What to Trust? A Trust-aware Knowledge-guided Method for Zero-shot Object State Understanding in Videos

Yayun Qi[1], **Xinxiao Wu**[1,2*]

[1]Beijing Key Laboratory of Intelligent Information Technology, School of Computer Science & Technology, Beijing Institute of Technology, China

[2]Guangdong Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, China

{qiyayun,wuxinxiao}@bit.edu.cn

Abstract

Object state understanding aims at recognizing the co-occurrence and transitions of multiple object states in videos. While learning from videos handles seen object states well, it struggles with novel ones. We address this task in a zero-shot setting by extracting state-specific knowledge from pre-trained models and using Vision-Language Models (VLMs) to verify whether such knowledge is visually grounded in videos. However, the extracted knowledge varies in its ability to distinguish states, and VLM observations are not always trustworthy. To address this issue, we propose a trust-aware knowledge-guided method to model knowledge trustworthiness and emphasize highly discriminative knowledge that VLMs can reliably observe. Specifically, we collect spatial knowledge for each object state from retrieved images and cues generated from a Large Language Model, then use VLMs to vote on each knowledge element by scoring its visual consistency with the video. In addition to a single scene, temporal dependencies of object states across scenes are also captured using a generative VLM. Under spatial and temporal constraints, we propose an adaptive knowledge refinement module that iteratively updates knowledge reliability weights to achieve a global consensus in object state inference across the video. Finally, object states are inferred by combining the refined weights with VLM voting results. Experiments on two datasets demonstrate the effectiveness of our method.

Introduction

Object states refer to the physical or functional conditions of objects. An object can exhibit multiple states, some of which may co-occur simultaneously. For example, an egg can be both raw and whisked. Changes between different states reflect how the object has been transformed or manipulated. A comprehensive understanding of object states is crucial for downstream tasks, such as procedural planning (Bi, Luo, and Xu 2021; Niu et al. 2024) and mistake detection (Flaborea et al. 2024; Lee et al. 2024; Peddi et al. 2024) in instructional videos.

Object state recognition requires reasoning over both the visual characteristics of the object and the contextual information within the video, such as object shape, co-occurring items, and operations applied to the object. Existing methods (Alayrac et al. 2017; Souček et al. 2022, 2024; Xue, Ashutosh, and Grauman 2024; Tateno et al. 2025) typically learn state-related patterns from videos that contain specific target object states. However, due to the long-tail distribution of object states, it is impractical to collect relevant videos and retrain recognition models once a novel object state is encountered. This motivates us to investigate on a zero-shot setting, where a method can be directly applied to recognize any object states without requiring additional training data. Under this setting, the main challenge lies in how to acquire state-specific knowledge and use it for recognition.

Considering that pre-trained models encode extensive knowledge from large-scale data, they are appropriate providers of state-specific knowledge. Moreover, the multimodal reasoning capabilities of VLMs make them effective observers for building connections between the acquired knowledge and video content. While this paradigm is promising, it raises two key concerns in practice. First, regarding the acquired knowledge, their efforts differ in distinguishing object states. Second, concerning VLM observations, they are not always reliable, especially when the visual information is subtle.

To address the above concerns, we explicitly model knowledge trustworthiness instead of assuming all knowledge contributes equally to object state recognition. We propose a trust-aware knowledge-guided method to emphasize knowledge that is both discriminative and reliably observable by VLMs. Moreover, our method introduces a voting strategy that uses generative VLMs to score the consistency between textual knowledge and video content. This strategy avoids the reliance on conventional discriminative VLMs for cross-modal alignment, which suffer from bag-of-words behaviors (Thrush et al. 2022; Yuksekgonul et al. 2022; Tang et al. 2023) and sparse attention to input text (Paiss, Chefer, and Wolf 2022; Paiss et al. 2023). By incorporating these VLM voting results and refined knowledge reliability weights, our method enables more reliable and accurate object state understanding.

Specifically, we construct spatial knowledge for each object state by collecting its representative visual cues, including anchor images retrieved from the web and descriptions (*i.e.,* appearances, co-occurring items, and typical operations) generated by an LLM. This knowledge is crucial

*Corresponding author.

for distinguishing states in a scene. Additionally, we extract temporal dependencies of object states across adjacent scenes based on video topics and scene context. Then, we measure the consistency between the spatial knowledge and the video content. For the visual ones, we take the visual embedding similarities between them and scenes given by a discriminative VLM as the voting results. For the textual ones, we prompt a generative VLM to retrieve these visual elements in scenes, and gather the word probabilities of positive responses as the voting results. A novel adaptive knowledge refinement module then alternates between estimating state posteriors using current knowledge reliability weights, and updating these weights by maximizing their alignment with the impact of knowledge elements in current spatial and temporal consistent estimations. Finally, the refined weights and voting results are combined to infer the object states.

The main contributions of this paper are three-fold:

- We propose a trust-aware knowledge-guided method for zero-shot object state understanding. It infers object states based on the visual consistency between state-specific knowledge and the video as voted by VLMs, where knowledge trustworthiness is explicitly modeled to weight the impact of corresponding voting results.
- We propose an adaptive knowledge refinement module that iteratively updates knowledge reliability weights to emphasize knowledge both informative for distinguishing states and reliably observable by VLMs.
- Extensive experiments on two datasets demonstrate that our method outperforms both zero-shot baselines and training-based methods.

