RETHINKING JEPA: COMPUTE-EFFICIENT VIDEO SSL WITH FROZEN TEACHERS

Xianhang Li, Chen Huang, Chun-Liang Li, Eran Malach, Josh Susskind, Vimal Thilak, Etai Littwin · ICLR 2026 · Apple

Abstract

SALT (Static-teacher Asymmetric Latent Training)

We introduce **SALT** (Static-teacher Asymmetric Latent Training), a video self-supervised learning approach that freezes the target encoder after an initial pixel-prediction stage, creating an asymmetric dynamic where only the student continues to learn.

This counterintuitive design proves **remarkably effective**. SALT achieves state-of-the-art performance on frozen-backbone video benchmarks and significantly outperforms V-JEPA2 as a robotics world model encoder—89% vs 66% success rate at the 1B scale with consistent positive scaling. A frozen target appears to force more robust, action-relevant representations, offering a new perspective on self-supervised video representation learning.

Objective

Why SALT?

V-JEPA is a self-distillation approach that uses a dynamic teacher via EMA to learn strong off-the-shelf video representations. EMA is used to avoid representation collapse.

However, using self-distillation involves a delicate balancing act between the student and the teacher during training and is inefficient for scaling models, forcing architectural parity between student and teacher.

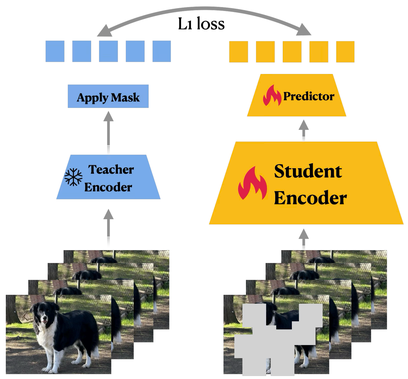
Our Method

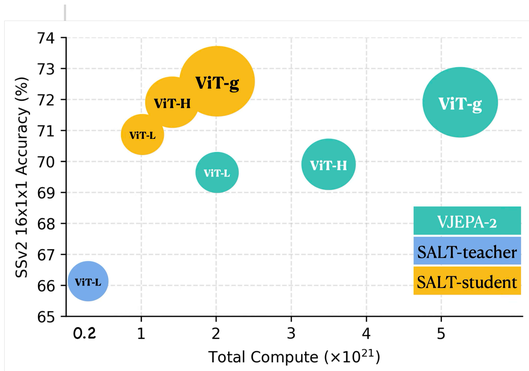
SALT consists of two stages
(i) train a target encoder witha simple pixel-reconstruction objective under V-JEPA masking (V-Pixel) then
(ii) freeze teacher and train a student to predict the teacher's latents on masked regions.

Surprisingly, breaking teacher-student architectural parity allows the student to vastly outperform its frozen teacher, suggesting that static targets provide a **more effective learning signal** than co-evolving ones.

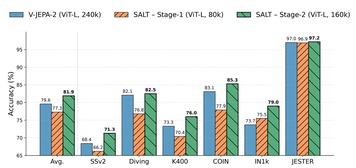
Results and Analysis

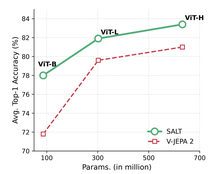
Improved Compute Efficiency Over V-JEPA 2





Systematic comparsion with SoTA models

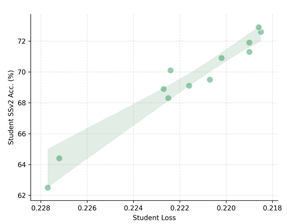


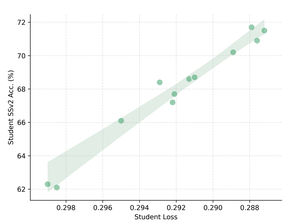


Method	Param.	Pretraining Dataset	Teacher (Params)	Total Compute	# Seen Samples	SSv2	K400
VideoMAEv2 (Tong et al., 2022)	1B	UnlabeledHybrid-1.4M	N/A	2.2	1.6B	56.1	82.8
PEcoreG (Bolya et al., 2025)	1.9B	MetaCLIP-5B (Xu et al.)	N/A	—	86B	55.4	88.5
InternVideo2-1B (Wang et al.)	1B	IV-25.5M	InternVL-6B + VideoMAEv2-g	—	—	67.3	87.9
VideoPrism (Zhao et al., 2024)	1B	VT-36M	Stage-1-ViT-g (1.0B)	—	2.0B	68.5	87.6
DINOv2 (Oquab et al., 2024)	1.1B	LVD-142M	EMA teacher (1.1B)	—	1.9B	50.7	83.6
SigLIP2 (Tschannen et al.)	1.2B	WebLI-10B (Chen et al.)	EMA teacher (1.2B)	—	40B	49.9	87.3
V-JEPA 2 ViT-L (Assran et al.)	300M	VM-22M	EMA teacher (300M)	1.9	0.7B	73.7	85.1
V-JEPA 2 ViT-H (Assran et al.)	600M	VM-22M	EMA teacher (600M)	3.5	0.7B	74.0	85.3
V-JEPA 2 ViT-g (Assran et al.)	1B	VM-22M	EMA teacher (1B)	5.3	0.7B	75.3	86.6
V-JEPA 2 ViT-L	300M	V-3.6M	EMA teacher (300M)	1.4	0.7B	68.2	83.8
V-JEPA 2 ViT-H	600M	V-3.6M	EMA teacher (600M)	2.6	0.7B	73.4	84.6
SALT ViT-L	300M	V-3.6M	—	1.2	0.7B	74.9	85.4
SALT ViT-H	600M	V-3.6M	—	1.5	0.7B	75.4	86.0
SALT ViT-g	1B	V-3.6M	SALT-ViT-L (300M)	1.9	0.7B	76.2	86.8
SALT ViT-G	2B	V-3.6M	—	2.6	0.7B	76.1	87.2

