

VIDEO2LORA: Parametric Video Internalization for Vision-Language Models

Anonymous ACL submission

Abstract

Processing video in vision-language models is expensive: each frame occupies hundreds of tokens, and inference cost scales with every frame and every repeated query. We introduce **VIDEO2LORA**, a method for parametric video internalization. A perceiver hypernetwork reads the intermediate representations produced layer-by-layer as a frozen VLM encodes a video, and generates a Low-Rank Adaptation (LoRA) adapter in a single forward pass. Unlike standard LoRA fine-tuning, which requires iterative gradient updates, VIDEO2LORA predicts these weights directly from the video. Trained for SmolVLM2 500M and 2.2B on video summarization and captioning, VIDEO2LORA enables the same frozen VLM to answer queries from the adapter alone, with *zero visual tokens* in its context at query time. VIDEO2LORA is statistically non-inferior and equivalent to direct video-in-context inference across all five captioning benchmarks at both model scales, and across seven of eight video question answering benchmark-scale pairings. Although trained only on 12 frames at 384px, it remains stable up to 1,024 frames and 1024px, where direct video-in-context inference often degenerates. Across this sweep, it reduces answer-time visual-token load by up to 1,500 \times and query TTFT by 6–80 \times , while preserving video-faithful outputs. We also find that independently generated adapters for non-overlapping video segments can compose in rank space, suggesting a path toward chunked long-video internalization.

1 Introduction

Video understanding in VLMs is built on a token-heavy abstraction: frames are encoded as visual tokens and concatenated into the model’s context window. Each frame at standard resolution contributes hundreds of visual tokens (Liu et al., 2024; Shang et al., 2025); even short clips of a few dozen

frames generate tens of thousands of tokens before any text query is added, and memory and latency scale with every frame and every query. Past a capacity threshold, this bottleneck does not produce gracefully degraded outputs: VLMs generate incoherent or repetitive text unrelated to the video (Chen et al., 2025b; Zhang et al., 2024). The context window (the model’s fixed token capacity) is therefore the fundamental bottleneck for video understanding, and it is re-encountered on every query over the same video.

Much work aims to fit more video into the context window. Frame subsampling (Zhang et al., 2023) discards frames to meet a token budget, sacrificing temporal coverage. Visual token compression methods (Shang et al., 2025; Li et al., 2025) prune or merge spatial tokens before the language backbone, reducing per-frame cost without discarding entire frames. Long-context architectures (Chen et al., 2025b; Zhang et al., 2024) scale the context window itself through sequence parallelism and position encoding modifications. Streaming methods (Qian et al., 2024) process video incrementally, maintaining a compact memory buffer in lieu of full context retention. Each approach reduces the burden without resolving it: visual tokens remain in context at query time, every query re-incurs the encoding overhead, and all approaches eventually encounter the same capacity ceiling. The capacity ceiling is not a constraint to manage: it is a constraint to eliminate.

We take a fundamentally different approach. Rather than compressing visual information to fit within the context window, we eliminate it from the query entirely, encoding the video into the model’s parameters before any query is issued. The video is stored as a LoRA adapter (Hu et al., 2022); subsequent queries are answered by a frozen base model with those adapter weights, with no visual tokens in context. Prior work has shown that feedforward hypernetworks (Ha et al., 2017; Charakorn et al.,

2026) can produce LoRA adapters from *text documents*, enabling a frozen LLM to answer queries about a document with no text tokens in context. Extending this paradigm to video introduces qualitatively harder challenges: the token volume per example is orders of magnitude larger, making iterative per-example optimization computationally impractical; the compression is cross-modal, requiring visual semantics to be expressed as perturbations to a language model’s parameter space; and the visual distribution varies along a resolution axis with no textual analog.

Main Result: We introduce **VIDEO2LORA**, a framework for parametrically internalizing videos into a frozen vision-language model (VLM). Given a video, a perceiver hypernetwork (Jaegle et al., 2021) processes the layer-wise hidden states of the frozen VLM encoder and generates LoRA adapter weights in a single forward pass. The generated adapter is then attached to the same frozen VLM, enabling it to answer questions about the video without requiring visual tokens in the context window. During training, both the VLM encoder and the answering model remain frozen; only the hypernetwork is optimized using cached teacher-generated captions and summaries as supervision. We train and evaluate VIDEO2LORA on SmolVLM2 500M and 2.2B (Marafioti et al., 2025). Our novel contributions include:

- **First parametric video internalization.** A Perceiver hypernetwork that converts a video into a LoRA adapter in a single forward pass, enabling a frozen VLM to answer queries with no visual tokens in context. We demonstrate feasibility across 2.2B and 500M model scales.
- **Strong performance on captioning and video question answering.** Statistical non-inferiority and equivalence to direct video-in-context inference across all five captioning benchmarks at both model scales (ActivityNet Captions, PLM-RDCap, PLM-RCap, VDC, CaReBench) and across seven of eight video question answering benchmark-scale pairings (NExT-QA, ActivityNet-QA, PLM-SGQA, VidCapBench).
- **Efficiency, generalization and emergent compositionality.** Although trained only on 12 frames at 384px, VIDEO2LORA remains stable up to 1,024 frames and 1024px, where

direct video-in-context inference often degenerates. It reduces answer-time visual-token load by up to 1,500× and query TTFT by 6–80×, while preserving video-faithful outputs. Compared to KV caching and token-compression techniques, we show that video internalization via Video2LoRA preserves performance across token budgets, is faster to process, and has the lowest time to first token. We further observe that adapters generated independently for non-overlapping video segments can compose in rank space, suggesting a path toward chunked long-video internalization.

2 Related Work

2.1 Efficient Video Understanding

Most efficient video-understanding methods reduce the number or cost of visual tokens while still keeping visual information in the model context. Frame subsampling (Zhang et al., 2023) lowers temporal coverage to fit a token budget; visual-token compression (Shang et al., 2025; Li et al., 2025) prunes or merges spatial tokens; long-context video models (Chen et al., 2025b; Zhang et al., 2024) extend the usable context window; and streaming methods (Qian et al., 2024; Zhang et al., 2023) maintain compact memory across time. These approaches improve scalability, but the language model still conditions on visual tokens at query time. VIDEO2LORA is orthogonal: it converts the video into adapter weights once, then answers later queries without visual tokens in context.

2.2 Parametric Knowledge Compression

Parameter-efficient methods such as LoRA, prefix tuning, and prompt tuning store task information in small learned updates rather than full model parameters (Hu et al., 2022; Li and Liang, 2021; Lester et al., 2021). More recent work moves instance-level context into compact representations, including gist tokens (Mu et al., 2023), hypernetwork-based editing (Mitchell et al., 2022; Ha et al., 2017), and deep context distillation (Caccia et al., 2025). Closest to our setting, Doc-to-LoRA maps text documents into LoRA adapters using a feedforward hypernetwork (Charakorn et al., 2026). VIDEO2LORA extends this idea from text to video, where the hypernetwork must compress high-volume visual context into language-model adapter weights and generalize across frame count

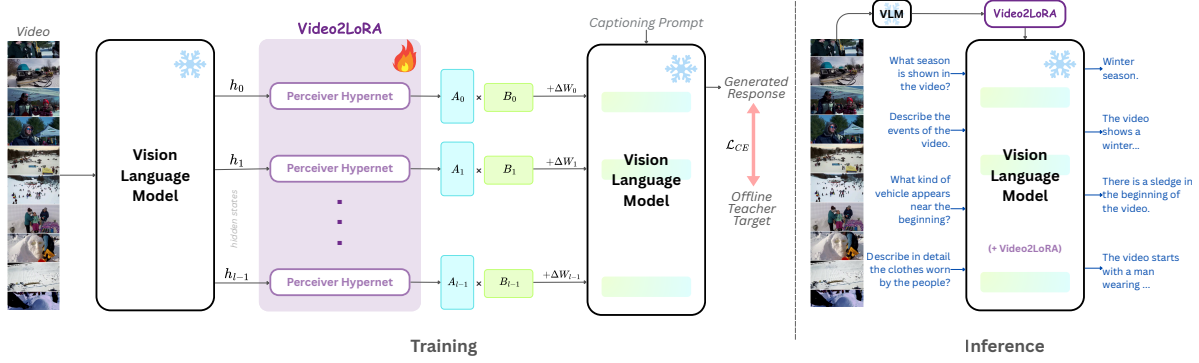


Figure 1: **VIDEO2LORA overview.** *Training (left):* A frozen VLM encodes the input video into hidden states. The trainable VIDEO2LORA hypernetwork reads these states and generates LoRA adapter weights in a single forward pass. The adapter-augmented frozen VLM is trained against teacher-generated targets. *Inference (right):* Given a new video, VIDEO2LORA generates the LoRA adapter once. The frozen VLM, augmented with this adapter, answers arbitrary text queries without visual tokens. Per-query cost is independent of video length.

184 and resolution.

185 3 VIDEO2LORA

186 VIDEO2LORA converts a video into a video-
 187 specific LoRA adapter in a single forward pass.
 188 A frozen VLM encodes the video into layer-wise
 189 hidden states, and a trainable Perceiver hypernet-
 190 work maps these states into LoRA weights. At
 191 inference time, the generated adapter is attached
 192 to the frozen answer model, which answers down-
 193 stream text prompts without receiving any visual
 194 tokens in its context.

195 3.1 Problem Formulation

196 Let v denote a video, i an internalization instruc-
 197 tion, p a downstream text prompt, and y the target
 198 response. We assume a frozen vision-language en-
 199 coder E , a frozen answer model F , and a trainable
 200 hypernetwork H_ϕ . The method is defined as:

$$201 \mathbf{C} = E(v, i), \quad (1)$$

$$202 \theta(v) = H_\phi(\mathbf{C}), \quad (2)$$

$$203 p_\phi(y | p, v) = F(y | p; \theta(v)). \quad (3)$$

204 Here, \mathbf{C} denotes video-conditioned hidden states
 205 and $\theta(v)$ denotes the generated LoRA adapter. The
 206 answer model receives the text prompt p and the
 207 adapter $\theta(v)$, but not the video tokens. During
 208 training, only ϕ is updated; both E and F remain
 209 frozen.

210 3.2 Video Encoder

211 We use a frozen SmolVLM2 model (Marafioti et al.,
 212 2025) as the video encoder. Given a sampled video

213 and the internalization instruction, we collect the
 214 text-side hidden states from each transformer layer:

$$215 \mathbf{C} = \text{stack}(\mathbf{h}_0, \mathbf{h}_1, \dots, \mathbf{h}_{L-1}) \in \mathbb{R}^{L \times S \times D}, \quad (4)$$

216 where L is the number of layers, S is the fused
 217 sequence length, and D is the hidden dimension.
 218 Keeping the layer dimension allows the hypernet-
 219 work to generate layer-indexed adapters instead of
 220 using a single pooled video vector for all layers.

221 3.3 Perceiver Hypernetwork

222 The hypernetwork maps \mathbf{C} into LoRA weights
 223 for selected linear modules of the frozen answer
 224 model. We use a Perceiver-style resampler archi-
 225 tecture (Jaegle et al., 2021). For each layer slice
 226 $\mathbf{C}_\ell \in \mathbb{R}^{S \times D}$, an encoder resampler attends from
 227 learned latent queries to the video-conditioned hid-
 228 den states, producing a fixed-size representation. A
 229 decoder resampler then uses one output query for
 230 each target module and LoRA rank direction.

231 For batch size B , number of target modules M ,
 232 rank R , and latent size Z , the hypernetwork output
 233 has shape

$$234 \mathbf{O} \in \mathbb{R}^{B \times L \times M \times R \times Z}. \quad (5)$$

235 A shared projection head maps each rank latent to
 236 the two LoRA factors:

$$237 \mathbf{A}_{\ell,m} \in \mathbb{R}^{R \times d_{\text{in}}}, \quad (6)$$

$$238 \mathbf{B}_{\ell,m} \in \mathbb{R}^{R \times d_{\text{out}}}.$$

239 where ℓ indexes the transformer layer and m in-
 240 dexes the target linear module. The generated fac-
 241 tors are scaled by learned multipliers, with the \mathbf{A}
 242 scale initialized to one and the \mathbf{B} scale initialized to
 243 zero. Thus, the generated adapter initially produces
 no perturbation to the frozen model.

3.4 Dynamic LoRA Injection

For a frozen linear layer with weight $\mathbf{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, we use the standard LoRA factorization (Hu et al., 2022). Under the row-vector implementation convention, the frozen layer computes $\mathbf{x}\mathbf{W}^\top$. The generated adapter adds:

$$\Delta\mathbf{y} = s(\mathbf{x}\mathbf{A}_{\ell,m}^\top)\mathbf{B}_{\ell,m}, \quad (7)$$

where s is the fixed LoRA scaling factor. The full adapted forward pass is:

$$\mathbf{y} = \mathbf{x}\mathbf{W}^\top + s(\mathbf{x}\mathbf{A}_{\ell,m}^\top)\mathbf{B}_{\ell,m}. \quad (8)$$

Each example receives its own generated adapter, so the LoRA weights are conditioned on the input video rather than shared across all videos.

3.5 Training Objective

We train the hypernetwork with teacher-forced cross-entropy over response tokens:

$$\mathcal{L}(\phi) = - \sum_t \log p_\phi(y_t | y_{<t}, p, \theta(v)). \quad (9)$$

The answer model receives only the downstream text prompt and the generated adapter during this loss computation.

4 Experimental Setup

4.1 Models and Training

We evaluate two SmolVLM2 model scales: 500M and 2.2B. For each scale, the video encoder and answer model are initialized from the same frozen backbone. Only the VIDEO2LORA hypernetwork is trained. Training uses 12 uniformly sampled frames at 384px longest-edge resolution (constrained by compute). We apply generated LoRA adapters to the MLP down_proj modules with rank $R = 16$. We train on video spans derived from FineVideo (Farré et al., 2024). The span mixture contains single-scene spans, adjacent multi-scene spans, and full-video spans, sampled in a 60/30/10 ratio. FineVideo metadata is used only to define spans; the final training targets are cached offline teacher generations from a frozen SmolVLM2 teacher conditioned on the sampled video frames and downstream prompt. Audio is excluded throughout. The hypernetwork is trained with teacher-forced cross-entropy over response tokens, while the answer model receives only the text prompt and generated adapter. Further details on training can be found in the appendix.

4.2 Evaluation Benchmarks

We evaluate captioning on ActivityNet Captions (Krishna et al., 2017), PLM-RDCap (Cho et al., 2025), PLM-RCap (Cho et al., 2025), VDC (Chai et al., 2025), and CaReBench (Xu et al., 2025); and video QA on NExT-QA (Xiao et al., 2021), ActivityNet-QA (Yu et al., 2019), PLM-SGQA (Cho et al., 2025), and VidCapBench (Chen et al., 2025a).

To scale LLM Judge evaluation, we fix the number of samples from each benchmark to 500. VDC and CaReBench use 500 examples per subset/style. VidCapBench has multiple QA pairs corresponding to each video, therefore we fixed the number of videos to 100 and obtained 1,523 QA pairs corresponding to it. For all benchmarks, the direct baseline and VIDEO2LORA use the same videos, prompts, references, frame sampling, and decoding configuration.

4.3 Metrics and Statistical Testing

We report two quality metrics. First, we compute token-level F1 between the generated output and the reference answer or caption. Second, we use an LLM judge to score output quality on a 1–5 scale, which is linearly rescaled to $[0, 1]$. We use Qwen3-30B (Yang et al., 2025) as our judge model, with a constrained rubric. Human study on this metric for a subset of 200 examples (100 captioning + 100 QA) reveals strong correlation with human judgements, with Spearman $\rho = 0.823$ for metric fidelity.

We estimate 95% confidence intervals using paired bootstrap resampling. For statistical measures, NI (Non-inferiority) and Eq (Equivalence) we use a margin of 0.05 for token-F1 and 0.15 for rescaled judge score.

5 Results

5.1 Captioning

VIDEO2LORA passes both non-inferiority and equivalence on **all 10 benchmark–scale combinations** under the LLM judge and all 10 under token-F1 (Table 1). For SmolVLM 2.2B, VIDEO2LORA recovers **91.9%** of the base model’s judge score, while for SmolVLM 500M, it recovers **84.2%**.

Per-benchmark analysis. Recovery rates at 500M span 79–96%, with compact clip-aligned benchmarks (PLM-RCap, PLM-RDCap) easiest to internalize and temporally dense benchmarks

LLM Judge																
Benchmark	SmolVLM 500M						SmolVLM 2.2B									
	Base	V2L	Δ	CI Eq NI			Base	V2L	Δ	CI Eq NI						
ActivityNet Captions	0.428	0.356	-0.072	[-0.104, -0.041]			Y	Y	0.576	0.492	-0.084	[-0.113, -0.057]			Y	Y
PLM-RDCap	0.308	0.263	-0.045	[-0.069, -0.021]			Y	Y	0.326	0.316	-0.010	[-0.032, +0.012]			Y	Y
PLM-RCap	0.252	0.242	-0.011	[-0.031, +0.009]			Y	Y	0.270	0.287	+0.017	[+0.001, +0.034]			Y	Y
VDC (aggregate)	0.515	0.406	-0.108	[-0.118, -0.098]			Y	Y	0.539	0.511	-0.028	[-0.037, -0.019]			Y	Y
CaReBench	0.334	0.278	-0.056	[-0.067, -0.045]			Y	Y	0.437	0.369	-0.068	[-0.078, -0.058]			Y	Y
<i>Average</i>	0.367	0.309	-0.058	[-0.078, -0.039]			Y	Y	0.430	0.395	-0.035	[-0.052, -0.018]			Y	Y

Token F1																
Benchmark	SmolVLM 500M						SmolVLM 2.2B									
	Base	V2L	Δ	CI Eq NI			Base	V2L	Δ	CI Eq NI						
ActivityNet Captions	0.236	0.243	+0.007	[+0.002, +0.012]			Y	Y	0.263	0.256	-0.007	[-0.012, -0.002]			Y	Y
PLM-RDCap	0.189	0.198	+0.009	[+0.005, +0.013]			Y	Y	0.198	0.207	+0.009	[+0.005, +0.013]			Y	Y
PLM-RCap	0.177	0.203	+0.026	[+0.021, +0.031]			Y	Y	0.199	0.204	+0.005	[+0.001, +0.010]			Y	Y
VDC (aggregate)	0.315	0.288	-0.027	[-0.030, -0.025]			Y	Y	0.297	0.304	+0.007	[+0.003, +0.010]			Y	Y
CaReBench	0.295	0.275	-0.020	[-0.023, -0.017]			Y	Y	0.292	0.279	-0.013	[-0.015, -0.010]			Y	Y
<i>Average</i>	0.243	0.242	-0.001	[-0.005, +0.003]			Y	Y	0.250	0.250	+0.000	[-0.004, +0.004]			Y	Y

Table 1: Comparison of the base model with video and VIDEO2LORA generated adapters, across captioning benchmarks using LLM Judge scores and Token F1. We report mean scores, the paired difference Δ (V2L – Base), 95% confidence intervals, and the statistical equivalence (Eq) and non-inferiority (NI) criteria.

