PREDICTION CAN BENEFIT UNDERSTANDING: VISUAL PREDICTION AS COMPLEMENTARY SUPERVISION FOR VIDEO LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning to predict how the visual world evolves is a powerful source of self-supervision, yielding strong video representations and world models that support planning and control. Video large language models (video LLMs), however, are trained almost entirely with language supervision; in the standard recipe, the video serves only as input and is never itself predicted. Prior work has unified visual generation and understanding in single models, yet whether learning to predict the visual world benefits understanding has not been systematically reported. In this work, we systematically investigate how visual prediction and understanding interact within a video LLM, asking whether learning to predict the visual world improves understanding. During model training, we remove a segment from a video and train the model to predict the visual features of that segment from the rest of the video and a description of the segment. A counterpart trained on the same data with language supervision alone serves as the control. At all three scales, from 0.8B to 4B, learning to predict improves every video understanding and temporal grounding benchmark we evaluate, with gains at 0.8B and 2B of up to 3.3 points on MVBench, 3.3 points on TempCompass and 3.4 points in mean grounding mIoU. Controlled ablations at multiple levels confirm that the gain is robust across design choices and data sources. These results show that visual prediction can serve as a general supervision paradigm for video LLMs, complementing language with a signal drawn from the visual world itself.

1 INTRODUCTION

Advances in visual generation and world models have made learning to predict the visual world one of the most powerful forms of self-supervision (Bruce et al., 2024). In self-supervised video representation learning, masked video modelling and joint-embedding prediction train models to predict masked or future video content, and the resulting representations support video understanding and anticipation (Tong et al., 2022; Assran et al., 2025). The same predictive knowledge can even drive decisions, letting agents plan by imagining the outcomes of their actions (Hafner et al., 2025) and improving robot manipulation (Wu et al., 2024). Visual prediction is thus a broadly useful training signal, and because its target is the video itself, it reaches visual content that no annotation describes.

Video large language models (video LLMs) have advanced rapidly in video understanding (Qwen Team, 2025; Zhang et al., 2024). Their training signal, however, is almost entirely linguistic. In the prevailing recipe, from multimodal pre-training to post-training, the loss is computed on text tokens only, and the video enters solely as context for predicting a textual response (Zhu et al., 2025; Wang et al., 2025b). Visual processing is therefore supervised only indirectly, through the text. Evaluations have also found that video LLMs perceive temporal change poorly (Liu et al., 2024b) and trail humans by a wide margin on fine-grained temporal questions (Cai et al., 2024), which suggests that this language-only training paradigm still leaves considerable room for improvement.

Whether visual prediction can improve understanding has not been studied sufficiently. Unified models already perform visual understanding and generation within one network, for images and increasingly for video (Xie et al., 2025; Deng et al., 2025; Wu et al., 2025), and other works add

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.

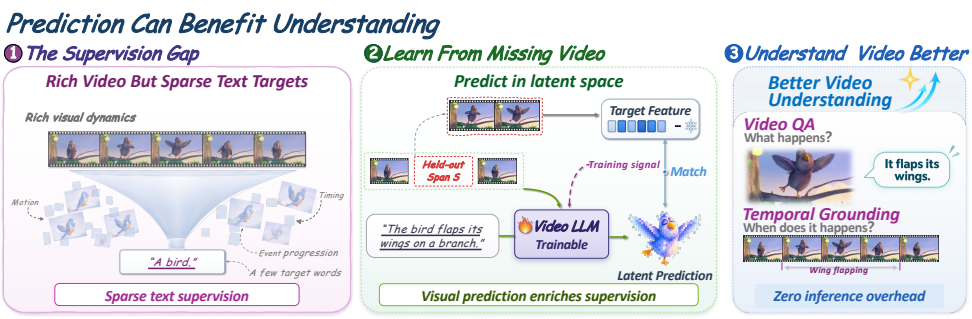

Figure 1: **Prediction can benefit understanding.** Text targets cover only a few words of a rich video. Predicting the latent features of a held-out span from the remaining video and its description adds a visual training signal, which improves video question answering and temporal grounding with no inference overhead.

feature-reconstruction or prediction objectives during training (Wang et al., 2025b; Yang et al., 2025; Zhu et al., 2026). These works show that one model can be trained to acquire both abilities, but how the two abilities relate to and influence each other has not been examined systematically. For video LLMs, whether learning to predict the visual world improves understanding is therefore an important question that calls for systematic study.

In this paper, we systematically study whether visual prediction improves understanding in video LLMs. The study spans three model scales and covers both general video question answering and temporal grounding, comparing each model with a language-only counterpart trained on the same data. **Our central finding is that prediction can benefit understanding.** Supplementing training with a visual prediction loss improves video understanding for models of every size we study; at 0.8B and 2B it raises MVBench by up to 3.3 points and TempCompass by up to 3.3 points, while temporal grounding improves by up to 3.4 points in mean mIoU. Our models also match or surpass published video LLMs built on larger backbones, often more than twice our size, and further tuned with specialised post-training. Visual prediction thus complements language as a source of supervision for video LLMs.

Our contributions are threefold. **(i)** To our knowledge, we present the first systematic study of whether visual prediction improves understanding in video LLMs, with controlled comparisons across model scales on both general understanding and temporal grounding. **(ii)** We propose a training paradigm that turns general video–language data into a visual prediction objective, so that one model learns from language targets and visual targets at once. **(iii)** The paradigm adds no module to the model and leaves inference unchanged, and ablations at multiple levels demonstrate its effectiveness. More broadly, these results point toward unifying video understanding and world modelling within a single model.

2 Preliminaries

A video–language example consists of a video V and text y about it, such as a question with its answer, a caption, or a description of an event with the span it occupies. A frozen visual encoder maps V to visual tokens $\mathbf{z} = f_v(V)$, and a causal language model g_θ processes the visual tokens followed by the text.

Language supervision. The text loss treats the video as context and the text as the prediction target,

$$\mathcal{L}_{\mathrm{CE}} = -\sum_i \log g_\theta\big(y_i \,\big|\, \mathbf{z},\, y_{<i}\big), \tag{1}$$

summed over the target positions. The visual tokens receive gradient only through the words they help predict.

Visual prediction supervision. Visual predictive learning instead hides a part $\mathcal{S}$ of the visual input and predicts its representation from the available context c,

$$\mathcal{L}_{\mathrm{pred}} = \ell\big(\hat{\mathbf{z}}_{\mathcal{S}}(c;\theta),\ \mathrm{sg}(\mathbf{z}_{\mathcal{S}})\big), \tag{2}$$

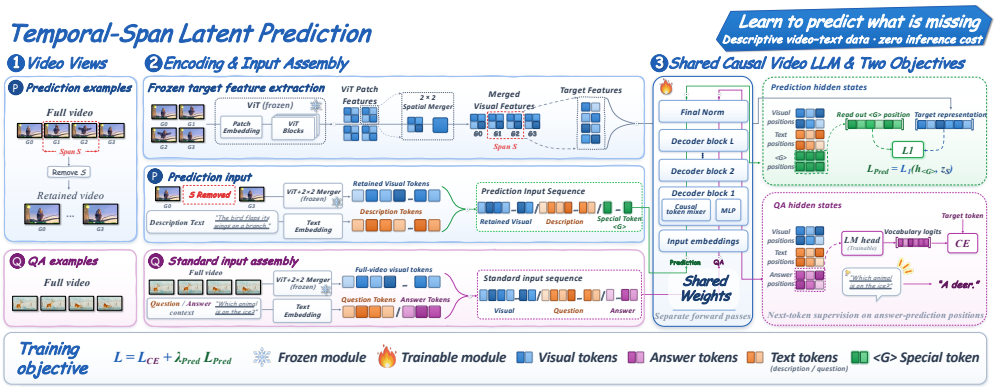

Figure 2: **Temporal-span latent prediction.** The same LLM is trained with cross-entropy on question answering and, on prediction examples, predicts the frozen-encoder features of a removed span $\mathcal{S}$ at special tokens <G>, one per missing visual token, read out at half depth.

where $\mathbf{z}_{\mathcal{S}}$ are the frozen features of the hidden part, ℓ a regression loss and sg a stop-gradient.

