InnerJEPA: Masked Visual Latent Prediction for Video Understanding and Generative Transfer

Anonymous authors
Paper under double-blind review

Abstract

Video multimodal large language models (MLLMs) are typically instruction-tuned with answer-token supervision, leaving visual prediction implicit. We introduce InnerJEPA, a training-only objective that uses the frozen vision encoder as a target encoder and the causal language model as a masked visual predictor. Complete encoded time slices are replaced by a learned mask vector, and a lightweight head reconstructs their clean features. Naive cosine regression produces nearly parallel predictions in the anisotropic target space. Target centring, leading-direction removal, and variance–covariance regularisation counter this degeneracy: prediction effective rank increases from 14 to 332 and mean pairwise cosine decreases from 0.908 to 0.160. In a controlled Qwen3.5-0.8B comparison, InnerJEPA improves matched SFT by 2.09 accuracy points averaged over four video-QA benchmarks and by 2.34 mIoU points across five temporal-grounding benchmarks. Under matched alignment-MLP and DiT-LoRA adaptation, CLIP score rises from 0.2329 to 0.2449 on 553 images generated from GenEval prompts; a fixed 20-prompt blinded pilot favours InnerJEPA on 14 prompts, with five ties. These results support collapse-resistant latent prediction as auxiliary supervision for video understanding and generative transfer, without adding inference-time components to the video MLLM.

1 Introduction

Video multimodal large language models (MLLMs) such as Qwen-VL and LLaVA-Video (Bai et al., 2025; Zhang et al., 2024) are commonly instruction-tuned through answer-token supervision. Visual information is therefore rewarded indirectly through language, motivating auxiliary reconstruction and representation objectives (Wang et al., 2025a; Yoon et al., 2025; Jain et al., 2024).

Joint-embedding predictive architectures (JEPAs) predict hidden representations instead of pixels (Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025). A video MLLM already supplies two useful components: its frozen vision encoder defines visual targets, and its causal language model produces contextual states at every input position. InnerJEPA reuses them by masking entire encoded time slices and reconstructing their clean features from the corresponding language-model outputs. A shared two-layer head and one mask vector are the only additions; both are discarded at inference. Targets and visible inputs share one video encoding.

The challenge is making this supervision informative. Frozen visual features can share a dominant direction, allowing cosine regression to decrease without distinguishing visual states. In our diagnostic, cosine-only predictions have effective rank 14 in 1024 dimensions and mean pairwise cosine 0.908. We centre targets, remove their leading principal direction, and apply VICReg-style variance and covariance penalties (Bardes et al., 2022). The complete objective reaches rank 332 and pairwise cosine 0.160. These diagnostics measure prediction diversity; downstream evaluation tests whether it is useful.

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.

In a controlled Qwen3.5-0.8B comparison, InnerJEPA improves matched SFT on all four video-QA benchmarks and four of five temporal-grounding sets, with mean gains of 2.09 accuracy points and 2.34 mIoU points. A separate Qwen3.5-2B study provides scaling context. Under identical alignment-MLP and DiT-LoRA adaptation settings, the learned representation also improves CLIP score on 553 GenEval generations. Frame-order perturbations reveal greater robustness but do not establish fine-grained order dependence.

Our contributions are: (1) a masked visual prediction objective that reuses the MLLM's encoder and causal backbone with no inference overhead; (2) a diagnosis of prediction collapse and a target-conditioning and variance–covariance remedy; and (3) controlled video-understanding gains and a matched generative-transfer probe. Appendix E provides a conceptual comparison with related architectures.

2 Related Work

Visual latent prediction. I-JEPA, V-JEPA, and V-JEPA 2 predict masked representations with momentum targets (Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025). NEPA studies causal next-patch prediction (Xu et al., 2025), while SALT examines frozen teachers (Li et al., 2026). VideoMAE and MaskFeat motivate temporal masking to prevent easy interpolation (Tong et al., 2022; Wei et al., 2021). Masked visual modelling also appears in video–language pre-training (Fu et al., 2021; Li et al., 2022; Cheng et al., 2022; Li et al., 2023). Our objective instead supplements answer-token cross-entropy inside a causal video MLLM.

Auxiliary supervision in MLLMs. ROSS, OLA-VLM, and VIRAL add reconstruction or feature-alignment objectives (Wang et al., 2025a; Jain et al., 2024; Yoon et al., 2025). Ross3D predicts VAE latents from masked views (Wang et al., 2025b); Latent Denoising corrupts visual inputs and recovers frozen features from intermediate LM states (Parikh et al., 2026). JARVIS and Orca use latent prediction with additional target or readout components (Caffagni et al., 2025; Orca Team, Beijing Academy of Artificial Intelligence, 2026), while PEARL and LLM-JEPA explore hidden-state prediction in language-oriented settings (Adhikari & Lapata, 2026; Huang et al., 2025). Closest to our design, Latent Denoising also removes auxiliary components at inference. InnerJEPA focuses on whole-slice masking, targets from the MLLM's own frozen visual encoder, and prediction from causal final-layer states. Its specific contribution is diagnosing and correcting degeneracy in that frozen target space; auxiliary loss reduction alone need not imply useful representations (Sengupta, 2026; Zhang et al., 2026). Table 10 in the appendix summarises the architectural distinctions.

