gavaldà, R., Gama, J., Adilova, L., Krishnamurthy, Y., Ferreira, P. M., Malerba, D., Medeiros, I., Ceci, M., Manco, G., Masciari, E., Ras, Z. W., Christen, P., Ntoutsi, E., Schubert, E., Zimek, A., Monreale, A., Biecek, P., Rinzivillo, S., Kille, B., Lommatzsch, A., and Gulla, J. A. (eds.), *ECML PKDD 2020 Workshops*, pp. 417–431, Cham, 2020. Springer International Publishing. ISBN 978-3-030-65965-3.

Oikarinen, T., Das, S., Nguyen, L. M., and Weng, T.-W. Label-free concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FlCg47MNvBA.

Pareek, P. and Thakkar, A. A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. *Artificial Intelligence Review*, 54(3): 2259–2322, 2021.

Prasse, K., Knab, P., Marton, S., Bartelt, C., and Keuper, M. DCBM: Data-efficient visual concept bottleneck models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=BdO4R6XxUH.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Rao, S., Mahajan, S., Böhle, M., and Schiele, B. 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*, pp. 444–461, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72979-9. doi: 10.1007/978-3-031-72980-5_26. URL https://doi.org/10.1007/978-3-031-72980-5_26.

Saha, A., Gupta, S., Ankireddy, S. K., Chahine, K., and Ghosh, J. Exploring explainability in video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 8176–8181, June 2024.

Sawada, Y. and Nakamura, K. Concept bottleneck model with additional unsupervised concepts. *IEEE Access*, 10: 41758–41765, 2022.

Schrodi, S., Schur, J., Argus, M., and Brox, T. Selective concept bottleneck models without predefined concepts. *Transactions on Machine Learning Research (TMLR)*, May 2025. URL http://lmb.informatik.uni-freiburg.de/Publications/2025/SAB25.

Singh, G., Wang, Y., Yang, J., Ivanovic, B., Ahn, S., Pavone, M., and Che, T. Parallelized spatiotemporal slot binding for videos. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=KpeGdDzucX.

Soomro, K., Zamir, A. R., and Shah, M. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision*, 2(11):1–7, 2012.

Sun, C.-E., Oikarinen, T., Ustun, B., and Weng, T.-W. Concept bottleneck large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=RC5FPYVQaH.

Tong, Z., Song, Y., Wang, J., and Wang, L. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, 2015.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, J., Xia, L., and Wen, X. Cmf-transformer: cross-modal fusion transformer for human action recognition. *Mach. Vision Appl.*, 35(5), August 2024. ISSN 0932-8092. doi: 10.1007/s00138-024-01598-0. URL https://doi.org/10.1007/s00138-024-01598-0.

Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., and Qiao, Y. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14549–14560, 2023.

Wang, X., Yan, Y., Tang, P., Bai, X., and Liu, W. Revisiting multiple instance neural networks. *Pattern recognition*, 74:15–24, 2018.

Wu, C., Parbhoo, S., Havasi, M., and Doshi-Velez, F. Learning optimal summaries of clinical time-series with concept bottleneck models. In Lipton, Z., Ranganath, R., Sendak, M., Sjoding, M., and Yeung, S. (eds.), *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pp. 648–672. PMLR, 05–06 Aug 2022. URL https://proceedings.mlr.press/v182/wu22a.html.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., and Yatskar, M. 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*, pp. 19187–19197, 2023.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, October 2023.

Zhang, R., Madumal, P., Miller, T., Ehinger, K. A., and Rubinstein, B. I. Invertible concept-based explanations for cnn models with non-negative concept activation vectors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11682–11690, 2021.

A. Implementation Details

A.1. Hyperparameter Settings

Unless noted otherwise, all experiments in the main paper follow these default hyperparameters:

- **Training.** 100 epochs, batch size 32, AdamW with learning rate 10^{-3} and weight decay 10^{-2}.
- **Pooling.** Log-sum-exp pooling with fixed temperature $\tau = 1.0$. Although LSE permits tuning of sharpness, we kept τ constant for unbiased comparisons.
- **Regularization.** Both the ℓ_1 penalty on classifier weights and the activation sparsity penalty are set to 10^{-3}.
- **Architecture.** One Transformer layer with per-channel (diagonal) attention; classifier weights constrained to be nonnegative. As stated in Section 2, the diagonal attention faces the trade-off: interpretability through concept isolation versus representational power. Thus, additional depth yields diminishing returns under strict concept isolation, motivating future work on selectively enabling cross-concept interactions.
- **Data.** Window size of 16 frames per temporal unit. For HMDB51 and Breakfast we report results on split *s1*, for UCF101 on *testlist01*, and for HMDB51 on *split1*. Class weighting is applied to mitigate imbalance.

Dataset-specific deviations: for SSv2 and UCF101 we use learning rate 10^{-4}, ℓ_1 penalty 10^{-4}, and sparsity penalty 10^{-4}; for SSv2, the non-negativity constraint in AdamW is disabled and batch size 128. HMDB51 uses a shorter window size of 8. For Breakfast, the window size is increased to 32 to account for longer videos.

All experiments were run with a random seed of 42 for reproducibility. Section C further analyzes the sensitivity to random seed choice.

These values serve as the baseline configuration; modifications for ablation studies are detailed in Section 3.2.1.

A.2. Computation Time and Complexity

Table 5 reports GPU memory (MB), epoch time (s) and throughput (samples/s) for HMDB51 and Breakfast using the two ablated clip backbones. Results are shown for diagonal attention enabled (On) and disabled (Off).

Table 5. **Complexity overview.** GPU memory, epoch time and throughput for RN/50 and B32 backbones with diagonal attention On/Off.

Dataset	Backbone	Setting	GPU memory (MB)	Epoch time (s)	Samples / s
HMDB51	RN/50	On	1723	0.91	4084
		Off	517	0.97	4063
	B32	On	2289	0.90	4108
		Off	1064	0.95	4071
Breakfast	RN/50	On	10634	4.26	454
		Off	736	0.51	3461
	B32	On	11463	4.28	445
		Off	1565	0.53	3445
Average T (train set)				HMDB51: 12.5	Breakfast: 65.4

For short videos (small T) epoch times remain nearly unchanged while GPU memory differs substantially. For longer sequences, as in Breakfast, disabling diagonal attention reduces memory use considerably; the runtime and memory gap grows with sequence length since complexity increases with T (see Sec. 2) if $H << C$. The runtime across backbones is similar in magnitude; for example the best-performing perception encoder in MoTIF required 4.55 s. Video embedding time is excluded because embeddings are reusable across models and vary with the embedding backbone.

A.3. Backbone Integration

Image-based. Image–text models such as CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023) embed single images rather than videos. To adapt them, we divide each video into windows of F frames and randomly select one frame per

window. This frame is embedded and serves as the window representation. Random sampling ensures variation in which part of the window is captured.

Video-based. The Perception Encoder (Bolya et al., 2025) is explicitly tuned to aggregate information across frames, producing a pooled embedding for each window. Unlike the image-based backbones, we therefore use multiple frames per window. In our experiments, we consistently used eight frames per window, except for HMDB51where the window size itself was eight, so only four frames were sampled.

A.4. Space-Time Attention Extension

While the standard MoTIF architecture applies temporal attention independently for each concept, we extend it with a factorized space-time attention mechanism (Bertasius et al., 2021) that enables concept interactions while preserving interpretability. This design is also related to recent axial-style, modular video binding approaches that preserve slot structure across time (Singh et al., 2024).

The *CBMTransformerST* architecture factorizes attention into two sequential components: spatial attention across concepts at each time step, followed by the original per-channel temporal attention.

PerTimeSpatialBlock. This module computes attention across concepts at each time step independently. Given input features $X \in \mathbb{R}^{B\times T\times C}$, spatial attention produces attention scores $W_s \in \mathbb{R}^{B\times T\times C\times C}$ where $W_s[b, t, i, j]$ represents the attention weight from concept i to concept j at time step t.

Unlike temporal attention, which maintains concept isolation, spatial attention allows concepts to interact within each temporal frame. To preserve concept interpretability while enabling controlled spatial interaction, we employ three identity-preserving mechanisms:

- **Identity bias**: We add a bias term $\alpha_I = 1.0$ to the diagonal of the attention score matrix, encouraging self-attention and reducing cross-concept mixing.
- **Spatial gating**: The spatial attention output is scaled by a gating factor $\beta_s \in [0, 1]$ before being added to the residual connection: $X' = X + \beta_s \cdot \text{SpatialAttn}(X)$. With $\beta_s = 0.1$ (default), the spatial branch contributes a 0.1-scaled residual update, limiting but not eliminating cross-concept mixing.
- **Per-channel FFN**: The feed-forward network in the spatial block uses per-channel convolutions (grouped by C), ensuring no cross-concept mixing occurs in the FFN, matching the design of the temporal block.

These mechanisms ensure that concepts remain separable and interpretable: the spatial mixing is controlled and identity-preserving, the subsequent per-channel temporal attention maintains diagonal structure, and the final concept activations $Z_{t,c}$ and spatial attention maps W_s enable attribution to individual concepts while revealing their interactions.

SpaceTimeBlock. The block applies spatial attention first to enable concept interactions at each time step, followed by per-channel temporal attention that maintains diagonal structure.

This factorization reduces computational complexity from $\mathcal{O}(T^2C^2)$ for full space-time attention to $\mathcal{O}(TC^2 + CT^2)$, while producing separate interpretable attention maps for spatial ($W_s \in \mathbb{R}^{B\times T\times C\times C}$) and temporal ($W_t \in \mathbb{R}^{B\times C\times T\times T}$) components.

Exemplary Explanations. MoTIF with CBMTransformerST (MoTIF-ST) applies two sequential attention mechanisms—spatial followed by temporal—yielding two complementary attention matrices for interpretation.

- The spatial attention block captures concept–concept interactions within each frame. It indicates which concepts influence a given concept and highlights additional concepts that are relevant at that moment. This provides a structured view of how concepts relate to each other spatially.
- The temporal attention block captures when a concept becomes important. It identifies the time steps at which a specific concept contributes most to the final prediction, thereby revealing the temporal structure of the activity.

Figure 4 shows two examples from Breakfast and UCF101, including their attention matrices and the corresponding explanations.