3 Temporal-Span Latent Prediction

Language supervision trains the model to predict the next text token (Equation 1). We pair it with a visual counterpart, a continuous temporal-span visual latent prediction objective in which the model predicts the latent representation of a removed span of the video from the rest of the video and the language that describes it (Figure 2). Where masked modelling reconstructs discrete visual tokens or random patches, the unit of prediction here is a continuous span of the video, and the objective is used only during training.

3.1 Video Representation and Temporal Groups

We keep the video encoding of the backbone. The frozen visual encoder and merger map the video to visual tokens in the hidden space of the LLM, of dimension d, organised into temporal groups G_t that each cover a short interval of the video, and we write $\mathbf{z}_t$ for the tokens of G_t. Groups are the temporal unit of prediction, so a target span $\mathcal{S}$ is a contiguous set of groups, and M denotes the number of visual tokens it contains.

3.2 Training Examples and Input Construction

Both kinds of training example pass through the same model with different inputs. Language-supervised examples keep their usual format and contribute only the text loss of Equation 1, so the descriptions also remain language targets as in the baseline. A prediction example provides a description q and the span $\mathcal{S}$ it refers to. The content of $\mathcal{S}$ is removed before visual encoding, and the input to the LLM is

$$[\mathbf{z}_t]_{G_t \notin \mathcal{S}} \quad q \quad \texttt{<G>}_1 \cdots \texttt{<G>}_M, \tag{3}$$

with M prediction tokens appended after the description. The prediction tokens share one learnable embedding and are placed at the positions of the missing visual tokens, so that $\texttt{<G>}_m$ stands for the m-th of them. The description need not come from temporal grounding annotations; a caption of the video can take its place, with the span to remove sampled at random (Section 4.6).

3.3 Prediction, Target and Loss

Prediction. We read the hidden states of the M prediction tokens from a hidden layer of the LLM, $\mathbf{H}_{\texttt{<G>}} = [\mathbf{h}_1, \ldots, \mathbf{h}_M] \in \mathbb{R}^{M \times d}$, where $\mathbf{h}_m$ is the prediction of the m-th missing visual token. The prediction is read directly from the LLM, with no prediction head.

Target. The target comes from the model's own frozen visual modules. We encode the complete video with the frozen encoder and merger, take the tokens of $\mathcal{S}$ and order them as the prediction

tokens are ordered, which gives $\mathbf{Z}_{\mathcal{S}} = [\mathbf{z}_1, \ldots, \mathbf{z}_M] \in \mathbb{R}^{M \times d}$ in one-to-one correspondence with $\mathbf{H}_{<\text{G}>}$. Because the encoder attends only within a temporal group and $\mathcal{S}$ is removed before encoding, no information passes between the target and the retained input (Appendix E).

Loss. Predictions and targets are each normalised with layer normalisation over the feature dimension, and the loss is their ℓ_1 distance averaged over tokens and dimensions,

$$\mathcal{L}_{\text{pred}} = \frac{1}{Md} \sum_{m=1}^{M} \left\| \text{LN}(\mathbf{h}_m) - \text{LN}\big(\text{sg}(\mathbf{z}_m)\big) \right\|_1, \tag{4}$$

which instantiates Equation 2. Minimising $\mathcal{L}_{\text{pred}}$ requires the model to infer the visual content of the missing span from its temporal context and its description, which grounds the shared representation in what the video shows and strengthens the model's intrinsic understanding of the video. The text loss reaches the visual tokens only through the words of the target, whereas this objective supervises the visual content directly, including what the text leaves unsaid.

3.4 Joint training and inference

Language-supervised examples contribute the cross-entropy $\mathcal{L}_{\text{CE}}$, and prediction examples contribute $\mathcal{L}_{\text{pred}}$. Both kinds are sampled in the same training run, and their gradients update the shared model under

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{pred}} \, \mathcal{L}_{\text{pred}}, \tag{5}$$

where λ_{pred} weights the prediction term; we set $\lambda_{\text{pred}} = 1$ from the ablation in Section 4.4. Each example takes its own forward pass, and prediction examples replace language-modelling examples within a batch, so the objective adds no forward passes. The visual encoder and the merger stay frozen throughout, and only the LLM and the embedding of the prediction token are updated. At inference no content is removed and no prediction token is added, so the model keeps its original inference procedure and cost.

4 Experiments

We test whether visual latent prediction can act as supervision complementary to text generation and improve the video understanding of a video LLM. We train Qwen3.5 (Qwen Team, 2026) at 0.8B, 2B and 4B and evaluate general video understanding and temporal grounding. The main comparison pairs Base+SFT+Pred with Base+SFT at each scale. All ablations are run at 0.8B. The experiments address five questions. **RQ1**: does learning to predict visual latents improve the video understanding of the model? **RQ2**: which layer of the LLM should the visual prediction supervise? **RQ3**: how does the weight of the prediction loss affect joint training? **RQ4**: what does the language in the prediction view contribute to the gain? **RQ5**: does visual latent prediction extend to ordinary video–text data?

4.1 Setup

Backbones and arms. We train at three scales, Qwen3.5-0.8B, 2B and 4B, with the vision encoder and merger frozen and the LLM trainable. Each scale has four rows. The released **Base** and **Instruct** checkpoints, evaluated as they are, serve as reference points, and the other two rows are trained by us from the Base checkpoint with the same data, data order and schedule. **Base+SFT** uses cross-entropy only, and **Base+SFT+Pred** adds $\mathcal{L}_{\text{pred}}$ (Section 3). Both arms use the same global batch size of 128 and are evaluated at the same step, 9000, and hence after the same number of forward passes. Base+SFT+Pred fills one quarter of each batch with prediction examples; Base+SFT fills these slots with language-modelling samples, so the baseline receives about one third more text-supervised passes. The comparison thus holds compute fixed and, if anything, favours the baseline. The objective only needs a text loss; we study it in supervised fine-tuning (SFT) with learning rate 2×10^{-6} and at most 128 frames (Appendices D and G).

Evaluation. We evaluate video understanding on MVBench, TempCompass, NExT-QA and STAR (Li et al., 2023a; Liu et al., 2024b; Xiao et al., 2021; Wu et al., 2021), and temporal grounding on Charades-STA, YouCook2, ActivityNet-Captions, QVHighlights and HC-STVG-v2 (Gao et al.,

Table 1: **Main results** (%). Grey rows: released checkpoints. Subscripts: difference to Base+SFT. †: TimeLens split.