Subset	SmolVLM 500M		SmolVLM 2.2B	
	Base	V2L (Δ)	Base	V2L (Δ)
Short caption	0.629	0.535 (-0.094)	0.556	0.579 (+0.022)
Detailed caption	0.476	0.401 (-0.074)	0.526	0.463 (-0.063)
Camera	0.310	0.131 (-0.178)	0.478	0.392 (-0.085)
Background	0.642	0.523 (-0.117)	0.588	0.606 (+0.018)
Main object	0.517	0.442 (-0.075)	0.546	0.514 (-0.032)

Table 2: VDC results broken down by caption style.

Subset	SmolVLM 500M		SmolVLM 2.2B	
	Base	V2L (Δ)	Base	V2L (Δ)
Caption	0.418	0.324 (-0.094)	0.465	0.400 (-0.065)
Events	0.201	0.169 (-0.032)	0.340	0.267 (-0.073)
Objects	0.368	0.327 (-0.043)	0.457	0.392 (-0.065)
Spatial caption	0.424	0.329 (-0.095)	0.519	0.426 (-0.094)
Temporal caption	0.260	0.242 (-0.018)	0.404	0.360 (-0.045)

Table 3: CaReBench results broken down by subset.

(VDC, ActivityNet Captions) hardest. Scale narrows this spread considerably: at 2.2B the floor rises to 85% and the ceiling breaks above the base, with PLM-RCap *surpassing* the base outright (CI entirely above zero) and PLM-RDCap reaching de-facto equivalence (CI straddling zero). The benchmarks most sensitive to scale—particularly VDC, where the gap contracts fourfold—are those requiring compression of visually diverse, longer-form descriptions; benchmarks with consistently structured references recover well at both scales.

Token F1. Token-F1 provides independent reference-based corroboration: the mean paired delta is -0.001 at 500M and 0.000 at 2.2B. VIDEO2LORA exceeds base on 3 of 5 benchmarks at 500M (ActivityNet Captions $+0.007$, PLM-RDCap $+0.009$, PLM-RCap $+0.026$) and 2 of 5 at 2.2B. The PLM-RCap result at 500M is notable: $+0.026$ ($+14.7\%$; CI $[+0.021, +0.031]$) with no token-level supervision.

5.2 Fine-Grained Captioning

Tables 2 and 3 break VDC and CaReBench into caption styles and semantic dimensions.

VDC Four of five VDC styles maintain 81–85% recovery at 500M: short (85.1%, $\Delta = -0.094$), detailed (84.2%, $\Delta = -0.074$), background (81.5%, $\Delta = -0.117$), main object (85.5%, $\Delta = -0.075$). *Camera* captions are the outlier: At 500M, VIDEO2LORA achieves only 42.3% recovery ($\Delta = -0.178$; base 0.310, V2L 0.131), as cinematographic attributes—shot framing, viewpoint, and camera motion—are difficult to encode as weight perturbations at this scale. At 2.2B, VIDEO2LORA recovers 82.0% ($\Delta = -0.085$), a gain of $+39.7$ pp. VIDEO2LORA recovers **82.0%** ($\Delta = -0.085$), a gain of **+39.7 pp**—the largest single-dimension scale improvement in the fine-grained evaluation—This suggests that part of the camera-description gap is capacity-related, although targeted camera-motion supervision or

LLM Judge												
Benchmark	SmolVLM 500M						SmolVLM 2.2B					
	Base	V2L	Δ	CI	Eq	NI	Base	V2L	Δ	CI	Eq	NI
NExT-QA (open)	0.501	0.547	+0.046	[+0.007, +0.084]	Y	Y	0.597	0.610	+0.013	[-0.022, +0.048]	Y	Y
ActivityNet-QA	0.524	0.541	+0.016	[-0.031, +0.064]	Y	Y	0.627	0.531	-0.096	[-0.144, -0.049]	Y	Y
PLM-SGQA	0.390	0.317	-0.074	[-0.113, -0.034]	Y	Y	0.493	0.295	-0.198	[-0.236, -0.161]	-	-
VidCapBench	0.502	0.451	-0.050	[-0.071, -0.030]	Y	Y	0.551	0.475	-0.076	[-0.096, -0.055]	Y	Y
<i>Average</i>	0.487	0.460	-0.027	[-0.043, -0.011]	Y	Y	0.562	0.477	-0.085	[-0.101, -0.069]	Y	Y

Token F1												
Benchmark	SmolVLM 500M						SmolVLM 2.2B					
	Base	V2L	Δ	CI	Eq	NI	Base	V2L	Δ	CI	Eq	NI
NExT-QA (open)	0.129	0.068	-0.061	[-0.076, -0.046]	-	-	0.140	0.076	-0.063	[-0.079, -0.048]	-	-
ActivityNet-QA	0.197	0.023	-0.174	[-0.199, -0.149]	-	-	0.149	0.013	-0.136	[-0.156, -0.117]	-	-
PLM-SGQA	0.081	0.225	+0.145	[+0.131, +0.158]	-	Y	0.092	0.203	+0.111	[+0.098, +0.124]	-	Y
VidCapBench	0.216	0.209	-0.007	[-0.019, +0.004]	Y	Y	0.196	0.218	+0.022	[+0.010, +0.033]	Y	Y
<i>Average</i>	0.156	0.131	-0.024	[-0.041, -0.008]	Y	Y	0.144	0.128	-0.017	[-0.032, -0.002]	Y	Y

Table 4: Comparison of the base model with video and VIDEO2LORA generated adapters, across video question answering benchmarks using LLM Judge scores and Token F1. We report mean scores, the paired difference Δ (V2L – Base), 95% confidence intervals, and the statistical equivalence (Eq) and non-inferiority (NI) criteria.

adaptive rank may still be needed. At 2.2B, two styles exceed the base outright: short captions (104.1%, $\Delta = +0.022$) and background (103.1%, $\Delta = +0.018$).

CaReBench Temporal captioning is best-recovered at both scales (500M: 93.1%, $\Delta = -0.018$; 2.2B: 89.1%, $\Delta = -0.045$); objects follow (500M: 88.9%; 2.2B: 85.8%). Holistic captioning and spatial description are hardest at 500M (77.5% and 77.6%), but scale closes the gap strongly: holistic reaches 86.0% (+8.5 pp) and spatial 82.1% (+4.5 pp) at 2.2B. The events dimension inverts: recovery falls from 84.1% (500M) to 78.5% (2.2B, -5.6 pp) as the 2.2B base improves substantially on event enumeration (base: 0.201 \rightarrow 0.340), raising the compression target beyond the adapter’s fixed rank.

5.3 Video Question Answering

VIDEO2LORA is trained exclusively on captioning; video QA is entirely a zero-shot transfer task. The LLM judge passes non-inferiority and equivalence on **7 of 8** benchmark–scale combinations (Table 4).

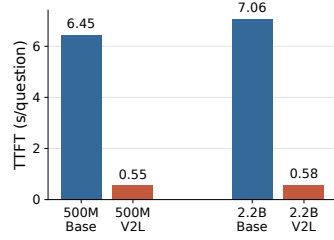
Per-benchmark judge analysis. Across the four QA benchmarks, VIDEO2LORA matches or exceeds the base on two of four at 500M and one of four at 2.2B, with NExT-QA being the standout: VIDEO2LORA *surpasses* the base at both scales,

with the 500M CI lying entirely above zero. The single failure—PLM-SGQA at 2.2B—is instructive rather than representative. The same benchmark passes comfortably at 500M points does not point to a fundamental limitation of parametric QA internalization.

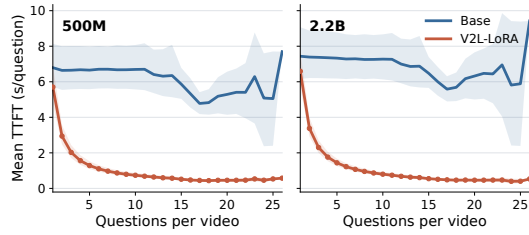
Token-F1 and the verbosity effect. Token-F1 diverges from the judge on short-answer QA. Token-F1 exposes a strong format mismatch on short-answer QA. This does not necessarily imply semantic failure, but it shows that captioning-trained Video2LoRA tends to produce more verbose answers than the direct baseline. On ActivityNet-QA, VIDEO2LORA token-F1 falls to 12% of base at 500M (0.023 vs. 0.197) and 9% at 2.2B (0.013 vs. 0.149); on NExT-QA it is 53%—yet both pass the judge test. The base VLM gives short, often one-to-three-word answers; VIDEO2LORA, trained on captioning, generates verbose summaries. Token-F1 is penalised by both the length mismatch and paraphrase variation, while the judge evaluates semantic correctness independently of response length. Two contrasts support this interpretation: PLM-SGQA—with longer, descriptive references—reverses direction entirely (500M: $\Delta = +0.145$; 2.2B: $\Delta = +0.111$); VidCapBench reaches near-parity ($\Delta = -0.007 / +0.022$).

5.4 Frame and Resolution Generalization

VIDEO2LORA checkpoints were trained with uniform sampling at 12 frames and 384px res-

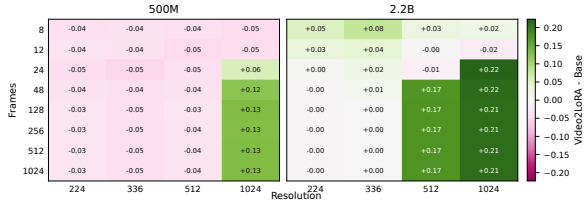


(a) Single-question average TTFT, with time taken to internalize the video accounted.

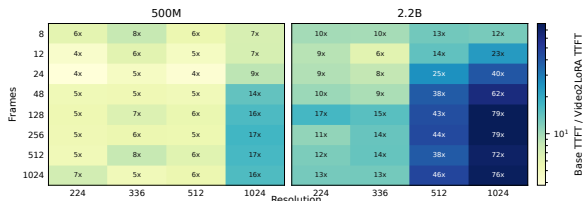


(b) Amortized TTFT per question vs. number of questions per video (shaded band = bootstrap 95% confidence interval).

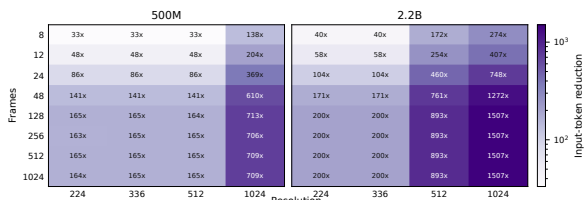
Figure 2: Inference efficiency on VidCapBench, comparing the base model and VIDEO2LORA.



(a) Change in mean Token-F1 from replacing in-context video tokens with VIDEO2LORA.



(b) Query-time TTFT speedup of VIDEO2LORA over the base video-in-context model.



(c) Input-token reduction achieved by VIDEO2LORA during answering.

Figure 3: Scaling behavior on VDC background captioning across frame count and spatial resolution.

olution. We test out-of-distribution scaling on VDC background captioning by sweeping {8, 12, 24, 48, 128, 256, 512, 1024} frames and {224, 336, 512, 1024} resolution for both 500M and 2.2B models. We compare video-in-context inference with VIDEO2LORA using Token-F1, query-time TTFT (Time to First Token), and input-token reduction during answering (Fig. 3).

Despite being trained at a single frame count-resolution setting, VIDEO2LORA remains stable across the sweep. For 500M, VIDEO2LORA is close to the base model overall, with an average Token-F1 change of -0.012 . At 1024px

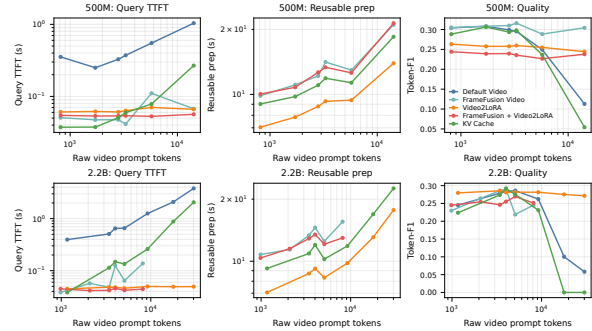


Figure 4: Efficiency comparison across video-token budgets. Columns report query TTFT, reusable preprocessing cost, and Token-F1.

and high frame counts, however, VIDEO2LORA outperforms the base model by $+0.12$ to $+0.13$ Token-F1. This large gain is partly because direct video-in-context inference becomes unstable in this regime: the base model often degenerates into repetitive or gibberish generations when significantly large number of visual tokens are supplied. The efficiency gains grow with video scale. VIDEO2LORA reduces query TTFT by a geometric mean of $6.7\times$ for 500M and $20.1\times$ for 2.2B, with maximum speedups of $17.2\times$ and $79.1\times$, respectively (Fig. 3b). This is explained by the token compression in Fig. 3c: VIDEO2LORA reduces answer-time input tokens by $150\times$ for 500M and $302\times$ for 2.2B on average, reaching $713\times$ and $1507\times$ at the largest settings, since it passes **zero tokens** during inference.

5.5 Inference Efficiency

VidCapBench is a natural setting for evaluating inference efficiency because each video is associated with multiple questions: in our evaluation split, 100 videos produce 1,523 total queries, or 15.23 questions per video on average. This matches the intended use case of VIDEO2LORA: the video is processed once to produce a video-specific LoRA,

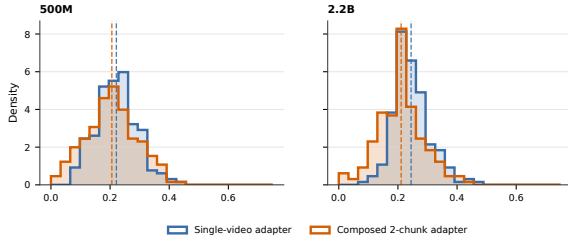


Figure 5: Two-chunk adapter composition on VDC.

and the adapter is reused for all subsequent questions about the same video. Thus, unlike direct in-context video inference, which repeatedly pays the cost of encoding and conditioning on the video, VIDEO2LORA pays a one-time setup cost and amortizes it over repeated queries. Figure 2 shows this amortization effect on both the 500M and 2.2B backbones. Averaged over all VidCap-Bench queries, VIDEO2LORA reduces TTFT from 6.45s to 0.55s for the 500M model, an $11.75\times$ speedup, and from 7.06s to 0.58s for the 2.2B model, a $12.11\times$ speedup (Figure 2a). The prefix-amortization curve in Figure 2b shows that after 5 questions, amortized TTFT drops to 1.29s for 500M and 1.44s for 2.2B; after 10 questions, it falls to 0.74s and 0.80s, respectively.

Figure 4 further studies different video inference strategies on 640 samples with varying token counts (by doing using a resolution, frame count grid). We compare Video2LoRA and the default setting with, FrameFusion (Fu et al., 2025) (a token compression technique, compression factor 4), and KV caching. We also use FrameFusion with VIDEO2LORA, to show VIDEO2LORA is compatible with existing token compression techniques. Across token budgets, VIDEO2LORA is the only method that provides all three properties needed for repeated video querying: (1) query TTFT stays nearly constant and low as video tokens grow, (2) reusable preparation is competitive or fastest and much cheaper than KV caching at scale, and (3) output quality remains stable as token count increases. In contrast, the default baseline, token compression results and KV caching scale with token counts. Together, these results show that VIDEO2LORA converts video conditioning from a repeated per-query overhead into a reusable video-specific computation.

5.6 Chunk Composition

Video2LoRA internalizes a video by generating a LoRA adapter from its visual context. Although

the model is trained to produce adapters for single video contexts, the adapter representation admits a simple test-time composition operation: independently internalize two temporal chunks of the same video, concatenate the resulting LoRA ranks, and decode from the composed adapter. We evaluate whether this operation produces coherent video-level generations, rather than degenerate text or captions tied to only one chunk.

We use the VDC short-caption and detailed-caption subsets, with 100 videos from each subset. Each video is split into two equal temporal halves. We compare two conditions: *single-video adapter*, where the full video is internalized as one adapter, and *composed two-chunk adapter*, where the two halves are internalized independently and the resulting adapters are composed before generation. Both conditions use 12 frames per adapter and the same text prompt. Figure 5 shows the resulting token-F1 score distributions against the VDC reference captions. The composed adapter remains close to the single-video adapter at both model scales. For VIDEO2LORA at 500M, the composed adapter retains 93.1% of the single-video adapter’s mean token-F1, with a mean score of 0.206 compared to 0.221. At 2.2B, it retains 86.2%, with a mean score of 0.211 compared to 0.245.

6 Conclusion

We introduced **Video2LoRA**, showing that parametric video internalization is achievable: a Perceiver hypernetwork converts a video into a LoRA adapter in a single forward pass, enabling a frozen VLM to answer queries with no visual tokens in context. Trained only on captioning, Video2LoRA is statistically non-inferior and equivalent to direct video-in-context inference across all five captioning benchmarks at both 500M and 2.2B scales, and transfers zero-shot to video QA on seven of eight benchmark-scale pairings. It remains stable at 1,024 frames where direct inference degenerates, achieves 6–76 \times lower query latency with up to 1,500 \times fewer answer-time tokens, and supports rank-space adapter composition for long-video internalization without dedicated training. Across token budgets, VIDEO2LORA uniquely combines near-constant query TTFT, scalable preprocessing costs below KV caching and token compression, and stable output quality at longer contexts.