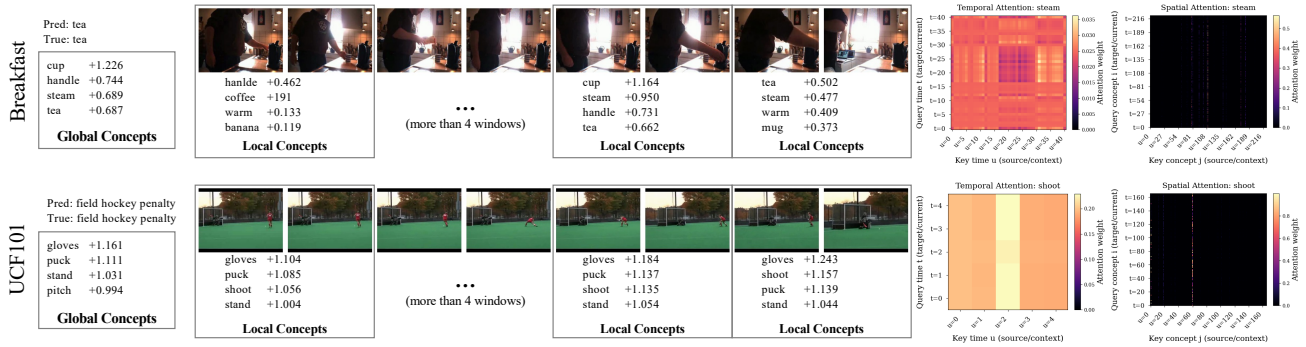


Figure 4. **MoTIF-ST explanations.** Correctly classified examples from Breakfast and UCF101, illustrating MoTIF with a space–time transformer (ViT-L/14) and the corresponding temporal and spatial attention matrices.

B. Additional Experiments

B.1. Where *motifs* help to understand Misclassifications

In Figure 5, we illustrate representative failure cases for each dataset using models trained with ViT-L/14. All corresponding videos, including the full set of temporal concepts across all windows, will be provided in the supplementary material.

Figure 5. **MoTIF explanations.** Example videos from all datasets with incorrect classifications.

The first example, from *Pirates of the Caribbean*, shows Barbossa eating an apple before throwing it away. Our model predicts *eat*, while the ground truth is *throw*. Global and local concept attributions reveal that the concept *eat*, triggered by the apple, was most strongly activated. The corresponding attention map illustrates that early query frames attend to early key frames, indicating that the model anchors the decision on the moment where the apple is clearly visible and eaten. This concentrated attention explains why the model emphasizes *eat* over *throw*.

In the second example from UCF101, MoTIF detects correct concepts, such as *basketball* and *scoreboard*. However, these were insufficient to discriminate between the actions *basketball dunk* and *basketball*, leading to misclassification. Since the

background frame remains nearly constant and only the players on the court are moving, the attention map for *scoreboard* shows a uniform distribution across time steps, with no clear temporal anchors. This indicates that the concept is consistently present.

The third example, from SSv2, highlights the dataset's inherent difficulty. Unlike the correctly classified case in Figure 12, MoTIF predicts *tearing sth just a little bit* instead of the ground truth *tearing sth into two pieces*. Concept activations focus primarily on hand movements, such as *crumple*, together with the *paper*. The attention map shows that the concept *crumple* receives strongest attention from the later key frames (u=2, u=3) across several query times while other frames receive lower, more diffuse weights.This diffuse attribution explains why MoTIF captures the general action but fails to resolve the fine-grained distinction required by SSv2.

Finally, in the Breakfast dataset, MoTIF correctly identifies and attends to concepts relevant for *egg*. Yet, the dataset contains two distinct egg-related actions, *friedegg* and *scrambledegg*, which leads to misclassification. Nevertheless, all identified concepts correspond to actions of a person cooking with eggs. Furthermore, the attention map for *grill* reveals a broad and diffuse distribution across many time steps, indicating that the concept is persistently active throughout the sequence rather than concentrated in a few decisive moments. This persistent but unspecific attention explains why MoTIF captures the general presence of egg-related cooking but fails to distinguish between the two fine-grained classes.

Although MoTIF does not always predict the correct class, its structure makes the reasoning process transparent by decomposing decisions into concepts and their temporal interactions. In this work, our emphasis was on the architecture rather than on designing the most suitable concept sets. We expect that future advances in concept extraction for video data will further improve performance, complementing MoTIF's interpretability with stronger predictive accuracy.

B.2. Diagonal vs. Full Attention

We illustrate why full attention produces non-interpretable results by revisiting previously shown examples. In Figure 1, replacing diagonal attention with full attention shifts the most important concepts to *frown* (activation 4.796) and *face* (1.282), while all others fall below 0.01 and are omitted. Although these concepts appear as locally relevant, they do not correspond to the depicted action (class *bow*). This indicates that concept mixing occurs: full attention entangles channels, creates arbitrary concepts that are not visually apparent and thus undermine interpretability.

B.3. Interpreting Temporal Dependencies

An attention map visualizes how a concept channel distributes its focus across time. Each entry $W_{c,t,u}$ encodes the attention weight between query time step t (vertical axis) and key/value time step u (horizontal axis). A bright cell at position (t, u) indicates that the activation of concept c at time t strongly attends to the representation of the same concept at time u. Diagonal patterns suggest that the concept mainly attends to itself at the same or nearby frames, while vertical stripes show that many query frames refer back to the same key frame, indicating the presence of a temporal anchor. In contrast, diffuse or uniform maps imply that the concept is expressed consistently across time rather than being tied to specific moments. Thus, by inspecting these maps, one can infer whether a concept is localized, persistent, or temporally linked to particular frames within the sequence.

B.4. Synthetic Temporal Data

The synthetic experiment is designed to isolate temporal reasoning by removing all appearance cues. Each "video" is a sequence of length-N (matching the Breakfast distribution), where each frame is a 1024-dimensional vector. We define five temporal classes: (1) ascending, (2) descending, (3) U-shape, (4) inverted-U, and (5) periodic oscillation. Class identity is determined only by the temporal trajectory of 30% of the dimensions; the remaining dimensions are filled with noise. Thus, the task can only be solved by learning the temporal order—static appearance information is uninformative by design.

To examine whether the model can learn interpretable temporal structure, we provide 37 artificial temporal pattern concepts. Each concept represents a prototypical analytic temporal signal (ascending, descending, S-curve, exponential growth/decay, chirp, sawtooth, step-up/down, periodic-fast, oscillating-decay, double-peak, plateau, etc.). All embeddings encode temporal shape only, without semantics.

The training setup matches our real pipelines, allowing a direct test of MoTIF's temporal inductive biases. MoTIF achieves 86.97% accuracy on the original (ordered) sequences but only 21.06% when frames are randomly permuted (chance 20%).

This 4.1× drop shows that MoTIF relies on temporal ordering rather than static frame statistics. An image-level CBM baseline ("Global CBM"), which collapses all frames into a single vector, reaches only 35.5%, confirming that methods without temporal modeling cannot solve this task.

C. Additional Ablations

For completeness, we report further ablation studies that complement the main paper (Section 3.2.1).

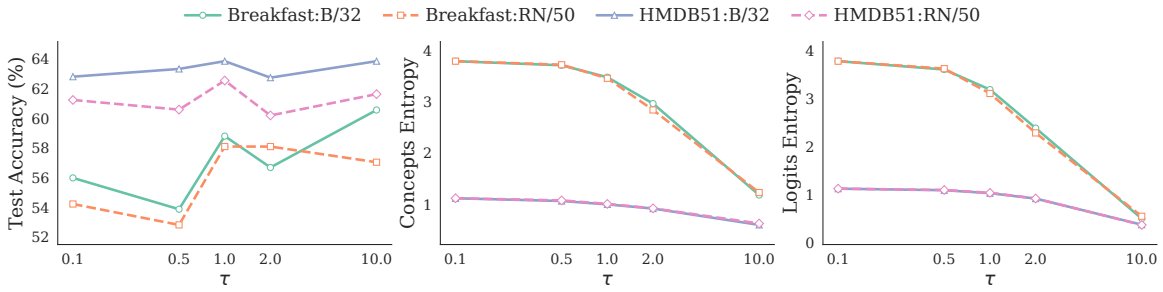
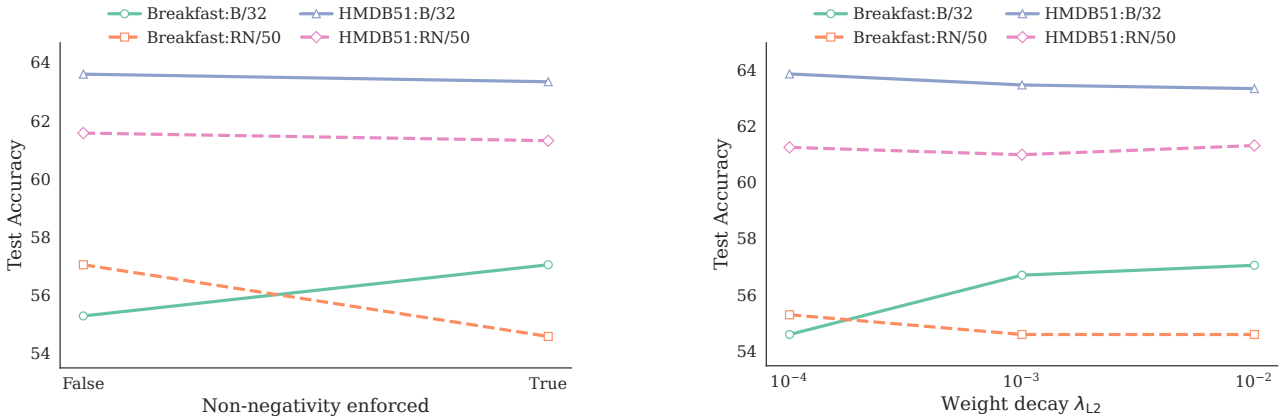


Figure 6. **Effect of log-sum-exp temperature τ on accuracy and entropy.** Accuracy is stable across τ, while the entropy of the temporal weighting changes systematically with the pooling sharpness.

Temperature τ. We vary the log-sum-exp pooling temperature τ to assess its effect on both accuracy and the sharpness of the time-importance distribution. Sharpness is quantified by the entropy of the softmax weights, computed either at the concept or at the logits level. Experiments show (see Figure 6) that accuracy varies only slightly across all tested values of τ, indicating robustness of predictive performance. Under this parameterization, smaller values of τ yield sharper, lower-entropy temporal attributions, whereas larger values produce smoother, more diffuse weighting across time. Hence, τ provides a controllable parameter: small values produce sharp, low-entropy explanations, whereas large values result in smoother, higher-entropy attributions, while accuracy remains relatively stable. Thus, this parameter can be tuned by the user.



(a) **Non-negativity.** Test accuracy with and without enforcing non-negativity.

(b) **Weight decay.** Test accuracy across different weight decays.

Figure 7. **Architectural choices.** Effects of non-negativity and weight decay.

Nonnegativity. The non-negativity constraint on classifier weights improves interpretability by ensuring class predictions are explained solely by positive concept evidence (Zhang et al., 2021). We enforce non-negativity of classifier weights W by projection after each update. Results are shown in Figure 7a. This constraint tends to cause minor accuracy reductions for RN/50 on both datasets, while for B/32 on Breakfast we observe a slight improvement. Given the small fluctuations inherent in training, these effects should be interpreted as indicative rather than strictly conclusive. Overall, the effect on test accuracy is negligible. We therefore enable it for all experiments except SSv2, where modeling negative concepts is required to capture fine-grained actions.

