Q-VISTA: SEMANTIC IDENTITY AND QUERY-LOCALIZED TRANSITION DYNAMICS FOR GENERALIZED VIDEO SKILL DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalized category discovery (GCD) organizes known and novel categories using partial supervision. Generalized Video Skill Discovery (GVSD) extends this capability to videos and requires preserving not only semantic identity, but also the localized interactions that define an operation. The same object can participate in different skills, yet global video aggregation can dilute the regions where their defining changes occur. To address this problem, we propose Q-VISTA, which combines *Semantic Identity* with *Query-Localized Transition Dynamics*. Following a *where-before-how* design, recurrent Action Queries first collect regional trajectories, and a multiscale transition operator then describes their evolution. A parallel semantic branch retains global context, while nonlinear fusion brings both representations into a shared discovery space after temporal modeling. This representation supports both non-parametric clustering and a jointly learned parametric classifier without requiring region annotations. Experiments across six human-action and robot-skill protocols demonstrate improved domain-level discovery over SimGCD for both variants under a shared frozen backbone, with pronounced gains in discovering unlabeled multi-step task categories alongside labeled atomic skills.

1 INTRODUCTION

When a recognition model is applied to real-world environments, it often encounters categories beyond those labeled during training. For example, a robot may observe an unfamiliar manipulation task involving objects and actions it has already seen. Recognizing familiar categories or merely detecting novelty is therefore insufficient; the model must also organize unlabeled observations into meaningful categories. Novel category discovery approaches this problem by transferring knowledge from labeled classes to disjoint unlabeled classes (Han et al., 2019; 2020; Fini et al., 2021). Generalized Category Discovery (GCD) considers a more realistic setting, jointly recognizing known categories and discovering novel ones within a mixed unlabeled pool (Vaze et al., 2022; Cao et al., 2022). Advances in contrastive learning, parametric classifiers, prototypes, and information maximization have strengthened image-based discovery within this framework (Pu et al., 2023; Wen et al., 2023; Chiaroni et al., 2023; Zhao et al., 2023). Extending this capability to videos would help organize human activities and robot demonstrations without requiring every action or skill to be manually labeled. We study this setting as Generalized Video Skill Discovery (GVSD).

For the robot in this example, discovering the unfamiliar task requires more than recognizing the objects involved. Real-world activities unfold continuously rather than as isolated images: a skill can depend on how an object changes state, where it is moved, or how several interactions are ordered. Figure 1 illustrates this distinction. V1 and V2 both place an item in a caddy but manipulate different objects, whereas V2 and V3 manipulate the same granola bar but place it at different destinations. Neither shared objects nor similar movements alone determine skill identity. Extending GCD to videos therefore requires more than replacing an image encoder with a video encoder. The discovery representation must preserve semantic identity—the participating objects, actors, and scenes—together with the localized changes that distinguish one operation from another.

Modern video architectures capture rich appearance and motion information (Simonyan & Zisserman, 2014; Feichtenhofer et al., 2019; Zhou et al., 2018; Bertasius et al., 2021; Arnab et al., 2021; Tong

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

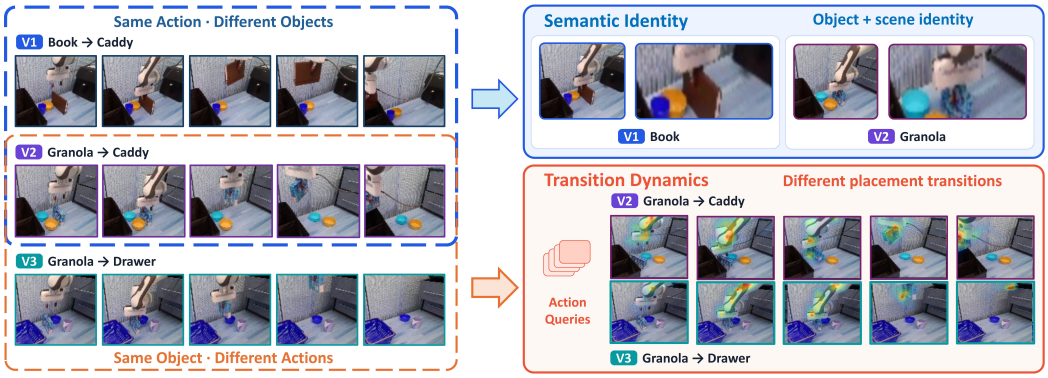


Figure 1: **Why video skill discovery requires dual representation.** V1–V3 are distinct operations: V1/V2 exhibit similar placement dynamics but manipulate different objects, whereas V2/V3 share the object but differ in destination and transition. Semantic Identity preserves object and scene evidence; Action Queries localize the regions used to construct Transition Dynamics.

et al., 2022), but discovery also depends on how this evidence is aggregated. Global pooling mixes persistent scenes with the small interaction regions where an operation changes state. This raises a natural question: how can discovery preserve localized changes together with their semantic context? We separate regional selection from global aggregation, describing the selected trajectory before combining it with the video's semantic identity.

Motivated by this observation, we propose Q-ViSTA with two jointly learned representations. *Semantic Identity* preserves global video context, while *Query-Localized Transition Dynamics* captures changes in selected regional observations. Recurrent Action Queries collect trajectories without boxes, masks, or part labels, and a multiscale operator describes their evolution before fusion with semantic context. The same queries serve known and novel categories without class-specific region supervision. Our atomic-to-composite protocol makes this design particularly relevant: labeled atomic skills anchor the discovery of unlabeled composite task categories. In summary, our main contributions are as follows:

- We formulate the representational gap between image GCD and GVSD: video categories depend jointly on persistent semantic identity and localized state transitions.
- We introduce Q-ViSTA, which combines global Semantic Identity with a *where-before-how* pathway that learns trajectories and multiscale transitions without region supervision.
- Across six human-action and robot-skill protocols, both discovery endpoints improve domain-level averages over SimGCD; a pronounced advantage occurs when distinct multi-step robot skills form the unlabeled novel category set.

2 Related Work

Generalized category discovery. Category discovery has progressed from disjoint known-to-novel transfer (Han et al., 2019; 2020; Fini et al., 2021) to mixed known and novel pools (Vaze et al., 2022; Cao et al., 2022). Contrastive objectives, prompting, information maximization, and neighborhood structure refine discovery features (Pu et al., 2023; Zhang et al., 2023; Wang et al., 2024; Chiaroni et al., 2023; Zhao et al., 2023; Rastegar et al., 2024); SimGCD learns parametric assignments (Wen et al., 2023). NC-GCD (Han et al., 2025) aligns supervised and unsupervised features with fixed equiangular tight frame (ETF) prototypes (Papyan et al., 2020; Han et al., 2026b). Language-assisted discovery (Zheng et al., 2024) includes GET (Wang et al., 2025), which synthesizes text embeddings from CLIP features (Radford et al., 2021), and AL-GCD (Han et al., 2026d), which transfers labeled knowledge through analogical learning (Gentner, 1983; Han et al., 2026c). CMS (Choi et al., 2024) refines features using mean shift (Comaniciu & Meer, 2002; Han et al., 2026a). APL (Dai et al., 2025) learns consistent image parts with shared queries, DINO priors, and an all-min contrastive objective. Whereas APL targets cross-image part correspondence, Q-ViSTA constructs transitions along video-conditioned regional trajectories before semantic fusion. MCCL brings spatiotemporal

consistency and memory to video discovery (Zhang et al., 2025), complementing our focus on regional evidence for discovering categories in unlabeled video collections.

Video representation and temporal evidence. Two-stream, I3D, non-local, temporal-shift, and SlowFast networks exploit appearance, motion, and temporal-rate information (Simonyan & Zisserman, 2014; Carreira & Zisserman, 2017; Wang et al., 2018; Lin et al., 2019; Feichtenhofer et al., 2019). Temporal relation networks and transformers model multiscale space–time structure (Zhou et al., 2018; Bertasius et al., 2021; Arnab et al., 2021; Liu et al., 2022; Fan et al., 2021); masked autoencoders learn from unlabeled videos (Tong et al., 2022; Wang et al., 2023). Structured cross-modal alignment also supports text–video retrieval (Wang et al., 2026). Action Transformer and Temporal Query Networks retrieve relevant context (Girdhar et al., 2019; Zhang et al., 2021), while DETR, Slot Attention, TubeR, and TrackFormer use queries to bind or propagate dense observations (Carion et al., 2020; Locatello et al., 2020; Zhao et al., 2022; Meinhardt et al., 2022). Q-ViSTA uses regional selection to construct explicit transitions for discovery, jointly organizing semantic identity and changing interactions rather than introducing query aggregation in isolation.

3 Q-ViSTA: Identity and Query-Localized Dynamics

3.1 Problem formulation and overview

In GVSD, we consider labeled videos from known classes together with unlabeled videos that contain both known and novel categories in the same discovery pool:

$$\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u, \qquad \mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^{N_l}, \quad \mathcal{D}_u = \{x_i\}_{i=1}^{N_u}, \quad \mathcal{C}_l \subseteq \mathcal{C}_u. \tag{1}$$

The objective is to organize the mixed unlabeled pool using known-class labels as anchors. GVSD retains this transductive GCD setting while addressing video categories whose identity depends on both semantic context and localized state changes. Novel videos participate in learning without their category labels.

As shown in Figure 2, Q-ViSTA learns complementary representations of global context and localized change. For each augmented view, a frozen VideoMAE encoder produces patch tokens $\boldsymbol{X}_i \in \mathbb{R}^{T\times P\times D_0}$. The Semantic branch aggregates these tokens globally to preserve identity, while Action Queries collect regional trajectories for subsequent transition modeling. The two branches are optimized jointly and fused only after their respective representations have been constructed. This design gives the discovery objective access to both the video context and the selected regional changes, without forcing either through the other's aggregation rule. Distinct aggregation pathways and learning objectives provide complementary cues without an independence constraint.

3.2 Semantic identity

The Semantic branch preserves the global context needed to interpret an interaction. We average the frozen tokens over space and time, normalize the resulting descriptor, and project it to a normalized embedding $\boldsymbol{s}_i$. By retaining participant and scene information, this branch complements the localized trajectory with cues about what is being manipulated and where the operation occurs. Its discovery objective jointly organizes these contextual cues across videos.

The Semantic objective combines cross-view InfoNCE over all videos (Chen et al., 2020) with supervised contrastive learning over labeled classes (Khosla et al., 2020):

$$\mathcal{L}_{\text{semantic}} = \mathcal{L}_{\text{GCD}}(\boldsymbol{s}), \qquad \mathcal{L}_{\text{GCD}}(\boldsymbol{e}) = (1-\lambda)\mathcal{L}_{\text{InfoNCE}}(\boldsymbol{e}) + \lambda\mathcal{L}_{\text{SupCon}}(\boldsymbol{e}), \tag{2}$$

with $\lambda = 0.50$. Cross-view agreement organizes the entire discovery pool, while supervised contrast anchors the known-class portion of that geometry throughout joint optimization.

3.3 Where: query-localized action trajectories

Q-ViSTA initializes four video-conditioned Action Queries from mean appearance and temporal change. They retrieve evidence from frozen patch tokens, concentrating small interaction regions into trajectories that remain available for subsequent temporal modeling.

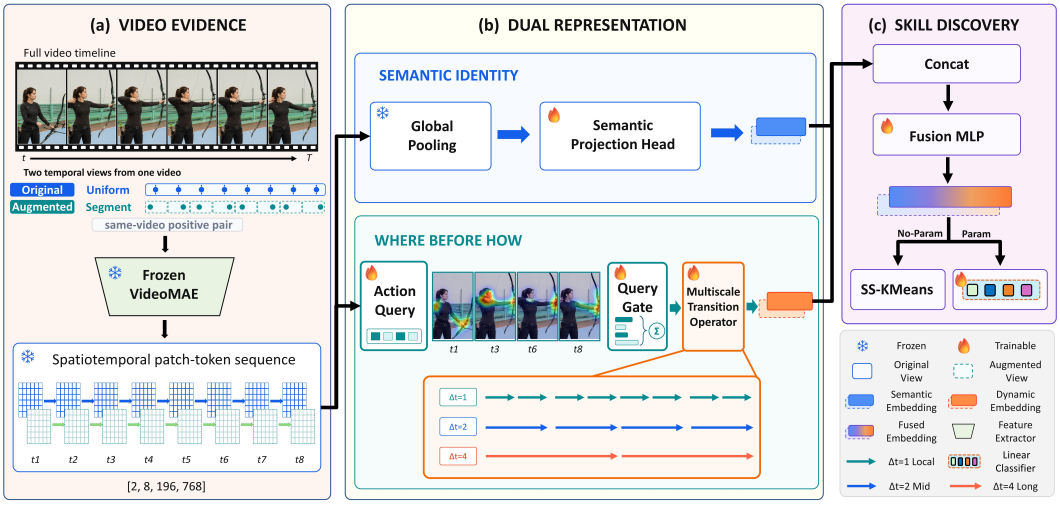


Figure 2: **Q-ViSTA overview.** A frozen VideoMAE supplies Semantic Identity and Query-Localized Transition Dynamics. Their fused embedding supports either non-parametric SS-KMeans or an explicit linear classifier for direct parametric category assignment.