Related Work

Understanding Object States

Object state understanding has received growing attention in recent years, as it captures fine-grained semantics essential for video analysis. Early studies typically assume that each video contains a single object state change, structured as a transition from an initial object state, through a manipulating action, to an end state. Based on this causal structure, many methods (Alayrac et al. 2017; Souček et al. 2022, 2024) jointly model state and action recognition. Given videos for each object state, Alayrac *et al.* (Alayrac et al. 2017) cluster object appearances into two state classes while identifying actions that temporally separate them, thereby using their complementarity to find a unified solution. Similarly, Souček *et al.* (Souček et al. 2022) train separate state and action classifiers for each object state under the guidance of temporal causality, which is later extended to a multi-task setting (Souček et al. 2024) to learn shared representations across different object states.

To enhance generalization, Xue *et al.* (Xue, Ashutosh, and Grauman 2024) introduce object-agnostic state prediction that enables the transfer of state transition understanding from seen objects to unseen ones, such as from cutting an apple to cutting a pear. To capture object state co-occurrence, Tateno *et al.* (Tateno et al. 2025) reformulate the task as a more challenging multi-label classification setting, enabling each video segment to be associated with multiple object states. Moreover, they use LLMs to generate pseudo-labels from video narrations to supervise model training.

Despite these advances, existing methods still require collecting videos and retraining to recognize novel object states, including the object-agnostic method (Xue, Ashutosh, and Grauman 2024). Different from them, we eliminate this need by acquiring knowledge about object states from pre-trained models for zero-shot object state understanding.

Probing the Reliability of Pre-trained Models

With the rapid progress of pre-trained models, both VLMs and LLMs have demonstrated impressive performance in various downstream tasks. However, recent probing studies reveal that the outputs of these models are not always reliable. A key issue is hallucination (Zhang et al. 2023; Liu et al. 2023a), where the generated responses deviate from the input data or instructions. Meanwhile, discriminative VLMs pre-trained using contrastive objectives like CLIP (Radford et al. 2021) are prone to exhibiting bag-of-words behaviors (Thrush et al. 2022; Yuksekgonul et al. 2022; Tang et al. 2023) and often attend to a sparse set of input texts during cross-modal alignment (Paiss, Chefer, and Wolf 2022; Paiss et al. 2023). Yuksekgonul *et al.* (Yuksekgonul et al. 2023) specifically attribute these limitations of discriminative VLMs to their retrieval-based training objectives and biases present in current vision-language datasets. Tschannen *et al.* (Tschannen et al. 2023) further show that generative VLMs trained using visual captioning as objectives outperform CLIP-style models in zero-shot classification tasks that require sensitivity to visual attributes.

Inspired by these insights, we explicitly model the trustworthiness of knowledge by considering both the quality of the knowledge itself and the reliability of VLMs used to observe the knowledge. We also introduce a novel voting strategy based on the word probabilities from generative VLMs, rather than relying on discriminative VLMs.

Our Method

For zero-shot object state understanding, the goal is to predict the object states $\mathcal{S}_t$ at each time step t in a given video V, without relying on additional training data. In this paper, we focus on the challenging multi-state setting, where multiple object states may co-occur simultaneously. Our method comprises two main stages: (1) knowledge-guided voting and (2) knowledge trustworthiness modeling. The second stage includes temporal state dependency extraction and adaptive knowledge refinement, which iteratively performs state posterior estimation and knowledge reliability weight update, as illustrated in Fig. 1.

Knowledge-guided Voting

The knowledge-guided voting stage begins with collecting state-specific knowledge from external sources, followed by employing VLMs to vote on each knowledge element by measuring its visual consistency with the video segments.

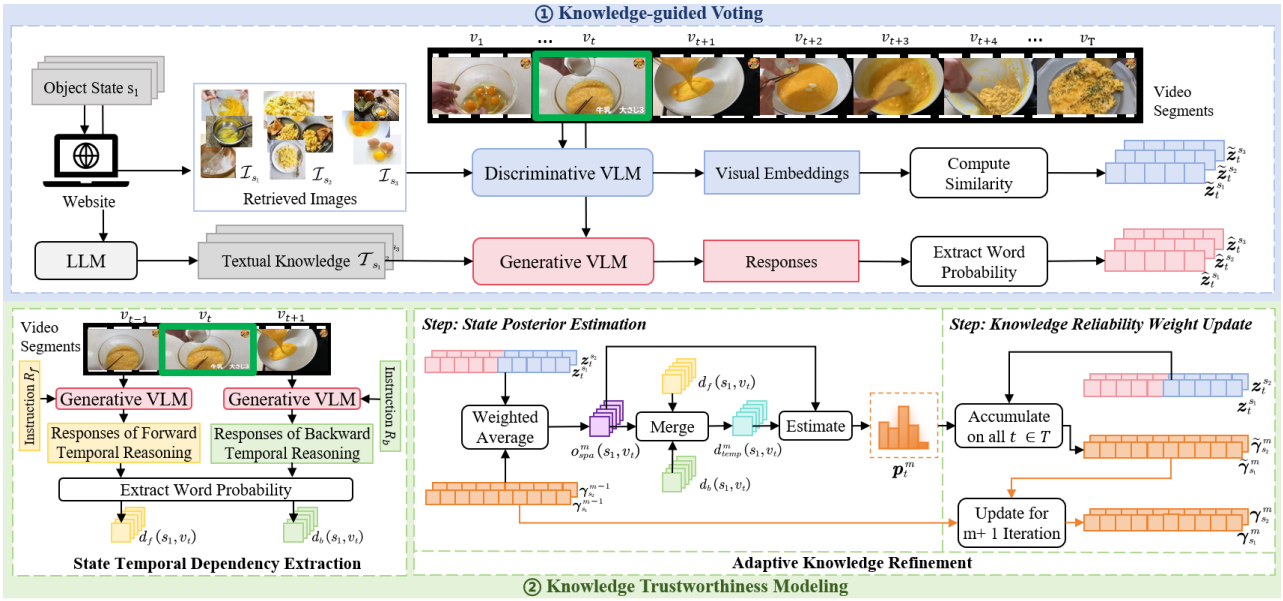

Figure 1: An overview of our method.