Weight decay. Figure 7b reports test accuracy across different weight decay values for the AdamW optimizer. Performance remains largely stable, similar to the non-negativity constraint, except for Breakfast using B/32. We therefore fix the weight decay to 10^{-2} for all experiments.

Per-concept affine transformation. We optionally insert a per-concept affine transformation between the *per-channel temporal self-attention* and the classification head. While not strictly required, this block sharpens activations by rescaling and shifting each concept dimension before the nonlinearity. Figure 9 shows the effect on Breakfast and HMDB51: enabling the affine transformation yields a slight increase in test accuracy and a consistent reduction in both logits and concept entropy, indicating sharper and more decisive concept activations, especially on the Breakfast dataset.

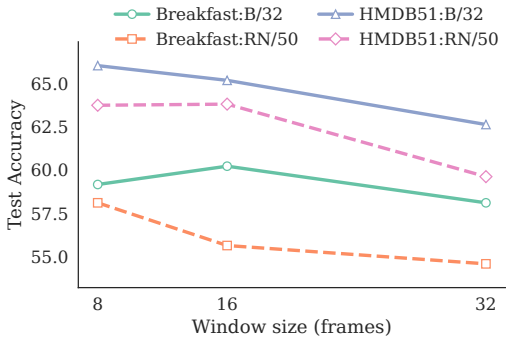


Figure 8. **Window size influence.** Test accuracy across different window sizes.

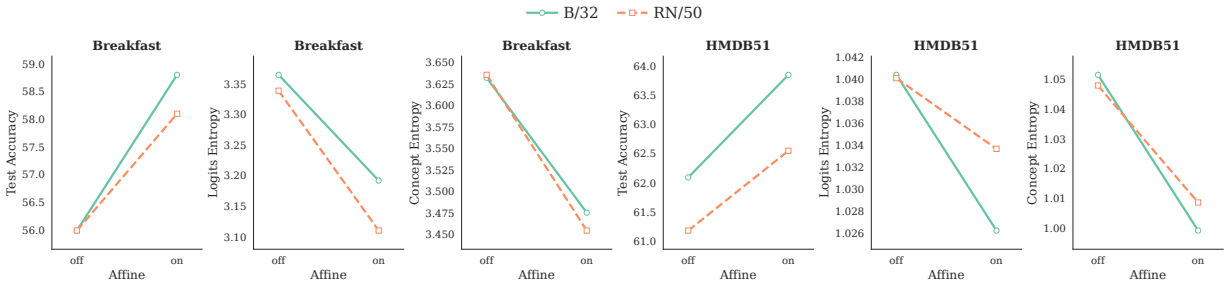


Figure 9. **Effect of the per-concept affine transformation.** Accuracy improves marginally, while entropy in both logits and concept activations decreases, suggesting that the affine block stabilizes and sharpens the CBM's internal representations.

Classifier sparsity. The sparsity penalty on classifier weights has a pronounced impact on test accuracy, as shown in Figure 10a. Larger values of λ_{ℓ_1} consistently reduce accuracy. We therefore set $\lambda_{\ell_1} = 10^{-3}$ in most experiments, as it offers a reasonable trade-off between regularization and performance.

Activation sparsity. The activation sparsity penalty shows little variation in test accuracy (see Figure 10b) across different values of λ_{sparse}, except for Breakfast. We attribute this to the comparatively long video sequences in that dataset. Accordingly, we use $\lambda_{\text{sparse}} = 10^{-3}$ for all experiments, except for Breakfast, where we set it to 10^{-4}.

Learning rate. The learning rate has a strong effect on test accuracy, as shown in Figure 10c. For Breakfast and HMDB51, we set it to 10^{-3}, while for UCF101 and SSv2, we use 10^{-4}, which yielded better performance, an effect was not seen in these ablations.

Window size. We evaluate MoTIF with varying temporal input lengths to study robustness to sequence duration and efficiency trade-offs. While increasing the number of frames provides more temporal context, it also raises memory requirements and may not yield consistent accuracy gains. For comparability, we fix the batch size to 8 across all ablations, ensuring that changes in performance are solely attributable to window size rather than training dynamics. The results in Fig. 8 highlight that optimal window size depends on both dataset characteristics and backbone choice.

Random seed. Since most experiments were run with a fixed seed of 42 for reproducibility, we additionally ablate the effect of varying the random seed (39,40,41,42,43) for both MoTIF and the global CBM. As Figure 11 illustrates, the influence of the seed depends on the dataset. For HMDB51 the effect is negligible, whereas for Breakfast — particularly with the RN/50 backbone — we observe noticeably higher variance. Never-

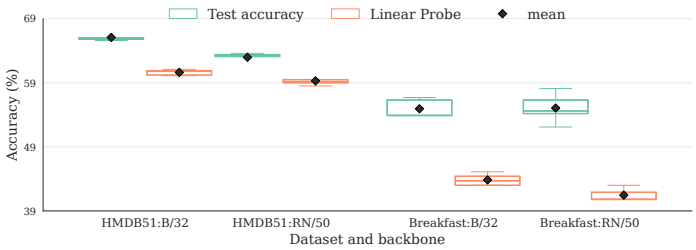
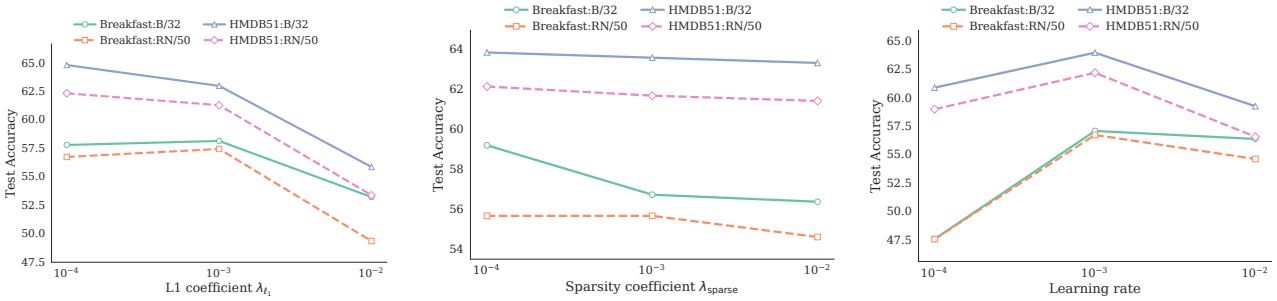


Figure 11. **Random seed.** Box plot of ablated datasets with five different seeds.



(a) **Classifier sparsity.** Test accuracy with different λ_{ℓ_1}.

(b) **Activation sparsity.** Test accuracy across different λ_{sparse}.

(c) **Learning rate.** Test accuracy with different learning rates.

Figure 10. **Architectural choices.** Effects of classifier sparsity, activation sparsity, and learning rate.

theless, MoTIF consistently outperforms the global CBM. Moreover, the figure indicates that the reported numbers in Table 1 are conservative: even higher scores are achievable, but we deliberately refrain from reporting maxima and instead provide a representative overview.

Additionally, since embeddings depend on the random choice of a representative image per window (previously fixed at seed 42), we re-embedded each dataset with five seeds. For each embedding variant we trained and evaluated MoTIF, the global CBM , and the zero-shot runs using a fixed training seed of 42 to isolate the effect of image-selection randomness. Table 6 reports the mean and standard deviation across embeddings (seeds 39–43). Std values are small overall; the largest observed variability is for Breakfast with RN/50 (std = 1.1).

Table 6. **Random seed on embeddings.** Mean and standard deviation (%) for methods on HMDB51 and Breakfast with two backbones.

Dataset (backbone)	Method	Mean (%)	Std (%)
HMDB51 (RN/50)	MoTIF	63.8	0.3
	Global CBM	60.2	0.3
	Zero Shot	29.4	0.2
HMDB51 (B32)	MoTIF	66.5	0.3
	Global CBM	61.0	0.6
	Zero Shot	38.7	0.3
Breakfast (RN/50)	MoTIF	53.2	1.1
	Global CBM	41.1	1.1
	Zero Shot	18.1	0.7
Breakfast (B32)	MoTIF	58.3	0.8
	Global CBM	43.9	0.6
	Zero Shot	25.8	0.2

C.1. Accuracies Testsplits

Tables 7–9 report Top-1 accuracies (%) on the test splits. Mean and standard deviation are computed across the shown splits (Breakfast: 4 splits; HMDB51 and UCF: 3 splits). For each backbone we list the MoTIF result (bold), the corresponding global CBM, and the zero-shot baseline. The reported standard deviation quantifies variability between test splits.

[1]from (Liu et al., 2021) with eight clusters and all frames (cumulative)

Table 7. **Top-1 accuracies (%) on Breakfast test splits.** Mean and standard deviation across splits.

Model	s1	s2	s3	s4	Mean	Std
CLIP RN/50 (MoTIF)	55.6	47.4	47.0	61.2	52.8	6.9
Global CBM (RN/50)	43.3	25.7	32.5	44.4	36.5	9.0
Zero-shot (RN/50)	17.3	19.2	16.0	21.9	18.6	2.6
CLIP B/32 (MoTIF)	56.3	44.9	51.3	61.0	53.4	6.9
Global CBM (B/32)	42.6	27.5	31.6	47.0	37.2	9.1
Zero-shot (B/32)	24.6	19.8	22.0	26.4	23.2	2.9
CLIP L/14 (MoTIF)	71.5	62.3	66.7	76.8	69.3	6.2
CLIP L/14 (MoTIF-ST)	73.2	64.8	65.0	80.8	73.5	12.2
Global CBM (L/14)	61.3	44.7	48.7	66.4	55.3	10.2
Zero-shot (L/14)	32.4	28.3	26.5	37.0	31.1	4.7
SigLIP L/14 (MoTIF)	76.1	62.1	73.1	82.7	73.5	8.6
Global CBM (SigLIP L/14)	59.9	44.7	53.3	70.6	57.1	10.9
Zero-shot (SigLIP L/14)	28.9	18.0	20.9	26.6	23.6	5.0
PE L/14 (MoTIF)	87.3	74.7	83.1	89.4	83.6	6.5
PE L/14 (MoTIF-ST)	86.2	74.7	82.3	91.2	83.6	7.0
Global CBM (PE L/14)	81.0	58.7	72.0	79.9	72.9	10.3
Zero-shot (PE L/14)	40.5	36.6	36.8	51.6	41.4	7.0

Table 8. **Top-1 accuracies (%) on HMDB51 test splits.** Mean and standard deviation across splits.