3 Degeneracy in a Frozen Target Space

The geometry of a frozen target space determines whether feature reconstruction is meaningful. Let $\mathbf{z} \in \mathbb{R}^d$ denote one output token of the frozen vision encoder and merger. Such features need not be isotropic: a shared component can dominate their directions even when the residual components retain useful content. Under a plain cosine objective, a predictor can therefore reduce its loss by reproducing the shared direction instead of distinguishing visual states.

Measuring collapse. For a centred prediction matrix $\mathbf{Y}$ with singular values $\{\sigma_j\}$, we use the entropy effective rank

$$r_{\text{eff}}(\mathbf{Y}) = \exp\left(-\sum_j p_j \log p_j\right), \qquad p_j = \frac{\sigma_j}{\sum_k \sigma_k}. \tag{1}$$

We report it together with mean pairwise cosine similarity. A constant or nearly constant predictor has low effective rank and pairwise cosine near one; a diverse representation has higher rank and lower pairwise cosine.

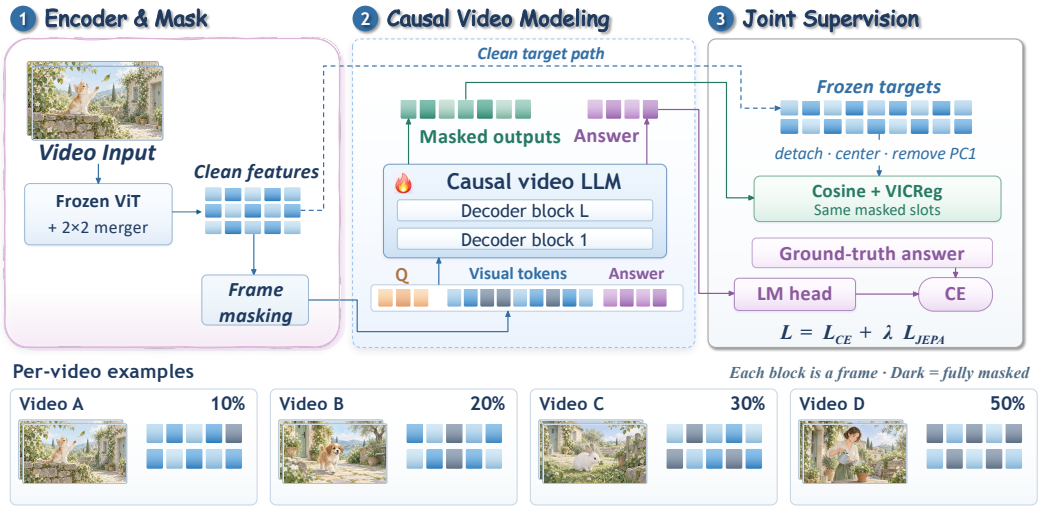


Figure 1: **InnerJEPA.** A sampled subset of one-second slices is replaced by `[MASK]` in the language-model input; the output at each hidden slot predicts that slot's clean encoder feature through a shared head. Targets are detached, centred and stripped of their leading principal direction; the head and `[MASK]` are discarded at inference.

Cosine regression takes the shortcut. On the Qwen3.5-0.8B diagnostic run, cosine-only predictions occupy just 14 effective dimensions out of 1024 and have mean pairwise cosine 0.908. The loss can thus decrease while the outputs remain nearly parallel. This observation is important: a low reconstruction loss is not evidence that the model has learned temporal state.

Removing the shared direction is necessary but not sufficient. We first detach and centre the clean targets, then project out their leading principal direction. This prevents the easiest constant-direction match while preserving the remaining target variation. The predictor can still collapse on its own, so we additionally impose VICReg variance and covariance penalties (Bardes et al., 2022): the variance term prevents dimensions from becoming constant and the covariance term discourages redundant dimensions. With the complete objective, effective rank rises from 14 to 332 and mean pairwise cosine falls from 0.908 to 0.160. These diagnostics establish that the predictions become substantially more diverse; downstream evaluation is still required to determine whether that diversity is useful.

4 InnerJEPA: Masked Latent Reconstruction Inside the LM

InnerJEPA adds one training-only objective to standard video instruction tuning and removes it at inference (Figure 1). The objective predicts frozen vision-encoder features from the language model's own outputs at hidden positions, so the model's frozen encoder serves as the target encoder and the language model as the predictor; no second encoder, teacher or architectural change is introduced.

4.1 Setup and notation

A video is sampled at 2 fps; with the encoder's temporal patch size of 2, every pair of frames forms one one-second *slice*, indexed $t = 1, \ldots, T$. The frozen vision encoder and merger f_v map slice t to P tokens $\mathbf{z}_t \in \mathbb{R}^{P \times d}$, and the processor writes a timestamp text before each slice. The encoder attends within a slice only, so $\mathbf{z}_t$ contains no information from other slices. The LM is causal; we write $\mathbf{o}_{t,p} \in \mathbb{R}^d$ for its last-layer output at slice t, position p.