State-specific Knowledge Collection We exploit knowledge related to the spatial characteristics of each object state from two external sources, *i.e.,* images stored on the web and an LLM. The images provide intuitive visual evidence about what an object looks like in a particular state, while the LLM offers empirical textual descriptions about the environmental context typically associated with that state.

For each object state $s \in \mathcal{S}$, we retrieve a set of images $\mathcal{I}_s = \{I_1^s, I_2^s, \cdots, I_N^s\}$ from the Bing website by using s as the keyword. The retrieval is based on images instead of videos because keyword-matched videos are often noisy, which may contain entire tutorial sequences or focus on deviated topics, rather than precisely depicting the target object state. For example, a query for "fried egg" retrieves full tutorials on cooking fried eggs from scratch, while "raw egg" returns videos that discuss the benefits of eating raw eggs.

Besides the visual cues, we prompt an LLM to generate textual descriptions $\mathcal{T}_s$ about objects and actions commonly observed simultaneously with state s, and also the representative appearances of s. These descriptions provide textual knowledge containing explicit contextual cues and high-level semantic cues related to s.

Finally, the retrieved image set $\mathcal{I}_s$ and the LLM-generated descriptions $\mathcal{T}_s$ together constitute the state-specific spatial knowledge $\mathcal{K}_s$, providing multimodal information essential for recognizing object states in a video.

Voting with VLMs Given the collected knowledge, we perform knowledge-guided voting via VLMs to support subsequent estimations of the posterior distributions of object states appearing in a video segment. Firstly, we split the video V into segments $\{v_1, \cdots, v_T\}$ based on scene changes, where each segment contains a semantically coherent visual scene. Inspired by voting-based ensemble, each knowledge element then acts as an independent "voter" that provides soft evidence based on its consistency with the visual content of segments. Specifically, the voting strategies differ depending on the knowledge modalities.

Voting strategy for visual knowledge: The visual knowledge $\mathcal{I}$ captures representative visual characteristics associated with specific object states. If similar visual patterns are observed in a video segment, it indicates a higher probability that the corresponding state is present. Therefore, we compute the visual similarity between the retrieved images and the video segment as the voting score.

Given a video segment v_t, we extract a frame f_t from the center of this segment and encode it with CLIP to obtain the frame embedding $\phi(f_t)$. Moreover, for each object state s, we obtain the embeddings for each retrieved image in $\mathcal{I}_s$ from CLIP, denoted as $\{\phi(I_1^s), \cdots, \phi(I_N^s)\}$. After that, we compute the cosine similarity between each pair of frame embedding and image embedding, formulated as

$$\text{score}_V(I_k^s, v_t) = \cos(\phi(I_k^s), \phi(f_t)). \quad (1)$$

This score reflects how visually consistent a video segment is with the prototypical appearance of state s. For each state s, we concatenate all $\text{score}_V(I_k^s, v_t)$ for $I_k^s \in \mathcal{I}_s$ to form the voting score vector for the visual knowledge, denoted as $\widetilde{\boldsymbol{z}}_t^s$.

Voting strategy for textual knowledge: The textual knowledge $\mathcal{T}$ provides high-level semantic cues describing contextual elements commonly associated with specific object states. The presence of such elements suggests a higher probability that the corresponding object state appears in the video segment. To quantify this, we design a strategy that employs generative VLMs to assess the visual consistency of each knowledge element with the video segment, thereby deriving the voting scores for textual knowledge.

Benefiting from pre-training on visual captioning tasks, generative VLMs like VideoLLaMA2 (Cheng et al. 2024) exhibit strong cross-modal alignment capabilities. However,

their output format is discrete with deterministic one-hot predictions, which hinders them from reflecting the cross-modal consistency in a fine-grained and quantitative format.

To overcome this limitation, we take into account the word-level probability distribution produced during the decoding process of generative VLMs. These probabilities offer soft evidence that reflects the confidence of the model across candidate words. Specifically, we instruct the VLM to retrieve the knowledge element from a video segment and respond with a single word, "yes" or "no". The instruction-following ability reduces the solution space to merely two single words.