Model	Video understanding				Temporal grounding (mIoU)					
	MVBench	TempC.	NExT-QA	STAR	Charades†	YouCook2	ANet†	QVH†	HC-STVG	Avg
Qwen3.5-0.8B										
Qwen3.5-0.8B-Base	50.7	60.2	62.1	51.7	25.3	30.2	27.6	41.0	19.7	28.7
Qwen3.5-0.8B-Instruct	51.4	61.0	63.3	52.8	33.3	31.7	29.5	43.9	22.7	32.2
Base + SFT	56.9	61.9	76.9	69.0	45.1	53.0	41.6	54.4	23.1	43.4
Base + SFT + Pred	59.6 $_{\uparrow 2.7}$	65.2 $_{\uparrow 3.3}$	78.1 $_{\uparrow 1.2}$	72.0 $_{\uparrow 3.0}$	48.2 $_{\uparrow 3.1}$	55.0 $_{\uparrow 2.0}$	44.7 $_{\uparrow 3.1}$	57.5 $_{\uparrow 3.1}$	26.6 $_{\uparrow 3.5}$	46.4 $_{\uparrow 3.0}$
Qwen3.5-2B										
Qwen3.5-2B-Base	59.0	68.7	74.3	61.4	36.0	39.2	36.2	49.2	17.7	35.7
Qwen3.5-2B-Instruct	59.0	69.9	73.6	64.4	37.6	41.1	39.5	50.4	23.2	38.3
Base + SFT	61.7	71.0	79.9	74.0	47.6	57.0	48.5	56.4	23.5	46.6
Base + SFT + Pred	65.0 $_{\uparrow 3.3}$	72.4 $_{\uparrow 1.4}$	82.5 $_{\uparrow 2.6}$	76.6 $_{\uparrow 2.6}$	52.5 $_{\uparrow 4.9}$	62.2 $_{\uparrow 5.2}$	50.4 $_{\uparrow 1.9}$	59.3 $_{\uparrow 2.9}$	25.7 $_{\uparrow 2.2}$	50.0 $_{\uparrow 3.4}$
Qwen3.5-4B										
Qwen3.5-4B-Base	66.3	73.4	78.5	66.7	39.9	47.5	39.3	51.2	18.3	39.3
Qwen3.5-4B-Instruct	67.0	75.2	80.1	68.6	42.0	49.8	48.8	64.5	26.7	46.4
Base + SFT	67.9	75.8	83.2	72.1	42.2	49.8	39.7	52.3	20.6	40.9
Base + SFT + Pred	70.4 $_{\uparrow 2.5}$	77.3 $_{\uparrow 1.5}$	83.7 $_{\uparrow 0.5}$	74.3 $_{\uparrow 2.2}$	44.3 $_{\uparrow 2.1}$	50.3 $_{\uparrow 0.5}$	49.0 $_{\uparrow 9.3}$	59.0 $_{\uparrow 6.7}$	26.1 $_{\uparrow 5.5}$	45.7 $_{\uparrow 4.8}$

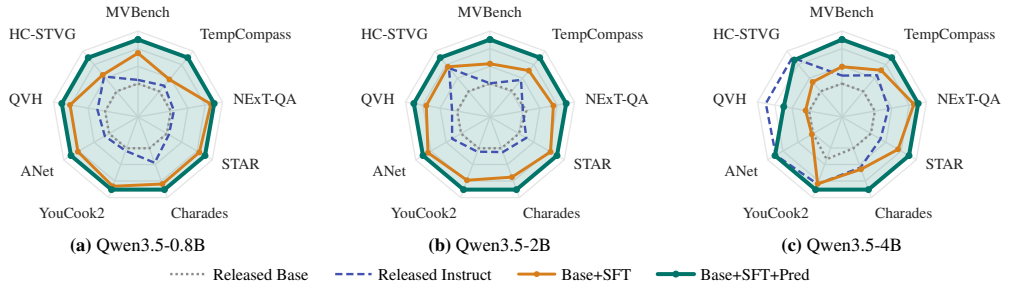

Figure 3: **Table 1 as radar charts.** Each axis has its own linear scale.

2017; Zhou et al., 2018; Krishna et al., 2017; Lei et al., 2021; Tang et al., 2021), using the TimeLens splits (Wang et al., 2026b) for Charades-STA, ActivityNet and QVHighlights. Grounding is reported as mIoU, with recall in Appendix B.

4.2 RQ1: Does learning to predict visual latents improve video understanding?

Learning to predict missing visual representations improves video understanding. The matched results in Table 1 and Figures 3 and 4 show that, once the visual latent prediction loss is added to the standard text generation objective, Base+SFT+Pred improves the corresponding standard SFT baseline on multiple downstream tasks at all three Qwen3.5 scales, 0.8B, 2B and 4B, with gains in both video question answering and temporal grounding. The benefit of visual latent prediction therefore reproduces across model capacities and is not confined to a particular scale.

The comparison is paired within each scale. All three scales use one data configuration, the same evaluation checkpoint at step 9000 and one evaluation protocol, with no setting tuned for a particular scale, so we measure the contribution of visual prediction mainly through the paired comparison between Base+SFT+Pred and Base+SFT at the same scale. Because all scales share one learning rate and data budget, the 4B model moves least from its base during SFT; its mean grounding mIoU rises by only 1.6 points over the released Base, against 10.9 at 2B and 14.7 at 0.8B, and stays below the released Instruct model. Even so, the prediction objective still improves the corresponding SFT baseline across the 4B benchmarks.

Comparison with published video LLMs. Beyond the gains over Base+SFT, our models are competitive with or better than published video LLMs that rely on specialised designs, reinforcement-learning post-training or larger backbones (Table 2). Our 2B model, for example, exceeds the publicly reported results of TimeLens-7B on all three TimeLens benchmarks. Visual prediction supervision therefore contributes substantially, enough for our models to rival larger models trained with more extensive post-training or reinforcement learning.

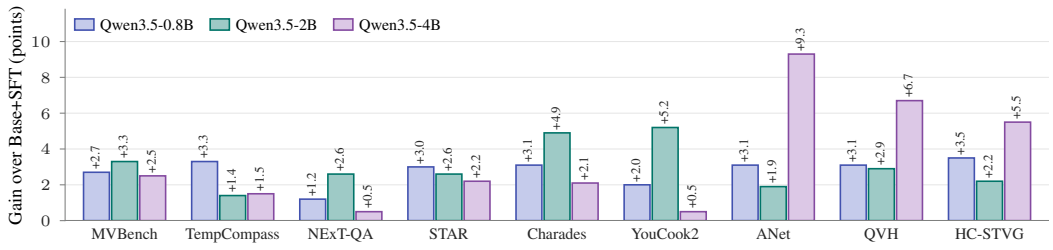

Figure 4: **Gain of Base+SFT+Pred over Base+SFT** on each benchmark (points).

Table 2: **Comparison with published video LLMs** (%). Published results are taken from the respective papers (Wang et al., 2025a; Yu et al., 2023; Yang et al., 2026a; Du et al., 2026a; Li et al., 2026; Shi et al., 2024; Feng et al., 2025; Hu et al., 2026; Jiang et al., 2025; Du et al., 2026b; Wang et al., 2026b; Liu et al., 2025b). †: TimeLens split. Best in **bold**, second underlined; "–": not reported. Appendix C extends the TempCompass comparison.