At tubelet t, each query matches projected patch keys and pools the original VideoMAE values. The attended observation is projected to the Query width and updates the recurrent state:

$$
\begin{aligned}
a_{t,k,p} &= \mathrm{Softmax}_p\left(\boldsymbol{q}_{t,k}^{\top}\boldsymbol{k}_{t,p}/\sqrt{d_q}\right), \\
\boldsymbol{u}_{t,k} &= \sum_{p=1}^{P} a_{t,k,p}\boldsymbol{v}_{t,p}, \qquad \widehat{\boldsymbol{q}}_{t+1,k} = \mathrm{GRU}(\boldsymbol{q}_{t,k}, W_U\boldsymbol{u}_{t,k}), \\
\boldsymbol{q}_{t+1,k} &= \mathrm{Norm}(\boldsymbol{q}_{t,k} + \rho(\widehat{\boldsymbol{q}}_{t+1,k} - \boldsymbol{q}_{t,k})).
\end{aligned}
\tag{3}
$$

Matching uses a compact query–key space, while pooled values preserve regional content. In Eq. 3, GRU takes the previous hidden state first and the projected observation second; $\rho = \sigma(\rho_{\mathrm{logit}})$ bounds its learned update, and Norm denotes L2 normalization. Recurrence links the next regional selection to the preceding observation without requiring fixed image coordinates or labeled regions. Thus *where before how* specifies an architectural dependency: explicit transition tokens are built from selected regional observations. The frozen encoder and video-conditioned initialization already supply temporal context, retaining appearance and interaction cues within each regional trajectory.

3.4 How: Soft Query Aggregation and Transition Operator

All four trajectories remain available, but their relevance changes with the local action state. At each tubelet, a learned scorer therefore lets them compete softly:

$$
g_{t,k} = \mathrm{Softmax}_k(r(\boldsymbol{u}_{t,k})), \qquad \widetilde{\boldsymbol{x}}_t = \sum_{k=1}^{4} g_{t,k}\boldsymbol{u}_{t,k}. \tag{4}
$$

This All-4 aggregation preserves the full Query bank while adapting its contribution at each tubelet. A diversity regularizer $\mathcal{L}_{\mathrm{query}}$ penalizes cosine overlap between flattened spatiotemporal attention maps, encouraging complementary coverage of action-bearing regions across tubelets and views.

The aggregated trajectory is converted to a contextualized state sequence $\boldsymbol{h}_{1:T}$. Q-ViSTA describes a transition at offset δ by pairing its endpoints with their difference, interaction, and temporal location:

$$
\boldsymbol{r}_t^{(\delta)} = \left[\boldsymbol{h}_t; \boldsymbol{h}_{t+\delta}; \boldsymbol{h}_{t+\delta} - \boldsymbol{h}_t; \boldsymbol{h}_t \odot \boldsymbol{h}_{t+\delta}; p_t^{(\delta)}\right]. \tag{5}
$$

The endpoints retain before-and-after context; their difference describes the direction of feature-space change, and their Hadamard product captures coordinate-wise interactions. Relative temporal position identifies the phase of the sampled sequence. Together, these components describe complementary aspects of change without requiring physical state annotations.

Separate encoders summarize local, intermediate, and longer-range changes at offsets $\{1, 2, 4\}$. Each scale scores its own Operator representation, $e_i^{(s)} = f_{\text{route}}(\boldsymbol{o}_i^{(s)})$. Let $\Delta e_i^{(s)} = e_i^{(s)} - e_i^{(1)}$ for $s \in \{2, 4\}$. The routing probabilities and bounded residual strength are

$$\boldsymbol{\pi}_i = \text{Softmax}\left(\frac{[0, \Delta e_i^{(2)}, \Delta e_i^{(4)}] + \boldsymbol{b}_{\text{prior}}}{\tau_{\text{scale}}}\right), \qquad \gamma = \gamma_{\max}\sigma(\gamma_{\text{logit}}). \tag{6}$$

The shortest offset anchors the mixture. Intermediate and longer offsets contribute bounded residual evidence, allowing broader temporal context to supplement local changes without replacing them:

$$\begin{aligned} w_i^{(1)} &= 1 - \gamma(\pi_i^{(2)} + \pi_i^{(4)}), \qquad & w_i^{(s)} &= \gamma\pi_i^{(s)},\ s \in \{2, 4\}, \\ \boldsymbol{o}_i &= \text{Norm}\left(\sum_{s\in\{1,2,4\}} w_i^{(s)}\boldsymbol{o}_i^{(s)}\right). \end{aligned} \tag{7}$$

The non-negative weights sum to one. With $\gamma_{\max} = 0.50$, the shortest offset retains at least half of the mixture. Routing uses only dynamic embeddings; semantic context enters at fusion. Appendix E.2 reports learned routing probabilities and effective mixture weights.

The action objective combines Operator supervision and cross-view consistency with state-change alignment and next-state reconstruction conditioned on the full-video Operator:

$$\begin{aligned} \mathcal{L}_{\text{operator}} = 0.25\mathcal{L}_{\text{op-view}} + 0.05\mathcal{L}_{\text{transition}} \\ + 0.50\mathcal{L}_{\text{dynamics}} + 0.10\mathcal{L}_{\text{op-sup}}. \end{aligned} \tag{8}$$

We define $\mathcal{L}_{\text{action}} = \mathcal{L}_{\text{operator}} + 0.01\mathcal{L}_{\text{query}}$. The dynamic term enforces full-video-conditioned consistency; it is not used as a causal future-prediction objective. Appendix C defines each term.

3.5 Semantic–dynamic fusion

Fusion combines the completed Semantic and Transition representations through a nonlinear head, preserving their distinct aggregation paths while learning a shared space for category discovery:

$$\boldsymbol{z} = \text{Norm}(F_{\text{fuse}}[\text{sg}(\boldsymbol{s}); \text{Norm}(W_O\boldsymbol{o})]), \tag{9}$$

where sg denotes stop-gradient. The nonlinear F1 head maps $512 \to 1024 \to 256$, with objective $\mathcal{L}_{\text{fused}} = \mathcal{L}_{\text{GCD}}(\boldsymbol{z})$. Detachment preserves a dedicated Semantic learning signal: fusion uses the embedding without redirecting its gradients. The Semantic head learns through its contrastive objective, while the dynamic and fusion modules also receive the fused discovery signal. All components participate in a single unified joint-learning process.

3.6 Joint objective and category discovery

The complete objective jointly learns the representations used for discovery. Including Query diversity in the action objective gives three complementary terms: one organizes global identity, one learns localized dynamics, and one shapes their fused category space:

$$\mathcal{L} = \mathcal{L}_{\text{semantic}} + \mathcal{L}_{\text{action}} + \mathcal{L}_{\text{fused}}. \tag{10}$$

At inference, we average the two views separately in Semantic and Operator space before applying F1. **Q-ViSTA** uses SS-KMeans with labeled assignments preserved (Vaze et al., 2022). **Q-ViSTA-Param** adds a jointly learned, weight-normalized linear classifier with one logit per category and a SimGCD-style classification objective (Wen et al., 2023). Both variants retain all three representation losses; classification additionally updates the Query, Operator, and fusion modules. Appendix C details the objective. Evaluation supplies the protocol's total category count to both variants. The four queries are reusable regional localizers, not class prototypes: they collect evidence before the embedding is assigned to categories. Known labels guide the discovery space without specifying where an interaction occurs. Consequently, changing the number of categories affects the assignment endpoint, but does not require a separate regional query for each known or novel category.

Table 1: **Human-action results (accuracy, %).** Q-ViSTA variants report mean ± sample standard deviation; best/second-best column means are **bold**/underlined.

Method	Venue	UCF101			SSv2			K400			Average		
		All	Old	New	All	Old	New	All	Old	New	All	Old	New
K-Means	–	20.34	16.84	22.09	6.00	6.21	5.90	5.91	6.33	5.70	10.75	9.79	11.23
SS-KMeans	–	21.70	22.52	21.28	6.37	6.75	6.18	5.98	6.44	5.74	11.35	11.90	11.07
GCD	CVPR'22	65.43	88.24	53.98	17.91	24.14	14.90	20.95	28.68	17.02	34.76	47.02	28.63
SimGCD	ICCV'23	68.90	80.80	62.93	20.26	29.96	15.58	21.01	29.91	16.49	36.72	46.89	31.67
CMS	CVPR'24	61.32	73.05	55.43	18.20	27.83	13.55	18.60	24.51	15.60	32.71	41.80	28.19
NC-GCD	NeurIPS'25	66.12	77.92	60.20	21.10	29.53	17.04	22.32	30.45	**18.20**	36.51	45.97	31.81
Q-ViSTA		68.72 ±0.86	**89.01** ±0.19	58.53 ±1.28	**24.79** ±0.28	**38.57** ±0.72	18.13 ±0.59	21.06 ±0.10	32.85 ±0.69	15.08 ±0.20	38.19 ±0.30	**53.48** ±0.49	30.58 ±0.22
Q-ViSTA-Param		**74.80** ±0.24	84.79 ±2.29	**69.79** ±1.45	24.22 ±0.86	33.34 ±1.49	**19.81** ±0.57	**24.01** ±0.17	**37.28** ±0.68	17.27 ±0.15	**41.01** ±0.31	51.80 ±0.53	**35.62** ±0.69

Table 2: **Robot-skill results (accuracy, %).** Q-ViSTA variants report mean ± sample standard deviation; best/second-best column means are **bold**/underlined.

Method	Venue	RoboCasa365 Atomic65			RoboCasa365 A→C			Mutex40			Average		
		All	Old	New	All	Old	New	All	Old	New	All	Old	New
K-Means	–	16.08	12.62	17.76	18.13	13.63	22.31	57.28	48.33	61.41	30.50	24.86	33.83
SS-KMeans	–	17.08	10.81	20.12	20.30	16.25	24.06	57.72	47.50	62.44	31.70	24.85	35.54
GCD	CVPR'22	61.24	75.31	54.42	69.11	79.20	59.74	70.35	76.11	67.69	66.90	76.87	60.62
SimGCD	ICCV'23	66.55	75.42	62.25	73.06	82.41	64.37	71.93	85.00	65.90	70.51	80.94	64.17
CMS	CVPR'24	54.24	67.69	47.73	60.61	75.29	46.97	70.18	68.06	**71.15**	61.68	70.35	55.28
NC-GCD	NeurIPS'25	61.90	72.38	56.82	71.72	79.82	64.20	69.65	76.67	66.41	67.76	76.29	62.48
Q-ViSTA		**73.41** ±0.61	**89.21** ±1.09	**65.75** ±1.13	80.56 ±0.62	92.14 ±0.69	69.81 ±1.23	**74.09** ±1.86	85.83 ±2.55	68.68 ±2.12	**76.02** ±0.75	**89.06** ±0.54	68.08 ±0.90
Q-ViSTA-Param		69.27 ±3.55	77.04 ±0.96	65.51 ±4.84	**83.96** ±0.71	**92.30** ±0.81	**76.21** ±0.63	71.35 ±3.29	**87.87** ±1.85	63.72 ±5.52	74.86 ±1.50	85.74 ±1.09	**68.48** ±2.03

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and protocols. We evaluate three human-action and three robot-skill protocols covering semantic, transition, and compositional evidence. UCF101 (Soomro et al., 2012) covers body motion and human–object actions, SSv2 (Goyal et al., 2017) emphasizes direction and state change, and K400 (Kay et al., 2017) broadens activity and scene diversity. RoboCasa365-Atomic65 (Nasiriany et al., 2026) uses 32 old and 33 novel classes. Atomic-to-Composite uses 65 old and 35 novel classes, following the pretrain100 list grouping. Mutex40 (Shah et al., 2023) evaluates real robot demonstrations. Old and novel categories coexist in each unlabeled pool, with half of the old-class videos labeled. Appendix B explains the grouping and exact sample counts used in each protocol.

Training and evaluation. All experiments run on NVIDIA RTX 3090 GPUs. Methods share the clips, two 16-frame views, frozen VideoMAEv2-Base encoder, labeled fraction, and category count (Wang et al., 2023). Q-ViSTA jointly trains its Semantic head, Action Queries, transition operator, and fusion head for 100 epochs with AdamW. All, Old, and New report accuracy on the complete unlabeled pool and its two partitions using one Hungarian mapping. Both Q-ViSTA variants report mean ± sample standard deviation over seeds 1, 2, and 3 using Known-K Final; Q-ViSTA-Param uses classifier-head accuracy. Baseline entries are final evaluations; RoboCasa SimGCD reports three-run classifier-head means. Appendix C details optimization, evaluation endpoints, and seed-level aggregation.

Evaluation questions. The experiments connect the representation to three questions: does the complete system improve discovery, does regional selection strengthen its observations, and does explicit transition encoding add value beyond temporal aggregation? Atomic-to-Composite tests

Table 3: **Domain averages (accuracy, %).** Protocols are equally weighted. Q-ViSTA variants report mean ± sample standard deviation. The last row reports Q-ViSTA-Param minus SimGCD in points.