7 Limitations

VIDEO2LORA demonstrates that video context can be internalized into generated adapter weights, enabling text-only querying after a one-time video processing step. Our current implementation trains a separate hypernetwork for each target VLM scale, and we evaluate it on the 500M and 2.2B SmolVLM2 backbones. Extending the same framework to additional VLM families, larger models, and shared or scale-transferable hypernetworks is an important direction for future work.

The present training setup uses captioning and summarization supervision. This makes transfer to video question answering a zero-shot setting, where answer style can differ from the direct video-in-context baseline. In particular, VIDEO2LORA sometimes produces more descriptive answers for short-answer QA, which can lower lexical-overlap metrics even when the answer is judged semantically appropriate. Future work can incorporate mixed captioning–QA supervision, answer-length control, or lightweight calibration for task-specific formats.

Because VIDEO2LORA converts a video into a compact adapter, the representation may emphasize high-level scene and event information over some fine-grained details. This is most relevant for tasks requiring precise camera, spatial, or object-level distinctions. Adaptive-rank adapters, richer internalization objectives, or more targeted supervision may improve preservation of these details.

Finally, our chunk-composition experiment is an initial two-chunk test. The result suggests that independently generated adapters can be combined in rank space, but the current operation does not explicitly model temporal order. More structured composition mechanisms and audio-visual internalization remain promising extensions.

8 Ethics Statement

Our research does not use any personally identifiable information (PII) and all datasets employed in this work are used in accordance with their respective licenses.

References

Lucas Caccia, Alan Ansell, Edoardo M. Ponti, Ivan Vulic, and Alessandro Sordani. 2025. Training plug-and-play knowledge modules with deep context distillation. *arXiv preprint arXiv:2503.08727*.

- Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D. Manning. 2025. AuroraCap: Efficient, performant video detailed captioning and a new benchmark. In *International Conference on Learning Representations*.
- Rujikorn Charakorn, Edoardo Cetin, Shinnosuke Uesaka, and Robert Lange. 2026. Doc-to-LoRA: Learning to instantly internalize contexts. *arXiv preprint arXiv:2602.15902*.
- Xinlong Chen, Yuanxing Zhang, Chongling Rao, Yushuo Guan, Jiaheng Liu, Fuzheng Zhang, Chengru Song, Qiang Liu, Di Zhang, and Tieniu Tan. 2025a. VidCapBench: A comprehensive benchmark of video captioning for controllable text-to-video generation. In *Annual Meeting of the Association for Computational Linguistics*.
- Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. 2025b. LongVILA: Scaling long-context visual language models for long videos. In *International Conference on Learning Representations*.
- Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, Muhammad Maaz, Yale Song, Tengyu Ma, Shuming Hu, Hanoona Rasheed, Peize Sun, Po-Yao Huang, Daniel Bolya, Suyog Jain, Miguel Martin, Huiyu Wang, Nikhila Ravi, Shashank Jain, Temmy Stark, and 10 others. 2025. PerceptionLM: Open-access data and models for detailed visual understanding. *arXiv preprint*.
- Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. 2024. Finevideo. <https://huggingface.co/datasets/HuggingFaceFV/finevideo>.
- Tianyu Fu, Tengxuan Liu, Qinghao Han, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, and Yu Wang. 2025. Framefusion: Combining similarity and importance for video token reduction on large vision language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22654–22663.
- David Ha, Andrew M. Dai, and Quoc V. Le. 2017. HyperNetworks. In *International Conference on Learning Representations*.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference*

667	<i>on Machine Learning</i> , volume 139 of <i>Proceedings of Machine Learning Research</i> , pages 4651–4664. PMLR.	719
668		720
669		721
670	Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In <i>IEEE International Conference on Computer Vision</i> .	722
671		723
672		724
673		725
674	Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In <i>Conference on Empirical Methods in Natural Language Processing</i> .	726
675		727
676		728
677		729
678	Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. 2025. TokenPacker: Efficient visual projector for multimodal LLM. <i>International Journal of Computer Vision</i> .	730
679		731
680		732
681		733
682	Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In <i>Annual Meeting of the Association for Computational Linguistics</i> .	734
683		735
684		736
685		737
686	Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> .	738
687		739
688		740
689		741
690	Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2025. SmolVLM: Redefining small and efficient multimodal models. <i>arXiv preprint arXiv:2504.05299</i> .	742
691		743
692		744
693		745
694		746
695		747
696		748
697		749
698	Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022. Fast model editing at scale. In <i>International Conference on Learning Representations</i> .	750
699		751
700		752
701		753
702	Jesse Mu, Xiang Lisa Li, and Noah D. Goodman. 2023. Learning to compress prompts with gist tokens. In <i>Advances in Neural Information Processing Systems</i> .	754
703		755
704		756
705	Rui Qian, Xiao-Wen Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. 2024. Streaming long video understanding with large language models. In <i>Advances in Neural Information Processing Systems</i> .	757
706		758
707		759
708		760
709		761
710	Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. 2025. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. In <i>IEEE International Conference on Computer Vision</i> .	762
711		763
712		764
713		765
714	Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. NExT-QA: Next phase of question-answering to explaining temporal actions. In <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> .	766
715		767
716		768
717		769
718		770
	Yifan Xu, Xinhao Li, Yichun Yang, Desen Meng, Rui Huang, and Limin Wang. 2025. CaReBench: A fine-grained benchmark for video captioning and retrieval. <i>arXiv preprint arXiv:2501.00513</i> .	
	An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. <i>Qwen3 technical report</i> . <i>Preprint</i> , arXiv:2505.09388.	
	Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. ActivityNet-QA: A dataset for understanding complex web videos via question answering. In <i>AAAI Conference on Artificial Intelligence</i> .	
	Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In <i>Conference on Empirical Methods in Natural Language Processing</i> .	
	Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024. Long context transfer from language to vision. <i>Transactions on Machine Learning Research</i> .	
	A LLM Judge Evaluation	
	We use an LLM judge for two purposes: reference-based quality scoring and reference-free output preservation. The judge is Qwen/Qwen3-VL-30B-A3B-Thinking-FP8, served locally with vLLM through an OpenAI-compatible API. For the main reported judge scores, we use text-only judging: the judge receives the task prompt, reference text, and model output, but no video frames. We set temperature to 0, use a maximum of 1024 output tokens for reference-based scoring, and request JSON-formatted outputs. For pure output similarity, we use the same judge with a maximum of 768 output tokens.	
	For reference-based quality, each candidate is scored independently against the reference. The judge is not shown model names. For auxiliary paired judgments, the direct baseline and VIDEO2LORA outputs are anonymized as Candidate A and Candidate B, and their order is randomized with a fixed seed. These paired judgments are used as an audit and are not the primary metric unless explicitly reported.	
	A.1 Reference-Based Captioning Judge	
	For captioning and description tasks, the judge measures semantic coverage of the reference caption. Extra details that are absent from the reference are	

771
772

not penalized unless they directly contradict the reference.

Captioning Judge Prompt

Task guidance: This is video caption/reference evaluation. Judge semantic coverage, factual precision, and task fit.

Prompt: {prompt}
Reference caption: {reference}
Candidate caption: {candidate}

Evaluate only semantic coverage of the reference caption for the requested caption task. Do not reward or penalize writing style, fluency, verbosity, or formatting, except when the candidate is invalid or impossible to understand. Extra details absent from the reference are not automatically wrong. Record notable extra details separately. Penalize only direct contradictions or missing reference facts. A direct contradiction requires that the candidate and reference cannot both be true.

Coverage scoring rubric: 5 = complete coverage; 4 = mostly complete; 3 = partial; 2 = weak overlap; 1 = minimal overlap; 0 = unrelated, invalid, nonsensical, or contradicts the main reference event.

Return only valid JSON with fields: coverage_score integer 0-5; coverage_label; covered_reference_facts; missing_reference_facts; direct_contradictions; extra_details; extra_details_type; rationale.

773
774
775
776
777
778

A.2 Reference-Based QA Judge

For QA tasks, the judge first extracts the answer implied by the model output and then compares it to the reference answer. This avoids over-penalizing verbose outputs that contain the correct answer.

QA Judge Prompt

Task guidance: This is video question answering. Judge semantic answer correctness. Allow paraphrases and indirect answers. For yes/no references, infer yes/no from the candidate if the candidate clearly implies it.

Question: {question}
Reference answer: {reference}
Candidate response: {candidate}

First extract the candidate's answer to the question, then compare that extracted answer to the reference. Do not require the candidate to be concise. A verbose response can be correct if it contains or clearly gives the answer. If the candidate describes the same scene but does not answer the requested attribute, action, location, or count, score it as low-to-partial rather than correct.

Scoring rubric: 5 = fully correct; 4 = correct main answer with minor missing specificity or harmless extra detail; 3 = partially correct; 2 = related but does not clearly answer; 1 = minimal overlap; 0 = contradiction, different answer, or invalid.

Return only valid JSON with fields: extracted_answer; score integer 0-5; answer_label; contains_answer; direct_contradiction; extra_details_affect_score;

779

missing_key_answer_parts; contradictory_parts; rationale.

780

B Evaluation Prompts and Task Templates

781
782

This appendix provides the exact evaluation prompts and task-specific templates used across all the benchmarks in our experiments.

783
784
785

B.1 Video Captioning and Description Benchmarks

786
787

Below are the prompts used to generate descriptions for whole videos, clips, and specific features (e.g., spatial layout, temporal progression, and cinematography style).

788
789
790
791

ActivityNet Captions

"Describe what is happening in this video."

792

PLM-RDCap

"Describe what happens in this video."

793

PLM-RCap

"Describe what happens in this video clip."

794

Video Description Corpus (VDC)

Short Caption:

"Summarize this video in one detailed sentence."

Detailed Caption:

"Describe this video in detail."

Camera:

"Describe the camera work, framing, and viewpoint in this video."

Background:

"Describe the background, setting, and environment in this video."

Main Object:

"Describe the main subject and its actions in this video."

795

CaReBench

Caption:

"Describe the video in as much useful

796

visual detail as possible. Include the main activity, visible people or objects, scene context, appearance, and any important visual details that help explain what is happening.”

Events:

“Describe the key visible events in chronological order. Include all important actions and changes you can observe, with enough detail to distinguish each event clearly.”

Objects:

“Describe the important visible objects and entities in the video in as much useful detail as possible. Include their appearance, location, and role in the scene when visible.”

Spatial Caption:

“Describe the spatial layout in as much useful detail as possible: where the people, objects, and scene elements are located, how they are positioned relative to each other, and what parts of the scene are in the foreground, background, left, right, center, above, or below.”

Temporal Caption: “Describe the temporal progression in as much useful detail as possible. Explain what happens over time, the order of visible actions, and how the scene or subjects change from the beginning to the end.”

B.2 Video Question Answering (QA) Benchmarks

For question answering tasks, templates are structured to format the inputs and instructions depending on whether choices are provided (offered options) or hidden.

NExT-QA

{question}
Answer only the question, in one sentence.

ActivityNet-QA

{question}
Answer only the question, in one sentence.

PLM-SGQA

{question}
Answer only the question, in one sentence.

VidCapBench

{question}
Answer only the question, in one sentence.

C Rank-Direction Ablation

C.1 Setup

We test whether different rank directions in a generated LoRA adapter contribute unequally to captioning performance. The ablation is run on 500 examples from the ActivityNet Captions evaluation split (Krishna et al., 2017), using the 2.2B VIDEO2LORA checkpoint. For each example, we generate the video-conditioned rank-16 adapter and decompose it into rank-slice pairs $\{(A_r, B_r)\}_{r=1}^{16}$, where $A_r \in \mathbb{R}^{1 \times d_{in}}$ and $B_r \in \mathbb{R}^{1 \times d_{out}}$. Under our row-vector implementation, rank slice r contributes

$$\Delta y_r = s(xA_r^\top)B_r.$$

We score each slice by the Frobenius norm product

$$\|A_r\|_F \cdot \|B_r\|_F.$$

We evaluate four selection strategies across budgets $k \in \{1, 2, 4, 8, 16\}$:

- **Top- k :** retain the k highest-scoring rank slices.
- **Bottom- k :** retain the k lowest-scoring rank slices.
- **Random- k :** retain k randomly selected slices, averaged over 3 seeds.
- **Remove-Top- k :** remove the k highest-scoring slices and retain the remaining $16 - k$.

We report Token-F1 against reference captions with 95% bootstrap confidence intervals over examples.

C.2 Analysis

Table 5 reports the numerical ablation results, and Figure 6 visualizes the same rank-pruning trajectories.

k	Top- k	Bottom- k	Random- k	Remove-Top- k
0 (Zero)	0.0561 [.052,.060]	0.0561	0.0561	0.0561
1	0.0894 [.083,.096]	0.0556 [.052,.060]	0.0709 [.068,.074]	0.1317 [.123,.141]
2	0.1097 [.102,.118]	0.0662 [.062,.071]	0.0712 [.069,.074]	0.1277 [.118,.137]
4	0.1196 [.111,.128]	0.0803 [.074,.086]	0.0991 [.095,.103]	0.1275 [.119,.137]
8	0.1264 [.118,.135]	0.1128 [.104,.121]	0.1215 [.117,.126]	0.1128 [.104,.121]
16 (Full)	0.1262 [.117,.136]	0.1262	0.1262	0.0561

Table 5: Token F1 scores under rank-direction ablation on ActivityNet Captions. Brackets denote 95% confidence intervals. *Full Adapter* ($k = 16$) and *Zero Adapter* ($k = 0$) serve as upper and lower baselines.

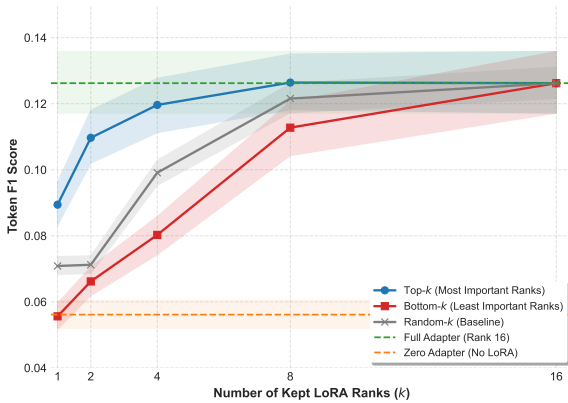


Figure 6: Rank-direction ablation on ActivityNet Captions. Top- k rank slices recover performance faster than random or bottom- k slices, suggesting that the Frobenius norm product is a useful heuristic for rank importance. The Remove-Top- k curve has a higher point estimate than the full adapter at small k , but this should be interpreted cautiously because confidence intervals overlap.

Rank directions are redundant but not exchangeable. The generated adapters are compressible along the rank dimension. Retaining the top-8 rank slices gives a Token-F1 of 0.1264, close to the full rank-16 adapter score of 0.1262. At $k = 4$, the top- k adapter reaches 0.1196, which is 94.8% of the full adapter’s absolute Token-F1 and recovers 90.6% of the improvement over the zero-adapter baseline. This suggests that much of the useful adaptation is concentrated in a subset of rank directions.

Norm product is a useful heuristic for rank importance. The Frobenius norm product separates useful from less useful directions. At $k = 1$, Top- k reaches 0.0894, while Bottom- k reaches 0.0556, slightly below the zero-adapter baseline of 0.0561. Random- k generally falls between Top- k and Bottom- k at matched budgets. Thus, high-

norm rank slices tend to be more useful, although the norm product should be treated as a heuristic rather than a complete causal explanation.

Removing the dominant direction has a higher point estimate. Removing the highest-norm rank slice gives a higher point estimate than the full adapter, increasing Token-F1 from 0.1262 to 0.1317. Removing the top four slices also remains close to the full adapter at 0.1275. Since the confidence intervals overlap, we treat this as suggestive rather than conclusive. One possible explanation is that the dominant direction captures a generic captioning prior, and removing it shifts generation toward more video-specific directions.

Rank ordering is stable across examples. The rank ordering is highly consistent across the 500 examples: rank direction R11 is the highest-scoring direction in all examples, while R7 is consistently among the lowest-scoring directions. This suggests that the hypernetwork learns a stable output coordinate system for rank directions, rather than assigning importance arbitrarily for each video.

D Interpreting Hypernetwork-Generated Adapters

D.1 Setup

We use two diagnostic interventions to study how generated adapters affect the frozen 2.2B answer model: layer-wise adapter removal and direct logit attribution. The experiments are run on CareBench diagnostic examples, including caption and spatial-caption prompts.

Each example is scored by teacher-forced log-probability under the frozen answer model with the generated adapter active. Since these diagnostics use open-ended reference strings, we score each reference string and use the highest-scoring

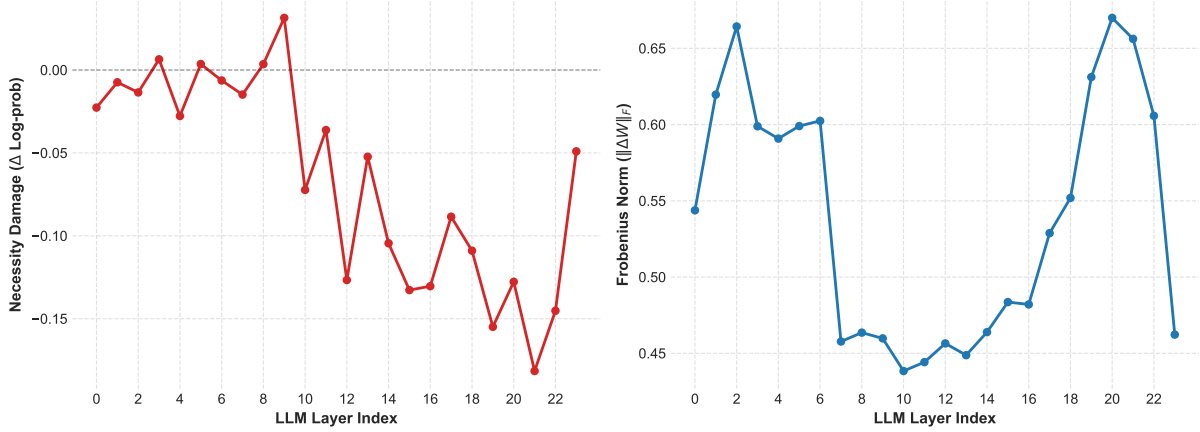


Figure 7: Layer-wise adapter-removal diagnostic. **Left:** signed removal effect from zeroing one layer’s generated LoRA update; negative values indicate that removing the layer lowers the score. **Right:** Frobenius norm $\|\Delta W\|_F$ of generated LoRA weights across layers.