4.2 Masked visual reconstruction

On every video sample the clean features $\{\mathbf{z}_t\}$ are computed once and cached. A subset $\mathcal{M}$ of slices is then hidden *after* the encoder: the P tokens of each slice in $\mathcal{M}$ are replaced in the LM input by one learned vector $\mathbf{m}$ (`[MASK]`); timestamps stay in place. The masking ratio is drawn per video from $\{5, 15, 30, 50\}\%$ (25% on average), slices are chosen at random, and the first slice is never masked. The output at each masked position passes through a shared two-layer MLP h_j (Linear–GELU–Linear) and must match the clean feature of the *same* position. Because the encoder is slice-local and the language model causal, this output is a function of the query text, the timestamps and the *earlier* slices only, so the objective is a causal prediction task and not bidirectional inpainting:

$$\mathcal{L}_{\text{JEPA}} = \ell\Big(\big\{\, \mathbf{y}_{t,p} = h_j(\mathbf{o}_{t,p}),\ \hat{\mathbf{z}}_{t,p} \,\big\}_{t\in\mathcal{M},\, p=1..P}\Big), \tag{2}$$

where $\hat{\mathbf{z}}$ is the processed target and ℓ the loss defined next.

Target processing and loss. Section 3 showed that frozen video features share a dominant mean direction onto which a plain regression collapses. Every target is therefore detached, centred by the mean of the N targets in the micro-batch, and projected off their leading principal direction, giving $\hat{\mathbf{z}}$. Predictions are compared by cosine and regularised, following VICReg (Bardes et al., 2022), so that they cannot collapse:

$$\begin{aligned}\ell\big(\{\mathbf{y}_i, \hat{\mathbf{z}}_i\}\big) = \frac{1}{N}\sum_i \Big(1 - \cos(\mathbf{y}_i, \hat{\mathbf{z}}_i)\Big) + \alpha_{\text{var}}\,\frac{1}{d}\sum_{j=1}^{d}\max\big(0, 1-\sigma_j(\tilde{\mathbf{Y}})\big) \\ + \alpha_{\text{cov}}\,\frac{1}{d}\big\|\operatorname{offdiag}\operatorname{Cov}(\tilde{\mathbf{Y}})\big\|_F^2,\end{aligned} \tag{3}$$

where $\tilde{\mathbf{Y}} = \sqrt{d}\cdot\operatorname{normalize}(\mathbf{Y})$ puts the predictions at unit RMS per dimension, σ_j is the standard deviation of dimension j across the N predictions, $\alpha_{\text{var}} = 1$ and $\alpha_{\text{cov}} = 0.04$. Section 3 quantifies what each term removes: centring and leading-direction removal eliminate the constant-prediction shortcut, and the two regularisers keep the predictions from degenerating.

4.3 Training objective and inference

The objective is

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda\,\mathcal{L}_{\text{JEPA}}, \tag{4}$$

where $\mathcal{L}_{\text{CE}}$ is the usual next-token cross-entropy on answer tokens, computed on the same masked sequence, and $\lambda = 0.1$. Only the head h_j and the `[MASK]` vector are added to the backbone; at inference both are discarded and no slice is masked, so the deployed model is the unmodified backbone with the unmodified inference cost.

No leakage. The masked content cannot reach the predictor through the input: the encoder attends within a slice, so no visible slice's features contain pixels of a masked one, and a masked slot holds only the constant $\mathbf{m}$. We verify the architectural property with tiny implementations of three Qwen-VL encoder families: perturbing future or hidden pixels leaves every earlier encoder and LM output bit-identical, while later outputs change (Appendix A).

5 Experiments

5.1 Setup

We compare Qwen3.5-0.8B-Base with a frozen vision tower under schedule-matched SFT and InnerJEPA training. SFT uses answer-token cross-entropy; InnerJEPA adds the masked latent objective with $\lambda = 0.1$. Both draw from the same mixture of LLaVA-Video-178K, MSR-VTT, OpenVid, ActivityNet VideoInstruct, and InterVid. Collapse and hyper-parameter ablations also use 0.8B models.

Table 1: **Clean-video QA and temporal-reasoning accuracy (%).** All rows use Qwen3.5-0.8B. SFT and InnerJEPA are schedule matched; Base and Instruct are same-scale reference checkpoints. The final column is the unweighted mean of four benchmarks.

Method	MVBench ($n = 4000$)	TempCompass ($n = 7540$)	NExT-QA ($n = 8564$)	STAR ($n = 7377$)	Mean
Base	50.65	60.16	62.00	51.72	56.13
Instruct	51.38	60.00	64.00	52.76	57.04
SFT	55.00	61.92	64.91	54.59	59.11
InnerJEPA	**56.07**	**63.64**	**67.36**	**57.68**	**61.19**
Δ vs. SFT	+1.07	+1.72	+2.45	+3.09	+2.09

Table 2: **Temporal grounding mIoU (%).** InnerJEPA improves four of five benchmarks, with gains on Charades, YouCook2, QVHighlights, and STVG; ActivityNet changes by only −0.45 points.