Method	Venue	Human Avg.			Robot Avg.			Overall Avg.		
		All	Old	New	All	Old	New	All	Old	New
K-Means	–	10.75	9.79	11.23	30.50	24.86	33.83	20.62	17.33	22.53
SS-KMeans	–	11.35	11.90	11.07	31.70	24.85	35.54	21.52	18.38	23.30
GCD	CVPR'22	34.76	47.02	28.63	66.90	76.87	60.62	50.83	61.95	44.63
SimGCD	ICCV'23	36.72	46.89	31.67	70.51	80.94	64.17	53.62	63.92	47.92
CMS	CVPR'24	32.71	41.80	28.19	61.68	70.35	55.28	47.19	56.07	41.74
NC-GCD	NeurIPS'25	36.51	45.97	31.81	67.76	76.29	62.48	52.13	61.13	47.14
Q-ViSTA		38.19 ±0.30	**53.48** ±0.49	30.58 ±0.22	**76.02** ±0.75	**89.06** ±0.54	68.08 ±0.90	57.11 ±0.45	**71.27** ±0.43	49.33 ±0.48
Q-ViSTA-Param		**41.01** ±0.31	51.80 ±0.53	**35.62** ±0.69	74.86 ±1.50	85.74 ±1.09	**68.48** ±2.03	**57.93** ±0.63	68.77 ±0.68	**52.05** ±0.75
Impro. over SimGCD		**+4.29**	**+4.91**	**+3.95**	**+4.34**	**+4.79**	**+4.30**	**+4.32**	**+4.85**	**+4.13**

unlabeled multi-step categories alongside labeled atomic skills. Its novelty is defined at the task-category level; the protocol does not isolate appearance from action order or primitive recombination.

4.2 Comparison with State-of-the-Art Methods

Tables 1 and 2 compare Q-ViSTA with K-Means (Lloyd, 1982), SS-KMeans and GCD (Vaze et al., 2022), SimGCD (Wen et al., 2023), CMS (Choi et al., 2024), and NC-GCD (Han et al., 2025); Table 3 summarizes domain averages. Methods share videos, the frozen VideoMAEv2 backbone, and supervision, but retain their discovery objectives and trainable architectures. Localized patch access is part of Q-ViSTA's design. Thus the comparison tests complete discovery systems under the same observed videos.

Human-action datasets. On UCF101, Q-ViSTA-Param achieves 69.79% novel-action accuracy, compared with 62.93% for SimGCD, the strongest prior entry on this split. On SSv2, whose labels emphasize directions and state changes, non-parametric Q-ViSTA achieves the strongest joint accuracy, while the parametric variant leads novel-category discovery. Q-ViSTA-Param also leads joint accuracy on K400. These results span object-linked activities, interaction-centered labels, and diverse activity contexts, demonstrating the applicability of the dual representation across human actions with different sources of category evidence. The two branches need not contribute equally to every category: global context can distinguish an object-centered activity, while local state changes distinguish interactions with similar participants. Combining these cues gives discovery a common representation across these different labeling conventions.

Robot-skill datasets. Q-ViSTA shows its clearest advantage on composite robot tasks. On Atomic-to-Composite, Q-ViSTA-Param achieves 83.96% joint accuracy, compared with 73.06% for SimGCD, the strongest prior entry. Novel-category accuracy reaches 76.21%, demonstrating useful organization of the 35 composite categories alongside 65 familiar task categories. Non-parametric Q-ViSTA leads joint accuracy on Atomic65 and Mutex40. The results establish value for both individual skills and complete operations involving familiar objects and manipulators. Their defining cues motivate retaining semantic context together with regional change, with class partitions held fixed across all compared methods. For demonstration collections, the distinction is practical: individual manipulations and complete tasks both need meaningful category groupings. The two RoboCasa protocols evaluate these complementary levels of organization within the same source collection.

Summary. Both variants improve human and robot All averages over SimGCD; Q-ViSTA-Param gives the strongest overall average with a 4.32-point gain. Composite tasks provide the clearest application-level advantage. The following controls examine the regional selection and transition encoding that support this shared discovery representation.

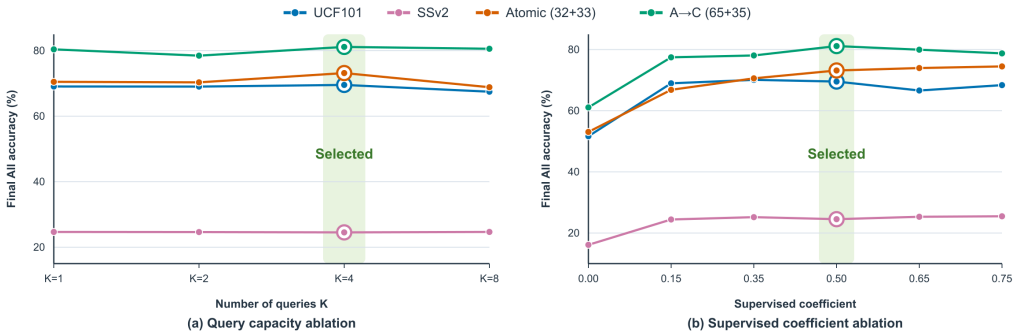


Figure 3: **Query capacity and supervision sensitivity.** Final All accuracy on UCF101, SSv2, Atomic (32+33), and A→C (65+35). The highlighted bands mark the shared settings, $K = 4$ and $\lambda = 0.50$; Appendix E reports the complete results for all evaluated protocols.

Table 4: **Core-component ablation (accuracy, %).** Sem., Gbl., Qry., and Multi. denote Semantic, global action pooling, Action Query, and multiscale description.

Components				UCF101			SSv2			Atomic65			Composite35		
Sem.	Gbl.	Qry.	Multi.	All	Old	New	All	Old	New	All	Old	New	All	Old	New
✓	×	×	×	63.59	89.26	50.71	19.06	29.86	13.85	65.80	81.62	58.12	70.13	80.46	60.54
✓	✓	×	×	66.88	89.09	55.74	21.09	35.93	13.93	67.59	87.75	57.82	77.46	87.71	67.94
✓	✓	×	✓	67.95	**90.52**	56.62	21.92	35.78	15.22	68.29	87.69	58.88	79.82	90.00	70.37
✓	×	✓	×	67.42	88.15	57.01	24.51	39.11	**17.46**	72.73	84.81	**66.88**	80.92	91.17	**71.40**
✓	×	✓	✓	**69.54**	88.93	**59.81**	**24.53**	**39.20**	17.45	**73.16**	**88.12**	65.91	**81.14**	**92.74**	70.37

4.3 Ablation of Core Components

Ablations use the main non-parametric configuration, with SS-KMeans applied to each embedding. Table 4 separates observation selection from temporal scale. Semantic remains active throughout; Gbl. denotes global action pooling. The single-scale setting encodes adjacent transitions, whereas the multiscale setting uses offsets of one, two, and four. Replacing global pooling with Action Query improves All accuracy on every protocol under either scale setting. On Atomic65, the single-scale contrast rises from 67.59% to 72.73% with transition construction fixed. Multiscale encoding further raises All accuracy to 73.16% with queries fixed. Regional selection provides the clearest component-level evidence. Auxiliary scales complement adjacent transitions, with effects that depend on the protocol and category partition.

A complementary control replaces explicit transition encoding with ordinary temporal aggregation. On Atomic65, explicit transition encoding achieves 73.16% All accuracy, compared with 70.08% for temporal attention, which already improves on mean pooling. The alternatives retain the same Query, state-encoder, multiscale, and fusion architectures, jointly optimized from a common initialization. The upstream state Transformer and Semantic branch remain active throughout. This within-framework control supports explicitly describing transitions beyond ordinary temporal attention. Appendix E.1 details the three encoders and their evaluation settings.

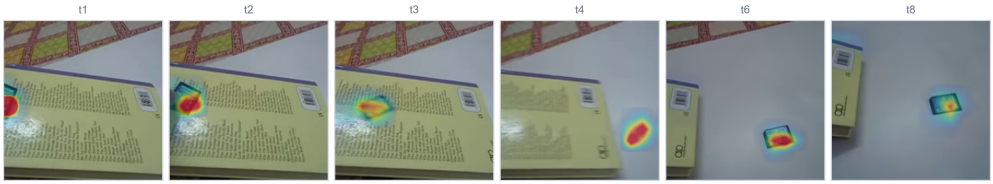
4.4 Query Capacity and Supervision Sensitivity

We examine query capacity and supervision using the non-parametric configuration. Four queries offer a compact shared setting, and one query already retains useful regional evidence. The UCF101 and SSv2 curves in Figure 3 remain comparatively stable. Four queries give the highest All accuracy on both RoboCasa protocols in the capacity sweep. Increasing the bank to eight does not consistently improve discovery, supporting a compact bank of reusable regional localizers.

Supervision improves all protocols relative to the zero-coefficient setting, with pronounced robot-skill gains. We retain the shared $K = 4$, $\lambda = 0.50$ defaults. Appendix E provides complete capacity and supervision results on all five protocols, including the old and novel partitions.



(a) Mutex40, novel: placing the granola bar in the top drawer.



(b) Something-Something V2, novel: placing a non-rollable object on a slope.



(c) UCF101, known: archery.

Figure 4: **Query-localized evidence across time.** One representative sequence per dataset; warmer colors indicate higher attention, with normalization shared within each video.

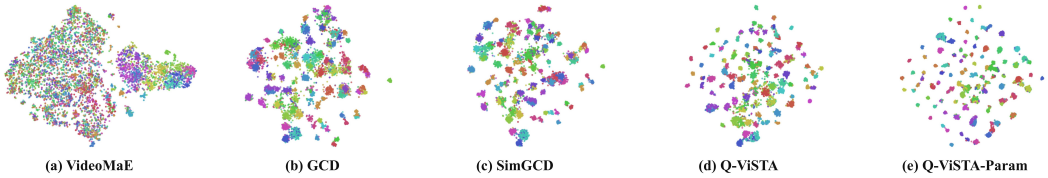


Figure 5: **Feature-space organization on Atomic-to-Composite.** Frozen VideoMAE, GCD, SimGCD, Q-ViSTA, and Q-ViSTA-Param use the same atomic and composite categories.

4.5 Qualitative visualization

Figure 4 shows Action Queries following manipulated objects in Mutex40 and SSv2 and the archer–bow interaction in UCF101. These sequences illustrate changing regional evidence alongside global context; Appendix H provides further known- and novel-category visualizations.

Figure 5 shows more distinct local groups for Q-ViSTA than the broad, overlapping structure of frozen VideoMAE, with particularly compact groups for Q-ViSTA-Param. This organization complements the composite-task results, illustrating how the dual representation organizes unfamiliar behaviors. Category discovery is evaluated in the original feature space, so the projection serves as a qualitative companion to the quantitative results on atomic and composite robot skills.

5 Conclusion

Q-ViSTA combines Semantic Identity with Query-Localized Transition Dynamics for GVSD. Its *where-before-how* design describes regional changes before fusion with global context. Experiments across six protocols support both clustering and parametric classification, with pronounced gains on unlabeled multi-step robot categories. Controlled pooling comparisons support regional observation selection. Together, these results establish a practical approach to organizing unfamiliar human actions and robot skills without region annotations, preserving both semantic context and localized interaction dynamics.

REFERENCES

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. ViViT: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning*, 2021.

Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. In *International Conference on Learning Representations*, 2022.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 2020.

João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4724–4733, 2017.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*, 2020.

Florent Chiaroni, Jose Dolz, Ziko Imtiaz Masud, Amar Mitiche, and Ismail Ben Ayed. Parametric information maximization for generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1729–1739, 2023.

Sua Choi, Dahyun Kang, and Minsu Cho. Contrastive mean-shift learning for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23094–23104, 2024.

Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002. doi: 10.1109/34.1000236.

Qiyuan Dai, Hanzhuo Huang, Yu Wu, and Sibei Yang. Adaptive part learning for fine-grained generalized category discovery: A plug-and-play enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25444–25453, 2025.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9284–9292, 2021.

Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2): 155–170, 1983. doi: 10.1207/s15516709cog0702_3.

Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 244–253, 2019.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "Something Something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

Jizhou Han, Shaokun Wang, Yuhang He, Chenhao Ding, Qiang Wang, Xinyuan Gao, SongLin Dong, and Yihong Gong. Consistent supervised-unsupervised alignment for generalized category discovery. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Jizhou Han, Chenhao Ding, SongLin Dong, Xinyuan Gao, Qiang Wang, Yuhang He, and Yihong Gong. Unleashing the potential of all test samples: Mean-shift guided test-time adaptation. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(8):11323–11335, 2026a. doi: 10.1109/TCSVT.2026.3676383.

Jizhou Han, Chenhao Ding, SongLin Dong, Yuhang He, Shaokun Wang, Qiang Wang, and Yihong Gong. GOAL: Geometrically optimal alignment for continual generalized category discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4565–4573, 2026b. doi: 10.1609/aaai.v40i6.42456.

Jizhou Han, Chenhao Ding, Yuhang He, SongLin Dong, Qiang Wang, Xinyuan Gao, and Yihong Gong. Learn by reasoning: Analogical weight generation for few-shot class-incremental learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(4):4842–4857, 2026c. doi: 10.1109/TCSVT.2025.3637903.