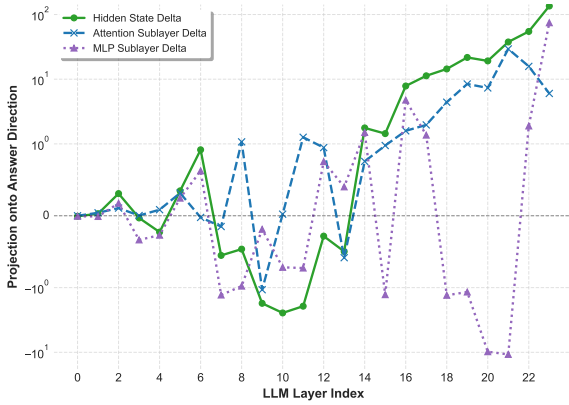


Figure 8: Direct logit attribution of adapter-induced representation shifts projected onto the diagnostic answer direction across 24 LLM layers. Later layers show the largest alignment with the answer direction, suggesting late-layer logit steering.

reference for the diagnostic. Candidate strings may contain multiple tokens, so we score a candidate string z by length-normalized teacher-forced log-probability:

$$\ell(z | p) = \frac{1}{|z|} \sum_{t=1}^{|z|} \log P(z_t | z_{<t}, p). \quad (10)$$

The scalar diagnostic score is therefore

$$\mathcal{S} = \max_{r \in \mathcal{R}} \ell(r | p), \quad (11)$$

where \mathcal{R} is the set of reference strings for the example.

For direct logit attribution, we need a direction in the output-embedding space. We use the

mean output embedding of the selected reference tokens and denote the normalized direction by \hat{d} . This gives a single diagnostic direction toward the reference answer/caption.

D.2 Layer-Wise Adapter Removal

For each transformer layer ℓ , we zero out only the generated LoRA update at that layer and recompute the diagnostic score. We report the signed removal effect

$$\text{Effect}_\ell = \mathcal{S}_{\text{without } \ell} - \mathcal{S}_{\text{full}}. \quad (12)$$

Negative values indicate that removing the layer lowers the score, so the layer’s adapter update is useful under this diagnostic. Values near zero indicate little measurable effect from removing that layer.

Figure 7 shows a mismatch between generated-weight norm and functional effect. Some early layers receive relatively large LoRA updates, but removing them changes the diagnostic score only weakly. In contrast, several later layers produce larger negative removal effects, indicating that their adapter updates matter more for the scored prediction. This suggests that the adapter is not used uniformly across the transformer stack: early updates may shape intermediate representations, while later updates appear more directly connected to the final answer/caption likelihood.

This also shows that Frobenius norm alone is not a complete measure of adapter importance. Large generated weights can be weakly causal under this intervention, whereas smaller or

comparable later-layer updates can have stronger effects on the output score. We therefore interpret the result as a norm–function dissociation, not as a full causal explanation of the adapter mechanism.

D.3 Direct Logit Attribution

We next ask where the adapter-induced representation shift becomes aligned with the diagnostic target direction. Let

$$\Delta x_\ell = x_\ell^{\text{adapter}} - x_\ell^{\text{base}}$$

denote the residual-stream shift at layer ℓ , and let Δa_ℓ and Δm_ℓ denote the corresponding attention and MLP sublayer shifts. We project these shifts onto the diagnostic answer direction:

$$\begin{aligned} \text{DLA}_\ell &= \Delta x_\ell \cdot \hat{d}, \\ \text{DLA}_\ell^{\text{attn}} &= \Delta a_\ell \cdot \hat{d}, \\ \text{DLA}_\ell^{\text{MLP}} &= \Delta m_\ell \cdot \hat{d}. \end{aligned} \quad (13)$$

Figure 8 shows that the adapter-induced shift is weakly aligned with the diagnostic direction in early and middle layers, but becomes much more aligned in later layers. This matches the layer-removal result: the adapter’s effect becomes most visible close to the output logits.

The sublayer breakdown suggests that both attention and MLP components contribute to this late-stage steering. Rather than claiming that the generated adapter implements a specific memory mechanism, we interpret the pattern more conservatively: VIDEO2LORA appears to induce representation changes throughout the network, but the changes most directly aligned with the target answer/caption emerge in later layers.

E Training Details

Table 6 summarizes the main training configuration for the two VIDEO2LORA model scales. In both runs, only the hypernetwork parameters are trained; the video encoder and answer model remain frozen.

Both models use rank-16 generated LoRA adapters, 12 uniformly sampled frames, a maximum video dimension of 384 pixels, Perceiver latent size 512, learning rate 1×10^{-4} , warmup ratio 0.03, and weight decay 0.01. The 500M model is trained for 9,000 steps on 4 A100 GPUs for 37 wall-clock hours, with per-device batch size 48 and gradient accumulation 2, giving an effective batch size of 384. The 2.2B model is trained for

Setting	500M	2.2B
Training steps	9,000	7,000
GPUs	4×A100	6×A100
Wall-clock training time	37 hours	201 hours
Per-device batch size	48	8
Gradient accumulation steps	2	5
Effective batch size	384	240
LoRA rank	16	16
Sampled frames	12	12
Max video dimension	384 px	384 px
Perceiver latent size	512	512
Learning rate	1×10^{-4}	1×10^{-4}
Warmup ratio	0.03	0.03
Weight decay	0.01	0.01

Table 6: Training configuration for the 500M and 2.2B VIDEO2LORA runs. Wall-clock training time reports elapsed training time, not total GPU-hours. Effective batch size is computed as number of GPUs \times per-device batch size \times gradient accumulation steps.

7,000 steps on 6 A100 GPUs for 201 wall-clock hours, with per-device batch size 8 and gradient accumulation 5, giving an effective batch size of 240.

F Additional Results

F.1 Distribution Plots

Figures 9 and 10 show the LLM-judge score distributions and per-example score differences. Figures 11 and 12 show the corresponding token-F1 distributions and differences.

F.2 Spider Plots

Figures 13 and 14 show the QA and captioning spider plots.

G Qualitative Examples

Qualitative examples are shown in Figure 15, Figure 16, Figure 17, Figure 18, Figure 19, Figure 20, Figure 21, Figure 22, Figure 24, Figure 25, Figure 26, Figure 27, Figure 28, Figure 29, Figure 30, Figure 31, Figure 23.

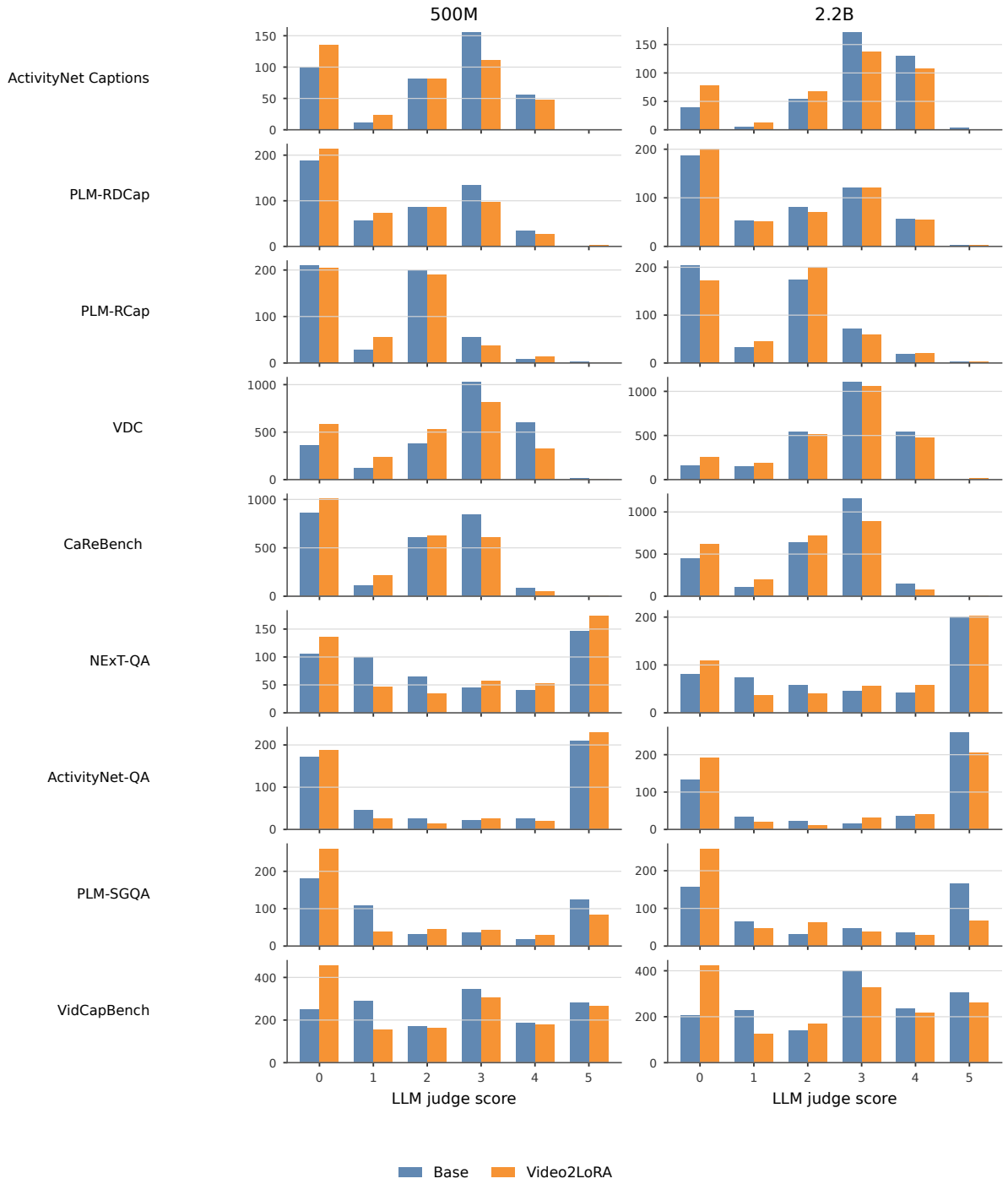


Figure 9: LLM-judge score distributions for the direct baseline and VIDEO2LORA.

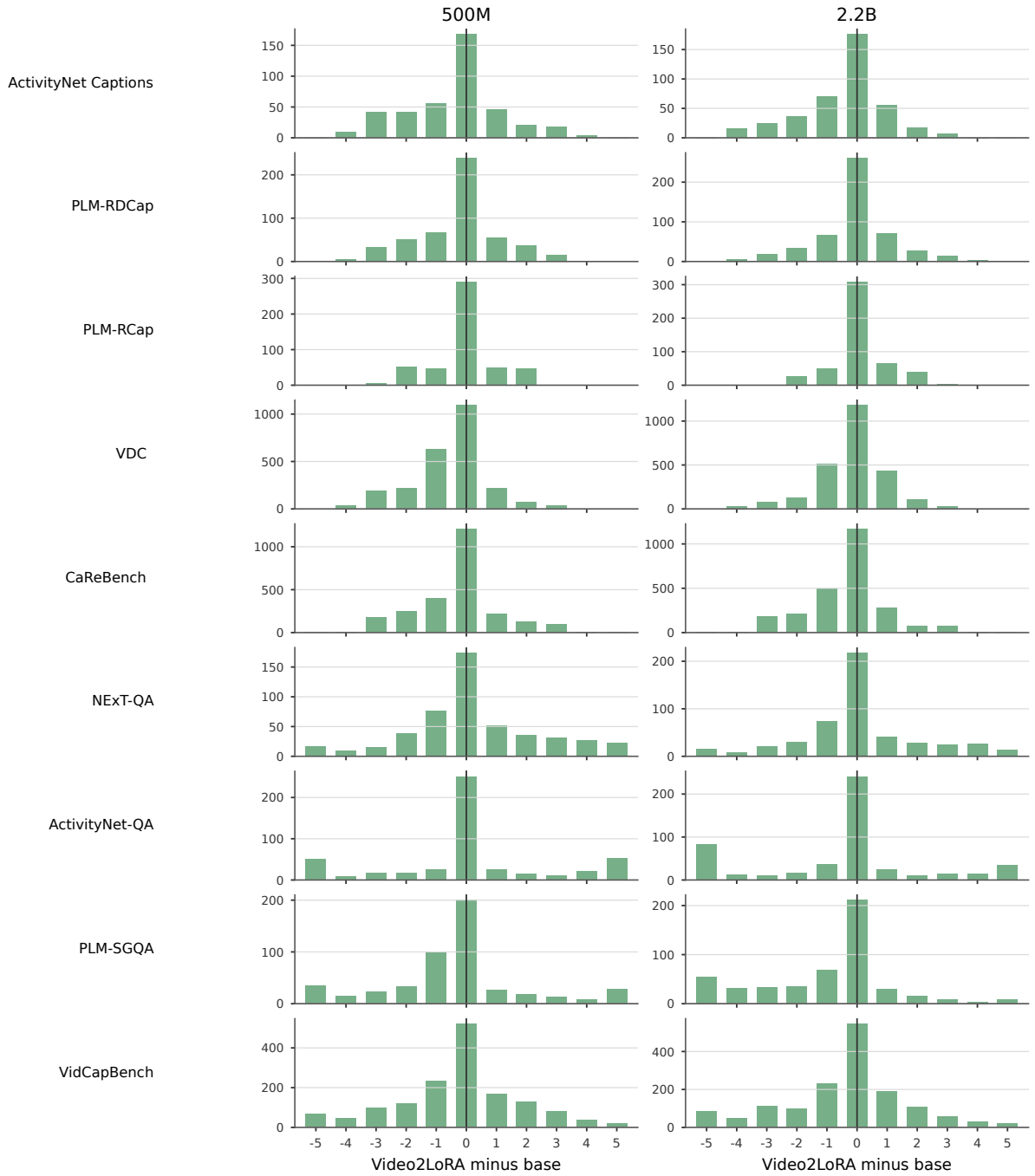


Figure 10: Per-example LLM-judge score differences between VIDEO2LORA and the direct baseline.

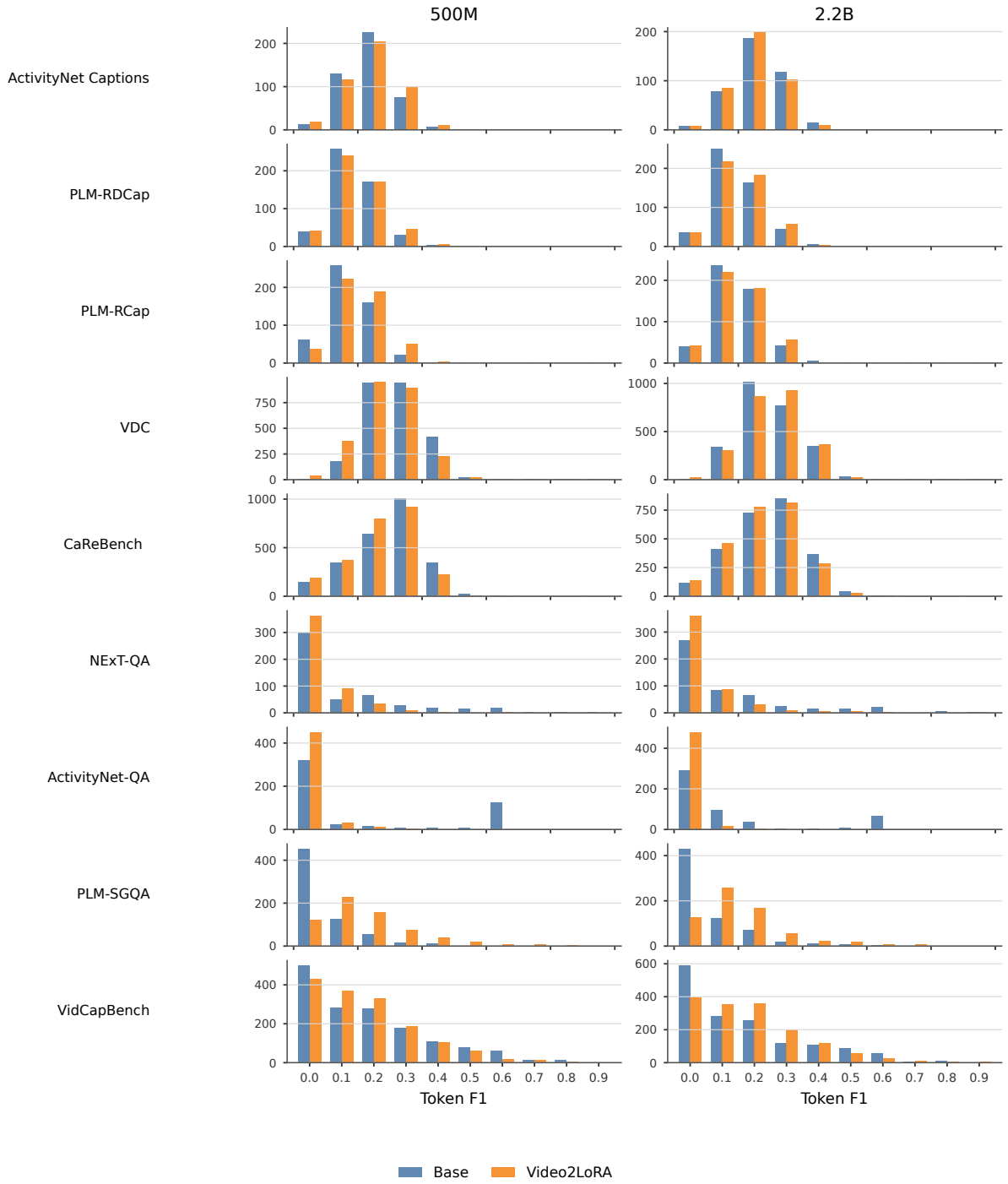


Figure 11: Token-F1 distributions for the direct baseline and VIDEO2LORA.