Method	Charades-TL ($n = 3363$)	YouCook2 ($n = 2994$)	ActivityNet-TL ($n = 4500$)	QVHighlights-TL ($n = 1541$)	STVG ($n = 2000$)	Mean
SFT	40.91	34.08	**41.59**	55.53	23.07	39.04
InnerJEPA	**45.12**	**36.10**	41.14	**56.70**	**27.80**	**41.37**
Δ	+4.21	+2.02	-0.45	+1.17	+4.73	+2.34

QA accuracy is evaluated on MVBench, TempCompass, NExT-QA, and STAR; grounding uses mIoU on the TimeLens variants of Charades-STA, YouCook2, ActivityNet, and QVHighlights, plus STVG. Paired runs share the scorer and evaluation protocol. Scores average three evaluation seeds for one fixed checkpoint per arm; table headers give scored sample counts. These averages do not measure variation across independent training runs. Prediction geometry is measured by entropy effective rank and mean pairwise cosine. Implementation settings and additional evaluation details appear in the appendix.

5.2 Main results

InnerJEPA improves all four QA benchmarks (Table 1), with gains of 1.07, 1.72, 2.45, and 3.09 points and an unweighted mean gain of 2.09 points. Grounding improves on four of five sets (Table 2), averaging +2.34 mIoU; ActivityNet decreases by 0.45 points. Thus, the QA gains coexist with improved aggregate localisation, although gains are not uniform across tasks. The separate Qwen3.5-2B InnerJEPA model improves six of seven reported columns over its 2B SFT reference, but STAR declines by 3.62 points. Its four-QA mean changes by −0.09 and its three-grounding mean by +1.28 (Table 3). These results are separate scaling evidence. Public compact-model scores are reported as protocol-dependent context in Table 11; neither table is part of the controlled 0.8B ablation.

5.3 Collapse ablation

At step 3000, cosine-only and no-variance predictions remain nearly parallel. Removing covariance yields low pairwise cosine but lower effective rank; omitting target centring retains a stronger shared direction. The full objective achieves rank 253.84 and cosine 0.179 at this matched step, reaching rank 332 and cosine 0.160 at step 5000. Thus, the abstract's endpoint is distinct from the step-matched ablation. Figure 2 shows the trajectories; geometry changes alone do not establish downstream usefulness.

Table 3: **Additional Qwen3.5-2B results.** QA columns report accuracy (%) and TL columns report mIoU (%). These runs are not paired with the main SFT/InnerJEPA comparison.

2B model	MVBench	TempCompass	NExT-QA	STAR	Charades-TL	ActivityNet-TL	QVHighlights-TL
Base	59.00	68.73	74.31	61.40	36.03	36.18	49.17
Instruct	59.03	69.93	73.56	64.36	37.55	39.53	50.35
SFT	59.20	70.97	74.32	66.99	43.64	47.14	57.57
InnerJEPA	59.70	71.67	76.37	63.37	46.50	47.30	58.40

Table 4: **Step-matched collapse diagnostics on Qwen3.5-0.8B.** The complete objective is the ablation reference; all variants use the same model scale. Every row is read at step 3000. Effective rank is measured in the 1024-dimensional prediction space.

Variant	Effective rank	Rank / 1024	Mean pairwise cosine
Cosine only	14.44	1.41%	0.908
Without variance	16.00	1.56%	0.964
Without covariance	129.85	12.68%	**0.109**
Without target centring	243.90	23.82%	0.301
Full objective	**253.84**	**24.79%**	0.179

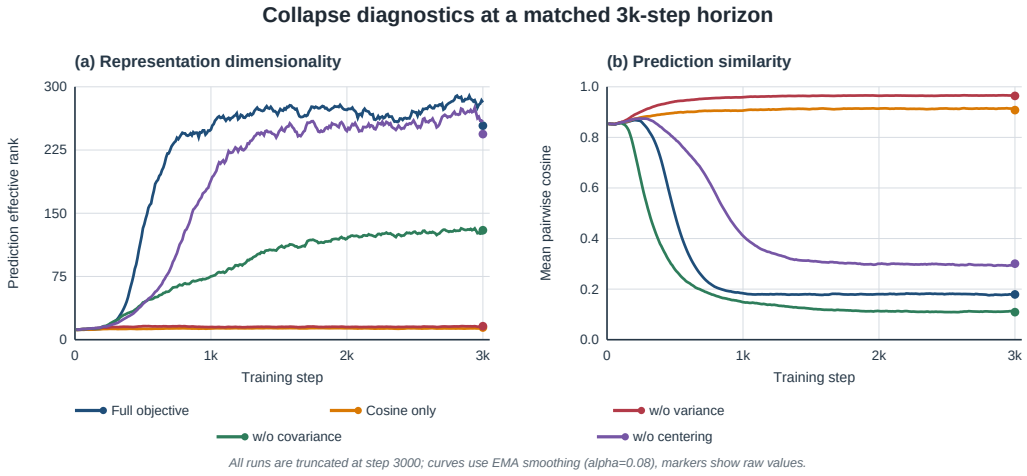


Figure 2: **Collapse diagnostics over training.** The full objective is the 0.8B reference and every ablation uses the same model scale. All curves are truncated at step 3000 and use EMA smoothing with coefficient 0.08; endpoint markers show unsmoothed values. Appended restart segments in the JSONL logs are excluded by selecting the longest monotonic step segment. Matching the observation step does not remove possible learning-rate-horizon differences, so these curves diagnose geometry rather than downstream quality.