Model	s1	s2	s3	Mean	Std
CLIP RN/50 (MoTIF)	64.1	62.3	62.1	62.8	1.1
Global CBM (RN/50)	58.6	60.1	59.2	59.3	0.8
Zero-shot (RN/50)	29.3	30.1	30.1	29.8	0.5
CLIP B/32 (MoTIF)	65.9	66.8	63.3	65.3	1.8
Global CBM (B/32)	62.0	63.0	59.8	61.6	1.6
Zero-shot (B/32)	38.4	37.9	38.0	38.1	0.3
CLIP L/14 (MoTIF)	73.8	73.9	72.2	73.3	1.0
CLIP L/14 (MoTIF-ST)	75.5	75.3	73.6	74.8	1.0
Global CBM (L/14)	68.5	68.8	67.9	68.4	0.5
Zero-shot (L/14)	45.8	45.6	45.6	45.7	0.1
SigLIP L/14 (MoTIF)	74.8	74.4	70.4	73.2	2.4
Global CBM (SigLIP L/14)	66.3	66.0	62.6	65.0	2.1
Zero-shot (SigLIP L/14)	48.4	50.0	49.5	49.3	0.8
PE L/14 (MoTIF)	79.9	79.3	79.6	79.6	0.3
PE L/14 (MoTIF-ST)	79.9	80.0	78.8	79.6	0.7
Global CBM (PE L/14)	74.0	75.0	74.1	74.4	0.6
Zero-shot (PE L/14)	56.7	57.3	56.2	56.7	0.6

Table 9. **Top-1 accuracies (%) on UCF test splits.** Mean and standard deviation across splits.

Model	s1	s2	s3	Mean	Std
CLIP RN/50 (MoTIF)	82.4	82.5	83.4	82.8	0.6
Global CBM (RN/50)	80.7	80.0	79.3	80.0	0.7
Zero-shot (RN/50)	56.5	56.9	58.3	57.2	0.9
CLIP B/32 (MoTIF)	84.4	85.7	86.7	85.6	1.2
Global CBM (B/32)	82.2	82.5	83.6	82.8	0.7
Zero-shot (B/32)	59.4	60.2	60.1	59.9	0.4
CLIP L/14 (MoTIF)	92.4	93.7	93.6	93.2	0.7
CLIP L/14 (MoTIF-ST)	92.9	94.4	94.0	93.8	0.8
Global CBM (L/14)	88.8	91.0	90.3	90.0	1.1
Zero-shot (L/14)	71.1	70.6	70.1	70.6	0.5
SigLIP L/14 (MoTIF)	93.3	93.8	94.9	94.0	0.8
Global CBM (SigLIP L/14)	90.0	91.0	90.5	90.5	0.5
Zero-shot (SigLIP L/14)	80.0	81.9	79.2	80.4	1.4
PE L/14 (MoTIF)	94.6	95.7	95.8	95.4	0.7
PE L/14 (MoTIF-ST)	95.7	96.4	96.9	96.3	0.6
Global CBM (PE L/14)	94.6	93.9	95.0	94.5	0.6
Zero-shot (PE L/14)	73.8	75.6	74.4	74.6	0.9

C.2. Complete Performance Comparison

Table 10 reports the complete evaluation using the VLM-based concept discovery setting. It compares zero-shot vision-language baselines, Global CBM variants, our MoTIF and MoTIF-ST models, and strong non-interpretable video models across all four datasets. This table serves as the main performance comparison, while Table 11 isolates the effect of replacing the VLM-derived concept set with the non-VLM concept set.

Table 10. **Performance comparison (% Top-1 accuracy).** Mean ± standard deviation on train-test splits on Breakfast Actions, HMDB51, UCF101, and SSv2 with VLM-based concept discovery.

Method	Breakfast	HMDB51	UCF101	SSv2
Zero-shot				
CLIP-RN/50 (Radford et al., 2021)	18.6 ± 2.6	29.8 ± 0.5	57.2 ± 0.9	0.8
CLIP-ViT-B/32 (Radford et al., 2021)	23.2 ± 2.9	38.1 ± 0.3	59.9 ± 0.4	0.9
CLIP-ViT-L/14 (Radford et al., 2021)	31.1 ± 4.7	45.7 ± 0.1	70.6 ± 0.5	0.9
PE-L/14 (Bolya et al., 2025)	41.4 ± 7.0	56.7 ± 0.6	74.6 ± 0.9	2.2
PE-G/14 (Bolya et al., 2025)	47.4 ± 5.4	60.7 ± 1.0	74.6 ± 0.9	2.2
Global CBM				
CLIP-RN/50 (Radford et al., 2021)	36.9 ± 7.7	61.8 ± 1.7	84.1 ± 0.8	17.0
CLIP-ViT-B/32 (Radford et al., 2021)	38.3 ± 9.5	62.6 ± 0.8	86.4 ± 0.6	18.2
CLIP-ViT-L/14 (Radford et al., 2021)	57.0 ± 8.0	71.0 ± 1.1	93.4 ± 0.7	22.0
PE-L/14 (Bolya et al., 2025)	72.4 ± 8.3	76.4 ± 0.8	96.3 ± 0.1	31.3
PE-G/14 (Bolya et al., 2025)	75.8 ± 7.1	77.8 ± 0.8	97.5 ± 0.4	33.6
MoTIF (ours)				
MoTIF (RN/50)	55.1 ± 7.1	66.2 ± 0.5	86.7 ± 0.6	20.0
MoTIF (ViT-B/32)	52.7 ± 5.8	68.5 ± 1.0	88.5 ± 0.6	20.7
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	<u>83.0</u> ± 0.6	98.0 ± 0.2	40.4
MoTIF-ST (PE-G/14)	<u>87.3</u> ± 7.1	82.1 ± 1.0	<u>98.4</u> ± 0.3	41.9
Non-interpretable video models				
TSM (Lin et al., 2019)	59.1[1]	73.5	95.9	61.7
No frame left behind (Liu et al., 2021)	62.0[1]	73.4[1]	<u>96.4</u>[1]	<u>62.7</u>[1]
VideoMAE V2 (Wang et al., 2023)	–	**88.1**	**99.6**	**76.8**

The corresponding results are reported in Table 11. Overall, both Global CBM and MoTIF exhibit slightly reduced performance compared to the VLM setting, which we attribute to the smaller and less complete concept set, limiting the amount of task-relevant information captured in the bottleneck.

Table 11. **Performance comparison (% Top-1 accuracy).** Mean ± standard deviation on train-test splits on Breakfast Actions, HMDB51, UCF101, and SSv2 with different CLIP-based backbones with the non-VLM-based concept set. We report seconds per training epoch next to the accuracy scores for all MoTIF variants.

Method	**Breakfast**	**HMDB51**	**UCF101**	**SSv2**
Zero-shot				
CLIP-RN/50 (Radford et al., 2021)	18.6 ± 2.6	29.8 ± 0.5	57.2 ± 0.9	0.8
CLIP-ViT-B/32 (Radford et al., 2021)	23.2 ± 2.9	38.1 ± 0.3	59.9 ± 0.4	0.9
CLIP-ViT-L/14 (Radford et al., 2021)	31.1 ± 4.7	45.7 ± 0.1	70.6 ± 0.5	0.9
SigLIP-L/14 (Zhai et al., 2023)	23.6 ± 5.0	49.3 ± 0.8	80.4 ± 1.4	1.3
PE-L/14 (Bolya et al., 2025)	41.4 ± 7.0	56.7 ± 0.6	74.6 ± 0.9	2.2
Global CBM				
CLIP-RN/50 (Radford et al., 2021)	36.5 ± 9.0	59.3 ± 0.8	80.0 ± 0.7	13.7
CLIP-ViT-B/32 (Radford et al., 2021)	37.2 ± 9.1	61.6 ± 1.6	82.8 ± 0.7	15.2
CLIP-ViT-L/14 (Radford et al., 2021)	55.3 ± 10.2	68.4 ± 0.5	90.0 ± 1.1	18.1
SigLIP-L/14 (Zhai et al., 2023)	57.1 ± 10.9	65.0 ± 2.1	90.5 ± 0.5	19.6
PE-L/14 (Bolya et al., 2025)	72.9 ± 10.3	74.4 ± 0.6	94.5 ± 0.6	25.5
MoTIF (ours)				
MoTIF (RN/50)	52.8 ± 6.9 (4.2)	62.8 ± 1.1 (0.9)	82.8 ± 0.6 (1.5)	16.0 (10.0)
MoTIF (ViT-B/32)	53.4 ± 6.9 (4.2)	65.3 ± 1.8 (0.9)	85.6 ± 1.2 (1.5)	17.5 (9.9)
MoTIF (ViT-L/14)	69.3 ± 6.2 (4.3)	73.3 ± 1.0 (0.8)	93.2 ± 0.7 (1.5)	20.4 (10.1)
MoTIF-ST (ViT-L/14)	71.1 ± 7.7 (7.2)	74.8 ± 1.0 (1.8)	93.8 ± 0.9 (3.3)	23.9 (26.8)
MoTIF (SigLIP-L/14)	73.5 ± 8.6 (4.2)	73.2 ± 2.4 (0.8)	94.0 ± 0.8 (1.5)	22.4 (9.8)
MoTIF (PE-L/14)	<u>83.6</u> ± 6.5 (4.3)	<u>79.6</u>± 0.3 (0.9)	95.4 ± 0.7(1.5)	30.0 (10.4)
MoTIF-ST (PE-L/14)	**84.1** ± 6.4 (7.3)	<u>79.6</u> ± 0.7 (1.8)	96.3 ± 0.6 (3.3)	35.1 (26.7)
Non-interpretable video models				
TSM (Lin et al., 2019)	59.1[1]	73.5	95.9	61.7
No frame left behind (Liu et al., 2021)	62.0[1]	73.4[1]	<u>96.4</u>[1]	<u>62.7</u>[1]
VideoMAE V2 (Wang et al., 2023)	–	**88.1**	**99.6**	**76.8**

D. Algorithms

This section summarizes MoTIF's procedures for training, test-time inference, and explanation. We adopt the notation from the main text: per-window concept activations $Z_{t,c}$, per-time logits ℓ_t, and log-sum-exp (LSE) pooling with temperature τ and optional mask m_t.

Algorithm 1 Training MoTIF (Moving Temporal Interpretable Framework)

Input: $\{(X^{(n)}, y^{(n)})\}_{n=1}^{N}$, concept bank $\mathcal{C}$
Initialize Transformer parameters; affine (γ, δ); classifier (W, b)
for each epoch **do**
  for each batch (x, y) **do**
    Temporal modeling: per-channel temporal self-attention $\rightarrow X_{t,c}^{(L)}$
    Affine & nonnegativity: $Z_{t,c} \leftarrow \mathrm{Softplus}(\gamma_c X_{t,c}^{(L)} + \delta_c)$
    Classification: $\ell_t \leftarrow W Z_{t,:} + b$
    Pooling: $\hat{\ell} \leftarrow \mathrm{LSE}_\tau(\{\ell_t\}, m)$
    Loss: $\mathcal{L} \leftarrow \mathrm{CE}(\hat{\ell}, y) + \lambda_{\ell_1} \|W\|_1 + \lambda_{\text{sparse}} \frac{1}{(\sum_t m_t)C} \sum_{t,c} m_t |Z_{t,c}|$
    Update all parameters with AdamW
    Optional: enforce nonnegativity $W \leftarrow \max(W, 0)$
  end for
end for
Output: trained MoTIF

Description. The training loop builds concept activations from window embeddings, refines them with per-channel temporal attention, applies a nonnegative affine projection, and classifies with a linear head pooled over time via LSE. The objective combines cross-entropy with sparsity regularizers; W can be projected to enforce nonnegativity.