Jizhou Han, Chenhao Ding, Yuhang He, Qiang Wang, Shaokun Wang, SongLin Dong, and Yihong Gong. Learning like humans: Analogical concept learning for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28734–28743, 2026d.

Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8400–8408, 2019.

Kai Han, Sylvestre-Alvise Rebuffi, Sébastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *International Conference on Learning Representations*, 2020.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, 2020.

Ji Lin, Chuang Gan, and Song Han. TSM: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7083–7093, 2019.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video Swin Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3202–3211, 2022.

Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28 (2):129–137, 1982. doi: 10.1109/TIT.1982.1056489.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, 2020.

Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixé, and Christoph Feichtenhofer. TrackFormer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8844–8854, 2022.

Soroush Nasiriany, Sepehr Nasiriany, Abhiram Maddukuri, and Yuke Zhu. RoboCasa365: A large-scale simulation framework for training and benchmarking generalist robots. *arXiv preprint arXiv:2603.04356*, 2026.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020. doi: 10.1073/pnas.2015509117.

Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptional contrastive learning for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7579–7588, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Sarah Rastegar, Hazel Doughty, and Cees G. M. Snoek. Learn to categorize or categorize to learn? self-coding for generalized category discovery. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Sarah Rastegar, Mohammadreza Salehi, Yuki M. Asano, Hazel Doughty, and Cees G. M. Snoek. SelEx: Self-expertise in fine-grained generalized category discovery. In *European Conference on Computer Vision*, pp. 440–458, 2024.

Rutav Shah, Roberto Martín-Martín, and Yuke Zhu. MUTEX: Learning unified policies from multimodal task specifications. In *Conference on Robot Learning*, 2023.

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Enguang Wang, Zhimao Peng, Zhengyuan Xie, Fei Yang, Xialei Liu, and Ming-Ming Cheng. GET: Unlocking the multi-modal potential of CLIP for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20296–20306, 2025.

Hongjun Wang, Sagar Vaze, and Kai Han. SPTNet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. In *International Conference on Learning Representations*, 2024.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. VideoMAE V2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Shaokun Wang, Weili Guan, Jizhou Han, Jianlong Wu, Yupeng Hu, and Liqiang Nie. StructAlign: Structured cross-modal alignment for continual text-to-video retrieval. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1800–1811, 2026. doi: 10.1145/3805712.3809704.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, 2018.

Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Temporal query networks for fine-grained video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4486–4496, 2021.

Jing Zhang, Nan Pu, Yuxiang Xie, Yanming Guo, Qianqi Lu, Shiwei Zou, Jie Yan, and Yan Chen. Video-based generalized category discovery via memory-guided consistency-aware contrastive learning. *arXiv preprint arXiv:2509.06306*, 2025.

Sheng Zhang, Salman Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Shahbaz Khan. PromptCAL: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3479–3488, 2023.

Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised gaussian mixture models for generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16623–16633, 2023.

Jiaojiao Zhao, Yanyi Zhang, Xinyu Li, Hao Chen, Bing Shuai, Mingze Xu, Chunhui Liu, Kaustav Kundu, Yuanjun Xiong, Davide Modolo, Ivan Marsic, Cees G. M. Snoek, and Joseph Tighe. TubeR: Tubelet transformer for video action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13598–13607, 2022.

Haiyang Zheng, Nan Pu, Wenjing Li, Nicu Sebe, and Zhun Zhong. Textual knowledge matters: Cross-modality co-teaching for generalized visual class discovery. In *European Conference on Computer Vision*, 2024.

Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *European Conference on Computer Vision*, 2018.

Appendix

A Reproducibility Statement

Section 3 specifies the model and learning objectives, and Section 4 describes the experimental protocols. Appendix B documents the datasets and category partitions, while Appendix C provides implementation details, gradient paths, optimization settings, and evaluation conventions. Appendix E supplies additional ablation results and their experimental settings. The supplementary materials include the Q-ViSTA source code, environment requirements, sample-level split records, and instructions for data preparation, feature extraction, training, evaluation, and result aggregation. An accompanying video illustrates Action Query behavior. These materials provide methodological and implementation references for reproducing and inspecting the main method.

B Dataset Details and Category Semantics

The GVSD protocol suite spans human activities, simulated manipulation, and real robot demonstrations. Each protocol defines one mixed discovery pool with labeled examples from old categories and unlabeled examples from both partitions. Table 5 reports the evaluated pool sizes: old-class totals include labeled and unlabeled videos, whereas every novel-class video is unlabeled. Tables 6 and 7 provide representative labels to clarify what distinguishes categories in each domain. The descriptions below relate these labels to the identity and transition evidence available in the videos.

B.1 Human-action protocols

UCF101. UCF101 contains 13,320 videos from 101 categories spanning sports, exercise, body motion, grooming, and human–object interaction. The split assigns 51 classes and 6,662 videos to the old partition, of which 3,319 videos are labeled and 3,343 remain unlabeled; the other 50 classes contribute 6,658 novel videos. Its diverse actors, viewpoints, objects, and scenes make semantic identity informative, while distinctions such as closely related ball sports and object-centered actions still require temporal evidence to separate visually related operations.

Something-Something V2. Our class-stratified 15% SSv2 subset contains 25,335 videos from 174 interaction categories. The 87 old classes contain 12,422 videos (6,187 labeled and 6,235 unlabeled), while the 87 novel classes contain 12,913 videos. Unlike datasets dominated by persistent appearance, SSv2 labels encode direction, destination, success, state change, and sometimes intent. Consequently, the same objects can support different categories depending on how their states evolve.

Kinetics-400. The class-stratified 15% K400 subset contains 36,178 videos from 400 activities. Its 200 old classes provide 18,190 videos (9,054 labeled and 9,136 unlabeled), and its 200 novel classes provide 17,988 videos. K400 introduces substantial variation in scene, viewpoint, actor, and recording style, complementing SSv2's transition-centric labels with broader activity contexts.

Category semantics across human-action protocols. Representative class names reveal why these datasets are complementary. UCF101 includes object-linked activities such as Archery alongside body-centered motion such as Baby Crawling. SSv2 uses object-agnostic templates: dropping an item into another and pushing it from right to left are distinguished by an interaction and its outcome. K400 covers a wider range of activity contexts, from cooking to sports. Consequently, the same discovery representation must accommodate labels supported by persistent appearance and labels that depend more directly on changing spatial relations. The old and novel examples in Table 6 illustrate the interaction and context cues defining these categories.

Class partitions and available supervision. Old and novel denote label availability within each protocol, not different visual domains. Labeled old-class examples anchor discovery, but the remaining old examples are evaluated together with novel videos in one unlabeled pool. A method must therefore organize both partitions without receiving a test video's membership. The equal numbers

of old and novel classes in SSv2 and K400 do not imply equal numbers of evaluated videos, since labeled old examples support learning and are excluded from all reported evaluation denominators.

Connection to robot-skill discovery. Human-action protocols vary the role of appearance and temporal evidence within activity recognition. The robot protocols add a complementary distinction in category structure: atomic manipulation, held-out multi-step compositions, and real object–destination tasks. We evaluate the same architecture and operating configuration separately on each protocol. This tests its applicability across these settings without assuming a model trained on one dataset is directly transferred to another. Atomic-to-Composite then makes the connection between familiar primitives and unfamiliar coordinated behaviors explicit in the old and novel class definitions.

B.2 Robot-skill protocols

RoboCasa365 task grouping. We construct the RoboCasa protocols from human demonstrations in the official pretrain100 task list (Nasiriany et al., 2026), retaining 100 demonstrations per task. Our 65/35 grouping follows the list's annotated sections: 65 atomic tasks and the remaining 35 tasks, comprising 16 seen-composite and 19 additional tasks. This grouping is a protocol choice, distinct from the task registry's storage taxonomy. In particular, *OrganizeMugsByHandle* appears in the list's atomic section but is registered and stored as a composite task. We include it in the 65-task atomic group for these experiments without changing its source videos, task identifier, or storage provenance. All 100 tasks remain in the same source collection.

RoboCasa365-Atomic65. This protocol contains 6,500 simulated kitchen demonstrations from the selected 65-task atomic group. The 32 old classes contribute 3,200 videos, evenly divided into 1,600 labeled and 1,600 unlabeled examples; the 33 novel classes contribute 3,300 videos. Thus accuracy is evaluated on 4,900 unlabeled videos. The old class IDs are $0, 2, \ldots, 62$; the novel IDs are $1, 3, \ldots, 63$ together with ID 64, *OrganizeMugsByHandle*. Many classes share a kitchen layout and end-effector but differ in the manipulated appliance, object, or target state, providing related visual contexts with distinct interaction and task labels.

RoboCasa365 Atomic-to-Composite. The compositional protocol contains 10,000 demonstrations from the same 100-task collection. All 65 tasks in the atomic group form the old partition, with 6,500 videos split into 3,250 labeled and 3,250 unlabeled examples. The remaining 35 composite tasks form the novel partition, with 3,500 unlabeled videos. The evaluated pool therefore contains 6,750 videos. Old IDs are $0, \ldots, 64$, and novel IDs are $65, \ldots, 99$. In contrast to Atomic65, *OrganizeMugsByHandle* is an old category here. The protocols are trained separately and differ in label visibility, not in that task's source demonstrations.

Our Old/New distinction is defined by category-label availability in discovery training, not by the official list's "seen composite" annotation. The 16 tasks under that heading therefore remain novel in A→C: their videos participate in unlabeled training, but their category labels are unavailable. These protocols use a subset of RoboCasa365 rather than its complete task vocabulary. Composite task labels define the discovery target under the selected object and scene coverage; the experiment does not hold appearance constant or require a verified atomic decomposition of every task.

Mutex40. Mutex40 contains 1,500 real-world robot demonstrations from 40 manipulation tasks. The 20 old classes contain 720 videos (360 labeled and 360 unlabeled), and the 20 novel classes contain 780 videos. Its labels jointly specify manipulated objects, destinations, and interaction sequences; repeated objects such as granola bars, bowls, and mugs appear across distinct tasks, directly exposing the complementary identity and transition evidence studied in the main paper.

Category semantics across robot-skill protocols. Atomic65 evaluates the 32-old/33-new partition of the 65-task atomic group. Composite35 denotes the 35 novel categories in the 65-old/35-new atomic-to-composite protocol; its All accuracy still covers the complete mixed unlabeled pool, not only those novel categories. Mutex40 extends evaluation to real demonstrations with repeated objects and destinations. These names specify category partitions, not different training architectures. Table 7 illustrates representative task labels within the resulting old and novel partitions.

Table 5: **Dataset statistics and GCD splits.** "Old videos" includes both labeled and unlabeled samples from old classes, whereas all new-class videos are unlabeled. Counts describe the discovery pools evaluated consistently across all experiments and reproduced baselines.

Dataset	Type	Classes			Videos		
		All	Old	New	All	Old	New
UCF101	Human action	101	51	50	13,320	6,662	6,658
SSv2 (15%)	Human action	174	87	87	25,335	12,422	12,913
K400 (15%)	Human action	400	200	200	36,178	18,190	17,988
RoboCasa365-Atomic65	Robot skill	65	32	33	6,500	3,200	3,300
RoboCasa365-A→C	Robot skill	100	65	35	10,000	6,500	3,500
Mutex40	Robot skill	40	20	20	1,500	720	780

Table 6: **Human-action class examples by partition.** Each column lists twelve names from the evaluated labels. Every class starts a separate line, with wrapped text continuing only that class.

Dataset	Old/known examples	New/novel examples
UCF101	ApplyEyeMakeup	ApplyLipstick
	Archery	BabyCrawling
	BalanceBeam	BandMarching
	BaseballPitch	Basketball
	BasketballDunk	BenchPress
	Biking	Billiards
	BlowDryHair	BlowingCandles
	BodyWeightSquats	Bowling
	BoxingPunchingBag	BoxingSpeedBag
	BreastStroke	BrushingTeeth
	CleanAndJerk	CliffDiving
	CricketBowling	CricketShot
SSv2	Pushing something from right to left	Putting something that can't roll onto a slanted surface, so it slides down
	Dropping something into something	Pretending to scoop something up with something
	Putting something onto something	Pushing something so that it falls off the table
	Dropping something onto something	Pretending to throw something
	Putting something behind something	Uncovering something
	Putting number of something onto something	Moving part of something
	Holding something	Putting something in front of something
	Holding something in front of something	Putting something on a surface
	Showing a photo of something to the camera	Dropping something next to something
	Lifting something up completely, then letting it drop down	Showing that something is empty
	Unfolding something	Holding something next to something
	Putting something similar to other things that are already on the table	Pretending to put something behind something
K400	eating spaghetti	brushing teeth
	hurdling	sword fighting
	changing oil	golf chipping
	playing paintball	yawning
	snowboarding	eating watermelon
	giving or receiving award	cooking egg
	dancing ballet	peeling potatoes
	passing American football (in game)	skipping rope
	golf driving	dining
	shooting basketball	making jewelry
	singing	cartwheeling
	punching bag	playing accordion

Table 7: **Robot-skill class examples by partition.** Each column lists twelve names from the evaluated labels. Every class starts a separate line, with wrapped text continuing only that class.