		Video understanding				Temporal grounding (mIoU)		
Method	Size	MVBench	TempC.	NExT-QA	STAR	Charades†	ANet†	QVH†
Grounded-VideoLLM	3.8B	60.0	–	–	–	30.0	30.0	33.4
SeViLA	4B	–	–	73.8	64.9	–	–	–
Mage-VL	4B	65.1	62.3	83.1	–	–	–	–
CRPO	4B	68.6	74.2	–	–	–	–	–
VideoChat3	4B	–	75.6	–	–	–	–	–
AoTD	7B	55.6	–	77.6	74.3	–	–	–
Video-R1	7B	63.9	73.2	76.5	–	–	–	–
PointThinker	7B	65.8	74.6	–	–	–	–	–
VideoP2R	7B	68.1	74.5	–	–	–	–	–
APPO	7B	64.6	–	79.2	–	–	–	–
TimeLens	7B	–	–	–	–	48.8	46.2	56.0
Oryx-1.5	32B	70.1	–	–	–	–	–	–
Ours (Qwen3.5-2B)	2B	65.0	72.4	82.5	**76.6**	**52.5**	**50.4**	**59.3**
Ours (Qwen3.5-4B)	4B	**70.4**	**77.3**	**83.7**	74.3	44.3	49.0	59.0

Prediction is a complementary learning signal. These results show that a prediction task targeting missing visual content can serve as an effective learning signal that complements text generation. With the same backbone and the same language-supervision data, requiring the model to also predict visual representations further improves its ability to answer video questions and to localise the relevant segments. Although the objective is used only during training, its benefit appears in the understanding tasks the model already performs; training needs no additional data, and inference needs no extra module or prediction step.

4.3 RQ2: Which layer should the visual prediction supervise?

The middle of the network is the most effective place to supervise. We read out the prediction tokens at depth $L/4$, $L/2$, $3L/4$ or at the last layer L, apply the same visual latent prediction objective, and keep every other training setting fixed. Among the four positions, $L/2$ gives the stronger results on most downstream tasks and the most balanced overall performance (Table 3), so we adopt it as the default. Although individual tasks peak at an earlier or a later position, the results show that the effect of visual prediction is sensitive to where the supervision is applied, and that this dependence is captured neither by "deeper is better" nor by "shallower is better".

A lower prediction loss does not mean a larger gain. To understand this further, we compare the training dynamics of the prediction objective at different depths. As Figure 5a shows, deeper read-outs reach a lower prediction loss for most of training and enter the low-loss regime earlier. This trend does not carry over to downstream performance; the deeper read-outs attain a lower prediction loss, yet their downstream results do not surpass those at $L/2$. Within the configurations compared, the level of the prediction loss alone is therefore not sufficient to judge the final downstream benefit, and the position at which the prediction supervision is applied matters as well.

This preference for the middle echoes PRe (Wang et al., 2026a), whose predictive regularisation of intermediate MLLM features works best at the middle layer and can hurt at the last, and related

Table 3: **Read-out depth** (Qwen3.5-0.8B). Subscripts: difference to Base+SFT. Avg: mean of the seven benchmarks.

Read-out	MVBench	TempC.	NExT-QA	STAR	Charades†	ANet†	QVH†	Avg
Base + SFT	56.9	61.9	76.9	69.0	45.1	41.6	54.4	58.0
1/4 depth	59.8 ↑2.9	64.3 ↑2.4	78.1 ↑1.2	72.3 ↑3.3	46.0 ↑0.9	44.5 ↑2.9	55.4 ↑1.0	60.1 ↑2.1
1/2 depth (ours)	59.6 ↑2.7	65.2 ↑3.3	78.1 ↑1.2	72.0 ↑3.0	48.2 ↑3.1	44.7 ↑3.1	57.5 ↑3.1	60.8 ↑2.8
3/4 depth	58.0 ↑1.1	63.2 ↑1.3	78.3 ↑1.4	72.0 ↑3.0	47.5 ↑2.4	43.5 ↑1.9	54.5 ↑0.1	59.6 ↑1.6
Last layer	58.5 ↑1.6	63.3 ↑1.4	78.4 ↑1.5	72.4 ↑3.4	47.7 ↑2.6	43.3 ↑1.7	55.2 ↑0.8	59.8 ↑1.8

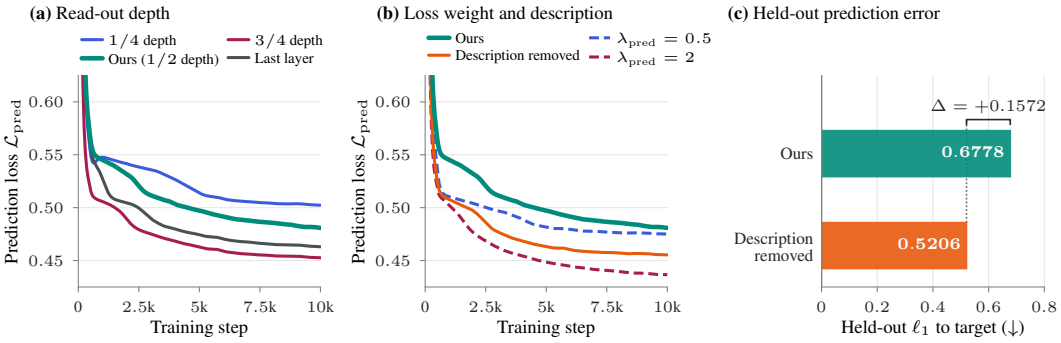

Figure 5: **Prediction loss and prediction error** (Qwen3.5-0.8B). (a, b) Smoothed training prediction loss over the first 10k steps; Ours: half depth, $\lambda_{\text{pred}} = 1$, with description. (c) ℓ_1 on 500 held-out Charades-TimeLens clips, without description; Δ: paired difference, 95% CI $[0.1552, 0.1592]$.

image-MLLM objectives (Yoon et al., 2025; Jain et al., 2024; Parikh et al., 2026); we observe it for the latents of a missing video span.

4.4 RQ3: How does the weight of the prediction loss matter?

A larger weight does not bring better understanding. We set the weight λ_{pred} of the visual prediction loss to 0.5, 1 and 2 and keep every other training setting fixed. As Table 4 shows, downstream performance differs clearly across the three weights, and increasing λ_{pred} does not bring a consistent improvement. The weight of the prediction loss and the gain in video understanding are thus not in a simple positive relation; among the settings tested, $\lambda_{\text{pred}} = 1$ gives the best overall performance and serves as the default. A weight that is too large lets the prediction objective compete with the text loss for the shared parameters, while one that is too small leaves the visual signal too weak to shape the representation used for understanding.

4.5 RQ4: What does the language in the prediction view contribute?

The text condition further strengthens the benefit of visual prediction for video understanding. We remove the textual description from the prediction branch while keeping the visual context and the original language supervision, with all other training settings unchanged; downstream evaluation still uses the complete video and the original questions. As Table 5 shows, the variant without text still outperforms standard SFT overall, and the full model improves further. With the temporal information of the target and the original language supervision retained, visual latent prediction thus provides effective complementary supervision even without the textual description in the prediction branch, and conditioning the prediction on the text increases this benefit further.

The additional gain is not reflected in a lower visual prediction error. Figure 5b shows that the variant without text reaches a lower prediction loss late in training, yet its downstream performance remains below the full model. We further run a diagnostic on 500 paired test samples from Charades-TimeLens, comparing the two models on identical inputs, the same videos with the same spans removed and no description, against the same teacher representations (Appendix F). The mean normalised ℓ_1 prediction error is 0.6778 for the model with text and 0.5206 for the model without it (Figure 5c). The training loss, measured in each model's own format (Figure 5b), and the

Table 4: **Loss weight** (Qwen3.5-0.8B). Notation as in Table 3.