Table 5: **Available downstream sensitivity results for 0.8B ablations (%).** QA columns report accuracy and temporal-localisation columns report mIoU. The best observed value in each column is bold.

Variant	STAR	MVBench	Charades-TL	ActivityNet-TL	QVHighlights-TL	YouCook2-TL
Mask ratio 0.1	**57.68**	53.89	41.4	**41.5**	**56.7**	31.8
$\lambda = 0.5$	57.64	53.71	**41.7**	41.0	56.0	**36.1**
$\lambda = 0.05$	57.11	53.97	41.1	40.7	55.5	31.9
Mask ratio 0.5	57.11	54.19	40.7	39.8	55.9	35.2
Grounding-oriented	55.87	54.34	40.9	40.0	55.5	27.2
Full objective	**57.68**	**54.50**	40.7	38.6	54.2	22.7

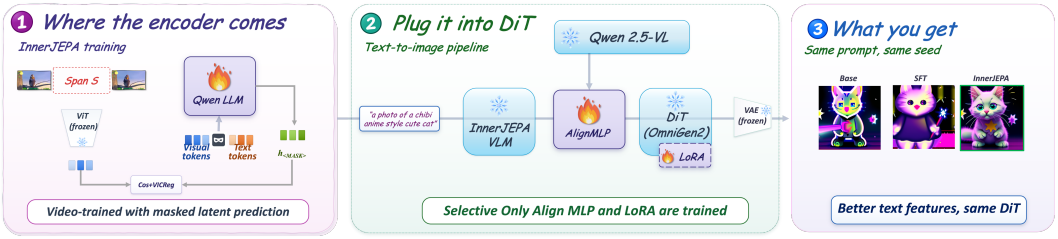


Figure 3: **Downstream text-to-image transfer pipeline.** Video training first produces the upstream VLM representation. For Base, SFT, and InnerJEPA alike, an identically configured `align_mlp_real` maps text features into the Qwen2.5-VL conditioning space expected by OmniGen2, while rank-8 LoRA adapts the same DiT attention projections. The source VLM and VAE remain frozen, and all adapter, optimisation, sampling, prompt, and seed settings are matched. Thus, the upstream VLM representation is the only experimental variable.

5.4 Masking and loss-weight sensitivity

Table 5 reports 0.8B variants with available, imperfectly matched training endpoints. No setting dominates: mask ratio 0.1 is strongest on ActivityNet-TL and QVHighlights-TL, while $\lambda = 0.5$ leads on Charades-TL and YouCook2-TL. The full objective leads on MVBench but is weaker on grounding at its earlier endpoint. These are sensitivity results rather than a controlled ranking.

5.5 Text-to-image transfer

For Base, SFT, and InnerJEPA, the same three-layer `align_mlp_real` maps text features into OmniGen2's Qwen2.5-VL conditioning space, with rank-8 LoRA on the same DiT attention projections. All branches share adaptation data, a 2500-step budget, optimiser, sampling, guidance, resolution, prompts, and seeds; source VLMs and VAEs are frozen. Only the upstream checkpoint changes, with adapters trained separately under this matched protocol (Figure 3).

On 553 images generated from GenEval prompts, CLIP score improves from 0.2329 for SFT to 0.2449 for InnerJEPA (+5.15% relative; Table 6). These are CLIP alignment scores, not GenEval compositional-accuracy scores. A fixed 20-prompt subset supplies detailed/short-prompt comparisons and blinded ratings of semantic fidelity, style adherence, and visual quality on a 1–5 scale. Two independently randomised A/B/C orderings yield 14 wins, five ties, and one loss against SFT. This session-fixed evaluator is an exploratory judge, not a versioned external API. Matched examples (Figure 4) show richer local texture and more distinct fine structures in several categories. Such examples and the small blinded pilot support the automatic result but do not establish universal image-quality gains or pixel-level world modelling.

Table 6: **Text-to-image transfer results.** CLIP uses ViT-L/14. The GenEval row covers 553 generated images; detailed/short-prompt CLIP and judge rows use the matched 20-prompt set. Judge scores use a 1–5 scale averaged over two independently order-randomised passes. Every branch uses the same `align_mlp_real`, rank-8 DiT LoRA, 2500-step adaptation budget, and generation configuration; only the upstream VLM checkpoint differs. Win/tie/loss compares InnerJEPA against SFT using a 0.15 overall-score margin.

Metric	Base+DiT	SFT+DiT	InnerJEPA+DiT
CLIP score (GenEval, $n = 553$)	0.2308	0.2329	**0.2449**
CLIP score (detailed prompts)	0.1976	0.2074	**0.2270**
CLIP score (short prompts)	0.1846	0.1891	**0.2145**
Semantic fidelity	3.44	3.43	**3.83**
Style adherence	3.39	3.36	**3.62**
Visual quality / artefacts	3.53	3.50	**3.92**
Overall mean	3.45	3.43	**3.79**
Win / tie / loss vs. SFT	–	–	**14 / 5 / 1**