Algorithm 2 Inference with MoTIF (test-time forward pass)

Input: video x, trained MoTIF
Compute concept activations from window embeddings
Apply per-channel temporal attention; affine + Softplus $\rightarrow$ nonnegative $Z_{t,c}$
Compute per-time logits $\ell_t \leftarrow W_k Z_{t,:} + b$
Aggregate with LSE pooling: $\hat{\ell} \leftarrow \mathrm{LSE}_\tau(\{\ell_t\}, m)$
Predict $y^* \leftarrow \arg\max_k \hat{\ell}_k$
Output: predicted label y^*

Description. Inference is a single forward pass: per-window logits are pooled with LSE to produce video-level logits, whose argmax yields the prediction.

Algorithm 3 Explanation with MoTIF (global, local, temporal views)

Input: video x, class k, trained MoTIF
Run forward pass to obtain $Z_{t,c}$, ℓ_t, and $\hat{\ell}$
Per-time contributions: $\mathbf{c}_t^{(k)} \leftarrow Z_{t,:} \odot W_{k,:}$
  $s_t^{(k)} \leftarrow \sum_c c_{t,c}^{(k)} + b_k$
Temporal importance: $\pi_t^{(k)} \leftarrow \mathrm{softmax}\big(s_t^{(k)}/\tau\big)$ over valid t (mask m_t)
Global attribution: $\bar{\mathbf{c}}^{(k)} \leftarrow \sum_t \pi_t^{(k)} \mathbf{c}_t^{(k)}$
Output: $\big(\bar{\mathbf{c}}^{(k)}, \{\pi_t^{(k)}\}_t, \text{attention maps}\big)$

Description. Explanations decompose the prediction into per-concept contributions and reweight them by a time-importance distribution that mirrors LSE pooling. This yields three complementary views: (1) global concepts via $\bar{\mathbf{c}}^{(k)}$, (2) local concepts at decisive windows (large $\pi_t^{(k)}$), and (3) temporal dependencies from per-concept attention maps.

Algorithm 4 VLM-based Concept Discovery for Video (objects + actions)

Input: training set $\mathcal{D}_{\text{train}}$, agent $\mathcal{A}$, windows k, frames/window l, max videos/class n, similarity threshold δ
Output: filtered concept set $\mathcal{C}$
Initialize candidate pool $\mathcal{C}_{\text{raw}} \leftarrow [\,]$
for each class y **do**
  Sample up to n videos $\{x\}$ of class y from $\mathcal{D}_{\text{train}}$
  for each video x **do**
    Split x into k windows and sample l frames per window
    Generate object and action concepts conditioned on y:
      $\mathcal{C}_x \leftarrow \mathcal{A}(\text{frames}, y)$
    Append $\mathcal{C}_x$ to $\mathcal{C}_{\text{raw}}$
  end for
end for
Embed all concepts using text encoder $\psi(\cdot)$
Filter $\mathcal{C}_{\text{raw}}$ by cosine similarity threshold δ to obtain $\mathcal{C}$
return $\mathcal{C}$

Description. VLM-based concept discovery decomposes videos into short temporal windows and leverages a vision–language agent to extract textual object and action concepts from sampled frames. Concepts are aggregated across windows and videos into a shared candidate pool and subsequently filtered using a similarity threshold to remove redundant or semantically overlapping entries. This results in a compact yet expressive concept set constructed in an unsupervised manner from the training data only, which serves as the bottleneck representation for MoTIF.

Visualized Pipeline. Figure 12 summarizes the MoTIF pipeline. Given an input video, temporal windows are first embedded using a frozen vision–language backbone. These embeddings are mapped to concept activations via cosine similarity to a predefined concept bank, yielding per-time, per-concept responses. MoTIF then applies per-channel temporal self-attention to model concept-specific dynamics independently over time. The resulting representations are passed through a nonnegative affine projection and a linear classifier, producing per-time logits that are aggregated into a video-level prediction using log-sum-exp pooling. Beyond prediction, the explicit concept bottleneck enables three complementary explanation views: global concept attributions aggregated over time, local concept activations at informative temporal segments, and temporal dependencies captured by the attention weights.

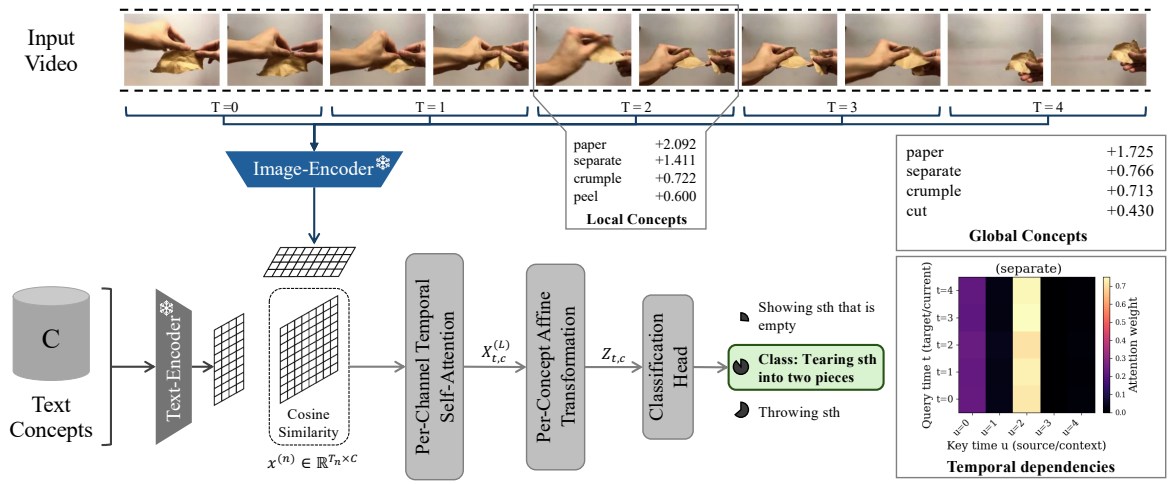


Figure 12. **MoTIF pipeline.** Videos are embedded with a vision–language backbone and mapped to concept activations via cosine similarity. Per-channel temporal self-attention models dynamics independently for each concept, followed by a nonnegative affine transformation and classification. MoTIF enables explanations across three views: global concepts, local concepts, and temporal dependencies. Sample frames from SSv2 (Materzynska et al., 2020) with MoTIF (ViT-L14).

E. Concept Sets

E.1. Concept Discovery with VLM Approach

This section evaluates concept-bank quality from several complementary but mostly indirect perspectives. Rather than directly annotating the semantic correctness of every discovered concept or per-window concept assignment, we study how downstream behavior changes when varying the discovery model, concept filtering threshold, concept source, and concept vocabulary. We view these analyses as targeted evidence about concept coverage, diversity, and temporal relevance.

Ablation on VLM Backbones. To study the impact of the underlying vision–language model, we repeat the concept discovery procedure using different Qwen 3 backbones (4B, 8B, and 30B parameters) on Breakfast and HMDB51. Table 12 reports the resulting mean classification accuracy together with the average number of discovered concepts. Note that the number of concepts can vary across splits, as concepts are extracted solely from the training videos.

Table 12. **Ablation on Qwen 3 backbones for VLM-based concept discovery.** Mean Top-1 accuracy and average number of discovered concepts (train split only), evaluated with different CLIP vision backbones.

Qwen	Vision	Breakfast		HMDB51	
	Backbone	Acc. (%)	#Concepts	Acc. (%)	#Concepts
Qwen 3 4B	RN-50	55.2	338	66.2	793
	ViT-B/32	53.3	240	68.6	618
Qwen 3 8B	RN-50	54.5	330	66.8	872
	ViT-B/32	54.1	238	68.6	662
Qwen 3 30B	RN-50	55.1	258	66.2	657
	ViT-B/32	52.7	193	68.5	511

Ablation on Similarity Threshold. We further ablate the effect of the similarity threshold used for concept filtering on Qwen 3 30B. While our main experiments use a threshold of 0.9, Table 13 reports results for thresholds of 0.8 and 0.7, together with the corresponding concept counts.

Table 13. **Ablation on concept similarity threshold.** Impact on concept count and downstream accuracy for different vision backbones.

Threshold	Vision Backbone	Breakfast		HMDB51	
		Acc. (%)	#Concepts	Acc. (%)	#Concepts
0.7	RN-50	39.5	24	56.1	51
	ViT-B/32	29.9	11	49.4	27
0.8	RN-50	51.5	88	63.8	217
	ViT-B/32	49.1	52	64.3	128
0.9	RN-50	55.1	258	66.2	657
	ViT-B/32	52.7	193	68.5	511

For a threshold of 0.9, the number of filtered concepts ranges from 1,748–2,325 on SSv2, 481–1,217 on HMDB51, 186–366 on BREAKFAST, and 729–2,075 on UCF101. The large variance is primarily driven by the embedding model and the size of the training splits. In this ablation, larger concept sets are associated with improved performance, suggesting that increased concept diversity can yield a more complete bottleneck representation.

Visual concepts via SAM3. Motivated by DCBM's use of segmentation-derived visual concepts (Prasse et al., 2025), we additionally replace the textual concept bank with visual concepts obtained from a promptable segmentation model (SAM3) (Carion et al., 2026). In this setting, the discovery pipeline proposes visual regions and short visual snippets rather than text labels, which are then embedded and used as the bottleneck vocabulary. Table 14 shows that this visual-concept variant remains viable across datasets and backbones, although it underperforms the textual concept bank in our current setup. We therefore treat it as a concept-quality robustness check showing that MoTIF is compatible with DCBM-style visual concept sources rather than as a new main result.

Ablation of Concept Types. To more directly analyze the role of different concept types, we conducted an additional ablation in Table 15 that separates the concept sets used by MoTIF into *object concepts*, *action concepts*, and their *combination*, and

Table 14. **Visual-concept robustness check with SAM3-style concepts.** Mean ± standard deviation across official splits for Breakfast, HMDB51, and UCF101; SSv2 is reported on its single official validation split.