Loss weight	MVBench	TempC.	NExT-QA	STAR	Charades†	ANet†	QVH†	Avg
Base + SFT	56.9	61.9	76.9	69.0	45.1	41.6	54.4	58.0
$\lambda_{\text{pred}} = 0.5$	60.0 ↑3.1	63.8 ↑1.9	78.1 ↑1.2	72.0 ↑3.0	48.2 ↑3.1	44.1 ↑2.5	54.3 ↓0.1	60.1 ↑2.1
$\lambda_{\text{pred}} = 1$ (ours)	59.6 ↑2.7	65.2 ↑3.3	78.1 ↑1.2	72.0 ↑3.0	48.2 ↑3.1	44.7 ↑3.1	57.5 ↑3.1	60.8 ↑2.8
$\lambda_{\text{pred}} = 2$	58.3 ↑1.4	63.3 ↑1.4	77.0 ↑0.1	70.9 ↑1.9	48.0 ↑2.9	42.1 ↑0.5	54.0 ↓0.4	59.1 ↑1.1

Table 5: **Description in the prediction view** (Qwen3.5-0.8B). Notation as in Table 3.

Prediction view	MVBench	TempC.	NExT-QA	STAR	Charades†	ANet†	QVH†	Avg
Base + SFT	56.9	61.9	76.9	69.0	45.1	41.6	54.4	58.0
Description removed	58.3 ↑1.4	65.3 ↑3.4	76.9 ±0.0	70.1 ↑1.1	47.9 ↑2.8	42.1 ↑0.5	55.9 ↑1.5	59.5 ↑1.5
With description (ours)	59.6 ↑2.7	65.2 ↑3.3	78.1 ↑1.2	72.0 ↑3.0	48.2 ↑3.1	44.7 ↑3.1	57.5 ↑3.1	60.8 ↑2.8

input-matched diagnostic (Figure 5c) agree; among the configurations compared, a lower auxiliary prediction error does not necessarily correspond to better downstream video understanding.

This observation is consistent with the loss-weight ablation of RQ3, where the size of the auxiliary prediction loss and downstream performance do not follow the same order across the settings of λ_{pred}. Taken together, the fitting accuracy of the auxiliary visual prediction task does not by itself determine its contribution to video understanding. Visual prediction acts on the shared LLM parameters as an auxiliary training objective, and downstream inference does not use the representations produced at the prediction tokens. Visual prediction supervision therefore benefits video understanding, and the size of this benefit depends not only on how well the prediction task is fitted but also on training settings such as the loss weight and the text condition, so it cannot be judged from the auxiliary prediction error alone.

4.6 RQ5: Does visual latent prediction extend to ordinary video–text data?

Visual latent prediction still improves video understanding on ordinary video–text data. The default scheme uses temporal grounding data and determines the segment to predict from the annotated start and end times of an event. We further ask whether the objective still brings a downstream benefit when the prediction task no longer relies on manually annotated event intervals. To this end, we replace the training data of the prediction branch with ordinary video–caption pairs and construct the prediction target by random sampling, which tests whether the objective extends to video–text data without event-level temporal annotation.

Specifically, for each video we randomly select a contiguous segment covering 25% of its duration as the prediction target, and its position may vary within the valid time range of the video. For a 16-second video, for example, the target segment is fixed at 4 seconds, and its start and end are not required to coincide with the boundaries of any event. The caption of the video serves as a video-level text condition and is not required to be aligned with the randomly selected segment at the event level. All other components follow the default scheme, including the extraction of teacher representations, the visual latent prediction loss and joint training.

As Table 6 shows, with the language supervision branch unchanged, building the visual prediction task from video–caption data still yields better video understanding than standard SFT. The prediction branch can therefore construct effective auxiliary prediction tasks from video–caption data without event-level temporal annotation, and its benefit is not confined to the default construction from grounding data. Because this experiment changes both the source of the prediction data and the construction of the target segment, we regard it mainly as a test of how far the data applicability of the objective extends, and not as a separate quantification of what event alignment contributes.

5 Related Work

World models and visual prediction. World models learn how an environment evolves and let agents act by imagining future observations (Hafner et al., 2025; Bruce et al., 2024; Maes et al., 2026), and large-scale video generative pre-training transfers to robot manipulation (Wu et al., 2024). For representation learning, masked modelling reconstructs hidden pixels or features (He et al., 2021; Wei et al., 2021; Tong et al., 2022), and joint-embedding predictive architectures pre-

Table 6: **Data for the prediction stream** (Qwen3.5-0.8B). Notation as in Table 3.

Prediction data	MVBench	TempC.	NExT-QA	STAR	Charades†	ANet†	QVH†	Avg
Base + SFT	56.9	61.9	76.9	69.0	45.1	41.6	54.4	58.0
Captions, random span	59.6 ↑2.7	63.9 ↑2.0	77.6 ↑0.7	71.9 ↑2.9	46.9 ↑1.8	44.1 ↑2.5	56.1 ↑1.7	60.0 ↑2.0
Grounded descriptions (ours)	59.6 ↑2.7	65.2 ↑3.3	78.1 ↑1.2	72.0 ↑3.0	48.2 ↑3.1	44.7 ↑3.1	57.5 ↑3.1	60.8 ↑2.8

dict hidden content in a learned feature space (Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025), with many recent variants (Xu et al., 2025; Li et al., 2025; Adhikari, 2026; Yu et al., 2025). Unified multimodal models go further and combine visual generation with understanding in one network (Xie et al., 2025; Deng et al., 2025; Wu et al., 2025). These works establish prediction as a strong training signal, yet they target the representation, the dynamics model or the generator and seldom ask what prediction contributes to a language model's understanding.

Visual objectives in multimodal LLM training. A growing line of work brings visual targets into the training of multimodal LLMs. Image MLLMs reconstruct, distil, align or denoise visual features at the hidden states of the LLM (Wang et al., 2025b; Jain et al., 2024; Yoon et al., 2025; Wang et al., 2025c; Parikh et al., 2026; Caffagni et al., 2025; Wang et al., 2026a), other models read predictions out of dedicated tokens (Chen et al., 2026; Adhikari & Lapata, 2026; Huang et al., 2025), and video models predict masked frames, future latents or the task outcome itself (Fu et al., 2021; Li et al., 2022; Cheng et al., 2022; Li et al., 2023b; Fu et al., 2022; Yang et al., 2025; Orca Team, Beijing Academy of Artificial Intelligence, 2026; Zhu et al., 2026; Jiang et al., 2026; Sun et al., 2026; Miao et al., 2026; Yang et al., 2026b). What these objectives do for understanding remains unclear. Reported gains concentrate on image perception, studies of unified models find small or architecture-dependent effects of generation on understanding (Wu et al., 2026), next-frame prediction during video instruction tuning leaves general benchmarks nearly unchanged (Yang et al., 2025), and auxiliary prediction losses can improve without any downstream benefit (Sengupta, 2026; Zhang et al., 2026). Unlike the next-frame target of Cambrian-S, which the adjacent frame largely reveals, our target is a removed multi-second span that cannot be copied from neighbouring frames.