Instead of relying on hard binary outputs, we extract the probability assigned to the generated word to derive a soft consistency score, which reflects the model's confidence in the presence of the knowledge element. If the model responds with "yes", the consistency score is set to the probability of generating the word "yes". Otherwise, if it responds with "no", indicating that the element is likely absent, the score is defined as the inverse probability of generating the word "no". In a summary, for a knowledge element $T_k^s \in \mathcal{T}_s$ and a video segment v_t, their consistency $\text{score}_T(T_k^s, v_t)$ can be derived from the VLM output y, formulated as

$$\begin{aligned} \alpha_\theta(v, x) &= \begin{cases} P_\theta(y \mid v, x), & y = \text{“yes”}, \\ 1 - P_\theta(y \mid v, x), & y = \text{“no”}, \end{cases} \\ \text{score}_T(T_k^s, v_t) &= \alpha_\theta(v_t, \mathcal{M}(T_k^s, R_v)), \end{aligned} \tag{2}$$

where $P_\theta(y \mid v, x)$ denotes the probability of the VLM to generate y based on model parameters θ, input visual data v, and instruction x. $\mathcal{M}(T_k^s, R_v)$ represents the instruction obtained by merging the knowledge element T_k^s into the instruction template R_v that contains the task definition and output format demand for the VLM. $\text{score}_T(T_k^s, v_t)$ reflects how well an element T_k^s typically associated with object state s can be visually grounded in v_t. For each state s, we concatenate all $\text{score}_T(T_k^s, v_t)$ for $T_k^s \in \mathcal{T}_s$ to form the voting score vector for the textual knowledge, denoted as $\hat{\boldsymbol{z}}_t^s$.

Knowledge Trustworthiness Modeling

After knowledge-guided voting, we introduce a reliability weight $\gamma_{s,i}$ to represent the trustworthiness of each state-specific knowledge $K_s^i \in \mathcal{K}_s$, and iteratively update these weights guided by the nature of spatial consistency and temporal coherence in object states. The voting results can be aggregated with the corresponding weights to infer the presence of object states. The following sections detail the implementation of knowledge trustworthiness modeling, which consists of a state temporal dependency extraction process and an adaptive knowledge refinement module.

State Temporal Dependency Extraction Since object states reflect physical conditions, their temporal transitions often follow implicit causal orders. Capturing such temporal dependencies is crucial for guiding state inference and constraining implausible inferences. In practice, beyond general patterns, the temporal dependency of states also varies with the video topic. For example, while an egg is unlikely to remain intact after being whisked, its specific transition sequence differs across videos. In an omelet tutorial, an egg typically transforms from raw to whisked, then to cooked. In contrast, a video about making a sunny-side-up egg may involve a direct transition from raw to cooked. Moreover, video editing frequently disrupts temporal continuity between adjacent scenes, necessitating video-specific temporal dependency extraction.

To model temporal dependencies between object states, we first infer the video topic C by querying a VLM to reason over the entire video V. Then, we introduce a bidirectional strategy to assess the plausibility of a state appearing in video segment v_t based on temporal coherence with adjacent segments. Specifically, we define a forward score $d_f(s, v_t)$ and a backward score $d_b(s, v_t)$, which reflect the plausibility of state s appearing in v_t from the perspective of its previous segment v_{t-1} and the subsequent segment v_{t+1}, respectively. Following a strategy similar to textual knowledge voting, both scores are obtained by prompting a generative VLM to reason about whether s is temporally coherent with each adjacent segment, formulated as

$$\begin{aligned} d_f(s, v_t) &= \alpha_\theta(v_{t-1}, \mathcal{M}(s, C, R_f)), \\ d_b(s, v_t) &= \alpha_\theta(v_{t+1}, \mathcal{M}(s, C, R_b)), \end{aligned} \tag{3}$$

where instructions R_f and R_b include both the task definition and output format demand for forward and backward reasoning, respectively.

Adaptive Knowledge Refinement Given the knowledge-guided voting results and state temporal dependencies, we propose an adaptive knowledge refinement module to model knowledge trustworthiness with iteratively updated reliability weights $\boldsymbol{\gamma}$. Starting from a uniform initialization of $\boldsymbol{\gamma}$, this module iteratively alternates between estimating the posterior distributions of object states in each video segment and updating the reliability weights until convergence.

State posterior estimation: For video segment v_t, we define $\boldsymbol{z}_t^s$ as the concatenation of the voting score vectors of visual and textual knowledge, *i.e.,* $\boldsymbol{z}_t^s = [\tilde{\boldsymbol{z}}_t^s, \hat{\boldsymbol{z}}_t^s]$. At the m-th iteration, we calculate the weighted average of voting scores over all knowledge elements $K_s^i \in \mathcal{K}_s$ using their reliability weights $\gamma_{s,i}^{m-1}$ from the previous iteration, denoted as $o_{spa}^m(s, v_t)$. It provides spatial consistency constraints for the estimation of state posteriors, calculated by

$$o_{spa}^m(s, v_t) = \frac{1}{|\mathcal{K}_s|} \sum_{i=1}^{|\mathcal{K}_s|} \gamma_{s,i}^{m-1} \cdot z_t^{s,i}. \tag{4}$$

To provide temporal coherence constraints, the obtained bidirectional state temporal dependencies $d_f(s, v_t)$ and $d_b(s, v_t)$ are merged with spatial cues $o_{spa}^m(s, v_t)$, which is based on the similarity between v_t and its adjacent segments v_{t-1} (from which $d_f(s, v_t)$ is derived) and v_{t+1} (from which $d_b(s, v_t)$ is derived). The similarity is given by $\text{sim}(v_i, v_j) = \frac{1}{\|\mathbf{z}_i' - \mathbf{z}_j'\|_2}$, where $\boldsymbol{z}_i'$ denotes the concatenation of the weighted version of all voting scores. Then, a factor w_i^t is computed to emphasize the contributions from segments more similar to v_t, while smoothing the impact of