Dataset	Old/known examples	New/novel examples
RoboCasa365-Atomic65	AdjustToasterOvenTemperature CheesyBread CloseBlenderLid CloseDishwasher CloseElectricKettleLid CloseFridgeDrawer CloseOven CloseToasterOvenDoor CoffeeSetupMug OpenBlenderLid OpenDishwasher OpenElectricKettleLid	AdjustWaterTemperature CloseCabinet CloseDrawer CloseFridge CloseMicrowave CloseStandMixerHead CoffeeServeMug LowerHeat MakeIcedCoffee NavigateKitchen OpenCabinet OpenDrawer
RoboCasa365 A→C	AdjustToasterOvenTemperature AdjustWaterTemperature CheesyBread CloseBlenderLid CloseCabinet CloseDishwasher CloseDrawer CloseElectricKettleLid CloseFridge CloseFridgeDrawer CloseMicrowave CloseOven	AddSweetener ArrangeBuffetDessert CleanBoard CleanMicrowave CoolBakedCake DateNight DeliverStraw FilterMicrowavableItem GetToastedBread KettleBoiling LoadDishwasher MeatSkewerAssembly
Mutex40	put_book_in_caddy_back put_granola_bar_in_caddy_back open_air_fryer put_pink_mug_on_white_plate put_granola_bar_in_basket pull_out_oven_tray put_blue_mug_on_oven_tray put_red_cup_in_caddy_back open_bottom_drawer put_yellow_bowl_on_oven_tray_push_in put_pink_mug_in_basket put_red_bowl_in_top_drawer_close_drawer	put_granola_bar_in_top_drawer open_air_fryer_put_blue_bowl_inside put_red_bowl_in_basket open_top_drawer_put_granola_bar_inside put_yellow_bowl_on_white_plate put_mac_cheese_box_in_basket put_bread_on_white_plate push_oven_tray_in put_hotdog_bowl_on_white_plate pull_oven_tray_put_red_bowl_on_it put_pink_mug_in_top_drawer_close_drawer put_blue_mug_on_white_plate

C Implementation and Reproducibility

Video views and patch cache. Each video is decoded once and sampled into two deterministic 16-frame views. The first uses linearly spaced frame indices from the beginning to the end of the decoded clip; the second partitions that frame sequence into 16 equal temporal bins and selects one seeded frame per bin while preserving order. Neither Q-ViSTA nor the compared methods receives video duration or timestamp metadata as an input feature. The first view uses a centered crop and the second a seeded crop after resizing the short side to 256 pixels; both are cropped to 224×224, normalized with ImageNet statistics, and never horizontally flipped because a flip can reverse direction-sensitive actions. We use the OpenGVLab VideoMAEv2-Base checkpoint pretrained self-supervised for 800 epochs on UnlabeledHybrid-1M (Wang et al., 2023); the encoder remains frozen for every compared method. With tubelet size two, it produces eight temporal positions and 14×14 patches, so the cache has shape $[N, 2, 8, 196, 768]$. The Semantic branch receives this cache's exact spatial–temporal mean; the transition branch instead traces local interactions directly from the patch tokens.

Splits. For human-action datasets and Mutex40, even-indexed classes are old and odd-indexed classes are novel. Atomic65 retains the 32 even-indexed old classes among IDs 0–63 and assigns

the other 33 classes, including ID 64, to the novel partition. Atomic-to-Composite uses all 65 atomic-group tasks as old and the remaining 35 tasks as novel. Within old categories, 50% of the selected videos are labeled. SSv2 and K400 use fixed class-stratified 15% subsets, and only successfully decoded K400 clips enter the pool. Sample selections, targets, labeled indicators, and class partitions remain fixed across compared methods and training seeds. Anonymized protocol records enumerate class order, sample identifiers, source labels, and labeled masks. For RoboCasa, original storage provenance is retained separately from the experimental 65/35 grouping, with 100 human demonstrations retained for each selected task.

Optimization. All experiments use NVIDIA RTX 3090 GPUs. All trainable components—the two-layer Semantic head, recurrent query generator, per-tubelet All-4 Query aggregator, multiscale transition operator, scale router, and nonlinear F1 fusion—are optimized jointly for 100 epochs. Four Action Queries, $\lambda = 0.50$, the classifier coefficient $\alpha = 0.35$, and the main loss weights are fixed in advance, rather than selected for individual evaluation pools. The sensitivity sweeps characterize these preset defaults; they do not select per-dataset operating points. InfoNCE and supervised-contrastive temperatures are 0.10 and 0.07. We use AdamW with batch size 256, weight decay 10^{-4}, and a fixed learning rate of 3×10^{-5}. Evaluation occurs every ten epochs; reported results use the final checkpoint from each of the three training seeds.

C.1 Tensor-level computation graph

For each view, the frozen cache provides $X \in \mathbb{R}^{8\times196\times768}$. The Semantic branch averages X over space and time, normalizes this mean, and applies a $768 \to 2048 \to 256$ GELU projection followed by normalization. The Action Query initializer concatenates mean appearance with the mean absolute adjacent-tubelet difference, giving a 1536-dimensional video descriptor; a $1536 \to 128 \to 4 \times 256$ MLP produces the initial queries. At every tubelet, queries attend to 256-dimensional projected keys but pool the original 768-dimensional values. A LayerNorm–linear map projects the pooled observation to 256 dimensions before the GRUCell update. The global update scalar is initialized to 0.10 and interpolates between the previous and candidate states before normalization. A learned scorer and softmax over all four regional observations then form one All-4 weighted observation per tubelet. Thus the complete Query bank remains available, with contributions adapting to the local action state in each view.

The aggregated 768-dimensional trajectory passes through a $768 \to 256 \to 256$ state projection, temporal position embeddings, and a one-layer, eight-head Transformer with 512-dimensional feed-forward blocks. For each offset $s \in \{1, 2, 4\}$, the transition token concatenates the two endpoint states, their difference, their Hadamard interaction, and normalized time position. Three non-shared one-layer scale encoders and learned temporal attention pooling produce one 256-dimensional Operator per offset. Each Operator predicts its own routing score. The routing temperature follows a cosine schedule from 1.0 to 0.5 over 30 epochs. The fixed bias is $\boldsymbol{b}_{\text{prior}} = \log[0.90, 0.05, 0.05]$, and the router's final linear layer is zero-initialized. The learned residual strength starts at 0.05 with $\gamma_{\max} = 0.50$, keeping the shortest-scale weight non-negative. Semantic features are not inputs to this router. The dynamics MLP has hidden width 512. Model, Semantic, and action-head dropout are 0.1, while Query aggregation and F1 dropout are zero. In Eq. 9, W_O is a learned, bias-free $256 \to 256$ map initialized to identity. We concatenate its normalized output with the normalized Semantic embedding and apply F1, aligning both paths in the final discovery representation.

C.2 Training objectives and gradient paths

For clarity, let $v \in \{1, 2\}$ index the two views, $\boldsymbol{o}_i^v$ denote the normalized Operator embedding, and $\boldsymbol{h}_{i,t}^v$ the contextualized state at tubelet t. The Operator-supervision term $\mathcal{L}_{\text{op-sup}}$ is exactly supervised contrastive loss on $\{\boldsymbol{o}_i^v : i \in \mathcal{D}_l\}$. Cross-view stability is

$$\mathcal{L}_{\text{op-view}} = \frac{1}{B}\sum_{i=1}^{B}\left[1 - \cos(\boldsymbol{o}_i^1, \boldsymbol{o}_i^2)\right]. \tag{11}$$

For adjacent normalized changes $\Delta \boldsymbol{h}^v_{i,t} = \mathrm{Norm}(\boldsymbol{h}^v_{i,t+1} - \boldsymbol{h}^v_{i,t})$, the transition term measures symmetric view agreement with a detached target in each direction:

$$\mathcal{L}_{\text{transition}} = \frac{1}{2B(T-1)} \sum_{i,t} \left[2 - \cos(\Delta \boldsymbol{h}^1_{i,t}, \mathrm{sg}(\Delta \boldsymbol{h}^2_{i,t})) - \cos(\mathrm{sg}(\Delta \boldsymbol{h}^1_{i,t}), \Delta \boldsymbol{h}^2_{i,t})\right]. \tag{12}$$

The two views span the same clip, so this objective aligns changes at corresponding relative positions, not verified frame-level events. Their sampled frames may differ within a temporal bin; the loss is a weak consistency constraint rather than an explicit temporal-registration procedure. With the normalization in Eq. 12, symmetric detachment preserves the ordinary cosine loss value while halving its gradient. It specifies the weighting of cross-view agreement, not a separate guarantee of optimization stability. The dynamics head receives the current state, the full-video Operator, and temporal position, and reconstructs the next normalized state under both augmented views:

$$\mathcal{L}_{\text{dynamics}} = \frac{1}{2B(T-1)} \sum_{i,v,t} \left[1 - \cos\big(D(\boldsymbol{h}^v_{i,t}, \boldsymbol{o}^v_i, p_t), \mathrm{sg}(\boldsymbol{h}^v_{i,t+1})\big)\right]. \tag{13}$$

Here D predicts a residual from the current state, Operator, and time position, adds it to the current state, and normalizes the sum. Because the Operator summarizes the complete clip, this term is a full-video-conditioned consistency objective rather than evidence of causal forecasting. It encourages the state and video representations to support a common reconstruction task. For Query diversity, let $\widehat{\boldsymbol{A}}_{i,k}$ be the normalized vectorization of Query k's complete spatiotemporal attention map in one view. We compute the following overlap separately for each view and average the two resulting losses:

$$\mathcal{L}_{\text{query}} = \frac{1}{BK(K-1)} \sum_{i} \sum_{k \neq l} \widehat{\boldsymbol{A}}^{\top}_{i,k} \widehat{\boldsymbol{A}}_{i,l}. \tag{14}$$

For the single-query ablation, the pair set is empty and $\mathcal{L}_{\text{query}} = 0$; Eq. 14 applies for $K > 1$. The action objective is therefore $\mathcal{L}_{\text{action}} = 0.25\mathcal{L}_{\text{op-view}} + 0.05\mathcal{L}_{\text{transition}} + 0.50\mathcal{L}_{\text{dynamics}} + 0.10\mathcal{L}_{\text{op-sup}} + 0.01\mathcal{L}_{\text{query}}$. The stop-gradient in Eq. 9 blocks the fused and classifier paths through the concatenated Semantic embedding. Because scale routing depends only on Operator representations, the Semantic head is trained directly by $\mathcal{L}_{\text{semantic}}$; the Query, transition, routing, and fusion modules receive their corresponding gradients within the same joint optimization.

Discovery endpoints. Q-ViSTA and Q-ViSTA-Param use the same dual-representation architecture and representation losses. Q-ViSTA has no classification head and applies fixed-K SS-KMeans to the final fused embedding. Q-ViSTA-Param adds one weight-normalized linear classifier after that embedding, with one output for every protocol category. Since its classification gradients update the Query, Operator, and fusion modules, the parameterized endpoint is a jointly optimized architectural variant rather than post-hoc relabeling of an identical embedding. For view logits $\boldsymbol{\ell}^1, \boldsymbol{\ell}^2$, it uses the following SimGCD-style classification objective over both views:

$$\mathcal{L}_{\text{cls}} = \alpha \mathcal{L}_{\text{CE}} + (1-\alpha)\big(\mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{ME}}\big), \qquad \mathcal{L}_{\text{Param}} = \mathcal{L} + \mathcal{L}_{\text{cls}}, \tag{15}$$

with $\alpha = 0.35$, fixed for every protocol. Cross-entropy averages the two labeled views. $\mathcal{L}_{\text{dist}}$ is symmetric cross-view distillation: each view predicts the detached soft distribution of the other view. Student temperature is 0.10; teacher temperature decreases from 0.07 to 0.04 over the first 30 epochs. $\mathcal{L}_{\text{ME}}$ is the KL divergence from the batch-mean prediction to the uniform category distribution. The classifier is learned with every other trainable component in the same joint optimization and predicts by logit argmax. Thus the head is the only structural addition, while its objective reshapes the shared fused representation and leaves all three representation losses active.

Controlled baseline reproduction. K-Means and SS-KMeans are applied directly to the globally pooled frozen VideoMAE views. Trainable reproduced baselines share the two-view sampling and frozen patch-token cache, retaining their discovery objectives and prototype or classifier parameterizations. For the reproduced SimGCD baseline, we report classifier-head accuracy from the final checkpoint after 200 training epochs. On Atomic65 and A→C65/35, these results are averaged over seeds 1, 2, and 3. Other baseline entries retain their single final evaluations. Q-ViSTA variants report 100-epoch final results. The comparison controls observations and backbone adaptation, not identical optimization budgets or trainable capacity, and compares complete learning and assignment systems.

Table 8: **Model cost and recognition accuracy including VideoMAE.** All accuracy (%) uses seed 1 Known-K Final; bold indicates the higher accuracy for each protocol.