What remains open. To our knowledge, no prior work has examined systematically, with matched data across model scales, whether learning to predict the visual world improves video LLMs on general question answering and temporal grounding. Meanwhile, descriptive video–language data such as captions and temporally grounded sentences enter training only as text to be generated (Huang et al., 2023; Ren et al., 2023; Qian et al., 2024; Huang et al., 2024; Guo et al., 2024b; Chen et al., 2024; Meinardus et al., 2024; Guo et al., 2024a; Wang et al., 2025a; Liu et al., 2024a; 2025a; Wang et al., 2026c), and whether such supervision helps general understanding is unsettled (Zeng et al., 2024; Wang et al., 2024; Yan et al., 2024). Our work takes a step toward both questions, turning such data into a visual prediction objective and studying what it contributes to understanding (Appendix A).

6 Limitations

Our study focuses on how visual prediction benefits understanding; whether stronger understanding in turn improves visual prediction or generation is a natural next question that we leave to future work. Our experiments are conducted within the Qwen3.5 family; the consistent gains from 0.8B to 4B indicate that the objective is not tied to a particular model size, and extending the study to other model families would further confirm its generality.

7 Conclusion

We presented, to our knowledge, the first systematic study of how visual prediction affects understanding in video LLMs. Learning to predict the visual world improves video understanding and temporal grounding at every scale we study; the strength of the prediction loss matters, as a weight that is too large or too small reduces the benefit; and the objective extends from temporally grounded data to ordinary video–caption pairs. These results establish visual prediction as a complementary source of supervision for video LLMs and lay a foundation for models that learn from language and from the visual world at once.

Ethics statement

This work studies visual prediction as a training objective for video LLMs on established public video benchmarks and does not introduce new human-subject data collection. All training and evaluation data are publicly available datasets, and all model architectures used in this work are open source. The experiments are intended for research and benchmark analysis rather than direct deployment in user-facing systems. Any deployment of video understanding systems should separately evaluate privacy, domain shift, calibration and potential performance disparities across user populations.

Reproducibility statement

All training and evaluation data used in this work are publicly available datasets. The training objective is specified in Section 3, the backbones and arms in Section 4.1, and the implementation details and data in Appendices G and H; Appendix E reports the check that no removed content reaches the predictor. We will release the code upon acceptance.

AI use statement

During preparation, the authors used AI assistants for language polishing, LaTeX formatting, consistency checking and coding assistance. AI tools did not generate experimental results or automatically validate references. All claims, analyses, citations and final decisions were manually verified and approved by the authors, who bear full responsibility for the manuscript.

References

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

Santosh Premi Adhikari. Factorized latent dynamics for video jepa: An empirical study of auxiliary objectives. *arXiv preprint arXiv:2605.17165*, 2026.

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, et al. Llava-onevision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 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.

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.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, et al. Genie: Generative interactive environments. In *ICML*, 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.

Mu Cai et al. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024.

Delong Chen, Mustafa Shukor, Théo Moutakanni, Willy Chung, Jade Yu, Tejaswi Kasarla, Yejin Bang, Allen Bolourchi, Yann LeCun, and Pascale Fung. Vl-jepa: Joint embedding predictive architecture for vision-language. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2512.10942.

Shimin Chen et al. Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability. *arXiv preprint arXiv:2411.18211*, 2024.

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.

Chaorui Deng, Deyao Zhu, Kunchang Li, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Dazhao Du, Jian Liu, Jialong Qin, et al. Learning spatiotemporal sensitivity in video llms via counterfactual reinforcement learning. *arXiv preprint arXiv:2605.21988*, 2026a.

Henghui Du, Chang Zhou, Xi Chen, and Di Hu. APPO: Attention-guided perception policy optimization for video reasoning. *arXiv preprint arXiv:2602.23823*, 2026b.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing video reasoning in MLLMs. In *NeurIPS*, 2025.

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

Tsu-Jui Fu et al. An empirical study of end-to-end video-language transformers with masked visual modeling. *arXiv preprint arXiv:2209.01540*, 2022. CVPR 2023.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. *arXiv preprint arXiv:1705.02101*, 2017.

Yongxin Guo et al. Trace: Temporal grounding video llm via causal event modeling. *arXiv preprint arXiv:2410.05643*, 2024a.

Yongxin Guo et al. Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. *arXiv preprint arXiv:2405.13382*, 2024b.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

Zhengdong Hu, Chao Wang, Fengyun Rao, Jing Lyu, Hehe Fan, and Yi Yang. Pointthinker: Point-incentivized parallel thinking for multimodal large language model. In *CVPR*, 2026.

Bin Huang et al. Vtimellm: Empower llm to grasp video moments. *arXiv preprint arXiv:2311.18445*, 2023.

De-An Huang, Shijia Liao, Subhashree Radhakrishnan, et al. Lita: Language instructed temporal-localization assistant. In *ECCV*, 2024. arXiv:2403.19046.

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

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

Tianxiang Jiang, Linquan Wu, Sheng Xia, Songze Li, Ziang Yan, Haoyu Yang, Yu Qiao, and Yi Wang. Imagine before you predict: Interleaved latent visual reasoning for video event prediction. *arXiv preprint arXiv:2606.05769*, 2026.

Yifan Jiang, Yueying Wang, Rui Zhao, Toufiq Parag, Zhimin Chen, Zhenyu Liao, and Jayakrishnan Unnikrishnan. VideoP2R: Video understanding from perception to reasoning. *arXiv preprint arXiv:2511.11113*, 2025.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017.

Jie Lei, Tamara L. Berg, and Mohit Bansal. QVHighlights: Detecting moments and highlights in videos via natural language queries. In *NeurIPS*, 2021.

Kunchang Li et al. Mvbench: A comprehensive multi-modal video understanding benchmark. *arXiv preprint arXiv:2311.17005*, 2023a.

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

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. *arXiv preprint arXiv:2509.24317*, 2025.

Xinhao Li, Yuhan Zhu, Xiangyu Zeng, et al. Videochat3: Fully open video mllm for efficient and generalist video understanding. *arXiv preprint arXiv:2607.14935*, 2026.

Ye Liu et al. E.t. bench: Towards open-ended event-level video-language understanding. *arXiv preprint arXiv:2409.18111*, 2024a.

Ye Liu et al. Videomind: A chain-of-lora agent for long video reasoning. *arXiv preprint arXiv:2503.13444*, 2025a.

Yuanxin Liu et al. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024b.

Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx MLLM: On-demand spatial-temporal understanding at arbitrary resolution. In *ICLR*, 2025b.

Lucas Maes et al. Leworldmodel: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

Boris Meinardus et al. Chrono: A simple blueprint for representing time in mllms. *arXiv preprint arXiv:2406.18113*, 2024.

Shangchen Miao, Ningya Feng, Jialong Wu, Ye Lin, Xu He, Dong Li, and Mingsheng Long. Jepa-vla: Video predictive embedding is needed for vla models. *arXiv preprint arXiv:2602.11832*, 2026.

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.

Long Qian et al. Momentor: Advancing video large language model with fine-grained temporal reasoning. *arXiv preprint arXiv:2402.11435*, 2024.

Qwen Team. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, 2026.

Shuhuai Ren et al. Timechat: A time-sensitive multimodal large language model for long video understanding. *arXiv preprint arXiv:2312.02051*, 2023.

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

Yudi Shi, Shangzhe Di, Qirui Chen, and Weidi Xie. Enhancing video-LLM reasoning via agent-of-thoughts distillation. *arXiv preprint arXiv:2412.01694*, 2024.