Backbone	**Breakfast**	**HMDB51**	**UCF101**	**SSv2**
RN/50	49.4 ± 3.8	57.5 ± 1.1	81.3 ± 0.6	16.5
ViT-B/32	50.6 ± 6.8	60.5 ± 1.5	84.6 ± 0.8	18.1
ViT-L/14	66.6 ± 5.3	67.2 ± 1.1	92.1 ± 1.1	22.2
PE-L/14	80.7 ± 5.3	74.5 ± 1.5	96.7 ± 0.2	32.5

evaluates them across the main datasets. While this remains a downstream evaluation rather than a direct semantic audit, it provides a more targeted analysis than the previous results, as it isolates the contribution of each concept type and, in particular, tests whether action concepts yield measurable benefits beyond object concepts.

Table 15. **Concept types ablation.** Performance of MoTIF when using different concept types across datasets and backbones.

Backbone	Concepts	Breakfast	HMDB	UCF	SSv2
L/14	Object	70.9 ± 7.2	73.4 ± 1.4	94.5 ± 0.7	24.1
L/14	Action	71.0 ± 7.6	77.1 ± 0.9	95.0 ± 0.4	24.6
L/14	Object + Action	71.0 ± 6.2	76.1 ± 0.5	94.8 ± 0.5	25.8
PE-L/14	Object	83.6 ± 7.4	80.0 ± 1.5	96.8 ± 0.3	33.8
PE-L/14	Action	85.1 ± 5.8	81.3 ± 0.5	97.0 ± 0.5	36.4
PE-L/14	Object + Action	83.6 ± 6.3	81.8 ± 0.6	97.0 ± 0.3	37.3

The results reveal a consistent pattern: action concepts are particularly beneficial on the more temporally demanding benchmarks. This is most evident on HMDB and SSv2, where action concepts outperform object concepts for both backbones. For example, with PE-L/14 on SSv2, performance increases from 33.8 with object concepts to 36.4 with action concepts, and further to 37.3 when object and action concepts are combined. A similar trend can be observed on HMDB, where the improvement from object to action concepts is also consistent across both backbones.

These findings are in line with the intended role of action concepts: they appear to capture information that goes beyond static object cues and is particularly useful when temporal structure matters. At the same time, the strongest overall results are typically obtained when object and action concepts are combined, suggesting that both concept types provide complementary information rather than redundant signals. This complementarity is further encouraged by our concept filtering procedure, which removes overly similar concepts and thus promotes a more diverse concept set.

E.2. LLM Concepts

In Table 16, we show the number and kind of concepts used for the construction of MoTIF for each dataset. The number and kind of concepts vary for each dataset, since we asked the LLM to create domain-specific concepts that are useful for the downstream classification task. As the tables indicate, the number of concepts is lower than the VLM-based one.

Table 16. **MoTIF concepts.** The textual concepts utilized for all experiments which are listed in this paper.

Set	Count	Concepts
Breakfast	223	add, adjust, apple, arrange, assemble, avocado, bacon, bagel, bake, balance, banana, batter, beat, bin, blend, blender, blow, boil, bottle, bowl, bread, brew, brush, burner, butter, button, carry, carton, catch, cereal, chair, cheese, chop, cinnamon, clap, close, coffee, colander, comb, container, cook, cookbook, cool, core, counter, cover, crack, croissant, cucumber, cup, cupboard, cut, cuttingboard, detergent, dish, drag, drain, drizzle, drop, dry, egg, faucet, fill, flame, flip, fold, fork, freezer, fridge, froth, frown, fruit, fry, garbage, gesture, granola, grate, grater, grill, grind, ham, handle, heat, herb, hide, honey, hood, ice, ingredient, insert, jar, juice, kettle, knife, knob, knock, ladle, laugh, leftover, lid, mash, measure, measuringcup, measuringspoon, milk, mix, mug, napkin, nod, onion, open, orange, oven, ovenmitt, pack, package, pan, pantry, pastry, peel, peeler, pick, pinch, pit, place, plate, plug, point, poke, pour, preheat, press, pull, push, put, reach, recipe, recycle, release, remove, reveal, rinse, roll, rotate, sausage, scale, scoop, scramble, scrub, seal, serve, serving, set, shake, shave, sieve, sink, sip, sit, slice, slide, smile, snap, soap, socket, sort, spatula, spin, sponge, spoon, spread, sprinkle, squeeze, stack, stand, start, steam, steep, stir, stirrer, stool, stop, stove, strawberry, sugar, switch, syrup, table, take, tamp, tap, taste, tea, thermometer, throw, tie, tilt, timer, toast, tomato, tongs, toss, towel, tray, turn, twist, uncover, unfold, unscrew, unstack, untie, unwrap, warm, wash, waste, water, wave, whisk, wipe, wring, yogurt, zest, zip
UCF101	166	aim, archer, archery, arena, arrow, athlete, balance, ball, bar, barbell, baseball, basket, basketball, bat, beam, bicycle, block, bounce, bow, bowl, boxing, breakdance, canoe, cap, catch, clap, climb, club, coach, control, court, cricket, curl, dance, dancer, deadlift, dismount, dive, dodge, dribble, dumbbell, enter, field, fight, flip, floor, frisbee, gallop, gloves, goal, goalpost, grab, grapple, grind, gun, gym, gymnast, handstand, hang, helmet, hit, hockey, hook, hoop, horse, hurdle, ice, instrument, jab, jersey, jump, kayak, kick, ladder, lane, lift, mat, microphone, mount, music, net, netting, opponent, pad, paddle, parry, pass, pedal, perform, pitch, platform, player, pool, press, puck, pull, push, racket, rail, raise, referee, reins, release, reload, ride, ring, rope, row, rower, rugby, run, sand, scoreboard, serve, sheet, shoot, shooter, sit, skateboard, skateboarder, skater, ski, skip, skis, smash, snow, snowboard, snowboarder, soccer, spike, spin, splash, sprint, squat, stadium, stage, stand, start, steer, stick, stop, strike, surf, surfboard, surfer, swim, swimmer, swing, sword, target, teammate, throw, timer, track, trampoline, tuck, turn, uniform, uppercut, volleyball, walk, wall, water, wave, wrestle, yoga
HMDB51	150	apply, around, backward, balance, ball, baseball, basketball, bat, bend, bicycle, block, blow, bottle, bounce, bow, brake, brush, button, carry, cartwheel, catch, chair, chew, climb, close, comb, crawl, cross, crouch, cup, dismount, dive, door, down, drag, dribble, drink, drop, eat, enter, exit, face, fall, fight, finish, flip, float, frisbee, from, frown, gallop, grab, hair, hand, hands, handstand, hat, head, headstand, high, hit, hop, horse, hug, jacket, jog, juggle, jump, kick, kiss, knock, laugh, leap, left, leg, lie, lift, line, look, low, makeup, mount, mouth, nod, object, off, on, open, pedal, point, pull, punch, push, put, racket, reach, release, ride, right, roll, room, run, serve, shake, shave, shirt, shoelace, shoot, sing, sip, sit, skate, skateboard, ski, sled, sleep, slide, smile, snowboard, somersault, spin, sprint, stand, start, steer, stretch, surface, swim, swing, sword, take, talk, teeth, tennis, throw, tie, toss, touch, turn, untie, up, utensils, wake, walk, wash, wave, with, words, yawn, zip
SSv2	284	accelerate, apple, arm, assemble, background, backpack, bag, balance, ball, banana, bend, bite, blow, book, bottle, bottom, bounce, bow, bowl, box, break, broken, can, cap, carrot, carry, catch, chair, chew, chop, chopstick, clap, clean, click, climb, close, closeable, cold, connect, container, cough, cover, crawl, crouch, crumple, cry, cucumber, cup, cut, dance, decelerate, dirty, disassemble, disconnect, door, downward, drag, drag mouse, draw, drink, drinkable, drop, dry, durable, eat, edible, empty, erase, face, fall, fasten, fill, finger, fixed, flatten, flip, floor, fold, fork, fragile, frown, fruit, full, gather, get up, grape, hand, heavy, hide, hold, hop, hot, insert, inside, juggle, jump, key, keyboard, kneel, knife, knock, laptop, laugh, lean, left, lid, lift, light, lock, loosen, mix, mouse, nod, object, open, openable, orange, other, outside, paint, paper, peel, pen, pencil, person, phone, plate, plug, point, pour, pourable, press, pretend to balance, pretend to block, pretend to bow, pretend to catch, pretend to catch fish, pretend to clap, pretend to clean, pretend to climb, pretend to close, pretend to cook, pretend to dance, pretend to dodge, pretend to draw, pretend to dribble, pretend to drink, pretend to drive, pretend to eat, pretend to fall, pretend to fire gun, pretend to honk, pretend to hug, pretend to jump rope, pretend to kick, pretend to kiss, pretend to load gun, pretend to lock, pretend to look around, pretend to measure, pretend to open, pretend to paddle, pretend to paint, pretend to play drums, pretend to play guitar, pretend to play piano, pretend to point, pretend to pour, pretend to pull, pretend to punch, pretend to push, pretend to read, pretend to row, pretend to salute, pretend to scroll, pretend to search, pretend to serve, pretend to shake hands, pretend to shoot arrow, pretend to shoot basket, pretend to sing, pretend to sleep, pretend to steer, pretend to steer wheel, pretend to stir, pretend to swing bat, pretend to swipe, pretend to throw, pretend to throw ball, pretend to type, pretend to unlock, pretend to use controller, pretend to wake, pretend to wave, pretend to weigh, pretend to write, pull, push, remote, remove, reveal, right, roll, rollable, rotate, rough, run, scatter, scoop, scroll, separate, shake, shake head, shelf, shout, sip, sit, sleep, slice, slide, smell, smile, smooth, snap, sneeze, speak, spill, spillable, spin, spin dance, spit, spoon, sprinkle, sprint, squeezable, stack, stackable, stand, start, stir, stop, stretch, stumble, surface, swing, swipe, table, tap, taste, tear, throw, tie, tighten, tilt, tomato, top, topple, touch, toy, turn off, turn on, type, uncover, unfasten, unfold, unlock, unplug, unstack, untie, unwrap, upward, vegetable, wake, walk, wall, wave, wet, whisper, window, wrap, write, yawn, zoom in, zoom out

E.3. Concept Set Variance

Concept set influence. To investigate the influence of the concept proposal set on the performance of MoTIF, we prompt GPT-5 five times, each time requesting a distinct set of candidate concepts (with some natural overlap across runs). Our experiments demonstrate that given the same prompt to the LLM, concept set affects test accuracy only moderately, with consistent trends across datasets and backbones (Fig. 13). In line with prior work on concept bottleneck models, more dataset-specific concepts tend to increase accuracy (Rao et al., 2024; Prasse et al., 2025; Schrodi et al., 2025). However, our main motivation here is not to optimize the quality of the concept proposals, but rather to demonstrate that our framework works broadly and robustly across datasets, backbones, and varying concept sets. In addition, we note that k—the number of concepts retained in the bottleneck—influences the achievable accuracy: very small k restricts expressive power, while very large k may introduce noise and redundancy if concepts are too similar or irrelevant. Table 18 lists the textual concepts used for the Breakfast ablation in Figure 13. We prompted the LLM five times to generate concepts for the CBM using the following prompt:

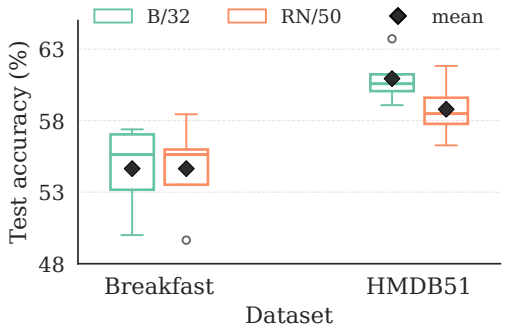


Figure 13. **Concept set influence.** Distribution of test accuracy across five different concept sets. The two dots indicate outliers within the interquartile range.

```
Create unique concepts (>100) for a concept-bottleneck model for
the dataset 'Breakfast Actions'. Return them in this format:
"prepare coffee, grind beans, ..."
```

Both the number and the variety of concepts vary across calls. Despite this variability, model performance remains stable, demonstrating MoTIF's robustness to different concept sets. The same prompting procedure was repeated for the other datasets (e.g. HMDB51, UCF101, Something-Something V2).