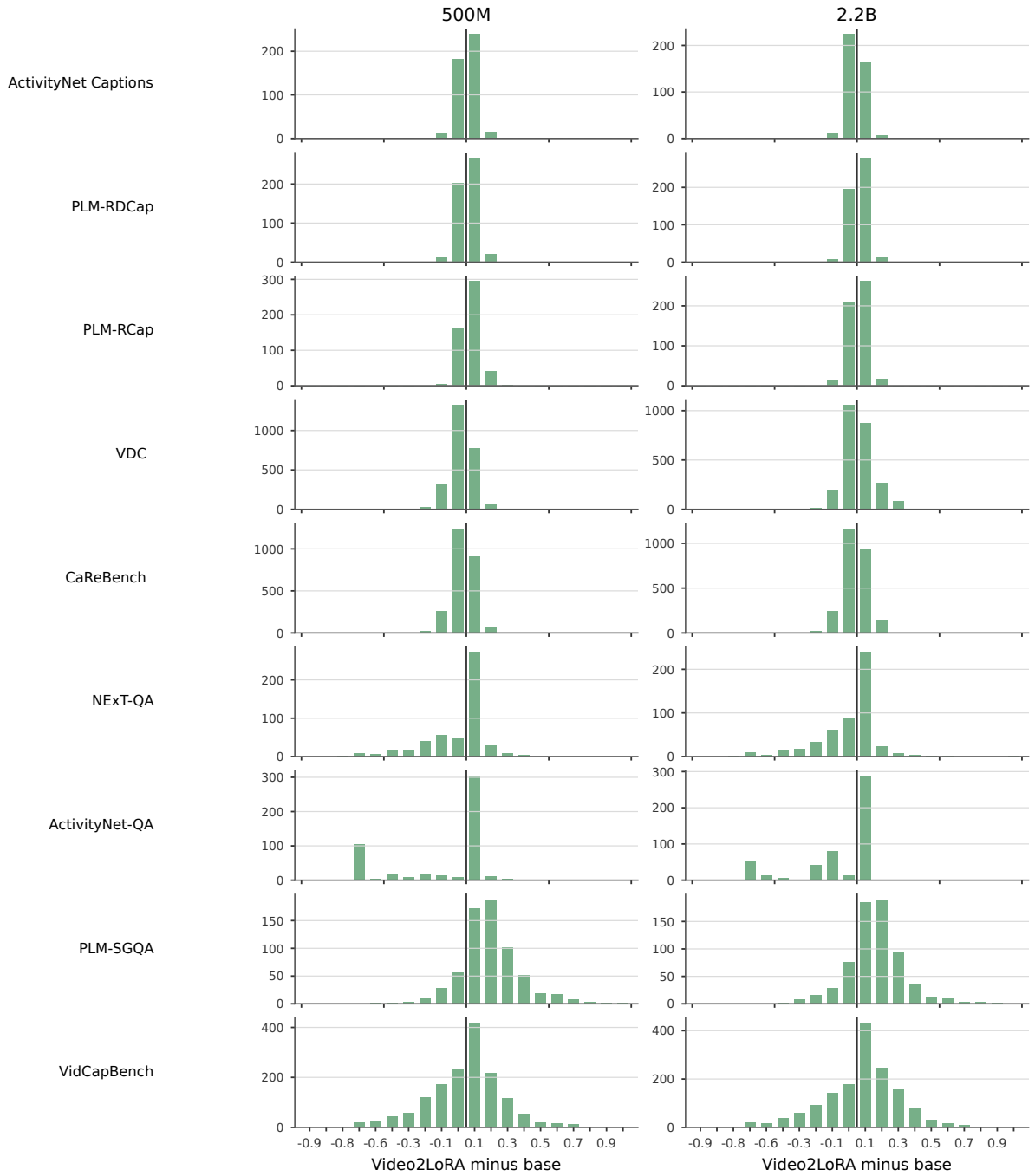


Figure 12: Per-example token-F1 differences between VIDEO2LORA and the direct baseline.

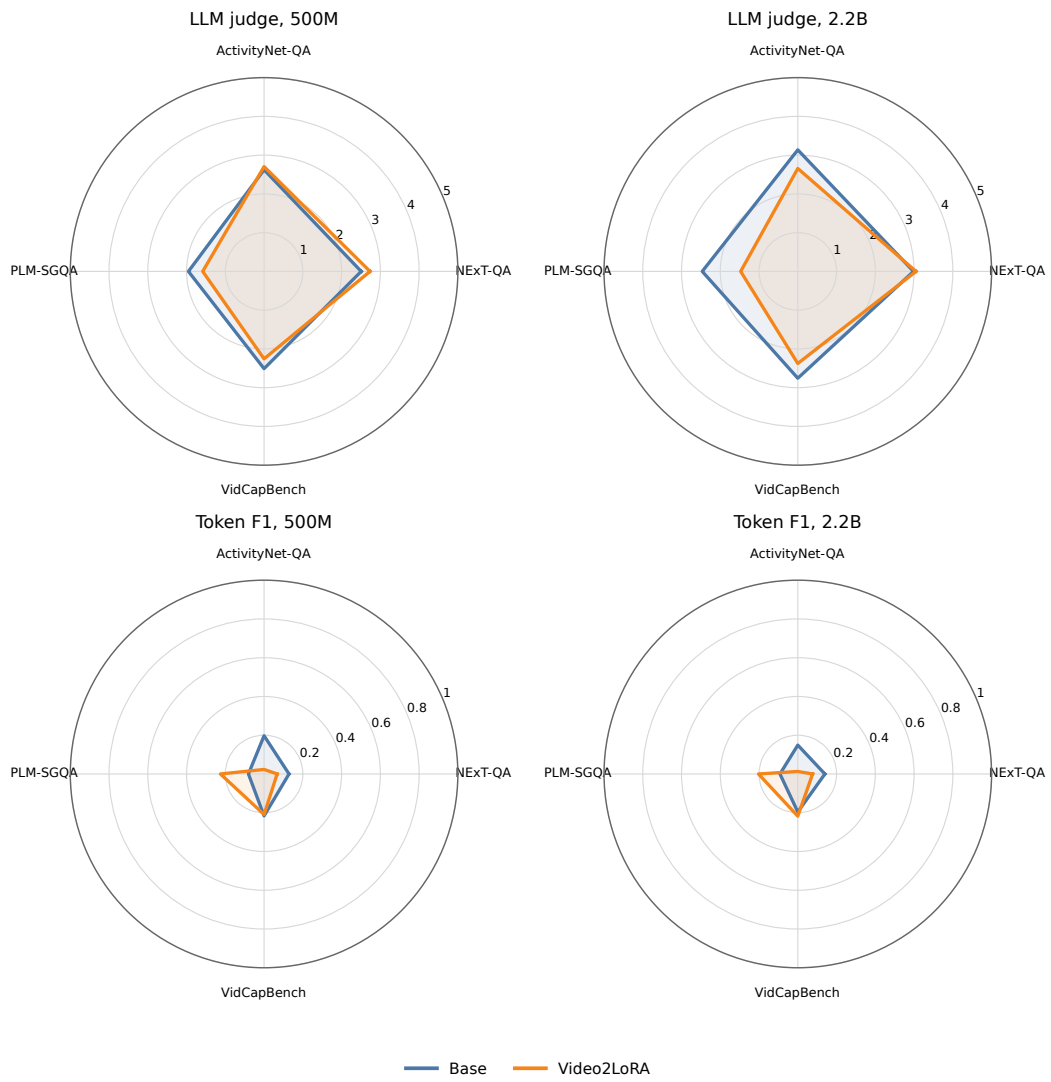


Figure 13: Spider plot for video question answering benchmarks.

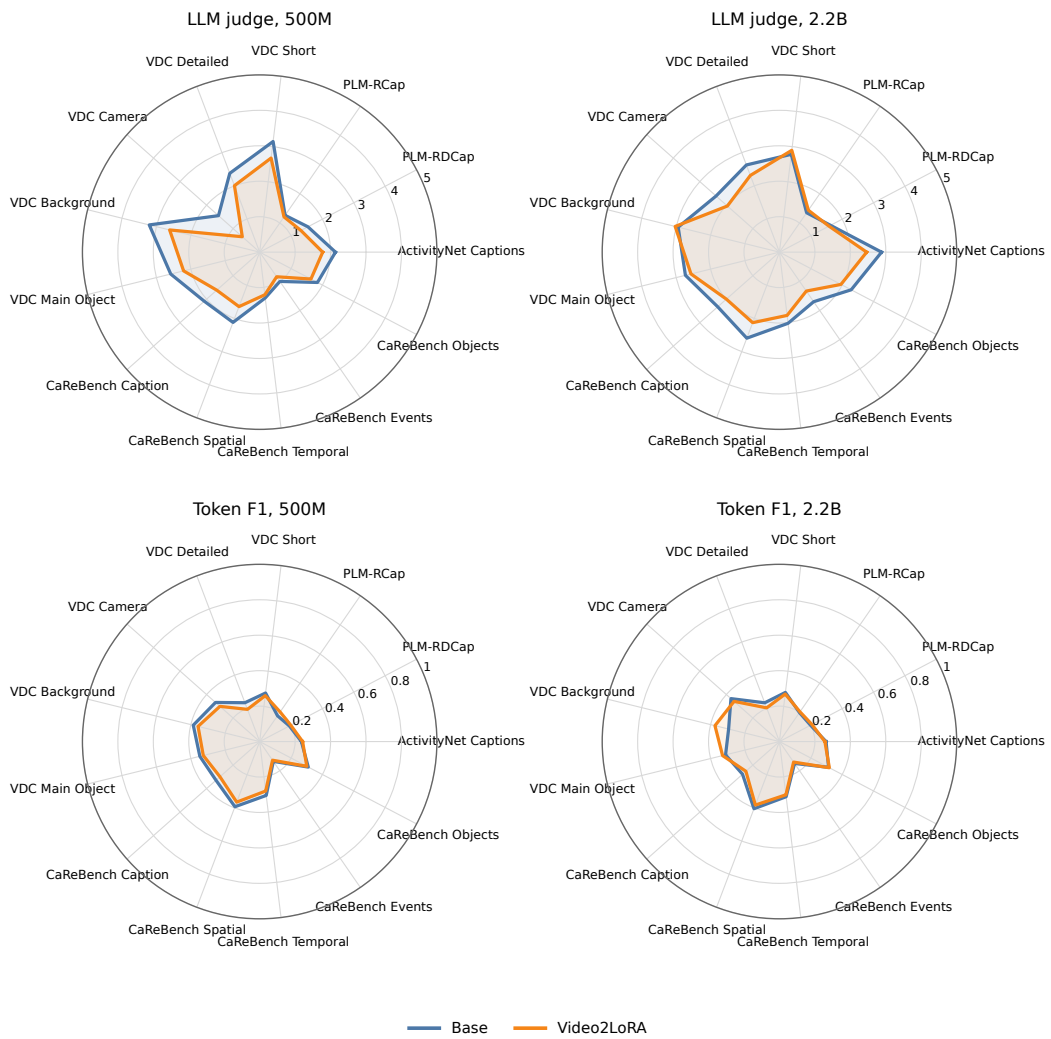




Figure 14: Spider plot for video captioning benchmarks.




what is on the right of the person in black clothes Answer in a single sentence.
Reference: window

<p>Base model</p> <p>A car.</p> <p>LLM Judge</p> <p>Justification: Candidate describes a car, not a window as required by the question.</p> <p style="text-align: right; border: 1px solid #ccc; border-radius: 5px; padding: 2px 5px;">0/5</p>	<p>Vid2LoRA</p> <p>The person in black clothes is standing in front of a sink with a green and white sign that reads "BREAKFAST" and "BREAKFAST" with a picture of a bowl of food.</p> <p>LLM Judge</p> <p>Justification: Candidate describes a white car but does not mention the window on the right.</p> <p style="text-align: right; border: 1px solid #ccc; border-radius: 5px; padding: 2px 5px;">1/5</p>
--	--




what is the person in the video doing Answer in a single sentence.
Reference: bodybuilding

<p>Base model</p> <p>The person is using a stationary bike.</p> <p>LLM Judge</p> <p>Justification: Candidate describes cycling, not bodybuilding.</p> <p style="text-align: right; border: 1px solid #ccc; border-radius: 5px; padding: 2px 5px;">2/5</p>	<p>Vid2LoRA</p> <p>A man is using an exercise bike in a gym.</p> <p>LLM Judge</p> <p>Justification: Candidate describes a different activity than the reference answer.</p> <p style="text-align: right; border: 1px solid #ccc; border-radius: 5px; padding: 2px 5px;">2/5</p>
--	--



what is the person in gray clothes doing Answer in a single sentence.
Reference: rubbing shoes

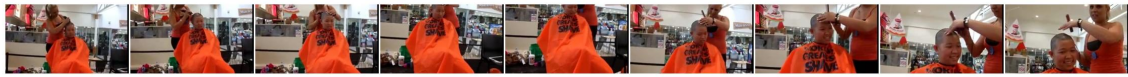
<p>Base model</p> <p>The person is holding a pair of red shoes.</p> <p>LLM Judge</p> <p>Justification: Candidate describes cleaning red boots instead of rubbing shoes.</p> <p style="text-align: right; border: 1px solid #ccc; border-radius: 5px; padding: 2px 5px;">3/5</p>	<p>Vid2LoRA</p> <p>A person in gray clothes is holding a bottle of shoe polish and a pair of shoes. They are also holding a red object, possibly a shoe brush. The person is talking to the camera and explaining something about the shoe polish and the shoes.</p> <p>LLM Judge</p> <p>Justification: The candidate describes the person holding a brush and demonstrating shoe-related products, implying shoe rubbing.</p> <p style="text-align: right; border: 1px solid #ccc; border-radius: 5px; padding: 2px 5px;">4/5</p>
--	---



what is the color of the biscuit Answer in a single sentence.
Reference: brown

<p>Base model</p> <p>Brown.</p> <p>LLM Judge</p> <p>Justification: Candidate directly states the color as Brown, matching the reference.</p> <p style="text-align: right; border: 1px solid #ccc; border-radius: 5px; padding: 2px 5px;">5/5</p>	<p>Vid2LoRA</p> <p>The biscuit is brown.</p> <p>LLM Judge</p> <p>Justification: The candidate clearly implies the biscuit color as golden brown.</p> <p style="text-align: right; border: 1px solid #ccc; border-radius: 5px; padding: 2px 5px;">5/5</p>
---	---

Figure 16: Qualitative examples from ActivityNetQA.



Describe the video in as much useful visual detail as possible. Include the main activity, visible people or objects, scene context, appearance, and any important visual details that help explain what is happening.

Reference: This video captures a woman giving a haircut to a child. The child is seated in a barber chair, draped in a bright orange cloth to prevent hair from falling onto his clothes, which feature black lettering across the chest area. The woman, standing behind him, is dressed in a rose-red sleeveless top and black shorts, with her blonde hair tied up in a high ponytail. In her right hand, she holds a black electric hair clipper. They are inside a barbershop, facing the camera. In the video, the woman begins cutting hair from the front right side of the child's head, her movements swift and precise as she works her way to the back. The camera then zooms in, showing her starting to trim the left side of his hair. Finally, she uses her right hand to tousle the hair at the front of the child's head to signify that the haircut is complete. The footage is shot from a third-person perspective, effectively showcasing the woman's rapid haircutting technique.

Base model

The video depicts a grooming session in a hair salon. The setting is a well-lit, modern salon with a variety of hair products and styling tools on display. A young boy is seated in a chair, receiving a haircut. The woman, wearing an orange cape, is carefully trimming his hair with a pair of scissors. The boy's hair is neatly cut, and the woman's cape is neatly tied around his shoulders. The background features a variety of hair products and styling tools, including a hair dryer and a hairbrush. The woman is also wearing a black tank top and a black cap.

LLM Judge

2/5

Justification: Candidate misses most reference facts and contradicts key details like using scissors instead of clipper. Included evidence: woman, haircut, seated, chair, orange. Missing evidence: captures, giving, child, barber, draped

Vid2LoRA

The video shows a woman in an orange shirt and a child in a black cape being shaved by a woman in a red shirt. The woman in the orange shirt is holding the child's head and the shaver, while the woman in the red shirt is holding the shaver and the child. The background is a well-lit room with various items and people in the distance. The woman in the orange shirt is wearing a black cape, and the child is wearing a black cape with a red and white design. The woman in the red shirt is wearing a red shirt with the words "WELCOME BACK".

LLM Judge

1/5

Justification: Candidate misidentifies subject and tool, missing all key reference details. Included evidence: woman, child, orange. Missing evidence: captures, giving, haircut, seated, barber



Describe the video in as much useful visual detail as possible. Include the main activity, visible people or objects, scene context, appearance, and any important visual details that help explain what is happening.

Reference: The video features a man shown in full face as he makes a phone call. Before the call connects, two lines of subtitles appear at the bottom of the screen, reading "618) 465-4545 if you ever need the time and temperature in Alton, Illinois." The man picks up a wired phone that is designed with a strong glossy finish and multiple facets in his right hand. He gazes into the camera and speaks into the microphone, then he purses his lips, blinks a few times, raises an eyebrow, and continues to speak. Afterward, he hangs up the phone. The video is shot from a first-person perspective, primarily aiming to promote and disseminate the phone number mentioned in the subtitles. It provides a detailed demonstration of the phone dialing service and its offerings through the man's actual phone call experience. Although the video is quite short, it effectively illustrates the convenience and practicality of calling to inquire about the time and weather.

Base model

The video features a person holding a phone to their ear, with a text overlay indicating that the person needs the time and temperature in Alton, Illinois. The person is wearing a dark-colored shirt and appears to be in a room with a light-colored wall and a window. The phone is held in their right hand, and the person's left hand is visible, holding a pen. The background includes a sign that reads "LET'S TALK," suggesting a casual or informal setting. The person's facial expression and body language suggest they are engaged in a conversation or presentation.

LLM Judge

3/5

Justification: Candidate covers basic phone call and text but misses key details about the man, phone, and purpose. Included evidence: phone, appear. Missing evidence: shown, full, face, makes, call

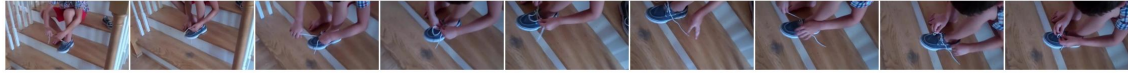
Vid2LoRA

The video shows a person holding a bottle of water in their hand, with the text overlay stating that if you are in the 20-25% of the population that is sweating in the summer, you should drink 1.5 liters of water per day. The person is wearing a white shirt and is in a room with a poster on the wall. The video does not show any significant changes or movements throughout the sequence.