Jingwen Sun, Wenyao Zhang, Zekun Qi, Shaojie Ren, Zezhi Liu, Hanxin Zhu, Guangzhong Sun, Xin Jin, and Zhibo Chen. Vla-jepa: Enhancing vision-language-action model with latent world model. *arXiv preprint arXiv:2602.10098*, 2026.

Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

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.

Enguang Wang, Qiang Wang, Yuanchen Wu, Ke Yan, Xinbin Yuan, Shouhong Ding, Xialei Liu, and Ming-Ming Cheng. Predictive regularization against visual representation degradation in multimodal large language models. In *CVPR*, 2026a.

Haibo Wang, Zhiyang Xu, Yu Cheng, et al. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 959–975, 2025a. arXiv:2410.03290.

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

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

Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, Jun Zhang, Ying Shan, and Limin Wang. Timelens: Rethinking video temporal grounding with multimodal llms. In *CVPR*, 2026b.

Xizi Wang et al. Timerefine: Temporal grounding with time refining video llm. In *WACV*, 2026c. arXiv:2412.09601.

Yueqian Wang et al. Hawkeye: Training video-text llms for grounding text in videos. *arXiv preprint arXiv:2403.10228*, 2024.

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

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B. Tenenbaum, and Chuang Gan. STAR: A benchmark for situated reasoning in real-world videos. In *NeurIPS Datasets and Benchmarks Track*, 2021.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, et al. Unleashing large-scale video generative pre-training for visual robot manipulation. In *ICLR*, 2024.

Penghao Wu, Haiwen Diao, Weichen Fan, et al. Uncovering understanding-generation synergy in native unified multimodal models: From representation, task to system. *arXiv preprint arXiv:2609.01607*, 2026.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, et al. VILA-U: a unified foundation model integrating visual understanding and generation. In *ICLR*, 2025.

Junbin Xiao et al. Next-qa: Next phase of question-answering to explaining temporal actions. *arXiv preprint arXiv:2105.08276*, 2021.

Jinheng Xie, Weijia Mao, Zechen Bai, et al. Show-o: One single transformer to unify multimodal understanding and generation. In *ICLR*, 2025.

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

Ziang Yan et al. Task preference optimization: Improving multimodal large language models with vision task alignment. *arXiv preprint arXiv:2412.19326*, 2024.

Senqiao Yang, Kaichen Zhang, Zhaoyang Jia, et al. Mage-vl: An efficient codec-native streaming multimodal foundation model. *arXiv preprint arXiv:2607.24904*, 2026a.

Shusheng Yang, Jihan Yang, Pinzhi Huang, et al. Cambrian-s: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, 2025.

Zhichao Yang, Yuanze Hu, Haojie Hao, Longkun Hao, Dongshuo Huang, Hongyu Lin, Gen Li, Lanqing Hong, Yihang Lou, Yan Bai, et al. Mirage: Mobile agents with implicit reasoning and generative world models. *arXiv preprint arXiv:2606.04627*, 2026b.

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.

Jiacan Yu, Siyi Chen, et al. Why and how auxiliary tasks improve jepa representations. *arXiv preprint arXiv:2509.12249*, 2025.

Shoubin Yu, Jaemin Cho, Prateek Yadav, et al. Self-chained image-language model for video localization and question answering. In *NeurIPS*, 2023. arXiv:2305.06988.

Xiangyu Zeng et al. Timesuite: Improving mllms for long video understanding via grounded tuning. *arXiv preprint arXiv:2410.19702*, 2024.

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.

Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018.

Jinguo Zhu, Weiyun Wang, Zhe Chen, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Xiaoyu Zhu, Xinke Deng, Suresh Taddewadikar, et al. Beyond visual cot: Internalized visual thinking for proactive video reasoning. *arXiv preprint arXiv:2608.15869*, 2026.

A Positioning

Table 7: **Positioning** of our objective among related objectives.

Method	Target space	Read-out	Infer. cost	Regime	Target defined by
I-JEPA / V-JEPA	own encoder (EMA)	patch	—	pre-train	appearance
NEPA	own encoder (co-trained)	patch	—	pre-train	appearance
Orca	frozen ViT	learnable query	0	cont. pre-train	appearance + event text
Ross3D / Latent Denoising	frozen VAE / CLIP	masked position	0	SFT	appearance
PEARL / LLM-JEPA	own hidden states	`[PRED]` token	0	SFT	trajectory
TimeSuite / Tarsier2	— (text)	text	0	SFT	language + time
E.T. Chat / TimeRefine	frame emb. / span	special token	head kept / dropped	SFT	language + time
Ours	frozen ViT	`<G>`, direct	0	SFT	**language + time**

B Full Temporal-Grounding Results

Table 1 reports mIoU; Table 8 gives R@0.5 and R@0.7 for the same rows.

Table 8: Temporal grounding, R@0.5 / R@0.7 (%). Notation as in Table 1.

Model	Charades†		YouCook2		ANet†		QVH†		HC-STVG	
	R@0.5	R@0.7	R@0.5	R@0.7	R@0.5	R@0.7	R@0.5	R@0.7	R@0.5	R@0.7
Qwen3.5-0.8B										
Qwen3.5-0.8B-Base	24.1	12.0	31.0	16.6	27.9	17.8	43.0	31.2	16.2	11.9
Qwen3.5-0.8B-Instruct	28.6	15.1	31.5	16.3	26.7	15.2	43.5	28.3	17.6	11.5
Base + SFT	50.9	25.8	60.0	38.3	44.8	29.4	58.3	43.0	18.9	11.9
Base + SFT + Pred	54.6 ↑3.7	28.9 ↑3.1	62.7 ↑2.7	41.3 ↑3.0	48.0 ↑3.2	32.3 ↑2.9	60.4 ↑2.1	46.5 ↑3.5	23.3 ↑4.4	16.5 ↑4.6
Qwen3.5-2B										
Qwen3.5-2B-Base	36.7	19.9	43.2	26.1	38.9	26.1	52.9	40.0	15.0	9.5
Qwen3.5-2B-Instruct	32.2	17.8	44.7	25.8	39.3	24.4	51.2	38.2	17.6	10.2
Base + SFT	52.2	28.2	65.2	43.4	52.8	35.0	59.1	45.3	18.2	11.0
Base + SFT + Pred	60.1 ↑7.9	31.6 ↑3.4	71.6 ↑6.4	50.2 ↑6.8	55.4 ↑2.6	37.8 ↑2.8	62.5 ↑3.4	47.2 ↑1.9	22.6 ↑4.4	15.0 ↑4.0
Qwen3.5-4B										
Qwen3.5-4B-Base	38.1	20.7	53.0	35.1	43.1	36.2	54.9	43.8	15.5	10.9
Qwen3.5-4B-Instruct	38.4	19.1	55.3	33.2	50.0	35.4	68.3	55.5	22.3	13.7
Base + SFT	41.7	22.4	56.3	37.9	42.0	29.5	54.8	43.5	17.3	12.2
Base + SFT + Pred	45.6 ↑3.9	25.5 ↑3.1	56.4 ↑0.1	37.6 ↓0.3	53.6 ↑11.6	37.4 ↑7.9	61.2 ↑6.4	49.2 ↑5.7	22.8 ↑5.5	16.1 ↑3.9

C Comparison with Public Models

Table 9 extends the comparison of Table 2 on TempCompass, multiple-choice split, to a wider set of published open video LLMs grouped by size. Published scores are copied from the cited tables and keep their own settings. Much of our absolute level comes from the backbone: the released Qwen3.5-4B-Instruct already scores 75.2, and the objective adds 3.3, 1.4 and 1.5 points over Base+SFT at the three scales.