E.4. Ablation of Concept Construction

To evaluate whether MoTIF can effectively ground explicitly temporal concepts, we constructed five concept sets for SSv2, the dataset in our benchmark that relies most heavily on temporal reasoning: (1) nouns only, (2) verbs only, (3) nouns combined with verbs, (4) the curated concept set used in the main paper, and (5) the union of all concepts (see Table 17 and Table 19). All variants were evaluated using MoTIF with and without space-time transformer architecture. Across all settings, MoTIF consistently outperforms Global CBM.

Table 17. **Concept-Set Ablation on SSv2.** Top-1 accuracy (%) with PE-L/14 evaluated on the four additional concept sets.

Concept Set	MoTIF	MoTIF-ST	Global CBM
(1) Nouns only	18.0	20.3	14.8
(2) Verbs only	24.4	28.6	21.2
(3) Nouns + Verbs	22.1	24.5	18.2
(4) All concepts	29.8	36.0	26.7
(5) Original set	30.0	35.1	25.5

We also ablate the effect of different concept prepThe ablation yields four observations:

- Across all concept sets, MoTIF improves over Global CBM, showing that MoTIF's temporal modeling reliably strengthens concept grounding regardless of the chosen vocabulary.
- Verb-only concepts (set 2) achieve better performance than the noun-only (set 1) counterpart, indicating that MoTIF is particularly effective at grounding action dynamics that require temporal structure.
- The curated concept set (set 4) yields the highest performance (36.0% - with MoTIF-ST), suggesting that a balanced vocabulary provides the best trade-off between coverage and specificity.

Table 18. **MoTIF Breakfast Concepts.** The textual concepts utilized for the ablation of concept influence.

Set	Count	Concepts
Breakfast Original	223	add, adjust, apple, arrange, assemble, avocado, bacon, bagel, bake, balance, banana, batter, beat, bin, blend, blender, blow, boil, bottle, bowl, bread, brew, brush, burner, butter, button, carry, carton, catch, cereal, chair, cheese, chop, cinnamon, clap, close, coffee, colander, comb, container, cook, cookbook, cool, core, counter, cover, crack, croissant, cucumber, cup, cupboard, cut, cuttingboard, detergent, dish, drag, drain, drizzle, drop, dry, egg, faucet, fill, flame, flip, fold, fork, freezer, fridge, froth, frown, fruit, fry, garbage, gesture, granola, grate, grater, grill, grind, ham, handle, heat, herb, hide, honey, hood, ice, ingredient, insert, jar, juice, kettle, knife, knob, knock, ladle, laugh, leftover, lid, mash, measure, measuringcup, measuringspoon, milk, mix, mug, napkin, nod, onion, open, orange, oven, ovenmitt, pack, package, pan, pantry, pastry, peel, peeler, pick, pinch, pit, place, plate, plug, point, poke, pour, preheat, press, pull, push, put, reach, recipe, recycle, release, remove, reveal, rinse, roll, rotate, sausage, scale, scoop, scramble, scrub, seal, serve, serving, set, shake, shave, sieve, sink, sip, sit, slice, slide, smile, snap, soap, socket, sort, spatula, spin, sponge, spoon, spread, sprinkle, squeeze, stack, stand, start, steam, steep, stir, stirrer, stool, stop, stove, strawberry, sugar, switch, syrup, table, take, tamp, tap, taste, tea, thermometer, throw, tie, tilt, timer, toast, tomato, tongs, toss, towel, tray, turn, twist, uncover, unfold, unscrew, unstack, untie, unwrap, warm, wash, waste, water, wave, whisk, wipe, wring, yogurt, zest, zip
Breakfast Set 2	140	adjust heat, arrange cutlery, bake bread, bake pastry, beat eggs, bend down, bite food, break chocolate, break egg shell, breakfast clock, bubbling liquid, butter toast, chair at table, chew food, chop onion, clean spoon, close carton, close cupboard, close drawer, close fridge, close jar, close microwave, close oven, close tap, cut banana, cut dough, cut sandwich, dice tomato, drip water, drizzle dressing, drizzle honey, drizzle oil, dry dish, dry hands, empty sink, family sitting, flip bread, flip toast, fold mixture, fold napkin, grab fork, grab knife, grab spoon, grate chocolate, grill sandwich, gulp drink, hold bowl, hold glass, hold plate, hold straw, knead dough, lean forward, lick spoon, mash egg, mash potato, melt chocolate, mix salad, mop spill, morning light, move chair, one person eating, open carton, open cupboard, open drawer, open fridge, open microwave, open oven, open tap, peel apple, peel banana, peel egg, peel orange, person standing at stove, place utensil, pour batter, pour carton, preheat oven, press button, put down bowl, put down glass, put down plate, reach cupboard, reach shelf, recycle carton, recycle glass, recycle paper, recycle plastic, rest dough, rinse cup, roast vegetable, roll dough, run water, separate yolk, set napkin, shape dough, shred lettuce, sip drink, sip straw, sit down, sizzling pan, slice cucumber, soap hands, spread batter, spreading butter, spreading jam, spreading topping, sprinkle cheese, sprinkle herbs, sprinkle spices, squeeze lemon, stack plates, stand up from chair, steam rising, stir chocolate, swallow food, take carton, take jar, take package, take utensil, tear package, throw trash, toast bun, toss salad, towel hands, turn knob, turn off kettle, turn off stove, turn on kettle, turn on stove, two people cooking, unwrap sandwich, wash dish, wash hands, water boiling, whisk whites, wipe counter, wipe knife, wipe plate, wrap sandwich, zest lemon
Breakfast Set 3	128	add cinnamon, add fruit topping, add granola, add honey, add ice to blender, add milk, add milk to cereal, add sugar, arrange cutlery, assemble sandwich, bake pastry, beat eggs, blend smoothie, boil potato, boil water, brew coffee, butter toast, check timer, chop herbs, chop onion, chop vegetables, clear table, close cupboard, close fridge, close jar, close oven, cook bacon, cook pancake, cook sausage, core apple, crack egg, cut sandwich, dice vegetables, drain bacon, drizzle honey, drizzle syrup, dry dishes, fill kettle, flip bacon, flip omelette, flip pancake, follow recipe, froth milk, fry egg, grate cheese, grill sandwich, grind coffee beans, heat pan, insert coffee pod, make omelette, mash avocado, mash potato, measure ingredients, mix batter, open cupboard, open egg carton, open fridge, open jar, open oven, operate espresso machine, pack leftovers, peel banana, peel potato, pick up spoon, pit avocado, place cup, place plate, pour batter, pour cereal into bowl, pour coffee into cup, pour hot water, pour milk, pour smoothie, pour syrup, pour yogurt into bowl, preheat oven, prepare coffee, pull espresso shot, put ingredient in fridge, put leftovers in fridge, read recipe, rinse fruit, scramble eggs, serve pancakes, set table, set timer, sip beverage, slice apple, slice avocado, slice bagel, slice banana, slice bread, slice cheese, slice cucumber, slice ham, slice orange, slice strawberries, slice tomato, spread butter, spread cream cheese, spread jam, spread peanut butter, sprinkle sugar, squeeze lemon, steam milk, steep tea, stir beverage, strain smoothie, take cup, take ingredient from fridge, take leftovers out fridge, take plate, toast bagel, toast bread, unscrew lid, unwrap bread, use fork, use knife, use measuring cup, use measuring spoon, use spatula, use tongs, use whisk, warm croissant, wash dishes, wash fruit, whisk eggs, wipe counter
Breakfast Set 4	175	add cereal, add cinnamon, add cocoa powder, add honey, add ice to blender, add milk, add milk to coffee, add pepper, add salt, add sugar to coffee, add sugar to tea, add toppings, adjust seasoning, arrange cutlery, assemble sandwich, beat eggs, blend smoothie, blow on hot food, boil potato, boil water, brew coffee, butter toast, carry plate to table, check food temperature, check timer, chop herbs, chop onion, chop tomato, clean blender, clean counter, clear table, close cupboard, close egg carton, close fridge, close jar, close microwave, close milk carton, close oven, cook bacon, cook sausage, crack egg, cut lemon, cut sandwich, dice vegetables, drain bacon, drizzle honey, drizzle syrup, dry dishes, dry hands, fill kettle, fill pot with water, flip bacon, flip omelette, flip pancake, follow recipe, froth milk, fry egg, grate cheese, grind coffee beans, hold pan lid, insert bread into toaster, insert coffee pod, make omelette, mash potato, measure ingredients, mix batter, open cupboard, open egg carton, open fridge, open jar, open microwave, open milk carton, open oven, open package, operate blender, operate espresso machine, pack leftovers, peel banana, peel orange, peel potato, pick up cup, pick up knife, pit avocado, place pan off stove, place pan on stove, place plate, pour batter, pour cereal into bowl, pour coffee, pour eggs into pan, pour from carton, pour hot water, pour milk into bowl, pour pancake batter, pour smoothie, pour syrup, pour tea, pour yogurt into bowl, preheat oven, prepare coffee, prepare tea, press coffee, pull espresso shot, put leftovers in fridge, reach for ingredient, read recipe, remove bread from toaster, remove lid from pot, rinse fruit, scoop butter, scramble eggs, search for ingredient, season food, serve omelette, serve pancakes, set table, set timer, sip beverage, sit down, slice apple, slice avocado, slice bagel, slice banana, slice bread, slice cheese, slice cucumber, slice fruit, slice ham, slice kiwi, slice orange, slice pancake stack, slice strawberries, slice tomato, spoon yogurt, spread butter, spread cream cheese, spread jam, spread peanut butter, sprinkle granola, sprinkle sugar, squeeze lemon, stand up, start microwave, steam milk, steep tea, stir coffee, stop microwave, strain smoothie, take leftovers out fridge, take plate, taste food, toast bagel, toast bread, turn off kettle, turn off stove, turn on kettle, turn on stove, unscrew lid, unwrap bread, use fork, use french press, use knife, use measuring cup, use oven mitts, use spatula, use spoon, use toaster, use tongs, use whisk, wash blender, wash dishes, wash fruit, wash hands, whisk eggs, wipe counter
Breakfast Set 5	161	add cinnamon, add fruit topping, add granola, add honey, add ice to blender, add milk, add milk to cereal, add nuts, add sugar, adjust seasoning, arrange cutlery, assemble sandwich, bake pastry, beat eggs, blend smoothie, blow on hot food, boil potato, boil water, brew coffee, butter toast, carry plate to table, check timer, chop herbs, chop onion, chop vegetables, clear table, close cupboard, close fridge, close jar, close milk carton, close oven, cook bacon, cook pancake, cook sausage, core apple, core pineapple, crack egg, cut parsley, cut pineapple, cut sandwich, dice vegetables, drain bacon, drain can, drizzle honey, drizzle syrup, dry dishes, dry hands, fill kettle, flip bacon, flip omelette, flip pancake, follow recipe, froth milk, fry egg, grate cheese, grill sandwich, grind beans, heat pan, insert coffee pod, juice orange, make omelette, mash avocado, mash potato, measure ingredients, mix batter, open cupboard, open egg carton, open fridge, open jar, open milk carton, open oven, open package, open tin, operate espresso machine, pack leftovers, peel banana, peel orange, peel potato, pick up cup, pick up spoon, pit avocado, place cup, place pan off stove, place plate, pour batter, pour cereal, pour coffee into cup, pour from carton, pour hot water, pour milk, pour smoothie, pour syrup, pour yogurt, preheat oven, prepare coffee, pull espresso shot, put ingredient in fridge, put leftovers in fridge, reach for ingredient, read recipe, remove lid from pot, rinse fruit, scramble eggs, seal container, serve pancakes, set table, set timer, sip beverage, sit down, slice apple, slice avocado, slice bagel, slice banana, slice bread, slice cheese, slice cucumber, slice ham, slice kiwi, slice lemon, slice melon, slice orange, slice pear, slice strawberries, slice tomato, spread butter, spread cream cheese, spread jam, spread peanut butter, sprinkle sugar, squeeze lemon, stand up, steam milk, steep tea, stir beverage, strain smoothie, take cup, take ingredient from fridge, take leftovers out fridge, take plate, taste food, toast bagel, toast bread, toast nuts, unscrew lid, unwrap bread, unwrap package, use fork, use knife, use measuring cup, use measuring spoon, use oven mitts, use spatula, use tongs, use whisk, warm croissant, wash dishes, wash fruit, wash hands, whisk eggs, wipe counter, zest lemon