Method	Apple		Egg		Flour		Shirt		Tire		Wire		Average	
	F1	mAP	F1	mAP	F1	mAP	F1	mAP	F1	mAP	F1	mAP	F1	mAP
LLAVA	0.34	—	0.29	—	0.35	—	0.28	—	0.47	—	0.27	—	0.33	—
CLIP	0.42	0.35	0.37	0.28	0.38	0.26	0.33	0.27	0.55	0.45	0.33	0.25	0.39	0.31
VideoCLIP	0.35	0.25	0.34	0.26	0.40	0.32	0.28	0.19	0.51	0.39	0.31	0.21	0.37	0.27
InternVideo	0.46	0.39	0.44	0.39	0.43	0.36	0.40	0.32	0.57	0.45	0.40	0.31	0.45	0.37
VideoLLaMA2	0.30	—	0.25	—	0.25	—	0.24	—	0.49	—	0.27	—	0.30	—
InternVL3	0.45	—	0.46	—	0.49	—	0.41	—	0.57	—	0.37	—	0.46	—
QWEN2.5VL	0.43	—	0.46	—	0.45	—	0.47	—	0.55	—	0.37	—	0.46	—
LNV-AoTD	0.24	—	0.26	—	0.26	—	0.23	—	0.48	—	0.25	—	0.29	—
Cap-LLM	0.37	—	0.32	—	0.37	—	0.41	—	0.55	—	0.31	—	0.39	—
LMOSA*	0.53	**0.50**	0.53	0.48	0.55	0.49	0.50	**0.45**	0.61	0.52	0.50	0.42	0.54	0.48
Ours	**0.57**	0.48	**0.55**	**0.49**	**0.57**	**0.53**	**0.56**	0.44	**0.66**	**0.58**	**0.57**	**0.50**	**0.58**	**0.50**

Table 1: Comparison results with zero-shot baselines and a training-based method on MOST, where the evaluation metrics (F1 and mAP) follow the original protocol of MOST. * denotes training using videos containing target object states.

segments that may have edits or content shifts, given by

$$w_i^t = \frac{\text{sim}(v_i, v_t)}{\text{sim}(v_{t-1}, v_t) + \text{sim}(v_t, v_{t+1})},$$
$$d_{\text{temp}}^m(s, v_t) = d_f(s, v_t)^{w_{t-1}^t} \cdot o_{spa}^m(s, v_t) \cdot d_b(s, v_t)^{w_{t+1}^t}. \tag{5}$$

Finally, we estimate the posterior probability $p_t^m(s)$ for v_t, indicating how likely this segment depicts state s, given by

$$p_t^m(s) = \lambda \cdot \frac{d_{\text{temp}}^m(s, v_t)}{\sum_{s'} d_{\text{temp}}^m(s', v_t)} + (1-\lambda) \cdot \frac{o_{spa}^m(s, v_t)}{\sum_{s'} o_{spa}^m(s', v_t)}, \tag{6}$$

where λ is a factor balancing spatial and temporal results.

Knowledge reliability weight update: Using the estimated $p_t^m(s)$, the reliability weight $\gamma_{s,i}^m$ of each knowledge element K_s^i is updated by maximizing its alignment with the current estimation. To prevent overly generic or easily misobserved knowledge elements from dominating the update, we introduce a global frequency penalty $g_{s,i} = \exp(-\xi \cdot f_{s,i})$, where $f_{s,i}$ is the average voting frequency of K_s^i, and ξ is a penalty factor.

At the m-th iteration, we accumulate evidence for each knowledge element K_s^i by weighting its observation $z_t^{s,i}$ at segment v_t with the posterior probability $p_t^m(s)$ of v_t containing state s. This quantifies how much K_s^i contributes to supporting the presence of state s. The updating process of knowledge reliability weights is given by

$$\widetilde{\gamma}_{s,i}^m = \frac{g_{s,i} \cdot \sum_{t=1}^T p_t^m(s) \cdot z_t^{s,i}}{\sum_{t=1}^T p_t^m(s)},$$
$$\gamma_{s,i}^m = \eta \cdot \widetilde{\gamma}_{s,i}^m + (1-\eta) \cdot \gamma_{s,i}^{m-1}, \tag{7}$$

where $\eta = \frac{1}{m+1}$ is a smoothing factor. This updating process promotes a gradual shift of the reliability weights toward more discriminative and trustworthy knowledge elements, enhancing state inference accuracy in later iterations.

Overall, the above two steps are alternately performed until convergence or a predefined maximum iteration M is reached (see Appendix Sec.1). Upon termination at the q-th iteration, the final state prediction is given by $p_t^q(s)$.

Experiments

Dataset and Evaluation Metric

We conduct experiments on the MOST dataset (Tateno et al. 2025), which follows a multi-label classification setting for co-occurring object states, and the ChangeIT dataset (Souček et al. 2022), which focuses on single state changes. Following their official evaluation protocols, we report F1-max and mean Average Precision (mAP) for MOST, and Precision@1 for ChangeIT. To better evaluate multi-label prediction quality on MOST, we further adopt Coverage Error (CoE) and Adaptive Top-K Accuracy (ATK). CoE measures the average number of top-ranked predictions required to cover all ground-truth labels, with lower values indicating better performance. ATK computes the macro-averaged F1 score over the top-K predictions, where K equals the number of ground-truth labels at specific time steps. Both metrics reflect how effectively the model prioritizes true object states among all candidates.