Method	UCF101 All ↑	RoboCasa Atomic65 All ↑	Mutex40 All ↑	Total params (M)	Trainable params (M)	Total dense GFLOPs /video
VideoMAE + Global Pooling + Semantic MLP	63.59	65.80	73.42	88.33	2.10	721.38
VideoMAE + Q-ViSTA	**69.54**	**73.16**	**76.23**	94.77	8.54	722.85

Computational costs are calculated for two views per video, each with 16 frames at 224×224 resolution, using 1 MAC = 2 FLOPs. Counts include the frozen encoder but exclude the pretraining reconstruction decoder, video decoding, disk reads, and clustering.

Evaluation and reporting. SS-KMeans uses the supplied category count, distinct from Query capacity, and preserves labeled assignments. Old-class centers start from their labeled members; novel centers use K-Means++. We select the minimum-inertia solution from ten seeded restarts, each capped at ten assignment–update iterations. Both endpoints use one unrestricted Hungarian mapping fitted on the complete unlabeled pool, allowing old and novel output indices to permute. All, Old, and New share this mapping, with partitions defined by ground-truth class membership. Labeled videos guide training and constrained clustering but do not enter accuracy denominators. Within each protocol, All weights Old and New by their unlabeled sample counts, not class counts. Evaluation-only matching aligns indices without changing predictions or supplying novel labels to training. This pool-level sample weighting differs from the equal-protocol averages below.

Known-K Final evaluates the final checkpoints for seeds 1, 2, and 3, using classifier-head accuracy for Q-ViSTA-Param. We report mean accuracy $\pm$ sample standard deviation with denominator $n - 1 = 2$ (`ddof=1`). Accuracy is in percent and standard deviation in points. Human and Robot averages weight their three datasets equally within each seed; Overall averages all six. Reported variability is computed from those three seed-level averages, not by averaging dataset standard deviations. Improvement entries are absolute point differences rather than relative percentages. Non-parametric Q-ViSTA retains representation losses and SS-KMeans; only Q-ViSTA-Param adds classification with its preset objective weight.

D Efficiency Analysis

Table 8 compares the pooled Semantic baseline with non-parametric Q-ViSTA, including the same frozen VideoMAE encoder in both systems. Q-ViSTA adds 6.44 M total parameters (7.29%) and a small additional dense computation of 1.47 GFLOPs per video (0.20%). The displayed All accuracies improve by 5.95, 7.36, and 2.81 points on UCF101, Atomic65, and Mutex40, respectively. These gains span human actions, simulated manipulation, and real robot skill discovery.

These figures describe a complete two-view dense forward pass with the frozen encoder included. The trainable heads contain 2.10 M and 8.54 M parameters, respectively; the common backbone dominates the full-model FLOPs. Training reads cached features, so the reported comparison concerns complete-model arithmetic rather than a matched head-training budget.

E Additional Ablation Studies

Tables 9 and 10 expand the main-text sensitivity curves with All, Old, and New accuracy on five protocols. Every ablation uses non-parametric Q-ViSTA: the active branches retain their representation losses, SS-KMeans assigns categories, and no classifier or $\mathcal{L}_{\text{cls}}$ is present. The default configuration is the same four-query, $\lambda = 0.50$ model used for non-parametric main results and rejection experiments. Ablations report seed 1 by default; main and rejection results summarize seeds 1, 2, and 3, all evaluated from the final checkpoint with the protocol category count supplied.

In the core-component study, the first row is Semantic-only. Gbl. supplies spatially averaged action observations and Qry. supplies query-weighted observations. The single-scale setting retains transitions between adjacent states, while the multiscale setting uses temporal offsets of one, two, and

Table 9: **Action Query capacity (accuracy, %).** Final-checkpoint All/Old/New results are reported for one seed on five protocols. The shared $K = 4$ configuration is shaded.

Configuration	UCF101			SSv2			Mutex40			Atomic65			Composite35		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
$K = 1$ All-1	69.06	**89.32**	58.89	**24.67**	38.33	18.07	74.12	78.61	72.05	70.49	87.06	62.45	80.40	91.78	69.83
$K = 2$ All-2	69.03	87.97	59.52	24.63	37.08	**18.62**	72.72	78.61	70.00	70.33	86.12	62.67	78.47	91.38	66.49
$K = 4$ **All-4**	**69.54**	88.93	**59.81**	24.53	**39.20**	17.45	76.23	**88.06**	70.77	**73.16**	**88.12**	**65.91**	**81.14**	**92.74**	**70.37**
$K = 8$ All-8	67.47	89.17	56.58	**24.67**	37.72	18.37	**76.58**	86.11	**72.18**	68.84	85.06	60.97	80.58	92.31	69.69

Table 10: **Supervised-coefficient sensitivity (accuracy, %).** Final-checkpoint All/Old/New results are reported for one seed on five protocols. The shared $\lambda = 0.50$ setting is shaded.

λ	UCF101			SSv2			Mutex40			Atomic65			Composite35		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
0.00	51.65	68.74	43.08	16.14	24.52	12.09	59.30	56.67	60.51	53.06	62.56	48.45	61.08	76.98	46.31
0.15	68.96	91.09	57.86	24.43	39.55	17.13	71.23	67.50	72.95	66.84	89.19	56.00	77.47	91.69	64.26
0.35	**70.09**	**91.33**	59.43	25.19	38.96	**18.55**	75.00	81.11	72.18	70.59	87.69	62.30	78.07	91.35	65.74
0.50	69.54	88.93	**59.81**	24.53	39.20	17.45	76.23	**88.06**	70.77	73.16	88.12	65.91	**81.14**	92.74	**70.37**
0.65	66.61	87.50	56.13	25.32	40.00	18.23	74.21	81.11	71.03	73.96	88.00	**67.15**	79.96	**92.86**	67.97
0.75	68.37	87.29	58.88	**25.48**	**41.06**	17.96	**78.77**	85.28	**75.77**	**74.49**	**92.19**	65.91	78.76	92.28	66.20

four. Rows two versus four and three versus five compare selection at fixed temporal scales. Rows two versus three and four versus five compare scales at a fixed selection rule. These seed-1 contrasts describe the observed component effects, whereas the complete-system tables estimate repeated-run performance. They are not significance tests across different component configurations.

Each sweep changes one operating parameter. Query capacity varies with $\lambda = 0.50$ fixed; the supervision coefficient varies with four queries fixed. The swept coefficient is λ in the Semantic and fused $\mathcal{L}_{\mathrm{GCD}}$ objectives, distinct from the parametric classifier's α. The shared reference in both tables is therefore an individual run of the main non-parametric configuration, not a different architecture or a three-seed average. This common reference makes the sensitivity trends interpretable alongside the component ablation without conflating them with the separately trained parametric variant.

Query capacity. Query count controls the number of regional trajectories, independently of the category count. Four queries give the highest All accuracy on UCF101, Atomic65, and Composite35 in this sweep, while SSv2 and Mutex40 favor other capacities. A shared small bank accommodates these varied category structures without dataset-specific query counts.

Supervision coefficient. The coefficient controls supervised contrast in the Semantic and fused embeddings. At $\lambda = 0$, those two terms vanish, but Operator-level supervised contrast and label-constrained SS-KMeans remain active. This is not label-free discovery. The sweep changes how strongly labeled classes shape the two embedding spaces while keeping the labeled fraction, category vocabulary, observations, and assignment procedure fixed.

E.1 Explicit transition encoding versus temporal aggregation

The component ablation separates regional selection from temporal scale. Here we ask a complementary question: once regional states are available, does explicitly describing their transitions improve discovery beyond ordinary temporal aggregation? We compare three operator encoders on UCF101, SSv2, and Atomic65 using the non-parametric endpoint, seed 1, and Known-K Final evaluation.

Controlled architecture. All three models retain frozen VideoMAE features, the Semantic head, four recurrent Action Queries, per-tubelet All-4 aggregation, the contextual state encoder, the three-scale residual mixture, and F1 fusion. Only the operator encoder after the contextualized state sequence is replaced at each scale. The unchanged modules share their architecture and parameter

Table 11: **Explicit transition encoding versus ordinary temporal aggregation (accuracy, %).** All methods use the non-parametric framework and seed 1. Within each dataset, bold identifies the highest accuracy separately for the All, Old, and New splits.

Dataset	Operator encoder	All	Old	New
UCF101	Mean pooling + MLP	65.94	**91.80**	52.96
	Temporal attention	68.60	87.71	59.01
	Transition encoding	**69.54**	88.93	**59.81**
SSv2	Mean pooling + MLP	23.32	39.79	15.36
	Temporal attention	24.35	**39.93**	16.83
	Transition encoding	**24.53**	39.20	**17.45**
Atomic65	Mean pooling + MLP	67.49	85.31	58.85
	Temporal attention	70.08	87.19	61.79
	Transition encoding	**73.16**	**88.12**	**65.91**

initialization across the three runs. Each model is then jointly optimized, so its learned Query trajectories can differ; the comparison does not freeze identical trajectories across alternatives. The original representation losses and non-parametric discovery procedure are unchanged.

Three temporal encoders. *Transition encoding* is the Q-ViSTA construction in Eq. 5: each token contains the two endpoint states, their difference, their elementwise interaction, and normalized temporal progress. With 256-dimensional states, the descriptor has $4 \times 256 + 1 = 1025$ entries. Projection, GELU, and LayerNorm map it to temporal tokens, followed by positional encoding, a Transformer, and learned attention pooling. The progress coordinate expresses position within the sampled sequence, not the video's duration or timestamp metadata.

Mean pooling + MLP replaces this dedicated transition encoder with a mean over the states at each scale, followed by a two-layer GELU MLP, dropout, and output normalization. It does not remove all temporal modeling: the shared upstream state encoder still contains a Transformer. *Temporal attention* instead projects individual states through an MLP and LayerNorm, adds positional encodings, and applies a Transformer with learned attention pooling and output normalization. This control preserves temporal attention without explicit endpoint differences or interactions.

Training and capacity control. The experiment scripts specify 100 epochs of joint optimization with AdamW, batch size 256, weight decay 10^{-4}, and a fixed learning rate of 3×10^{-5} without learning-rate warmup. They retain four queries, 256-dimensional state and Operator embeddings, offsets $\{1, 2, 4\}$, and $\lambda = 0.50$ for Semantic and fused discovery losses. Model and action-head dropout are 0.1; Query aggregation and F1 dropout are zero. The encoder implementation chooses each control MLP's hidden width to keep its parameter count within 5% of the corresponding transition encoder and rejects configurations outside that tolerance. This is a per-scale encoder-capacity criterion, not a FLOPs constraint. These are script settings, not an independent audit of executed runs: complete per-arm configurations and parameter-count records were not available.

What explicit transitions contribute. Table 11 shows the highest All and New accuracy for transition encoding on all three datasets. The clearest separation occurs on Atomic65: explicit encoding reaches 73.16% All accuracy, improving on ordinary temporal attention by 3.08 points. Its novel-skill accuracy also increases from 61.79% to 65.91%. Ordinary attention already improves on mean pooling, so the additional gain supports describing how localized states change, beyond allocating temporal attention to those states. All-accuracy gains over attention are 0.94 points on UCF101 and 0.18 points on SSv2, extending the same ordering to human actions. These single-seed results support the descriptor within the jointly learned framework; they do not estimate cross-seed uncertainty or statistical significance.

Table 12: **Learned routing probabilities on unlabeled videos (%).** We average the two views within each video, then average videos equally. Null means no auxiliary residual; it is not the final short-scale weight. Bold marks the preferred auxiliary route.

Dataset	Unlabeled videos	Null	Medium ($\pi^{(2)}$)	Long ($\pi^{(4)}$)
UCF101	10,001	0.69	34.30	**65.01**
SSv2	19,148	0.49	23.88	**75.63**
Atomic65	4,900	0.72	**51.84**	47.44

Table 13: **Effective coefficients before final normalization (%).** The learned γ is a raw scalar. Coefficients follow Eq. 7; video averaging and auxiliary totals use unrounded values.

Dataset	γ	Short ($w^{(1)}$)	Medium ($w^{(2)}$)	Long ($w^{(4)}$)	Medium + long
UCF101	0.0585	94.19	2.01	3.80	5.81
SSv2	0.0661	93.42	1.58	5.00	6.58
Atomic65	0.0541	94.63	2.80	2.57	5.37

E.2 Learned scale routing and effective mixture weights

We next inspect how the trained main model combines temporal scales. This diagnostic uses the original non-parametric Q-ViSTA final checkpoints at epoch 100 for seed 1, not the three temporal-control checkpoints. No parameters are updated: inference runs in evaluation mode with gradients disabled. Repeated forward passes on the same batch have a recorded maximum absolute fused-output difference of zero on all three datasets. This verifies repeatability of the diagnostic forward pass, not variability across independent training runs.

Routing is distinct from final mixing. The shared scoring network applies LayerNorm and a $256 \to 128 \to 1$ GELU MLP to each scale's Operator. Relative scores against the shortest offset produce the Null, medium, and long probabilities in Eq. 6; no Semantic features enter this decision. Equation 7 then scales the auxiliary probabilities by γ and assigns the remaining coefficient to the short-scale anchor. Equivalently, the pre-normalization mixture adds $\gamma\pi^{(2)}(\boldsymbol{o}^{(2)} - \boldsymbol{o}^{(1)})$ and $\gamma\pi^{(4)}(\boldsymbol{o}^{(4)} - \boldsymbol{o}^{(1)})$ to $\boldsymbol{o}^{(1)}$. Thus auxiliary routing preferences describe a controlled adjustment to the short-scale anchor, not replacement of that anchor by the preferred auxiliary scale.