Table 19. **MoTIF Breakfast Concepts.** Textual concepts used in the concept-set ablation.

Set	Count	Concepts
Nouns-Only Concepts	60	object, container, box, cup, bowl, plate, spoon, knife, fork, chopstick, pen, pencil, paper, book, phone, remote, laptop, keyboard, mouse, bag, backpack, toy, ball, fruit, apple, orange, banana, grape, vegetable, carrot, cucumber, tomato, bottle, can, lid, cap, key, lock, door, window, wall, floor, table, chair, shelf, hand, finger, arm, face, person, background, surface, inside, outside, top, bottom, left, right, upward, downward
Verbs-Only Concepts	135	push, pull, lift, drop, hold, carry, throw, catch, slide, drag, roll, spin, rotate, flip, fold, unfold, wrap, unwrap, tie, untie, fasten, unfasten, tighten, loosen, break, cut, slice, chop, tear, peel, crumple, flatten, bend, stretch, shake, stir, pour, scoop, sprinkle, stack, unstack, assemble, disassemble, open, close, lock, unlock, press, tap, swipe, scroll, zoom in, zoom out, point, touch, wave, clap, knock, snap, swing, juggle, bounce, balance, topple, insert, remove, fill, empty, mix, separate, spill, scatter, gather, cover, uncover, hide, reveal, lean, tilt, climb, crawl, jump, hop, walk, run, sprint, stumble, fall, get up, sit, stand, kneel, crouch, bow, dance, nod, shake head, smile, frown, laugh, cry, shout, whisper, speak, yawn, sneeze, cough, sleep, wake, eat, chew, bite, sip, drink, spit, blow, smell, taste, write, draw, erase, paint, type, click, drag mouse, plug, unplug, connect, disconnect, turn on, turn off, start, stop, accelerate, decelerate
Noun–Verb Concepts	104	lift the box, open the box, close the box, drop the box, push the box, pull the box, carry the cup, pour from the cup, fill the cup, empty the cup, hold the cup, rotate the cup, open the bottle, close the bottle, pour from the bottle, lift the bottle, drink from the bottle, cut the paper, fold the paper, tear the paper, crumple the paper, write on the paper, open the book, close the book, flip the book, read the book, drop the book, type on the keyboard, plug in the laptop, unplug the laptop, open the laptop, close the laptop, click the mouse, drag the mouse, press the key, turn on the phone, turn off the phone, swipe the phone, tap the phone, scroll on the phone, charge the phone, unlock the phone, lock the door, unlock the door, open the door, close the door, lift the chair, move the chair, sit on the chair, stand from the chair, throw the ball, catch the ball, bounce the ball, roll the ball, peel the banana, eat the banana, cut the apple, eat the apple, slice the cucumber, pour the water, stir the soup, mix the ingredients, chop the vegetables, open the can, close the lid, place the lid, remove the lid, stack the boxes, unstack the boxes, wrap the gift, unwrap the gift, tie the rope, untie the rope, press the button, flip the switch, turn the knob, insert the plug, remove the plug, connect the cable, disconnect the cable, shake the bottle, squeeze the bottle, pour the juice, stir the drink, draw on the paper, paint the wall, erase the drawing, fold the towel, open the backpack, close the backpack, pick up the bag, drop the bag, open the window, close the window, wipe the table, clean the floor, open the box with one hand, lift the object, move the object, drop the object, hold the object, place the object, throw the object, pick up the object
All Concepts	391	push, pull, lift, drop, hold, carry, throw, catch, slide, drag, roll, spin, rotate, flip, fold, unfold, wrap, unwrap, tie, untie, fasten, unfasten, tighten, loosen, break, cut, slice, chop, tear, peel, crumple, flatten, bend, stretch, shake, stir, pour, scoop, sprinkle, stack, unstack, assemble, disassemble, open, close, lock, unlock, press, tap, swipe, scroll, zoom in, zoom out, point, touch, wave, clap, knock, snap, swing, juggle, bounce, balance, topple, insert, remove, fill, empty, mix, separate, spill, scatter, gather, cover, uncover, hide, reveal, lean, tilt, climb, crawl, jump, hop, walk, run, sprint, stumble, fall, get up, sit, stand, kneel, crouch, bow, dance, spin dance, nod, shake head, smile, frown, laugh, cry, shout, whisper, speak, yawn, sneeze, cough, sleep, wake, eat, chew, bite, sip, drink, spit, blow, smell, taste, write, draw, erase, paint, type, click, drag mouse, plug, unplug, connect, disconnect, turn on, turn off, start, stop, accelerate, decelerate, pretend to push, pretend to pull, pretend to pour, pretend to eat, pretend to drink, pretend to throw, pretend to catch, pretend to type, pretend to swipe, pretend to scroll, pretend to climb, pretend to fall, pretend to hug, pretend to kiss, pretend to wave, pretend to play guitar, pretend to drive, pretend to steer, pretend to read, pretend to sleep, pretend to wake, pretend to write, pretend to draw, pretend to paint, pretend to clean, pretend to cook, pretend to stir, pretend to measure, pretend to weigh, pretend to look around, pretend to search, pretend to point, pretend to balance, pretend to open, pretend to close, pretend to lock, pretend to unlock, pretend to kick, pretend to punch, pretend to block, pretend to dodge, pretend to jump rope, pretend to row, pretend to paddle, pretend to shoot arrow, pretend to load gun, pretend to fire gun, pretend to throw ball, pretend to dribble, pretend to shoot basket, pretend to swing bat, pretend to serve, pretend to catch fish, pretend to steer wheel, pretend to honk, pretend to use controller, pretend to play piano, pretend to play drums, pretend to dance, pretend to sing, pretend to clap, pretend to salute, pretend to bow, pretend to shake hands, pretend to hug, pretend to kiss, lift the box, open the box, close the box, drop the box, push the box, pull the box, carry the cup, pour from the cup, fill the cup, empty the cup, hold the cup, rotate the cup, open the bottle, close the bottle, pour from the bottle, lift the bottle, drink from the bottle, cut the paper, fold the paper, tear the paper, crumple the paper, write on the paper, open the book, close the book, flip the book, read the book, drop the book, type on the keyboard, plug in the laptop, unplug the laptop, open the laptop, close the laptop, click the mouse, drag the mouse, press the key, turn on the phone, turn off the phone, swipe the phone, tap the phone, scroll on the phone, charge the phone, unlock the phone, lock the door, unlock the door, open the door, close the door, lift the chair, move the chair, sit on the chair, stand from the chair, throw the ball, catch the ball, bounce the ball, roll the ball, peel the banana, eat the banana, cut the apple, eat the apple, slice the cucumber, pour the water, stir the soup, mix the ingredients, chop the vegetables, open the can, close the lid, place the lid, remove the lid, stack the boxes, unstack the boxes, wrap the gift, unwrap the gift, tie the rope, untie the rope, press the button, flip the switch, turn the knob, insert the plug, remove the plug, connect the cable, disconnect the cable, shake the bottle, squeeze the bottle, pour the juice, stir the drink, draw on the paper, paint the wall, erase the drawing, fold the towel, open the backpack, close the backpack, pick up the bag, drop the bag, open the window, close the window, wipe the table, clean the floor, open the box with one hand, lift the object, move the object, drop the object, hold the object, place the object, throw the object, pick up the object, object, container, box, cup, bowl, plate, spoon, knife, fork, chopstick, pen, pencil, paper, book, phone, remote, laptop, keyboard, mouse, bag, backpack, toy, ball, fruit, apple, orange, banana, grape, vegetable, carrot, cucumber, tomato, bottle, can, lid, cap, key, lock, door, window, wall, floor, table, chair, shelf, hand, finger, arm, face, person, other, background, surface, inside, outside, top, bottom, left, right, upward, downward, hot, cold, wet, dry, clean, dirty, empty, full, broken, fixed, smooth, rough, heavy, light, fragile, durable, rollable, stackable, squeezable, pourable, spillable, openable, closeable, edible, drinkable