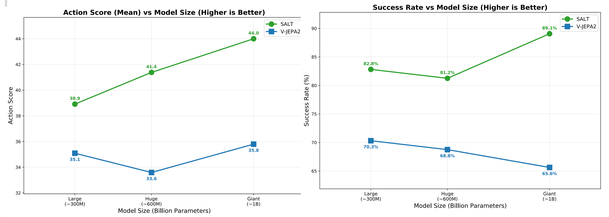
SALT **outperforms** SoTA models on Something-Something-V2, a temporal understanding task, and remains competitive on Kinetics-400, an appearance understanding task.

Interpretable Model Selection with SALT





World Modeling and Planning with SALT

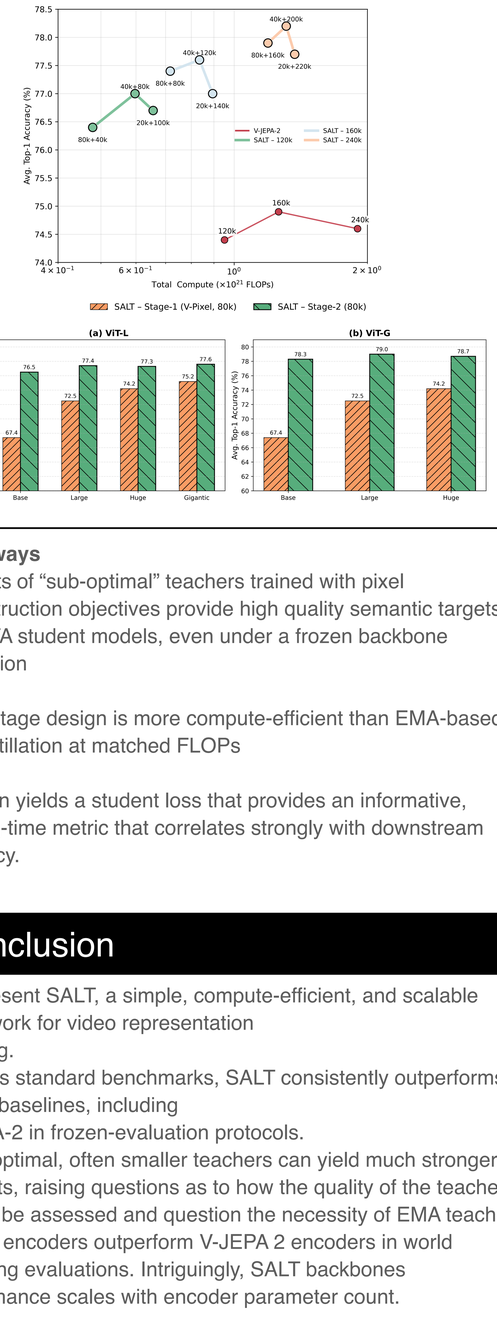


Metric	SALT Large	SALT Huge	SALT Giant	V-JEPA2	V-JEPA2	V-JEPA2
Action Score (mean)	38.91	41.38	44.00	35.09	33.59	35.79
Action Score (median)	49.68	53.62	53.26	47.06	42.53	48.79
End Distance (m)	0.06	0.06	0.06	0.08	0.10	0.09
Success Rate (%)	82.8	81.2	**89.1**	70.3	68.8	65.6

Setup
- Train world model predictors on the Franka robotics dataset using frozen SALT and V-JEPA2 encoders
- Three model sizes (300M-1B).
- Models are evaluated on action prediction quality via Cross-Entropy Method (CEM) planning.

Results
- SALT **consistently outperforms** V-JEPA2 across all metrics and model sizes.
- SALT shows **positive scaling** behavior, with the 1B model achieving 89% success rate compared to V-JEPA2's 66%.
- SALT also achieves 23% higher action scores and 33% lower end-effector error than V-JEPA2 at the 1B scale.

SALT Design Principles



Takeaways
- Latents of "sub-optimal" teachers trained with pixel reconstruction objectives provide high quality semantic targets for SoTA student models, even under a frozen backbone evaluation

- Two-stage design is more compute-efficient than EMA-based self-distillation at matched FLOPs

- Design yields a student loss that provides an informative, training-time metric that correlates strongly with downstream accuracy.

Conclusion

We present SALT, a simple, compute-efficient, and scalable framework for video representation
learning.
- Across standard benchmarks, SALT consistently outperforms strong baselines, including
V-JEPA-2 in frozen-evaluation protocols.
- Sub-optimal, often smaller teachers can yield much stronger students, raising questions as to how the quality of the teacher should be assessed and question the necessity of EMA teacher
- SALT encoders outperform V-JEPA 2 encoders in world modeling evaluations. Intriguingly, SALT backbones performance scales with encoder parameter count.

TM and © 2026 Apple Inc. All rights reserved.