Implementation Details

For state-specific knowledge, we retrieve the top-10 images for each object state from Bing to form $\mathcal{I}_s$, and use QWEN2.5-72B (Qwen et al. 2024) to provide textual knowledge $\mathcal{T}_s$. For knowledge-guided voting on textual knowledge, we adopt VideoLLaMA2 (Cheng et al. 2024) as the generative VLM and implement our voting strategy on it. The convergence condition of our adaptive knowledge refinement is defined as $\max_{s \in \mathcal{S}} \left(||\boldsymbol{\gamma}_s^m - \boldsymbol{\gamma}_s^{m-1}||_\infty \right) < 1 \times 10^{-5}$, with a maximum of $M = 20$ iterations. For MOST, we set the penalty factor $\xi = 2$ and balancing factor $\lambda = 0.3$ for eggs and shirts, $\xi = 1$ and $\lambda = 0.4$ for wires, while $\xi = 3$ and $\lambda = 0.2$ for others. For ChangeIT, we set $\xi = 1$ and $\lambda = 0.8$. Experiments are conducted on 4 NVIDIA A40 GPUs. All instructions are provided in Appendix Sec.3.

Comparison with Existing Methods

To evaluate the effectiveness of our method on multiple object state understanding, we compare it with several

Method	Apple		Egg		Flour		Shirt		Tire		Wire		Average	
	CoE↓	ATK	CoE↓	ATK	CoE↓	ATK	CoE↓	ATK	CoE↓	ATK	CoE↓	ATK	CoE↓	ATK
CLIP	3.68	0.39	3.38	0.34	2.33	0.23	2.65	0.26	5.16	0.31	3.56	0.16	3.46	0.28
VideoCLIP	3.01	0.44	4.18	0.30	1.73	0.40	3.82	0.18	5.01	0.28	4.81	0.23	3.76	0.31
InternVideo	3.30	0.39	3.78	0.32	2.50	0.27	4.09	0.18	5.09	0.35	4.83	0.23	3.93	0.29
Ours	**2.40**	**0.50**	**2.75**	**0.46**	**0.86**	**0.56**	**1.87**	**0.41**	**3.06**	**0.53**	**1.76**	**0.48**	**2.12**	**0.49**

Table 2: Comparison results with zero-shot baselines and a training-based method on MOST, where the evaluation metrics (CoE and ATK) are able to evaluate the quality of multi-label predictions.

Method	State Prec.@1	Action Prec.@1
CLIP	0.29	0.70
VideoCLIP	0.24	0.55
InternVideo	0.25	0.61
LookForTheChange	0.25	0.68
MultiTaskChange	0.22	0.62
VIDOSC	0.29	0.63
Ours	**0.37**	**0.73**

Table 3: Comparison results with VLM baselines, close-vocabulary methods, and the open-vocabulary method VIDOSC on the novel object split of ChangeIT.

VLMs, a baseline (Cap-LLM), an Agent-of-Thoughts Distilled VLM (LNV-AoTD) (Shi et al. 2025), and a state-of-the-art method (LMOSA) (Tateno et al. 2025). Specifically, VLMs perform the task either by computing similarities between sampled frames (or video segments) and object state categories (*e.g.,* CLIP, VideoCLIP (Xu et al. 2021), InternVideo (Wang et al. 2022)), or by directly predicting object states (*e.g.,* LLaVA (Liu et al. 2023b), VideoLLaMA2, InternVL3 (Zhu et al. 2025), QWEN2.5-VL (Bai et al. 2025), LNV-AoTD). Cap-LLM generates segment-level captions using VideoLLaMA2 and prompts LLaMA3-8B (Grattafiori et al. 2024) to infer object states from the captions. In contrast, LMOSA requires training on videos containing target object states. More results of our method using InternVL3 as the generative VLM are shown in Appendix Sec. 2.1.

For the MOST dataset, we report results using the original metrics F1-max and mAP in Table 1, and using the multi-label metrics CoE and ATK in Table 2. Note that CoE and ATK rely on ranking the predicted state scores and are therefore not applicable to methods that output binary predictions. It can be observed that our zero-shot method achieves a significant improvement over other zero-shot baseline methods. Compared with the training-based method LMOSA, our method also achieves the best performance (averaged across object categories) on all metrics.

We also compare our method with VIDOSC (Xue, Ashutosh, and Grauman 2024), a representative open-vocabulary method designed to recognize single object state changes. Although our method targets the more challenging multi-state recognition setting, we follow the task definition from VIDOSC and evaluate on the novel object state split of ChangeIT. Table 3 reports the performance of several methods, including VLM baselines, two closed-vocabulary methods (*i.e.,* LookForTheChange (Souček et al. 2022) and MultiTaskChange (Souček et al. 2024)), and VIDOSC that trains a unified model across all categories. Our method outperforms all baselines and existing methods, demonstrating superior generalization over the open-vocabulary method and strong effectiveness under the single-state transition setting.

Type	Method	CoE↓	ATK
Voting with Discriminative VLMs	CLIP Score	4.02	0.20
	InternVideo Score	3.55	0.25
Voting with Generative VLMs	Binary Answer	2.88	0.39
	Predicted Numbers	2.85	0.37
	Ours	**2.12**	**0.49**

Table 4: Comparison of applying different voting strategies for textual knowledge, averaged across object categories.

Comparison between Different Voting Strategies

To validate the effectiveness of our voting strategy for textual knowledge, we replace it with using discriminative VLMs to compute the similarity between textual knowledge and visual content, including image-level CLIP and video-level InternVideo. We also design two variants based on generative VLMs: (1) Binary Answer: we prompt VideoLLaMA2 with the same instruction as our strategy and assign a score of 1 to "yes" and 0 to "no". (2) Predicted Numbers: we prompt VideoLLaMA2 to directly output a consistency score between 0 and 1. As shown in Table 4, methods with generative VLMs perform better than discriminative ones, which indicates the advantages of generative VLMs in cross-modal alignment. Moreover, our method achieves the best performance among all variants, demonstrating the effectiveness of our knowledge voting strategy, especially the superiority of extracting word-level probabilities from generative VLMs to obtain soft consistency scores.