Dataset-dependent auxiliary context. Table 12 shows a preference for the long route on UCF101 and SSv2, whereas Atomic65 distributes auxiliary routing more evenly, with a modest preference for the medium route (51.84% versus 47.44%). The model therefore does not allocate auxiliary context identically across human actions and atomic manipulation. Table 13 clarifies the magnitude of this adjustment: the short-scale coefficient remains above 93%, with medium and long scales together contributing 5.37%–6.58% of the pre-normalization coefficient mass. Short-scale anchoring is part of the architecture; the learned allocation describes how the trained model supplements it. These weights quantify mixture coefficients, not causal contributions to accuracy or normalized feature magnitude.

Connection to Semantic Identity. After scale mixing, F1 concatenates the 256-dimensional Semantic and Operator embeddings. LayerNorm, a $512 \to 1024 \to 256$ GELU MLP with zero dropout, and final L2 normalization form the fused discovery representation. The Semantic input is detached on this fusion path, so the fused objective does not backpropagate through it; the Semantic head still learns from its own discovery objective. Together, these experiments connect improved transition encoding with a trained main model that supplements short-scale dynamics using dataset-dependent auxiliary context before fusion with semantic identity.

F Additional Comparison with MCCL

Comparison setting. Table 14 replaces VideoMAE with TimeSformer (Bertasius et al., 2021). MCCL's code has not been publicly released at the time of writing (Zhang et al., 2025), so we use its

Table 14: **TimeSformer comparison (accuracy, %).** †: results reported in MCCL (Zhang et al., 2025); ∗: training uses a parametric category classification head. Datasets are weighted equally in Average. Column-best and second-best values appear in **bold** and with underlining, respectively.

Backbone	Method	Venue	UCF101			SSv2			K400			Average		
			All	Old	New	All	Old	New	All	Old	New	All	Old	New
TimeSformer	SimGCD†∗	ICCV'23	63.93	82.65	54.70	11.36	14.27	9.93	21.57	25.15	19.79	32.29	40.69	28.14
	InfoSieve†	NeurIPS'23	58.02	69.90	52.60	11.10	12.92	10.22	21.28	22.28	20.78	30.13	35.03	27.87
	SPTNet†∗	ICLR'24	38.50	55.69	30.03	7.92	10.11	6.85	12.55	14.32	11.67	19.66	26.71	16.18
	SelEx†	ECCV'24	63.43	77.07	56.72	12.54	14.96	11.35	21.23	23.14	20.28	32.40	38.39	29.45
	MCCL†∗	arXiv'25	68.27	89.54	57.80	13.58	18.69	**11.57**	22.95	29.89	19.51	34.93	46.04	29.63
	Q-ViSTA		70.39	90.02	60.53	13.18	**20.66**	9.56	29.34	47.55	20.10	37.64	52.74	30.06
	Q-ViSTA-Param∗		**74.68**	**90.31**	**66.84**	**14.06**	19.27	11.54	**33.17**	**50.09**	**24.57**	**40.64**	**53.22**	**34.32**

Both Q-ViSTA variants use a frozen eight-layer TimeSformer initialized from public image-pretrained weights, not the MCCL authors' video-trained checkpoint.

reported results rather than rerunning its implementation. Both Q-ViSTA variants use a frozen eight-layer encoder initialized from public image-pretrained weights, not the MCCL authors' video-trained checkpoint. Published baselines include SimGCD (Wen et al., 2023), InfoSieve (Rastegar et al., 2023), SPTNet (Wang et al., 2024), SelEx (Rastegar et al., 2024), and MCCL. Daggers mark scores from the MCCL paper; asterisks indicate a parametric category head during training. MCCL uses 100 UCF101 categories rather than our 101, and different SSv2 and K400 selections. This comparison tests applicability beyond VideoMAE; the encoder family alone does not match the experimental conditions.

Transfer beyond VideoMAE. With the frozen image-initialized TimeSformer, non-parametric Q-ViSTA reaches 37.64% average All accuracy. MCCL reports 34.93% under its own setup; on K400 the corresponding entries are 29.34% and 22.95%. Our runs establish applicability of the same representation design beyond VideoMAE across human-action vocabularies. These cross-study figures provide context for the attained accuracy; the shared-VideoMAE experiments remain the controlled basis for comparing complete video discovery systems.

Effect of the parametric head. Q-ViSTA-Param achieves the highest All accuracy in the table on every dataset, averaging 40.64%. Compared with non-parametric Q-ViSTA, its average novel-class accuracy increases from 30.06% to 34.32%, alongside an increase for old classes. The category head thus provides additional value within the shared dual-representation framework. The within-Q-ViSTA comparison fixes our TimeSformer initialization and discovery pools, measuring classification training and assignment together. Comparison against dagger-marked methods provides cross-study context, with differences in observations and training kept explicit.

G Confidence-Guided Rejection Experiments

In practical video curation, discovered labels can initialize annotation rather than replace it. We evaluate confidence-guided selective pre-labeling using the same non-parametric Q-ViSTA configuration as the main results and default ablations, with Known-K Final results for seeds 1, 2, and 3. Within each seed, `cluster_margin` ranks clips by $d_2 - d_1$, the difference between the second-smallest and smallest squared Euclidean distances to cluster centers in normalized feature space. Low-confidence clips are routed to manual review, while the retained subset forms a candidate pre-label pool after an annotator names each cluster. At every rejection level, we preserve both the original SS-KMeans assignments and the Hungarian mapping fitted on the complete unlabeled pool. Confidence selection changes the evaluated subset, not the predicted cluster or matched category of any retained video.

Table 15 reports means and sample standard deviations across the three seeds. Coverage is the retained fraction, and Retained is the number of clips kept in each seed. Split accuracies are computed within each seed before aggregation; the proportions of retained old and novel clips can differ across seeds and rejection levels. The zero-rejection results exactly reproduce the non-parametric main-table entries for each protocol.

Table 15: **Confidence-guided selective pre-labeling on robot-skill protocols (accuracy, %).** Mean ± sample standard deviation over seeds 1, 2, and 3 using Known-K Final. Rejection routes low-confidence clips to manual review; retained counts are per seed. Bold marks the highest mean for each split within each protocol.

Rejection	Coverage	Retained	All	Old	New
Mutex40					
0%	100%	1,140	74.09 ±1.86	85.83 ±2.55	68.68 ±2.12
10%	90%	1,026	78.04 ±1.30	86.50 ±2.64	73.56 ±1.34
20%	80%	912	81.07 ±1.32	86.94 ±2.41	77.51 ±1.61
30%	70%	798	83.42 ±2.16	88.03 ±3.16	80.15 ±3.23
40%	60%	684	85.48 ±2.49	88.56 ±2.75	82.85 ±3.33
50%	50%	570	**86.73** ±2.99	**89.70** ±2.72	**83.63** ±4.57
RoboCasa365-Atomic65					
0%	100%	4,900	73.41 ±0.61	89.21 ±1.09	65.75 ±1.13
10%	90%	4,410	77.38 ±0.65	93.13 ±0.66	69.13 ±1.11
20%	80%	3,920	81.05 ±1.06	95.12 ±0.61	72.99 ±1.54
30%	70%	3,430	84.20 ±1.41	96.50 ±0.81	76.37 ±2.02
40%	60%	2,940	86.50 ±1.48	97.58 ±0.98	78.41 ±1.94
50%	50%	2,450	**88.61** ±1.40	**98.38** ±0.95	**80.21** ±1.95
RoboCasa365 A→C (65/35)					
0%	100%	6,750	80.56 ±0.62	92.14 ±0.69	69.81 ±1.23
10%	90%	6,075	85.48 ±0.53	94.70 ±0.63	76.01 ±1.10
20%	80%	5,400	89.46 ±0.71	96.60 ±0.55	81.17 ±1.18
30%	70%	4,725	92.21 ±0.33	97.80 ±0.17	84.63 ±0.85
40%	60%	4,050	94.03 ±0.40	98.56 ±0.11	86.59 ±1.18
50%	50%	3,375	**95.80** ±0.52	**99.10** ±0.20	**88.80** ±1.57

The operating point can be chosen according to the available review budget. At 50% coverage, Mutex40 retains 570 candidate pre-labels per seed at 86.73% mean accuracy. The same coverage yields 88.61% on Atomic65 and 95.80% on Atomic-to-Composite, while retaining 2,450 and 3,375 clips per seed, respectively. Mean accuracy increases at every rejection level on all three protocols, including their novel-class partitions. This consistent ordering supports cluster margin as a useful triage signal: manual review receives uncertain assignments while the retained candidates provide an increasingly accurate starting point for subsequent manual annotation.

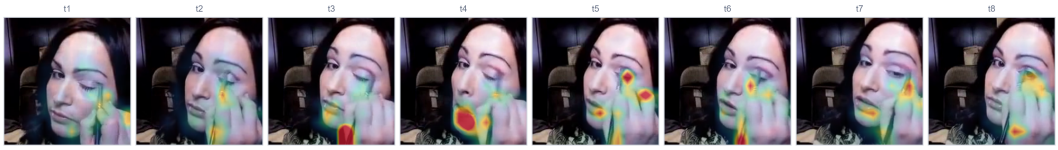
Operational interpretation. The experiment supports a candidate-label workflow with an explicit review budget. An annotator names each cluster from representative clips, checks the confidence-ranked candidates, and reviews the rejected clips individually. Coverage quantifies how much of the archive remains in the candidate pool; retained accuracy quantifies its label quality. These are selection-quality measurements, not measured annotation-time savings. Keeping both accepted and rejected assignments allows subsequent corrections to propagate through the curated collection.

H Additional Visualizations

We inspect regional attention and feature-space organization across human actions and robot skills. Attention sequences expose the collected interaction evidence, while feature plots show how known and novel categories are organized in the resulting representations.

H.1 Regional attention across human actions and robot skills

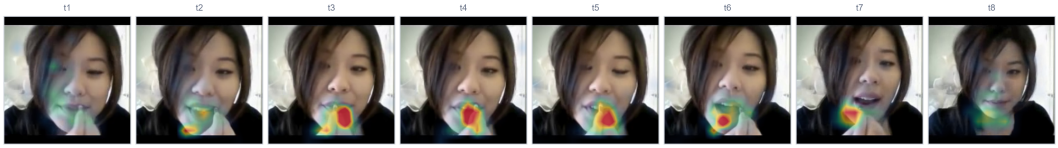
Figures 6–9 show sixteen complete eight-tubelet sequences: two known and two novel classes per dataset. Each row displays the per-tubelet All-4 weighted sum of Query attention, with warmer colors indicating stronger responses and normalization shared within the video. The selection covers distinct interaction regions in UCF101, SSv2, RoboCasa365, and Mutex40. These selected examples illustrate regional evidence, not measured localization accuracy.



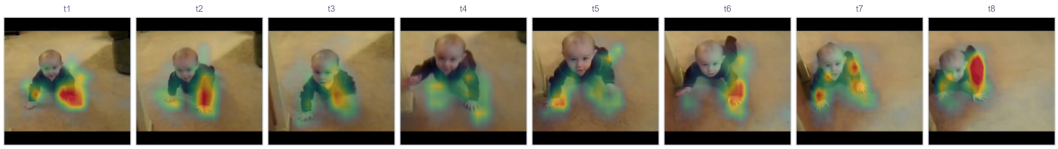
(a) Known: Apply Eye Makeup.



(b) Known: Archery.

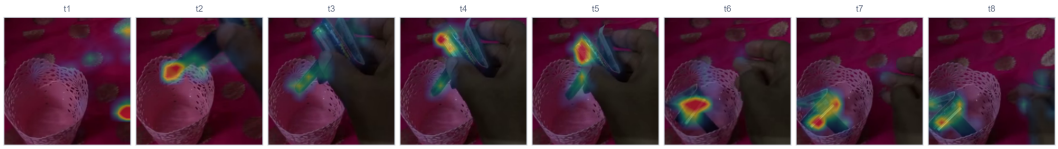


(c) Novel: Apply Lipstick.

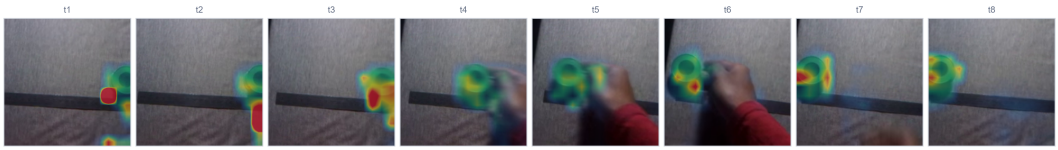


(d) Novel: Baby Crawling.