LLM Judge

2/5

Justification: Candidate misses key reference facts about the call sequence and promotional purpose. Included evidence: some coarse scene context. Missing evidence: shown, full, face, makes, phone



Describe the video in as much useful visual detail as possible. Include the main activity, visible people or objects, scene context, appearance, and any important visual details that help explain what is happening.

Reference: This video captures a little boy as he ties his own shoelaces. The boy is dressed in a red and blue plaid shirt paired with red shorts. He is sitting on a staircase, which has wooden steps and white railings. The boy has a blue sneaker on his left foot, while the right shoe rests beside him. In the video, the boy lowers his head to focus on tying his shoelaces. He grips both ends of the laces, trying to weave them around each other to make a knot. However, one of the laces falls to the ground. Afterward, he successfully ties both ends of the lace together and pulls tightly on them, securing the knot. At this point, the camera zooms in for a close-up of the boy's feet. Next, he forms a loop with the left lace and tucks the left lace through the loop. Finally, he tightens both ends, resulting in a neatly tied bow. The video is shot from a third-person perspective, showcasing the boy's still-developing skill in shoelace tying, but his patience is commendable.

Base model

The video features a young boy sitting on a wooden staircase, meticulously tying the laces of a blue and white sneaker. He begins by tying the laces around the shoe's tongue, ensuring they are securely fastened. The boy then ties the laces around the shoe's heel, making sure they are properly secured. The boy continues to tie the laces around the shoe's toe, ensuring they are securely fastened. The video captures the boy's careful attention to detail as he meticulously ties the laces of the sneaker.

LLM Judge

4/5

Justification: Captures main activity and setting but misses key details of shoelace tying process. Included evidence: captures, ties, blue, sitting, staircase. Missing evidence: little, shoelaces, dressed, plaid, shirt

Vid2LoRA

A young boy is sitting on the floor, wearing a plaid shirt and red shorts, and is tying his shoes. He is using a pair of blue and white sneakers. The boy is focused on his task, and the video shows him meticulously tying the laces of his shoes.

LLM Judge

4/5

Justification: Captures main action and setting but misses key details about clothing, shoe position, and lacing steps. Included evidence: blue, plaid, shirt, shorts, sitting. Missing evidence: captures, little, ties, shoelaces, dressed



Describe the video in as much useful visual detail as possible. Include the main activity, visible people or objects, scene context, appearance, and any important visual details that help explain what is happening.

Reference: This video depicts an elderly man meticulously cleaning water stains off a glass window using a squeegee. He is dressed in a gray-blue plaid shirt, and his hair has turned mostly white, suggesting he is quite advanced in age. In his right hand, he holds the squeegee, while in front of him stands a window with a black frame. Inside the room, there are several wooden desks, their surfaces a warm yellow hue, each equipped with a drawer. A whiteboard for writing hangs on the wall, and a water dispenser sits in the corner of the room. The elderly man has a red cord around his neck, from which hangs a black card holder. In the video, the gentleman is positioned outside the building, skillfully scraping from the upper left corner of the window toward the right. He moves down to the bottom, then creates a sweeping arc back to the lower left corner, followed by another arc over to the right, repeating this motion until all the water stains are thoroughly removed. The footage is captured from the right side of the elderly man as he demonstrates the technique for cleaning the window, explaining the process aloud, likely teaching the videographer how to effectively wipe away the water stains.

Base model

A man wearing a plaid shirt is standing in front of a glass door, holding a tool. He is using the tool to clean the glass door. The background shows a classroom with desks and chairs, and the man is wearing a red and white striped shirt. The video is set in a classroom environment, and the man is likely demonstrating or explaining something related to the glass door.

LLM Judge

3/5

Justification: Candidate covers basic action but misses key details and has minor contradiction. Included evidence: glass, using, plaid, shirt, hair. Missing evidence: depicts, elderly, meticulously, cleaning, water

Vid2LoRA


A man wearing a plaid shirt is standing in front of a window, looking out at a room with tables and chairs. He is holding a long, thin object, possibly a tool or a piece of equipment, and appears to be examining it closely. The room has a modern design with white walls and a ceiling with recessed lighting. The man seems to be in the process of inspecting or demonstrating something about the object he is holding.

LLM Judge

4/5

Justification: Covers main action and setting but misses key details about man's appearance and room features. Included evidence: window, plaid, shirt, hair. Missing evidence: depicts, elderly, meticulously, cleaning, water

Figure 17: Qualitative examples from CaReBench: Caption.



Describe the key visible events in chronological order. Include all important actions and changes you can observe, with enough detail to distinguish each event clearly.

Reference: Athlete grips javelin handle firmly; Athlete lifts javelin above shoulder; Athlete swings left arm; Athlete runs forward quickly; Athlete body leans back; Athlete stretches left hand forward; Athlete draws right hand back; Athlete powerfully drives right shoulder forward; Athlete applies force to javelin; Athlete propels javelin from shoulder

Base model

The video begins with a man in an orange shirt and black shorts running on a grassy field. He is seen running towards the camera, and the background features a lush green landscape with trees and a fence. The man then stops running and starts performing a jump rope exercise. He begins by jumping up and down, then transitions to a series of jumps and twists, showcasing his agility and coordination. The video continues with the man performing a series of jumps and twists, maintaining his energetic and focused demeanor.

LLM Judge 0/5


Justification: Candidate describes a completely different activity with no overlap with javelin throwing reference.
Included evidence: back
Missing evidence: athlete, grips, javelin, handle, firmly

Vid2LoRA

A man in an orange shirt and black shorts is running on a grassy field. He is performing a series of exercises, including jumping and running. He is also seen performing a series of exercises with a rope.

LLM Judge 0/5

Justification: Candidate describes running with stick, not javelin throwing sequence
Included evidence: some coarse scene context
Missing evidence: athlete, grips, javelin, handle, firmly



Describe the key visible events in chronological order. Include all important actions and changes you can observe, with enough detail to distinguish each event clearly.

Reference: Speaker delivers speech in crowd; Speaker clasps left hand around right wrist; Speaker separates hands and places on hips; Speaker turns body to the left; Speaker takes two steps; Speaker gestures upward with finger; Speaker spreads hands apart; Speaker resumes speech; Audience applauds continuously

Base model

The video begins with a man standing in front of a large crowd, addressing them with a serious expression. He is dressed in a dark sweater and jeans, and the background features a large banner with the text "TRUSTED LEADERSHIP for a STRONGER AMERICA" and "JEB2016.COM". The crowd is engaged, with some individuals clapping and others looking towards the man. The man then turns to face the camera, maintaining his serious expression. The scene transitions to a close-up of the man, who is now wearing a brown sweater and jeans.

LLM Judge 2/5


Justification: Candidate describes setting and audience but misses all speaker actions from reference.
Included evidence: crowd, turns
Missing evidence: speaker, delivers, speech, clasps, left

Vid2LoRA

The video shows a man in a brown sweater and blue jeans standing on a stage in front of a blue backdrop with the words "SERVICE TO EVERYONE" and the logo of the American Leadership in Leadership. He is speaking to an audience, gesturing with his hands as he talks. The audience is seated in rows of chairs, and there are several people standing in the background, some of whom are clapping. The man on stage is wearing a microphone, and there is a sign in front of him that reads "SERVICE TO EVERYONE." The video also shows a group of people standing in front of a blue

LLM Judge 2/5

Justification: Candidate describes setting and audience but misses most speaker actions
Included evidence: hand, hands
Missing evidence: speaker, delivers, speech, crowd, clasps



Describe the key visible events in chronological order. Include all important actions and changes you can observe, with enough detail to distinguish each event clearly.

Reference: Man waters plants in garden; Man maintains consistent posture; Man waters pot of flowers; Man turns off nozzle; Man looks at the camera; Man prepares to bend down

Base model

A man is standing in a garden, surrounded by various plants and flowers. He is holding a watering can and watering the plants. The garden is well-maintained and has a red fence in the background. The man is wearing a white shirt and shorts.

LLM Judge 3/5


Justification: Candidate covers basic watering action but misses key postures and actions from reference.
Included evidence: plants, garden, flowers
Missing evidence: waters, maintains, consistent, posture, turns

Vid2LoRA

A man is standing in a garden, holding a potted plant. He is wearing a white shirt and patterned shorts. The garden is filled with various plants and flowers, and there are red buckets and other gardening tools scattered around. The man appears to be tending to the plants, possibly watering them or pruning them.

LLM Judge 4/5

Justification: Candidate covers main watering action but misses key postural/camera actions.
Included evidence: plants, garden, flowers
Missing evidence: waters, maintains, consistent, posture, turns



Describe the key visible events in chronological order. Include all important actions and changes you can observe, with enough detail to distinguish each event clearly.

Reference: Little boy watering plants outdoors; Using watering can to pour water into flower pot; Shifting camera angle from side view to rear view; Tapping edge of flower pot a few times; Setting down watering can

Base model

A young child is seen watering plants in a garden. The child is wearing a white tank top and blue shorts. The garden is filled with various plants, including a large potted plant with green leaves and a smaller potted plant with yellow leaves. The child is using a blue watering can to water the plants. The video concludes with the child standing in the garden, surrounded by the plants and the watering can.

LLM Judge 3/5

Justification: Captures basic watering action but misses camera movement and tapping details.
Included evidence: watering, plants, using, water
Missing evidence: little, outdoors, pour, flower, shifting


Vid2LoRA

A young child is seen watering plants in a garden. The child is wearing a white tank top with yellow trim and blue shorts. The garden is filled with various plants and flowers, and the child is using a blue watering can to water the plants.

LLM Judge 4/5

Justification: Candidate covers main actions but misses camera shift and tapping; extra details are harmless.
Included evidence: watering, plants, using, water, flower
Missing evidence: little, outdoors, pour, flower, shifting, angle

Figure 18: Qualitative examples from CaReBench: Events.



Describe the important visible objects and entities in the video in as much useful detail as possible. Include their appearance, location, and role in the scene when visible.

Reference: There is a white long-sleeve shirt; There are brownish-yellow sandals; There are deep blue trousers; There is a gray lining inside the pants; There is a deep blue T-shirt; There are gray long pants; There is a brown foot bath basin; There are two slices of green lemon; There is a movable drainage outlet; There is a stainless steel faucet; There is a semi-transparent brown bowl; There is a square bottle containing foot soak liquid; There are light yellow walls; There are small tiles in black, white, and gray; There are brown wooden tiles.

Base model

The video depicts a series of scenes involving a person washing their feet in a large, dark basin. Initially, the person is seen standing in front of a mirror, holding a towel, and preparing to wash their feet. The person then proceeds to wash their feet, using a white cloth to clean them thoroughly. The scene transitions to a close-up of the person's feet being washed, with the cloth still in use. The person then stands in front of the mirror, holding the towel, and continues to wash their feet. The final scene shows the person standing in front of the mirror, holding the towel, and washing.

LLM Judge 0/5


Justification: Candidate describes different scene with conflicting details
Included evidence: white
Missing evidence: long, sleeve, shirt, brownish, yellow

Vid2LoRA

The video shows a person washing their feet in a bathtub filled with water. The person is wearing blue jeans and is seen lathering their feet with soap. The water is clear, and the person's feet are wet. The person is also wearing a blue shirt. The background includes a tiled wall and a wooden table. The person is seen lathering their feet with soap, and the video shows the person's feet being washed.

LLM Judge 0/5

Justification: Candidate describes bathtub scene contradicting reference's foot bath basin setting
Included evidence: shirt, blue
Missing evidence: white, long, sleeve, brownish, yellow



Describe the important visible objects and entities in the video in as much useful detail as possible. Include their appearance, location, and role in the scene when visible.

Reference: There is a young boy with light brown hair; There is a white long-sleeved shirt; There are pure white shorts; There is a pair of black long pants; There is a white sheet with brown floral patterns; There is a brown floor; There are white walls; There is a brownish-yellow wardrobe; There are two drawers at the bottom of the wardrobe; There are several pairs of scattered socks.

Base model

The video features a young child dressed in a white sweater and dark pants, standing in a room with a wooden floor and a bed with a white and red floral-patterned blanket. The child is seen adjusting a pair of black pants on the bed, which are neatly folded and placed on the bed. The child then proceeds to pick up a pair of black pants from the bed and places them on the floor, indicating a casual and playful activity. The child continues to pick up the black pants, moving them around the room, and eventually placing them on the bed, completing the task.

LLM Judge 2/5


Justification: The candidate covers some basic facts but misses most reference details and adds extra elements.
Included evidence: young, white, pair, black, pants
Missing evidence: light, brown, hair, long, sleeved

Vid2LoRA

A young boy is sitting on the floor in a room with a bed covered in a red and white patterned blanket. He is wearing a white sweater and black pants. The room has wooden floorboards and a wooden wall. The boy is holding a black bag and appears to be organizing or packing it. He then moves to a different part of the room, where he is sitting on a wooden floor. He is wearing a white sweater and black pants. The room has wooden floorboards and a wooden wall. The boy is holding a black bag and appears to be organizing or packing it. He then moves to a different part of the room.

LLM Judge 3/5

Justification: Candidate covers some reference facts but misses key elements and adds unsupported details.
Included evidence: young, white, black, pants
Missing evidence: light, brown, hair, long, sleeved



Describe the important visible objects and entities in the video in as much useful detail as possible. Include their appearance, location, and role in the scene when visible.

Reference: There is a young boy; There is a white tank top; There is a blue watering can; There are flowers in pots; There are vibrant green plants; There is a large purple bucket; There is a sunny ambiance; There is a cozy family yard; There is a balcony; There is an array of potted greenery.

Base model

A young child is seen watering plants in a garden. The child is wearing a white tank top and blue shorts. The garden is filled with various plants, including a large potted plant with green leaves and a smaller potted plant with yellow leaves. The child is using a blue watering can to water the plants. The video is set outdoors, with a concrete surface and a stone wall in the background.

LLM Judge 4/5


Justification: Candidate covers most reference facts but misses some details like white tank top and balcony.
Included evidence: young, white, tank, blue, watering
Missing evidence: flowers, pots, vibrant, purple, bucket

Vid2LoRA

A young child is seen watering plants in a garden. The child is wearing a white tank top with yellow trim and blue shorts. The garden is filled with various plants and flowers, and the child is seen watering the plants with a blue watering can.

LLM Judge 4/5

Justification: Captures most reference facts but misses some details like flowers in pots and a purple bucket.
Included evidence: young, white, tank, blue, watering
Missing evidence: pots, vibrant, green, large, purple



Describe the important visible objects and entities in the video in as much useful detail as possible. Include their appearance, location, and role in the scene when visible.

Reference: There is a person chopping green onions; There is a red outfit; There is a knife in the right hand; There is a yellow cutting board; There are four whole green onions.

Base model

A person is preparing a dish by chopping green onions on a wooden cutting board. The person uses a knife to cut the green onions into smaller pieces. The green onions are then placed on a plate.

LLM Judge 4/5

Justification: Candidate covers most reference facts but misses 'four whole green onions' and specifies wooden board instead of yellow.
Included evidence: chopping, green, onions, knife, cutting
Missing evidence: outfit, right, hand, yellow, four


Vid2LoRA

The video shows a person slicing green onions on a wooden cutting board. The person is wearing a red shirt and is using a knife to cut the onions. The onions are sliced into thin rounds, and the person is seen cutting them into smaller pieces. The video shows the person slicing the onions into smaller pieces, with the onions being cut into smaller rounds. The person is seen cutting the onions into smaller pieces, with the onions being cut into smaller rounds. The person is seen cutting the onions into smaller pieces, with the onions being cut into smaller rounds.

LLM Judge 5/5

Justification: The candidate covers all reference facts with extra harmless details.
Included evidence: green, onions, knife, cutting, board
Missing evidence: chopping, outfit, right, hand, yellow

Figure 19: Qualitative examples from CaReBench: Objects.



Describe the spatial layout in as much useful detail as possible: where the people, objects, and scene elements are located, how they are positioned relative to each other, and what parts of the scene are in the foreground, background, left, right, center, above, or below.

Reference: This video captures the process of a barber getting his head shaved from behind his head. In the video, the barber is facing away from the camera, revealing his nearly bald scalp with only short stubble visible, but there is a patch of longer black hair behind his left ear. The barber's right hand holds a silver razor as he stands behind him, helping to style his hair.

Base model

The video shows a person shaving their head with a razor. The person is wearing a white shirt and is using a metal razor to shave the top of their head. The background is a blurred image of a room with a window and a door. The scene is set in a room with a window and a door, and the person is using a metal razor to shave their head.

LLM Judge 3/5


Justification: Candidate misses key spatial details and has contradiction about baldness
Included evidence: head
Missing evidence: captures, process, barber, getting, shaved

Vid2LoRA

The video shows a close-up of a person's face being shaved. The person is using a razor to shave the person's face, which is positioned in the center of the frame. The person is wearing a white shirt and is using a black razor to shave the person's face. The background is blurred, but it appears to be an indoor setting. The person is using a comb to brush the person's hair, which is positioned in the background. The person is wearing a white shirt and is using a black razor to shave the person's face. The background is blurred, but it appears to be an indoor.

LLM Judge 0/5

Justification: Candidate describes salon haircut with scissors, contradicting reference's razor shave from behind.
Included evidence: shaved
Missing evidence: captures, process, barber, getting, head



Describe the spatial layout in as much useful detail as possible: where the people, objects, and scene elements are located, how they are positioned relative to each other, and what parts of the scene are in the foreground, background, left, right, center, above, or below.

Reference: This video showcases a scene of a man playing a two-tiered keyboard organ in a restaurant. The man is dressed in a black short-sleeved shirt on top and a pair of black shorts on the bottom, along with black athletic shoes. He is seated on a black and white striped chair. In front of the organ, the organ features two rows of keys, with a glowing purple frame and a white body, adorned with yellow icons and text on the side that reads "WERSL." Behind him stands a group of adults and children, while to the right, there are several brown tables and chairs. On the left side of the video, you can see a window draped with light gray curtains.