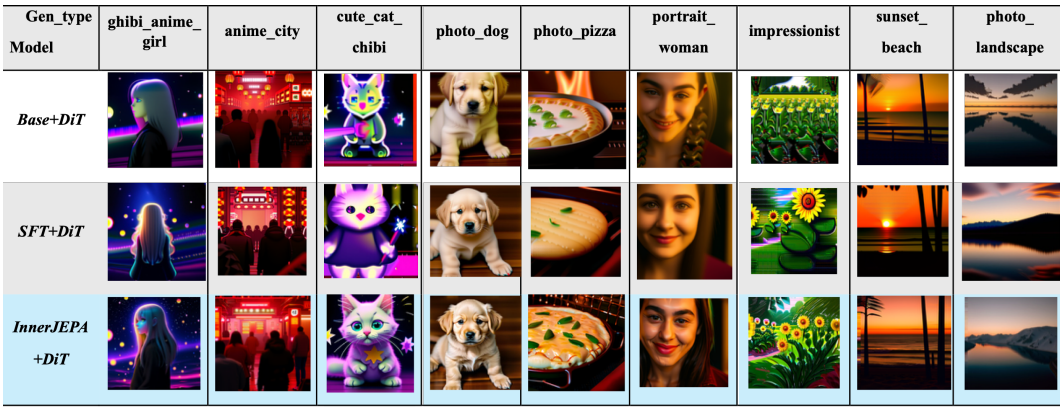


Figure 4: **Matched-prompt text-to-image comparison across nine content and style categories.** Columns correspond to prompt categories and rows compare Base+DiT, SFT+DiT, and InnerJEPA+DiT under matched prompts and random seeds. All three systems use the same `align_mlp_real` and rank-8 DiT-LoRA adaptation settings. Relative to the two controls, visual-assisted training generally yields richer local texture and more coherent fine detail, particularly in fur, facial features, food toppings, foliage, and scene boundaries, while preserving global prompt semantics. These samples come from the fixed 20-prompt GenEval subset and provide qualitative evidence; quantitative conclusions use the complete 553-image GenEval CLIP evaluation, with the identity-blinded judge retained on the 20-prompt subset.

Table 7: **Frame-order sensitivity (accuracy, %).** Deltas are measured relative to the same checkpoint under original frame order.

Model	Original	Shuffled	Reversed	Δ shuffled	Δ reversed
SFT	52.11	49.57	40.00	−2.54	−12.11
InnerJEPA	**54.30**	**54.49**	**49.41**	+0.19	−4.89

5.6 Frame-order sensitivity

We compare original, shuffled, and reversed frames on the same evaluation subset (Table 7). InnerJEPA exceeds SFT by 2.19, 4.92, and 9.41 points, respectively. Its reversal penalty is 4.89 points versus SFT's 12.11, supporting robustness to this perturbation. Shuffling changes its score by only +0.19 points, so the result does not establish fine-grained order dependence. Direction affects predictions, but appearance and unordered event content may still explain part of the gains.

6 Discussion and Limitations

The geometry diagnostics establish that predictions become more diverse; downstream gains establish their usefulness. Neither alone proves temporal reasoning. Frozen slice features capture appearance and coarse events but do not explicitly supervise event boundaries, and the unchanged shuffled-frame score limits claims about fine-grained order dependence. The frozen encoder also bounds the visual distinctions available to the objective.

The main comparison uses one 0.8B backbone and one data mixture. Scores average three evaluation seeds for one fixed checkpoint per arm, not independent training runs. The separate 2B results and public model comparisons are supporting context. Downstream sensitivity runs have unmatched endpoints. Generation uses identically configured but separately trained adapters; its 20-prompt blinded pilot uses a session-fixed evaluator rather than a versioned external API. Thus, generative transfer is diagnostic evidence, and broader training-seed, temporal-order, and generation evaluations remain necessary.

7 Conclusion

InnerJEPA reuses a video MLLM's frozen vision encoder and causal language model for masked latent prediction. Target conditioning and variance–covariance regularisation reduce collapse without inference overhead. The controlled 0.8B comparison improves mean QA accuracy by 2.09 points and grounding mIoU by 2.34 points over SFT. Matched DiT adaptation improves CLIP alignment on 553 GenEval generations, supporting generative transfer alongside task-dependent video-understanding gains.

Ethics Statement

This work uses existing video datasets and introduces no new human-subject data collection. Web videos and pretrained models may carry privacy risks, social biases, and copyrighted content; generated images may reproduce stereotypes or enable misleading media. Applications should respect dataset licences, privacy, and copyright, and disclose synthetic content. Our benchmarks and exploratory judge evaluation do not establish deployment safety.

AI Use Statement

Generative AI assisted manuscript organisation, literature discovery, drafting, language editing, and code development, and served as the disclosed pilot evaluator. The authors remain responsible for all claims, results, citations, and the final submission.

References

Ashutosh Adhikari and Mirella Lapata. Multimodal latent reasoning via predictive embeddings. *arXiv preprint arXiv:2604.08065*, 2026.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. *arXiv preprint arXiv:2301.08243*, 2023.

Mahmoud Assran et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Shuai Bai et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*, 2022.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Davide Caffagni, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, Pier Luigi Dovesi, Shaghayegh Roohi, Mark Granroth-Wilding, and Rita Cucchiara. Seeing beyond words: Self-supervised visual learning for multimodal large language models. *arXiv preprint arXiv:2512.15885*, 2025.