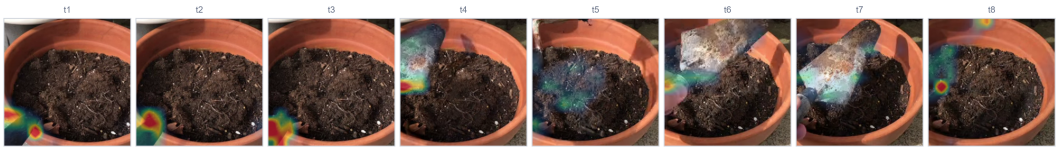
Figure 6: **UCF101 visualizations.** Two known and two novel classes.



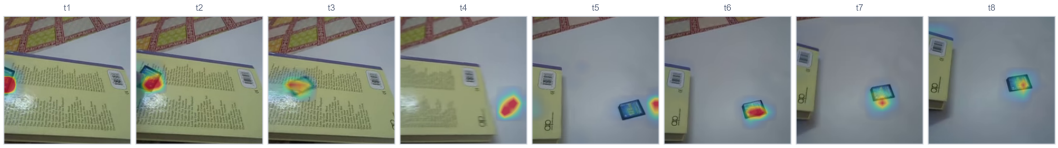
(a) Known: Dropping something into something.



(b) Known: Pushing something from right to left.



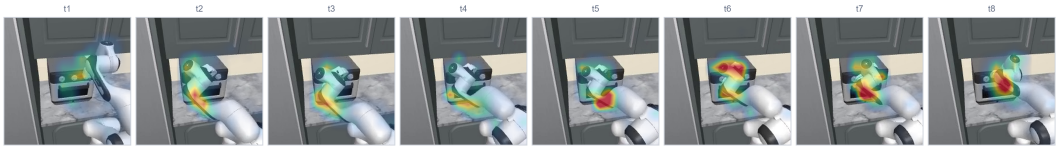
(c) Novel: Pretending to scoop something up.



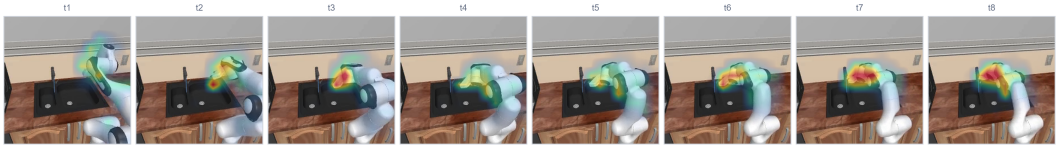
(d) Novel: Placing a non-rollable object on a slope.

Figure 7: **SSv2 visualizations.** Two known and two novel classes.

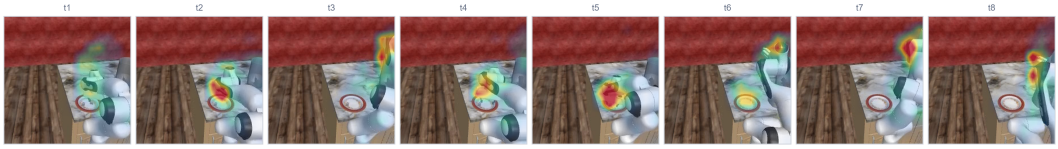
Human actions. Across UCF101, responses follow face–hand interactions in Apply Lipstick, the archer–bow relation in Archery, and broader body support in Baby Crawling. On SSv2, they follow manipulated objects through dropping, pushing, scooping, and sliding. These known and novel examples show the same query architecture collecting category-bearing interactions without assigning a separate Query to every known or novel category. Each displayed map combines the



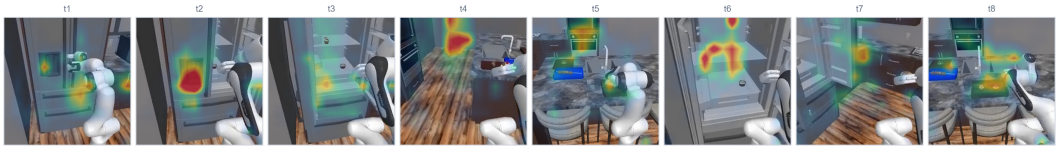
(a) Known: Adjust Toaster Oven Temperature.



(b) Known: Adjust Water Temperature.

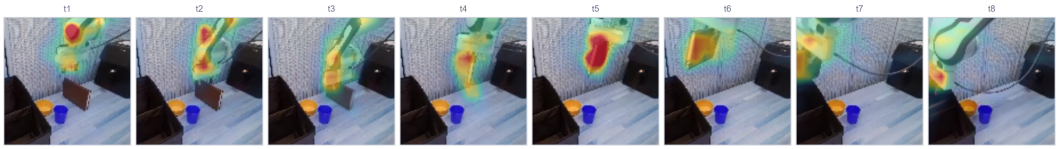


(c) Novel: Add Sweetener.

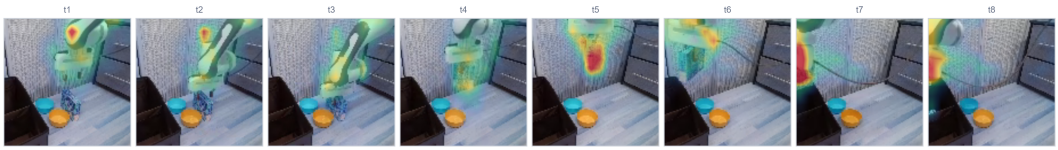


(d) Novel: Arrange Buffet Dessert.

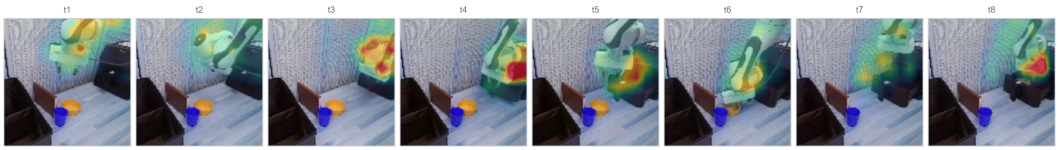
Figure 8: **RoboCasa365 visualizations.** Two known and two novel classes.



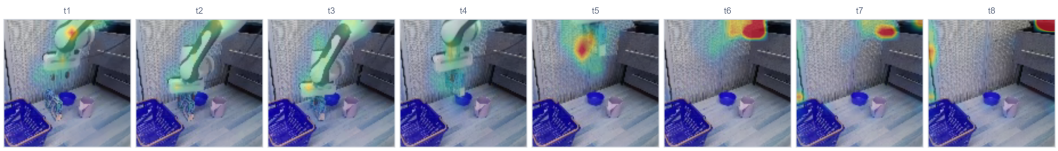
(a) Known: Put Book in Caddy Back.



(b) Known: Put Granola Bar in Caddy Back.



(c) Novel: Open Air Fryer and Put Blue Bowl Inside.



(d) Novel: Put Granola Bar in Top Drawer.

Figure 9: **Mutex40 visualizations.** Two known and two novel classes.

four Query responses. It therefore exposes the regional evidence passed to the next temporal module, rather than identifying a separate object track for every individual Query.

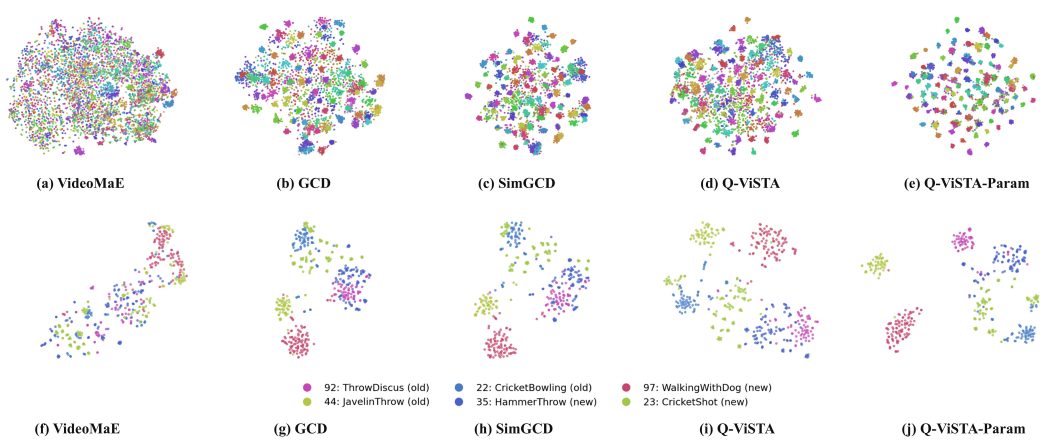


Figure 10: **UCF101 feature visualizations.** Top: all-category view. Bottom: three old and three new classes. The legend identifies each class and its status.

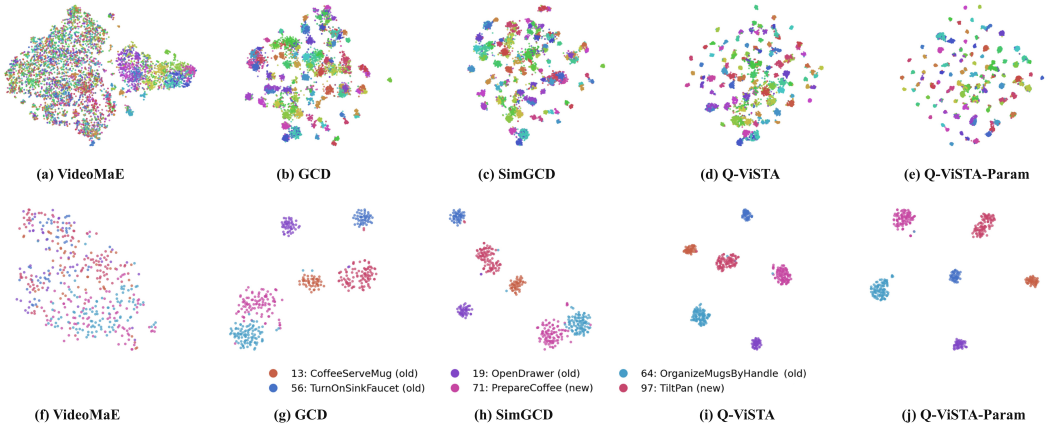


Figure 11: **Atomic-to-Composite feature visualizations.** Top: all-category view. Bottom: four old and two new classes. The legend identifies each class and its status.

Robot skills. In RoboCasa365 and Mutex40, attention follows manipulated objects, contact points, and successive destinations. These regional trajectories complement the Semantic branch's object and scene context, illustrating the evidence supplied to the transition pathway for known and novel task categories.

H.2 Feature-space organization of old and new categories

Figures 10 and 11 compare VideoMAE, GCD, SimGCD, Q-ViSTA, and Q-ViSTA-Param. The upper row shows all categories; the lower row isolates six selected classes, with identical selections across methods. UCF101 displays three old and three new classes, while Atomic-to-Composite displays four old and two new classes. The reduced view highlights within-class concentration and between-class overlap. These projections complement quantitative evaluation in the original feature space.

Human actions. The UCF101 selection contains the old classes ThrowDiscus, JavelinThrow, and CricketBowling, together with the new classes HammerThrow, WalkingWithDog, and CricketShot. In Figure 10, Q-ViSTA-Param forms concentrated groups for ThrowDiscus and WalkingWithDog, while non-parametric Q-ViSTA visibly separates WalkingWithDog from the selected sporting actions. The view illustrates distinct neighborhoods for familiar and unfamiliar activities.

Robot skills. The Atomic-to-Composite selection contains the old skills CoffeeServeMug, OpenDrawer, TurnOnSinkFaucet, and OrganizeMugsByHandle, together with the new skills PrepareCoffee

and TiltPan. In Figure 11, both Q-ViSTA variants form clearly separated groups for these six skills, including the new composite categories. Compared with the intermingled frozen features, this view makes the organization of individual robot skills easier to inspect in the dual representation. Attention sequences expose regional evidence; feature plots show how fusion organizes categories.

I Broader Impact and Ethics Statement

Benefits and interpretation. Q-ViSTA can support retrieval, curation, and candidate pre-labeling in partially labeled video archives. A curator can name clusters from representative clips and verify their members, retaining authority over the resulting semantic vocabulary. Cluster assignments remain hypotheses about shared behavior and should support correction as the collection evolves. Confidence-guided selection prioritizes review; it does not certify label correctness. Deployment should choose coverage on a held-out validation stream and preserve rejected clips for manual inspection and correction before candidate labels enter a curated training collection.

Data governance and privacy. Video can reveal faces, locations, routines, possessions, and other context beyond the action of interest. Collection and deployment should respect consent, licensing, and the original recording purpose, with access controls, retention limits, and removal procedures for private data. Discovered groups should not be used to infer sensitive attributes or expand surveillance beyond that purpose or the scope of informed consent.

Robot deployment boundary. The method groups demonstrations; it neither generates controls nor certifies execution safety. Any use of discovered skills in policies or curricula still requires trajectory validation, collision and hardware checks, task-specific review, human oversight, environment testing, and explicit fail-safe mechanisms before any physical deployment.

Dataset and representation scope. Discovered groups reflect the available video collection. Curation should cover relevant actors, environments, viewpoints, objects, and task variants. Reporting retained coverage together with group quality makes the scope of these observations explicit, helping users assess which data can support subsequent annotation and learning.

J AI Use Statement

AI tools assisted with language polishing. They helped refine grammar and sentence clarity. This assistance improved the fluency and consistency of the English prose.