Base model

The video shows a group of people gathered around a piano in a cozy, well-lit room. The piano is positioned in the center of the room, with a person playing it. The room has a warm and inviting atmosphere, with wooden floors and a tiled floor. The people are engaged in playing the piano, with some standing and others sitting. The scene includes a woman in a white shirt and jeans, a man in a black shirt and jeans, and a child in a red shirt and jeans. The video captures the joy and camaraderie of the group as they enjoy their time together.

LLM Judge 0/5


Justification: Candidate describes a living room with piano instead of restaurant with organ
Included evidence: playing, black, shirt
Missing evidence: tiered, organ, restaurant, dressed

Vid2LoRA

The video shows a man playing a keyboard in a room with a wooden floor and a window in the background. The room is decorated with a couch and a table with a vase of flowers. There are other people in the room, some of whom are standing and others are sitting. The people are engaged in conversation and appear to be enjoying the music.

LLM Judge 2/5

Justification: Candidate misses most reference facts and has multiple contradictions.
Included evidence: playing, keyboard
Missing evidence: tiered, organ, restaurant, dressed, black



Describe the spatial layout in as much useful detail as possible: where the people, objects, and scene elements are located, how they are positioned relative to each other, and what parts of the scene are in the foreground, background, left, right, center, above, or below.

Reference: In this video, four girls have neatly tied their hair up at the back of their heads, giving them a clean and tidy look. They are wearing matching blue sports vests and black athletic shorts, paired with black sneakers. The setting is an open space where they are performing a jump rope routine. The athletes present a unified appearance, and behind them, a large audience is seated. In the front row of the audience, there are athletes dressed in different colored outfits, while the spectators behind them are dressed in various styles, all intently watching the performance.

Base model

The video shows a group of people performing a synchronized dance routine in a gymnasium. The dancers are dressed in light blue shirts and black shorts, and they are performing a series of coordinated movements, including spins, jumps, and arm movements. The gymnasium has a green floor, and the audience is seated in the background, watching the performance. The scene is well-lit, and the dancers' movements are synchronized, creating a dynamic and engaging performance.

LLM Judge 4/5


Justification: Captures main event and most key details but misses some specific appearance and audience details.
Included evidence: back, blue
Missing evidence: four, girls, have, neatly, tied

Vid2LoRA

The video shows a group of young women in light blue shirts and black shorts performing a synchronized dance routine in a gymnasium. They are performing a series of synchronized movements, including arm movements and leg kicks, in front of a large audience. The audience is seated in the background, watching the performance. The gymnasium has a green floor and a large audience in the stands. The women are wearing black shorts and black leggings, and they are all wearing black shirts. The video captures the dancers' movements in detail, showing the synchronized movements and the audience's reactions.

LLM Judge 4/5

Justification: Covers main event and most key details but misses hair, sneakers, and spectator details while adding gymnasium and floor color.
Included evidence: back, blue
Missing evidence: four, girls, have, neatly, tied



Describe the spatial layout in as much useful detail as possible: where the people, objects, and scene elements are located, how they are positioned relative to each other, and what parts of the scene are in the foreground, background, left, right, center, above, or below.

Reference: This video showcases the process of a person chopping green onions. She is wearing a red outfit and holding a knife in her right hand. In front of her is a yellow cutting board that has four whole green onions on it.

Base model

The video shows a person chopping green onions on a wooden cutting board. The person's hands are seen chopping the green onions, with the knife slicing through the onion. The scene is set in a kitchen, with the person wearing a red shirt. The background includes a red wall and a white countertop. The video captures the action of chopping the green onions, with the person's hands and the knife visible in the foreground.

LLM Judge 4/5

Justification: Candidate covers main action and key elements but misses knife position and cutting board color.
Included evidence: chopping, green, onions, knife, hand
Missing evidence: process, outfit, right, front, yellow


Vid2LoRA

The video shows a person slicing green onions on a wooden cutting board. The person is wearing a red shirt and is using a knife to cut the onions. The onions are sliced into thin rounds, and the person is seen cutting them into smaller pieces. The video shows the person slicing the onions into smaller pieces, with the onions being cut into smaller rounds. The person is seen cutting the onions into smaller pieces, with the onions being cut into smaller rounds. The person is seen cutting the onions into smaller pieces, with the onions being cut into smaller rounds.

LLM Judge 4/5

Justification: Candidate covers main action and key elements but misses knife position and cutting board color.
Included evidence: green, onions, knife, cutting, board
Missing evidence: process, chopping, outfit, right, hand

Figure 20: Qualitative examples from CaReBench: Spatial Caption.



Describe the temporal progression in as much useful detail as possible. Explain what happens over time, the order of visible actions, and how the scene or subjects change from the beginning to the end.

Reference: This short film depicts the process of a Black male athlete performing the triple jump. We see him taking off in a swift run, his stride frequency increasing as his arms swing more vigorously in tandem. Upon reaching the first take-off line, he launches into his first jump, propelling his hands forward while pushing off the ground with his legs. After landing, he bends one knee, gracefully transitioning through the air as he crosses over to land successfully on the third line. Following this, he makes a powerful leap upward, bringing his feet together and drawing his upper and lower body closer until they form a compact shape, before ultimately landing safely in the sandpit.

Base model

The video begins with a man in a blue shirt and beige pants standing on a track, preparing for a race. He then runs down the track, passing by other athletes and spectators. The scene transitions to a group of athletes in blue and yellow uniforms preparing for a race, with some running and others sitting on the track. The video then shows a man in a red shirt and beige pants standing on a track, observing the athletes. The scene shifts to a man in a red shirt and beige pants standing on a track, observing the athletes. The video concludes with a man in a blue shirt and beige pants.

LLM Judge 0/5


Justification: Candidate describes shot put event instead of triple jump, contradicting reference.
Included evidence: athlete
Missing evidence: short, film, depicts, process, black

Vid2LoRA

The video shows a group of athletes in a large indoor sports facility, possibly a gymnasium, with a red floor and white lines marking the boundaries. The athletes are dressed in athletic gear, including shorts, t-shirts, and shoes, and are seen running, jumping, and performing various athletic activities. The facility is well-lit, and there are spectators seated in the background, watching the event. The athletes are seen in different positions, some standing, others running, and some jumping. The video captures the athletes in various stages of their performance, from standing to running, jumping, and performing various athletic activities. The facility is

LLM Judge 0/5

Justification: Candidate describes group scene instead of single athlete's triple jump.
Included evidence: short, athlete, performing, jump
Missing evidence: film, depicts, process, black, male



Describe the temporal progression in as much useful detail as possible. Explain what happens over time, the order of visible actions, and how the scene or subjects change from the beginning to the end.

Reference: This video showcases the daily activity of a middle-aged man watering plants in his garden. From the beginning to the end of the video, the man maintains a consistent posture as he waters a pot of flowers at his feet. In the final shot, he turns off the nozzle, looks at the camera, and makes a motion as if he is preparing to bend down.

Base model

The video begins with a person standing in a garden, holding a potted plant. They are surrounded by various plants and flowers, and the background features a red fence and a garden bed. The person is then seen watering the plants in the garden, which includes a large pot with red flowers and a smaller pot with green plants. The scene transitions to a close-up of the person's hands as they pick up a potted plant from the garden bed. The video then shows the person standing in the garden, holding a potted plant and looking at it. The final scene shows the person standing in the garden, holding

LLM Judge 3/5


Justification: Candidate covers basic activity but misses key temporal details and specific actions from reference.
Included evidence: watering, plants, garden, flowers, final
Missing evidence: daily, activity, middle, aged, beginning

Vid2LoRA

A man is standing in a garden, holding a potted plant. He is wearing a white shirt and light-colored shorts. The garden is filled with various plants and flowers, and there are red buckets and other gardening tools scattered around. The man appears to be tending to the plants, possibly watering them or pruning them.

LLM Judge 3/5

Justification: Candidate covers basic activity but misses key temporal details and specific actions.
Included evidence: watering, plants, garden, flowers
Missing evidence: daily, activity, middle, aged, beginning



Describe the temporal progression in as much useful detail as possible. Explain what happens over time, the order of visible actions, and how the scene or subjects change from the beginning to the end.

Reference: This video captures a scene of a little girl at home showcasing her skills on the violin. With her left hand holding the violin and her right hand controlling the bow, she moves the bow back and forth across the strings. In the end, she pulls the bow away, bringing her performance to a close.

Base model

The video begins with a young girl playing the violin in a cozy living room. She is dressed in a yellow skirt and a black top, and her hair is tied back. The room is decorated with pink balloons and a window with a view of a cityscape. The girl is focused on her music, and the camera captures her playing the violin with precision and grace. The scene develops as the girl continues to play, and the background remains consistent with the same room and decorations. The video concludes with the girl still playing the violin, maintaining the same setting and atmosphere.

LLM Judge 3/5


Justification: Captures basic violin playing but misses key temporal details and setting context
Included evidence: captures, girl, violin, back
Missing evidence: little, home, showcasing, skills, left

Vid2LoRA

A young girl is playing a violin in a room with a Christmas tree in the background. She is wearing a yellow dress and a black skirt. The room has a Christmas tree, a couch, and a window with a curtain. The girl is playing the violin with her hands on the strings.

LLM Judge 4/5

Justification: Covers main violin action but misses 'at home' and 'showcasing skills' while adding unsupported dress details.
Included evidence: girl, violin, hand, back, strings
Missing evidence: captures, little, home, showcasing, skills



Describe the temporal progression in as much useful detail as possible. Explain what happens over time, the order of visible actions, and how the scene or subjects change from the beginning to the end.

Reference: This video showcases a scene of a little boy playing the drums. While he is drumming, the boy holds a drumstick in each hand and rhythmically strikes the drum in front of him, crossing his arms as he plays. After a few beats, he raises his left hand to hit the cymbal, perfectly coordinating with his right hand on the drum. Then, he brings his left hand back down and continues to play the drum with both hands together.

Base model

The video begins with a person playing a drum set, wearing a hat and a bandana, and a drum set is visible in the background. The person is seen playing the drums, with the drum set positioned in front of them. The scene then transitions to a different setting where the person is now playing a keyboard. The video then shows the person playing a guitar, with the guitar placed on a stand. The video concludes with the person playing a guitar, with the guitar placed on a stand.

LLM Judge 3/5

Justification: Candidate covers basic action but misses key drumming details and temporal progression.
Included evidence: playing, drums, drum, front
Missing evidence: little, drumming, holds, drumstick, each


Vid2LoRA

The video shows a drummer playing a drum set in a room with a blue wall. The drummer is wearing a hat and a black shirt. The drum set includes a snare drum, a bass drum, and a hi-hat. The drummer is playing the drums with his hands and feet. The video does not show any significant changes in the scene or the drummer's actions over time.

LLM Judge 4/5


Justification: Candidate covers main drumming actions but misses background details.
Included evidence: playing, drums, hand, drum
Missing evidence: little, drumming, holds, drumstick, each

Figure 21: Qualitative examples from CaReBench: Temporal Caption.




where is the toy sitting at Answer with a short phrase.
Reference: chair

<p>Base model</p> <p>toy.</p> <p>LLM Judge 4/5</p> <p><small>Justification: The candidate clearly implies the toy is on a chair with extra harness detail.</small></p>	<p>Vid2LoRA</p> <p>The toy is sitting on the floor.</p> <p>LLM Judge 1/5</p> <p><small>Justification: Candidate says 'toy sitting at table' which contradicts reference 'chair'.</small></p>
--	--




why does the baby take the item from the lady in grey beside her Answer in a single sentence.
Reference: woman passed to baby

<p>Base model</p> <p>to eat it.</p> <p>LLM Judge 4/5</p> <p><small>Justification: Candidate implies the lady passed the item to the baby, matching reference.</small></p>	<p>Vid2LoRA</p> <p>The baby takes the item from the lady in grey beside her because the lady in grey is holding a piece of food in her hand.</p> <p>LLM Judge 3/5</p> <p><small>Justification: Candidate explains why baby takes item but misses key action of woman passing it.</small></p>
---	--



why is there a piece of tissue in front of the baby Answer in a single sentence.
Reference: to wipe baby after feeding him


<p>Base model</p> <p>To clean the baby's face.</p> <p>LLM Judge 3/5</p> <p><small>Justification: Candidate mentions cleaning but misses the specific purpose of wiping after feeding.</small></p>	<p>Vid2LoRA</p> <p>The video shows a baby being fed with a bottle, and a piece of tissue is placed in front of the baby.</p> <p>LLM Judge 4/5</p> <p><small>Justification: The candidate implies the tissue's purpose is to prevent messes after feeding.</small></p>
---	---



what does the man above hold onto while climbing up Answer with a short phrase.
Reference: rope

<p>Base model</p> <p>rope.</p> <p>LLM Judge 5/5</p> <p><small>Justification: Candidate directly states the answer 'rope' matching the reference.</small></p>	<p>Vid2LoRA</p> <p>The man holds onto a rope while climbing up.</p> <p>LLM Judge 5/5</p> <p><small>Justification: Candidate directly states 'rope' as the answer.</small></p>
--	---

Figure 22: Qualitative examples from NExT-QA.



What color is the trunk of the tree on the left side of the picture? Answer only the question, in one sentence.

Reference: Deep chestnut brown.

Base model

Green.

LLM Judge

4/5

Justification: Candidate implies brown but lacks 'deep chestnut' detail.


Vid2LoRA

The trunk of the tree on the left side of the picture is green.

LLM Judge

0/5

Justification: Candidate states trunk is green, directly contradicting reference's deep chestnut brown.



What kind of shoes is this person wearing? Answer only the question, in one sentence.

Reference: She was wearing a pair of canvas shoes with white soles and black and white checkered uppers.

Base model

Boots.

LLM Judge

1/5

Justification: Candidate says sandals but reference says canvas shoes.


Vid2LoRA

The person is wearing a blue and orange striped harness and a black and orange harness. They are also wearing a black shirt and a black hat.

LLM Judge

3/5

Justification: Candidate states black shoes but misses key details like canvas material and checkered uppers.



What color are the butterfly's wings? Answer only the question, in one sentence.

Reference: Orange, with black, yellow, blue, and white spots and black edging.

Base model

Orange and black.

LLM Judge

4/5

Justification: Candidate states 'Orange' which matches the reference's main color but omits additional details.

Vid2LoRA

The butterfly's wings are orange and black.

LLM Judge

4/5

Justification: Candidate states wings are orange but omits additional spot details.



After the vegetables, what does this person need to cut? Answer only the question, in one sentence.

Reference: Salmon.

Base model

A piece of salmon.

LLM Judge

5/5

Justification: Candidate directly states 'salmon' as the answer.

Vid2LoRA


The person needs to cut the vegetables on a cutting board.

LLM Judge

5/5


Justification: Candidate directly states 'The salmon' as the answer.

Figure 23: Qualitative examples from VidCapBench.




Do my arms look like the right forearm for these arms circles ?
Reference: Yes, your arms look like they're in the correct position for arm circles.

<p>Base model</p> <p>No.</p> <p>LLM Judge</p> <p>Justification: Candidate directly answers 'No' matching the reference answer.</p> <p>5/5</p>	<p>Vid2LoRA</p> <p>The arms are not the right forearm for these arms circles.</p> <p>LLM Judge</p> <p>Justification: Candidate directly contradicts the reference answer with 'No'.</p> <p>0/5</p>
--	---



Where is she in?
Reference: The girl is likely in a living room.


<p>Base model</p> <p>bedroom.</p> <p>LLM Judge</p> <p>Justification: Candidate states bedroom, reference says living room, direct contradiction.</p> <p>0/5</p>	<p>Vid2LoRA</p> <p>She is in a room with a white wall and a wooden door.</p> <p>LLM Judge</p> <p>Justification: The candidate describes a room with details that imply it's likely a living room.</p> <p>4/5</p>
--	---



Does my hand and knee positioning look like his?
Reference: Yes, your hand and knee positioning matches his.

<p>Base model</p> <p>Yes.</p> <p>LLM Judge</p> <p>Justification: Candidate directly confirms hand and knee positioning matches.</p> <p>5/5</p>	<p>Vid2LoRA</p> <p>Yes, the hand and knee positioning look like his.</p> <p>LLM Judge</p> <p>Justification: Candidate directly matches reference answer.</p> <p>5/5</p>
---	--

Figure 24: Qualitative examples from PLM SGQA.



Describe what happens in this video clip.

Reference: Out of frame

Base model

The video captures a bustling scene inside a vibrant toy store, showcasing the diverse array of products and the lively atmosphere. Initially, a woman in a black hijab and a man in a brown jacket are seen walking through the store, with the woman carrying a blue shopping cart. The store is well-lit, with a variety of colorful toys and merchandise displayed on shelves and in bins. The woman and the man are seen interacting with each other, possibly discussing the products or engaging in a conversation. As the video progresses, the focus shifts to a different section of the store, where a woman in a black hijab

LLM Judge 0/5


Justification: Candidate describes a toy store scene while reference states 'Out of frame'
Included evidence: some coarse scene context
Missing evidence: frame

Vid2LoRA

The video showcases a bustling toy store filled with various toys and games. The store is well-lit, with bright lights illuminating the shelves and the floor. The shelves are stocked with a variety of items, including toys, games, and other merchandise. There are people walking around the store, some of whom are browsing the shelves while others are looking at the products. The store is clean and organized, with a variety of products on display. The video captures the lively atmosphere of the store, with people moving around and interacting with the products.

LLM Judge 0/5

Justification: Candidate describes a toy store scene while reference states 'Out of frame'
Included evidence: some coarse scene context
Missing evidence: frame



Describe what happens in this video clip.

Reference: A boy reappears behind a woman walking back and forth near a wall.

Base model

The video captures a scene in a residential area, likely in India, where a woman in a vibrant green traditional dress is seen cleaning a dirty floor. The woman, dressed in a colorful, patterned dress with intricate designs, is seen walking towards a white plastic bag on the floor. She is accompanied by a child, dressed in a white t-shirt and blue shorts, who is also seen walking towards the bag. The woman is seen picking up the bag, which is placed on the floor, and then proceeds to clean the floor with a blue plastic brush. The floor is covered in dirt and debris, indicating a neglected or

LLM Judge 2/5


Justification: Candidate misses the boy's reappearance and position behind the woman.
Included evidence: woman, walking
Missing evidence: reappears, behind, back, forth, near

Vid2LoRA

In the video, a woman in a colorful sari is seen walking towards a man in a green shirt and shorts. They are in a room with a blue tarp and a bicycle in the background. The woman is carrying a blue bag and appears to be looking at the man. The man is standing in front of a white wall with a black bag on it. The woman then walks away from the man and the camera follows her as she moves towards a black bag on the floor.