Ablation Study

We conduct ablation studies by removing the following key components: knowledge-guided voting (denoted as Voting), state temporal dependency extraction (Temporal), and adap-

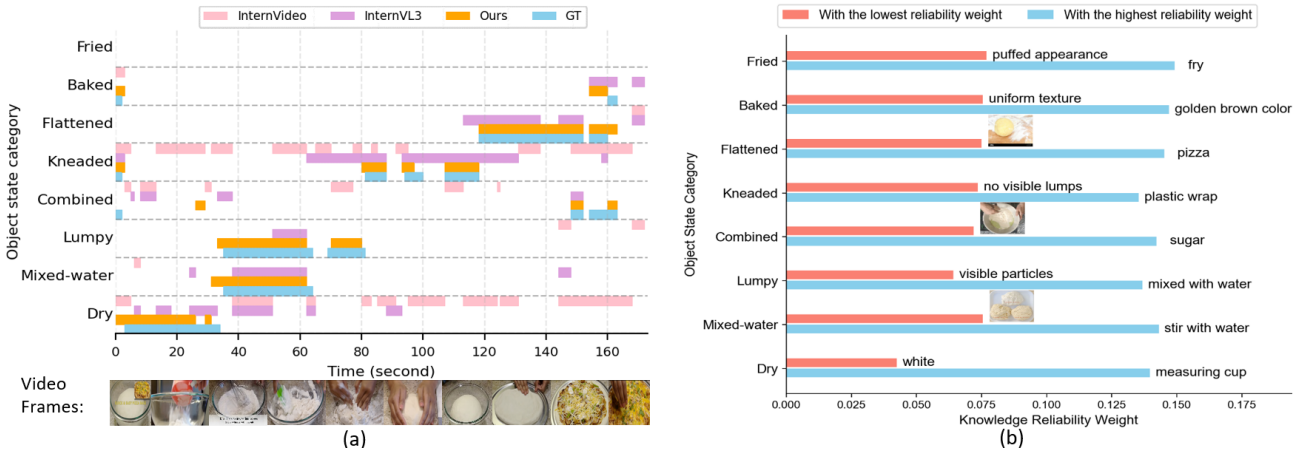

Figure 2: Qualitative results of inferred object states (a) and our knowledge reliability weights (b) on a flour-related video. For clarity, we abbreviate the "combined with other ingredient" state as "Combined", and "mixed with water" as "Mixed-water".

Voting	Temporal	Refine	CoE ↓	ATK
√	×	×	2.65	0.42
×	√	×	2.74	0.34
√	√	×	2.32	0.44
√	×	√	2.64	0.42
√	√	√	**2.12**	**0.49**

Table 5: Results of ablation studies. The results are averaged across object categories.

tive knowledge refinement (Refine). Notably, the removal of knowledge-guided voting also disables the adaptive knowledge refinement due to the absence of knowledge, and sets $o^m_{spa}(s, v_t)$ and w^t_i in Eq. (5) to 1. When adaptive knowledge refinement is removed, reliability weights $\boldsymbol{\gamma}$ are fixed to 1, and state inference equals the initial estimation $p^1_t(s)$.

Results on the MOST dataset are shown in Table 5, from which we have the following observations: (1) The removal of any components leads to a performance drop, which highlights the contribution of each component in zero-shot object state understanding. (2) The comparison of the first three rows indicates the benefit of jointly considering the spatial consistency within a segment and temporal coherence across segments. (3) For the comparison of the first, fourth, and last rows, it suggests the importance of introducing temporal constraints into the adaptive knowledge refinement module for effective knowledge trustworthiness modeling.

Qualitative Results

Fig. 2(a) shows the object states inferred by our method (in orange), InternVL3 (purple), and InternVideo (pink) on a tutorial video about making pizza with flour. Compared to VLM baselines, our method infers object states more consistent with the GT annotations (blue), highlighting its advantage in the challenging zero-shot object state understanding task, where even the advanced InternVL3 still struggles. Interestingly, our method identifies that flour is "combined with other ingredients" during the 25s-30s segment, which is not suggested by the GT. Upon checking the video, we find that this segment indeed shows a person mixing flour with salt and yeast, which is successfully captured by our method. Our method also meets challenges in the 150s-160s segment, where the dough is topped with vegetables and cheese but not yet baked. In this case, our method mistakenly infers a "baked" state, as this inference seems to be plausible from both the perspectives of spatial and temporal.

Fig. 2(b) further visualizes knowledge elements with the lowest and highest reliability weights as determined by our method. For example, although "white" is a valid visual characteristic of dry flour, it is a non-discriminative cue shared by many flour states, which is successfully captured by our method and assigned a lower weight. Similarly, misleading visual cues (*e.g.,* kneaded-like images for the flattened and mixed with water states) and elements hard to observe (*e.g.,* "no visible lumps") are also assigned lower weights. These results demonstrate that our method is effective in modeling knowledge trustworthiness. More qualitative results are shown in Appendix Sec.4.