Table 9: **TempCompass, multiple-choice split** (%). Published scores from VideoChat3[a] (Li et al., 2026), LLaVA-OneVision-2[b] (An et al., 2026) and PerceptionLM[c] (Cho et al., 2025).

Model	TempC.
Up to 3B	
InternVL2.5-1B	56.4[c]
Qwen2-VL-2B	62.3[c]
Qwen2.5-VL-3B	63.0[c]
PLM-1B	64.6[c]
PLM-3B	69.3[c]
Qwen3.5-0.8B-Instruct	61.0
Base + SFT	61.9
Base + SFT + Pred	65.2
Qwen3.5-2B-Instruct	69.9
Base + SFT	71.0
Base + SFT + Pred	72.4
About 4B	
InternVL2.5-4B	65.2[c]
InternVL3.5-4B	68.8[a]
Qwen3-VL-4B	70.8[a]
Molmo2-4B	72.8[a]
VideoChat3-4B	75.6[a]
Qwen3.5-4B-Instruct	75.2
Base + SFT	75.8
Base + SFT + Pred	77.3

Model	TempC.
7B–8B	
LLaVA-OV-1.5-8B	57.5[b]
Video-XL2-8B	58.5[a]
LLaVA-Video-7B	66.6[a]
VideoLLaMA3-7B	68.1[a]
VideoChat-Flash-7B	69.4[a]
InternVideo2.5-8B	70.1[a]
InternVL3.5-8B	70.4[b]
MiniCPM-V-4.5-8B	72.7[a]
PLM-8B	72.7[a]
Qwen3-VL-8B	74.3[b]
Eagle2.5-8B	74.4[a]
LLaVA-OV-2-8B	74.5[b]
Keye-VL-1.5-8B	75.5[b]

D Compute Matching

Compute is matched between the two arms (Section 4.1), and the baseline already sees more text-supervised passes (Table 11). Between its 4k- and 6k-step checkpoints, Base+SFT at 4B changes no benchmark by more than 0.4 points; additional language supervision does not close the gap, whereas the prediction objective does. At inference the prediction stream does not exist, and the deployed model has the architecture, parameter count and latency of its SFT counterpart.

E NO-LEAKAGE ASSERTION

Section 3.3 claims that no information passes between the removed span and the rest of the input. The claim rests on two structural facts: the vision encoder computes attention within a temporal group, so a group's features depend on its own pixels only; and the removed groups are not in the sequence at all. The second holds by construction; we verify the first with an executable check rather than by reading the source. On tiny randomly initialised vision towers of each encoder family, with the released patch, temporal-patch and merge settings, we encode eight groups and treat the fourth as removed. We then (i) replace the pixels of the removed group with noise and measure the change of the merged features of the retained groups, and (ii) replace the pixels of every retained group and measure the change of the removed group's features, which form the target. Both must be exactly zero: (i) shows that the retained input carries no removed pixels, and (ii) that the target carries no retained ones. The perturbed side must change, otherwise the perturbation never reached the encoder and the test is vacuous.

Table 10: Encoder locality: maximum absolute change of the merged features under the two perturbations.

	(i) removed group perturbed		(ii) retained groups perturbed	
Family	retained	removed	removed (target)	retained
Qwen3-VL	0.00e+00	1.10e-01	0.00e+00	1.20e-01
Qwen2.5-VL	0.00e+00	1.37e-01	0.00e+00	1.29e-01
Qwen3.5	0.00e+00	1.13e-01	0.00e+00	1.19e-01

F HELD-OUT PREDICTION ERROR

Figure 5c compares the model trained with the description and the model trained without it on 500 clips with an annotated span from the Charades-TimeLens test split, both at step 9000. Both models receive identical inputs, the same video with the same span removed and no description, and are compared with the same frozen targets of that span. We withhold the description so that the comparison is on identical inputs and the error reflects only what each model has learned to infer from visual context; providing it to one model alone would confound the comparison with the extra input. The comparison in each model's own training format is given by the training curves in Figure 5b, where the model trained with descriptions also attains the higher prediction loss late in training. The error is the normalised ℓ_1 of Equation 4, averaged over the tokens of the span and then over clips, and the figure reports the paired difference between the two models with its 95% confidence interval.

G IMPLEMENTATION DETAILS

Table 11 lists every setting of the runs in Table 1; all arms at a given scale share them, except that Base+SFT has no prediction stream. The prediction token <G> is a single row added to the embedding matrix, initialised at random like the vocabulary embeddings, and one copy is appended per missing visual token, at the spatio-temporal position of that token. Its hidden states at half the depth of the LM are compared to the target without any projection, because the merged visual features already live in the LM embedding space and the M read-outs match the M target tokens one to one.

Table 11: **Training settings of the two arms.** Both arms use the same steps, batch size and number of forward passes at every scale.

Setting	Base+SFT	Base+SFT+Pred
Backbones	Qwen3.5-0.8B, 2B, 4B (Base checkpoints)	
Trainable	LLM; vision encoder and merger frozen	LLM and the prediction-token embedding; vision encoder and merger frozen
Video input	2 fps, one group per second, at most 128 frames	
Learning rate	2×10^{-6}	
Data pool	572K language-modelling samples (including the grounding descriptions as text); 196K prediction samples	
Evaluated step, global batch	9000, 128	9000, 128
Batch composition	128 language-modelling	96 language-modelling + 32 prediction (3:1)
Language-modelling passes	≈1.15M (≈2.0 epochs)	≈864K (≈1.5 epochs)
Prediction passes	0	≈288K (≈1.5 epochs)
Total forward passes	≈1.15M	≈1.15M
Loss	$\mathcal{L}_{\text{CE}}$	$\mathcal{L}_{\text{CE}} + \lambda_{\text{pred}}\mathcal{L}_{\text{pred}}$, $\lambda_{\text{pred}} = 1$
Prediction target	–	merged tokens of the removed span from the frozen encoder and merger, detached
Read-out	–	M prediction tokens, one per missing visual token, at half depth, no projection head
Prediction loss	–	ℓ_1 between layer-normalised predictions and targets, averaged over tokens and dimensions

H DATA

The training mixture combines question–answer data from LLaVA-Video-178K (Zhang et al., 2024), YouCook2 (Zhou et al., 2018), NExT-QA (Xiao et al., 2021) and STAR (Wu et al., 2021) with temporal grounding data from TimeLens-100K (Wang et al., 2026b), Charades-STA (Gao et al., 2017) and YouCook2. These include the training splits of NExT-QA, STAR, Charades-STA and YouCook2; all evaluations use held-out test splits or the TimeLens re-annotated test sets, which do not overlap with the training data. The prediction stream draws on descriptive video–language data that is already part of the instruction-tuning mixture, so adding the objective leaves the visual data unchanged. The default source is temporally grounded descriptions: each sentence–span pair enters the prediction stream about 1.5 times over training (Table 11) and remains in the mixture in its text form, which is what the SFT baseline sees. The caption source of Section 4.6 consists of ordinary video–caption pairs from LLaVA-Video-178K (Zhang et al., 2024); the removed span is a contiguous window covering one quarter of the video at a uniformly random start, and the caption describes the whole video.