Feng Cheng et al. Vindlu: A recipe for effective video-and-language pretraining. *arXiv preprint arXiv:2212.05051*, 2022.

Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, et al. PerceptionLM: Open-access data and models for detailed visual understanding. *arXiv preprint arXiv:2504.13180*, 2025.

Tsu-Jui Fu et al. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021.

Hai Huang et al. Llm-jepa: Large language models meet joint embedding predictive architectures. *arXiv preprint arXiv:2509.14252*, 2025.

Hugging Face. SmolVLM2: Bringing video understanding to every device. https://huggingface.co/HuggingFaceTB/SmolVLM2-2.2B-Instruct, 2025. Model card, accessed September 2026.

Jitesh Jain et al. Elevating visual perception in multimodal llms with visual embedding distillation. *arXiv preprint arXiv:2412.09585*, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Kunchang Li et al. Unmasked teacher: Towards training-efficient video foundation models. *arXiv preprint arXiv:2303.16058*, 2023.

Linjie Li et al. Lavender: Unifying video-language understanding as masked language modeling. *arXiv preprint arXiv:2206.07160*, 2022.

Xianhang Li, Chen Huang, Chun-Liang Li, Eran Malach, Josh Susskind, Vimal Thilak, and Etai Littwin. Rethinking jepa: Compute-efficient video ssl with frozen teachers. 2026. arXiv:2509.24317.

Orca Team, Beijing Academy of Artificial Intelligence. Orca: The world is in your mind. *arXiv preprint arXiv:2606.30534*, 2026.

Dhruv Parikh, Jacob Fein-Ashley, Rajgopal Kannan, and Viktor Prasanna. Latent denoising improves visual alignment in large multimodal models. *arXiv preprint arXiv:2604.21343*, 2026.

Biswa Sengupta. Representation without reward: A jepa audit for llm fine-tuning. *arXiv preprint arXiv:2605.15394*, 2026.

Abdelrahman Shaker, Muhammad Maaz, Chenhui Gou, Hamid Rezatofighi, Salman Khan, and Fahad Shahbaz Khan. Mobile-VideoGPT: Fast and accurate model for mobile video understanding. *arXiv preprint arXiv:2503.21782*, 2025.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022.

Haochen Wang et al. Reconstructive visual instruction tuning. In *ICLR*, 2025a. arXiv:2410.09575.

Haochen Wang et al. Ross3d: Reconstructive visual instruction tuning with 3d-awareness. *arXiv preprint arXiv:2504.01901*, 2025b.

Chen Wei et al. Masked feature prediction for self-supervised visual pre-training. *arXiv preprint arXiv:2112.09133*, 2021.

Sihan Xu et al. Next-embedding prediction makes strong vision learners. *arXiv preprint arXiv:2512.16922*, 2025.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mPLUG-Owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024.

Heeji Yoon, Jaewoo Jung, Junwan Kim, Hyungyu Choi, Heeseong Shin, Sangbeom Lim, Honggyu An, Chaehyun Kim, Jisang Han, Donghyun Kim, Chanho Eom, et al. Visual representation alignment for multimodal large language models. *arXiv preprint arXiv:2509.07979*, 2025.

XiuYu Zhang, Junfeng Fang, and Zhenkai Liang. Cosine misleads: Auxiliary losses reshape vision language models, not their latents. *arXiv preprint arXiv:2606.05753*, 2026.

Yuanhan Zhang et al. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

A No-leakage Assertion

Table 8 reports the empirical assertion of Section 4.3 on tiny randomly initialized models of the three encoder families.

Table 8: Empirical no-leakage check on tiny randomly initialized models of each family (transformers 4.57.1 and 5.x). Future/masked-frame pixels are perturbed; "past" columns must be exactly zero, "future" columns must be non-zero (otherwise the perturbation did not reach the model).

Family	encoder past max $\lvert\Delta\rvert$	encoder future	LM past max $\lvert\Delta\rvert$	LM future
qwen3_vl	0.00e+00	1.38e-01	0.00e+00	4.96e+00
qwen2_5_vl	0.00e+00	1.32e-01	0.00e+00	4.42e+00
qwen3_5	0.00e+00	1.18e-01	0.00e+00	4.84e+00

B Statistics

The main 0.8B tables average three evaluation seeds for one fixed checkpoint per arm. The available per-example JSON archive belongs to a separate 2B experiment and is deliberately excluded from the 0.8B analysis. Consequently, the reported three-seed averages characterise evaluation-time variation but do not estimate uncertainty across independently trained checkpoints; we do not report paired significance tests without the matching 0.8B per-example outputs.

C Data Statistics

The training mixture draws from LLaVA-Video-178K, MSR-VTT, OpenVid, ActivityNet VideoInstruct, and InterVid. Evaluation uses the benchmark-specific files counted in Tables 1 and 2. We report only counts verified by the evaluation records and do not infer per-source training counts from dataset names.

D Hyper-parameters and Calibration

Table 9: Method-specific hyper-parameters fixed by the implementation.