Conclusion

We present a trust-aware knowledge-guided method for zero-shot object state understanding, which mitigates the training process on videos with target object states. Our method performs knowledge-guided voting, which employs VLMs to measure the consistency between state-specific knowledge and visual content. A novel adaptive knowledge refinement module iteratively refines knowledge reliability weights, allowing trustworthy knowledge elements to dominate the state estimation and support accurate object state inference. Extensive experiments on two datasets demonstrate that our method outperforms zero-shot baselines and even surpasses training-based methods. In the future, we plan to extend our method to more challenging scenarios, such as object state understanding in online videos and multi-object state understanding, with incremental and more structured knowledge integration.

Acknowledgments

This work was supported in part by the grants from the Natural Science Foundation of China (NSFC) under Grant No. 62072041, the Shenzhen Science and Technology Program under Grant No. JCYJ20241202130548062, the Natural Science Foundation of Shenzhen under Grant No. JCYJ20230807142703006, and the Key Research Platforms and Projects of the Guangdong Provincial Department of Education under Grant No. 2023ZDZX1034.

References

Alayrac, J.-B.; Laptev, I.; Sivic, J.; and Lacoste-Julien, S. 2017. Joint discovery of object states and manipulation actions. In *Proceedings of the IEEE International Conference on Computer Vision*, 2127–2136.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Bi, J.; Luo, J.; and Xu, C. 2021. Procedure planning in instructional videos via contextual modeling and model-based policy learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15611–15620.

Cheng, Z.; Leng, S.; Zhang, H.; Xin, Y.; Li, X.; Chen, G.; Zhu, Y.; Zhang, W.; Luo, Z.; Zhao, D.; et al. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Flaborea, A.; Di Melendugno, G. M. D.; Plini, L.; Scofano, L.; De Matteis, E.; Furnari, A.; Farinella, G. M.; and Galasso, F. 2024. PREGO: online mistake detection in PRocedural EGOcentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18483–18492.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Lee, S.-P.; Lu, Z.; Zhang, Z.; Hoai, M.; and Elhamifar, E. 2024. Error detection in egocentric procedural task videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18655–18666.

Liu, F.; Guan, T.; Li, Z.; Chen, L.; Yacoob, Y.; Manocha, D.; and Zhou, T. 2023a. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Niu, Y.; Guo, W.; Chen, L.; Lin, X.; and Chang, S.-F. 2024. SCHEMA: State CHanges MAtter for procedure planning in instructional videos. *arXiv preprint arXiv:2403.01599*.

Paiss, R.; Chefer, H.; and Wolf, L. 2022. No token left behind: Explainability-aided image classification and generation. In *European Conference on Computer Vision*, 334–350. Springer.

Paiss, R.; Ephrat, A.; Tov, O.; Zada, S.; Mosseri, I.; Irani, M.; and Dekel, T. 2023. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3170–3180.

Peddi, R.; Arya, S.; Challa, B.; Pallapothula, L.; Vyas, A.; Gouripeddi, B.; Zhang, Q.; Wang, J.; Komaragiri, V.; Ragan, E.; et al. 2024. CaptainCook4D: A dataset for understanding errors in procedural activities. *Advances in Neural Information Processing Systems*, 37: 135626–135679.

Qwen, A. Y.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 technical report. *arXiv preprint*.

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

Shi, Y.; Di, S.; Chen, Q.; and Xie, W. 2025. Enhancing Video-LLM Reasoning via Agent-of-Thoughts Distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 8523–8533.

Souček, T.; Alayrac, J.-B.; Miech, A.; Laptev, I.; and Sivic, J. 2022. Look for the change: Learning object states and state-modifying actions from untrimmed web videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13956–13966.

Souček, T.; Alayrac, J.-B.; Miech, A.; Laptev, I.; and Sivic, J. 2024. Multi-task learning of object states and state-modifying actions from web videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7): 5114–5130.

Tang, Y.; Yamada, Y.; Zhang, Y.; and Yildirim, I. 2023. When are Lemons Purple? The Concept Association Bias of Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 14333–14348.

Tateno, M.; Yagi, T.; Furuta, R.; and Sato, Y. 2025. Learning Multiple Object States from Actions via Large Language Models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 9555–9565. IEEE.

Thrush, T.; Jiang, R.; Bartolo, M.; Singh, A.; Williams, A.; Kiela, D.; and Ross, C. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5238–5248.

Tschannen, M.; Kumar, M.; Steiner, A.; Zhai, X.; Houlsby, N.; and Beyer, L. 2023. Image captioners are scalable vision learners too. *Advances in Neural Information Processing Systems*, 36: 46830–46855.

Wang, Y.; Li, K.; Li, Y.; He, Y.; Huang, B.; Zhao, Z.; Zhang, H.; Xu, J.; Liu, Y.; Wang, Z.; et al. 2022. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*.

Xu, H.; Ghosh, G.; Huang, P.-Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; and Feichtenhofer, C. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*.

Xue, Z.; Ashutosh, K.; and Grauman, K. 2024. Learning object state changes in videos: An open-world perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18493–18503.

Yuksekgonul, M.; Bianchi, F.; Kalluri, P.; Jurafsky, D.; and Zou, J. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*.

Yuksekgonul, M.; Bianchi, F.; Kalluri, P.; Jurafsky, D.; Zou, J.; et al. 2023. WHEN AND WHY VISION-LANGUAGE MODELS BEHAVE LIKE BAGS-OF-WORDS, AND WHAT TO DO ABOUT IT? In *11th International Conference on Learning Representations, ICLR 2023*. International Conference on Learning Representations, ICLR.

Zhang, M.; Press, O.; Merrill, W.; Liu, A.; and Smith, N. A. 2023. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.