Component	Setting
Vision tower	frozen
Video sampling	2 fps
Temporal patch size	2 frames (one-second slice)
Mask ratios	$\{5, 15, 30, 50\}\%$ sampled per video
Mask constraint	first slice always visible
Corruption location	after ViT and merger, before the LM
Prediction head	shared Linear–GELU–Linear MLP
Target gradient	stop-gradient
Target processing	batch centring and removal of PC_1
Similarity term	$1 - \cos(\mathbf{y}, \hat{\mathbf{z}})$
$\alpha_{\mathrm{var}}, \alpha_{\mathrm{cov}}$	1.0, 0.04
Auxiliary weight λ	0.1
Inference	no masks or auxiliary head

E Extended Comparisons and Transfer Protocol

Figure 5 contrasts the use of existing MLLM components with auxiliary architectures. Table 10 summarises related methods.

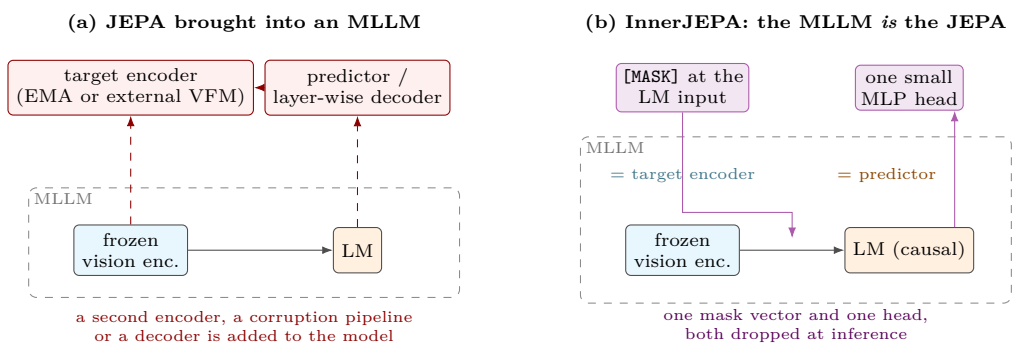


Figure 5: **A video MLLM already contains both halves of a joint-embedding predictive architecture.** (a) Existing instantiations often attach a target encoder, teacher, or decoder. (b) InnerJEPA uses the frozen vision encoder as the target encoder and the causal language model as the predictor, adding only a mask vector and a small MLP head during training.

Table 10: **Positioning.** Representative latent-prediction methods differ in target space, read-out, and training regime. InnerJEPA adds only a mask vector and a head, both removed at inference.

Method	Target space	Loss	Read-out	Added to the model	Infer. cost	Regime
I-JEPA / V-JEPA	own (EMA)	regression	patch	enc.+pred.+EMA	–	pre-train
NEPA	own (co-trained)	cosine regr.	patch	enc.+pred.	–	pre-train
Orca	frozen ViT	regression	query	queries + decoders	0	cont. pre-train
Ross3D / Latent Denoising	frozen VAE / CLIP	denoising / regression	mask pos. (mid-layer)	denoiser / decoder + teacher	0	SFT
PEARL	own hidden	SmoothL1	query	[PRED] tokens + head	0	SFT
InnerJEPA	frozen ViT (own)	$1 - \cos$ + var/cov	masked slot (last layer)	**mask vector + head**	0	SFT

E.1 Additional model comparisons

Table 3 includes InnerJEPA trained from Qwen3.5-2B. Relative to its 2B SFT reference, it improves MVBench, TempCompass, NExT-QA, Charades-TL, ActivityNet-TL, and QVHighlights-TL by 0.50, 0.70, 2.05, 2.86, 0.16, and 0.83 points, respectively; STAR declines by 3.62 points. The four-QA mean is nearly unchanged (-0.09), while mean grounding increases by 1.28 points. These runs are independent of the main 0.8B comparison.

Public model scores below retain their reported protocols and are contextual, not controlled baselines. Training data, frame sampling, prompts, and scoring differ; PLM-1B supplies a stronger reference alongside compact alternatives.

E.2 Generative transfer pipeline

All three upstream checkpoints use the same `align_mlp_real`, rank-8 DiT-LoRA adaptation, and training and generation settings described in Section 5.5. The source VLM

Table 11: **Public comparison at a similar model scale.** Values are taken from the corresponding papers or official model cards (Hugging Face, 2025; Li et al., 2024; Shaker et al., 2025; Ye et al., 2024; Cho et al., 2025). Total-parameter conventions and evaluation protocols differ across reports. Dashes denote unreported entries.

Model	Reported scale	MVBench	NExT-QA	STAR	TempCompass
SmolVLM2-500M	0.5B	39.73	–	–	–
LLaVA-OneVision-0.5B	1.0B total	47.2	57.2	–	–
mPLUG-Owl3	1B	47.1	65.3	–	–
Mobile-VideoGPT-0.5B	0.6B total	53.5	65.4	–	–
InnerJEPA	0.8B	56.07	67.36	57.68	63.64
PLM-1B	1B	70.1	80.3	83.7	64.6

and VAE remain frozen. AlignMLP reaches approximately 0.96 cosine alignment for all branches, which does not imply equivalent generation quality. Automatic CLIP evaluation uses 553 GenEval generations; detailed/short-prompt and blinded analyses retain a fixed 20-prompt subset. Per-prompt scores and anonymisation manifests support re-evaluation.