LLM Judge 2/5

Justification: Candidate describes unrelated scene details but misses key action of boy reappearing behind woman.
Included evidence: woman, walking, back, wall
Missing evidence: reappears, behind, forth, near



Describe what happens in this video clip.

Reference: A woman is facing towards the camera and is walking on a busy road then continues to exit the frame.

Base model

In the video, a person wearing a black hoodie with a pink elephant design and a black beanie is walking down a busy street. The street is lined with various shops and buildings, and there are people walking around, some of whom are carrying shopping bags. The person is seen holding a pink phone in their hand, possibly taking a photo or video. The video captures the person walking down the street, with the camera following them as they move forward.

LLM Judge 3/5


Justification: Captures walking on busy road and exiting frame but misses gender and facing camera
Included evidence: walking, busy
Missing evidence: woman, facing, towards, road, then

Vid2LoRA

The video shows a person walking down a busy street in a city. The person is wearing a black hoodie with a pink unicorn design and a black scarf. The street is lined with buildings and cars, and there are people walking in the background.

LLM Judge 4/5

Justification: Captures main action and setting but misses 'facing camera' detail.
Included evidence: walking, busy
Missing evidence: woman, facing, towards, road, then



Describe what happens in this video clip.

Reference: In this video, a dog can be seen playing with a balloon.

Base model

The video captures a playful interaction between a golden retriever and a green balloon. Initially, the dog is seen holding the balloon in its mouth, with its front paws resting on a wooden surface. The dog's attention is focused on the balloon, which is positioned close to its face. As the video progresses, the dog's mouth moves towards the balloon, and it begins to play with it by pulling it towards its mouth. The dog's body language suggests it is enjoying the toy, with its front paws extended towards the balloon. The scene then shifts to a close-up of the dog's face,

LLM Judge 5/5

Justification: Candidate captures all reference facts with extra harmless details.
Included evidence: seen, balloon
Missing evidence: playing


Vid2LoRA

In the video, a small dog is seen playing with a bunch of colorful balloons on a wooden floor. The dog is initially seen holding a green balloon, which is then removed from its mouth. The dog then picks up a red balloon and continues to play with it. The dog's movements are playful and energetic, with the balloons being moved around and the dog's head being turned in different directions. The background is a plain wooden floor, and there are no other objects or people visible in the scene. The dog's actions suggest that it is enjoying the activity and is having fun.

LLM Judge 5/5


Justification: Candidate captures all reference facts with extra harmless details.
Included evidence: seen, playing, balloon
Missing evidence: few obvious reference facts

Figure 25: Qualitative examples from RCAP.




Describe what happens in this video.
Reference: A man walks forward and then kneels down to pray.

<p>Base model</p> <p>The video captures a lively outdoor event, likely a festival or fair, with a large crowd gathered around a central structure. The structure is adorned with ornate decorations, including a large golden statue and a colorful floral arrangement. The crowd is diverse, with people dressed in various colors, indicating a festive atmosphere. The event is taking place in a bustling urban area, with buildings and vehicles visible in the background. The lighting is bright, suggesting it is either early morning or late afternoon. The video provides a comprehensive view of the event, showcasing the vibrant and colorful atmosphere.</p> <p>LLM Judge 0/5</p> <p><small>Justification: Candidate describes a market scene, not a man praying. Included evidence: some coarse scene context Missing evidence: walks, forward, then, kneels, down</small></p>	<p>Vid2LoRA</p> <p>The video showcases a vibrant and bustling festival scene, likely taking place in a large indoor venue. The atmosphere is lively, with people of various ages and genders gathered around a large, ornate structure adorned with golden decorations. The structure is surrounded by a crowd of people, some of whom are standing and others are seated, all engaged in the festivities. The environment is well-lit, with natural light illuminating the scene from above. The video captures the essence of a festive event, with the crowd's movement and the structure's intricate design creating a dynamic and engaging visual experience.</p> <p>LLM Judge 0/5</p> <p><small>Justification: Candidate describes a city street scene while reference describes a man praying Included evidence: some coarse scene context Missing evidence: walks, forward, then, kneels, down</small></p>
---	---



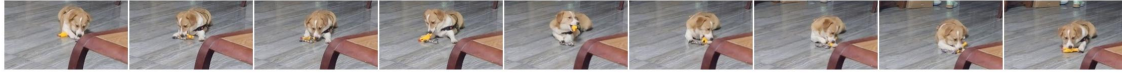
Describe what happens in this video.
Reference: A man enters the frame standing and holding a microphone when the camera pans to the right. He then puts the mic to his mouth and gestures with his left hand.

<p>Base model</p> <p>The video captures a lively and festive event, likely a birthday celebration, where a young child dressed in a pink dress and a child in a white dress are the main subjects. The child in the pink dress is seen holding a microphone, while the child in the white dress is holding a pink dress. The child in the pink dress is then seen dancing energetically, while the child in the white dress stands nearby, possibly observing or participating in the celebration. As the video progresses, the child in the pink dress is seen dancing energetically, while the child in the white dress continues to stand nearby, possibly observing or</p> <p>LLM Judge 2/5</p> <p><small>Justification: Candidate describes a child and stage scene, missing key reference actions. Included evidence: microphone, then Missing evidence: enters, frame, standing, when, pans</small></p>	<p>Vid2LoRA</p> <p>The video begins with a young boy in a yellow shirt and blue pants standing in front of a red curtain, holding a microphone. He appears to be singing or speaking into it. The scene then transitions to a different setting where a young girl in a pink dress is standing in front of a microphone, also holding a microphone. The background shows a stage with a red curtain and a few people standing around. The girl is singing or speaking into the microphone. The video then shows a young boy in a blue shirt and black pants standing in front of a microphone, holding a microphone. He appears to be singing or speaking into it. The background</p> <p>LLM Judge 3/5</p> <p><small>Justification: Candidate covers main man actions but misses camera movement and entry details. Included evidence: standing, microphone, then Missing evidence: enters, frame, when, pans, right</small></p>
---	---



Describe what happens in this video.
Reference: The video shows a boy holding a stick moving it on the grassy ground, he is moving and jumping towards the camera.


<p>Base model</p> <p>The video captures a young boy engaging in a playful activity on a grassy area. Initially, the boy is seen running towards a tree, which is partially visible in the background. As he approaches the tree, he bends down to pick up a stick, indicating a possible game or activity involving the stick. The boy then stands up and continues to pick up the stick, occasionally looking around and moving slightly. As the video progresses, the boy continues to pick up the stick, occasionally looking around and moving slightly. He then bends down again, picking up a stick and continuing to pick it up, maintaining a playful and energetic</p> <p>LLM Judge 4/5</p> <p><small>Justification: Candidate captures main action but misses 'jumping towards camera' detail. Included evidence: stick, moving, grassy, ground, towards Missing evidence: jumping</small></p>	<p>Vid2LoRA</p> <p>In the video, a young boy is seen running towards a tree in a grassy area. He is wearing a red shirt and blue shorts. As he approaches the tree, he stops and begins to climb it. The tree has a large, thick trunk and a few branches. The boy continues to climb the tree, eventually reaching the top. He then stands on the tree and looks around, possibly observing his surroundings. The video does not show any other people or animals in the area.</p> <p>LLM Judge 4/5</p> <p><small>Justification: Captures main actions but adds harmless extra details not in reference. Included evidence: grassy, towards Missing evidence: stick, moving, ground, jumping</small></p>
---	--



Describe what happens in this video.
Reference: A dog is playing on the floor with the toy.

<p>Base model</p> <p>The video captures a scene in a room filled with various items, including a wooden table, a green plant, and a pair of shoes. A small dog, likely a Pembroke, is seen lying on the floor, chewing on a yellow toy. The dog's attention is captured by the toy, which is positioned close to the camera. The room is well-lit, and the dog's actions suggest a playful and curious behavior. The video provides a glimpse into the dog's daily life, showcasing its playful and curious nature.</p> <p>LLM Judge 4/5</p> <p><small>Justification: Candidate covers all reference facts with extra harmless details. Included evidence: floor Missing evidence: playing</small></p>	<p>Vid2LoRA</p> <p>In the video, a small dog is seen playing with a yellow chew toy in a room with a tiled floor. The dog is brown and white, and it is holding the toy in its mouth. The room has a beige wall and a gray floor. The dog appears to be enjoying the toy and is moving around the room.</p> <p>LLM Judge 5/5</p> <p><small>Justification: Candidate fully covers reference facts with additional harmless details. Included evidence: playing, floor Missing evidence: few obvious reference facts</small></p>
---	---

Figure 26: Qualitative examples from RDCAP.



Describe the background, setting, and environment in this video.

Reference: The setting is a partially renovated interior space, likely a hallway or entryway, with a clean, light-colored wall that reflects natural light streaming in from an open door. The floor is tiled, featuring a mix of light and dark colors, and a ladder is positioned nearby, indicating ongoing work. Outside, a sunny day is evident, with a glimpse of parked cars and greenery visible through the doorway, contributing to a bright and inviting atmosphere. The sounds of distant traffic and the occasional rustle of leaves can be faintly heard, enhancing the sense of a lively neighborhood.

Base model

The video showcases a construction site where a person is working on a wall, using a hammer and a level to ensure the wall is level and the ceiling is properly aligned. The person is wearing a blue shirt and is seen using a hammer to tap the wall into place, indicating the process of installing a ceiling. The environment is a well-lit area with a window and a door visible in the background, suggesting that the construction is taking place in a residential or commercial building. The person is seen using a level to ensure the wall is level, and the camera provides a top-down view of the construction site, highlighting the

LLM Judge 2/5


Justification: Candidate only mentions basic room elements without covering key reference details.
Included evidence: light, wall
Missing evidence: setting, partially, renovated, interior, space

Vid2LoRA

A man is working on a wall in a residential building. He is using a tool to clean the wall, and he is also using a ladder to reach higher areas. The man is wearing a blue shirt and beige pants. The wall has a hole in it, and there is a window in the background.

LLM Judge 1/5

Justification: Candidate focuses on foreground action, missing all background details.
Included evidence: clean, wall
Missing evidence: setting, partially, renovated, interior, space



Describe the background, setting, and environment in this video.

Reference: The video is set in a well-lit bedroom with a simple, minimalist aesthetic. The bed is made with a white mattress and a patterned duvet cover featuring floral designs in shades of gray, yellow, and black. The walls are painted a soft color, and there are two decorative pillows resting against the headboard. The floor is visible, showing a hint of clutter with a blue bag and some items scattered around, suggesting a lived-in space. The atmosphere is calm and serene, with no noticeable sounds other than the soft rustling of fabric.

Base model

The video begins with a person folding a gray blanket, showcasing the texture and pattern of the fabric. The scene transitions to a close-up of a person's hands, which are seen folding a gray blanket with a floral pattern. The person then proceeds to fold a white blanket, which is placed on a bed. The focus shifts to a close-up of a person's hands folding a gray blanket, with the background showing a white bed with a floral pattern and a red wall. The person continues to fold the blanket, with the background remaining consistent. The video then shows a close-up of a person's hands folding

LLM Judge 2/5


Justification: Candidate misses most reference details and has direct contradictions.
Included evidence: white, floral, gray
Missing evidence: well, bedroom, simple, minimalist, aesthetic

Vid2LoRA

A person is folding a blanket on a bed. The blanket is white and has a pattern of small squares. The person is folding the blanket in a way that shows the pattern. The bed has a gray blanket and a white pillow. The person is also folding a gray blanket on the bed.

LLM Judge 2/5

Justification: Candidate misses most reference details and has contradictory pillow description.
Included evidence: white, gray
Missing evidence: well, bedroom, simple, minimalist, aesthetic



Describe the background, setting, and environment in this video.

Reference: The scene unfolds in a bustling city street at night, illuminated by a kaleidoscope of neon lights reflecting off the rain-drenched pavement. The atmosphere is alive with the sounds of distant traffic and the soft patter of rain, creating a serene yet vibrant ambiance. The street is lined with parked cars, and the glow of shop signs adds a warm, inviting feel to the otherwise cool, wet environment. The air is thick with the scent of rain and urban life, enhancing the sensory experience of the scene.

Base model

The video captures a nighttime scene on a wet street, illuminated by the glow of streetlights and the reflections of lights on the wet pavement. The scene is dominated by the silhouettes of two individuals walking side by side, each holding an umbrella, suggesting they are either walking or waiting for someone. The wet pavement reflects the vibrant colors of the night, creating a dynamic and dynamic visual effect. The environment is bustling with activity, with cars and buildings visible in the background, adding to the urban atmosphere. The video concludes with a close-up of the two individuals, emphasizing their silhouettes against the wet street, maintaining the focus

LLM Judge 4/5


Justification: Captures key setting elements but misses some atmospheric details.
Included evidence: bustling, street, night, illuminated, lights
Missing evidence: unfolds, city, kaleidoscope, neon, reflecting

Vid2LoRA

The video shows a nighttime scene with a wet street lit by streetlights. The street is filled with people walking and standing, some of whom are holding umbrellas. The sky is dark, and the overall atmosphere is one of a rainy night.

LLM Judge 4/5

Justification: Candidate covers key elements but misses some reference details like neon lights and shop signs.
Included evidence: street, night, lights, rain, atmosphere
Missing evidence: unfolds, bustling, city, illuminated, kaleidoscope



Describe the background, setting, and environment in this video.

Reference: The video captures a tranquil lakeside setting surrounded by dense, vibrant green foliage. The trees, rich in various shades of green, create a lush canopy overhead, while the water's surface mirrors the landscape, enhancing the sense of calm. The sky is partly cloudy, with soft light filtering through the leaves, casting gentle shadows on the water. The atmosphere is peaceful, with the occasional rustle of leaves and distant bird calls adding to the natural soundscape.

Base model

The video showcases a serene, natural setting with a tranquil lake surrounded by lush green trees. The sky is partly cloudy, and the light is soft and diffused, creating a calm and peaceful atmosphere. The reflection of the trees and sky on the water's surface adds depth and dimension to the scene. The overall environment is calm, with no visible human activity or movement, emphasizing the tranquility of the natural setting.

LLM Judge 5/5

Justification: Candidate covers all key reference facts with minor harmless extra detail about blue sky patches.
Included evidence: tranquil, setting, surrounded, green, trees
Missing evidence: captures, lakeside, dense, vibrant, foliage


Vid2LoRA

The video showcases a serene and picturesque scene of a forest with tall trees and a calm body of water. The water is a deep blue, reflecting the surrounding greenery and creating a mirror-like effect. The sky is overcast, casting a soft light over the scene. The trees are lush and green, with some displaying a mix of green and yellow leaves, indicating the changing seasons. The overall atmosphere is one of tranquility and natural beauty.

LLM Judge 5/5

Justification: Candidate covers all reference facts with additional harmless details.
Included evidence: tranquil, green, trees, lush
Missing evidence: captures, lakeside, setting, surrounded, dense

Figure 27: Qualitative examples from VDC Background.



Describe the camera work, framing, and viewpoint in this video.

Reference: The camera employs a series of close-up shots that capture the child's hands and facial expressions in detail, emphasizing their engagement with the clay. The angles are primarily eye-level, providing an intimate perspective of the child's activities. Smooth transitions between shots maintain a fluid narrative, while the depth of field blurs the background slightly, keeping the focus on the child and their creative process. The camera occasionally shifts to a wider shot to include the surrounding toys, enhancing the context of the playful environment.

Base model

The video features a child engaged in a creative activity, likely drawing or coloring, on a rug. The child is seated on the floor, wearing a light blue polo shirt and dark pants, with a soccer ball and other toys nearby. The child's hands are actively involved in the activity, using a variety of colored markers to create a colorful design on a white piece of paper. The camera captures the child's movements and the design being created, providing a clear view of the child's actions and the environment.

LLM Judge 3/5


Justification: Candidate misses key camera techniques and contradicts reference angles.
Included evidence: capture, hands, providing
Missing evidence: employs, series, close, shots, child

Vid2LoRA

A young girl is sitting on the floor in a living room, surrounded by a soccer ball and a colorful toy. She is wearing a light blue shirt and dark pants. The girl is holding a marker and drawing on a piece of paper.

LLM Judge 0/5

Justification: Candidate describes different activity and setting than reference's clay activity with camera work focus.
Included evidence: some coarse scene context
Missing evidence: employs, series, close, shots, capture



Describe the camera work, framing, and viewpoint in this video.

Reference: The camera employs a series of close-up shots that focus on the ornament, capturing its intricate details and the way it interacts with the light. The angles are slightly tilted, adding a dynamic feel to the composition. The camera transitions smoothly between different focal points, occasionally blurring the background to emphasize the ornament. There are subtle movements, such as a slow zoom in and out, enhancing the viewer's connection to the ornament's beauty. The use of depth of field creates a soft bokeh effect, further highlighting the ornament against the dark backdrop.

Base model

The video begins with a close-up shot of a Christmas ornament, showcasing its intricate design and sparkling sequins. The ornament is set against a blurred background of Christmas lights, creating a festive atmosphere. The camera then zooms in on the ornament, highlighting its detailed design and the sparkling sequins, emphasizing the ornament's craftsmanship. The focus remains on the ornament as the camera continues to zoom in, providing a detailed view of its intricate design and sparkling sequins. The background remains blurred, maintaining the festive ambiance. The camera then zooms out slightly, revealing more of the Christmas tree and the surrounding decorations, maintaining the

LLM Judge 2/5


Justification: Candidate misses most camera work details and contradicts movement claims.
Included evidence: close, focus, ornament, intricate, light
Missing evidence: work, employs, series, shots, capturing

Vid2LoRA

The video features a close-up of a Christmas ornament, which is a large, ornate, and detailed piece of jewelry. The ornament is made of a shiny, reflective material and is adorned with a pattern of small, sparkling beads. The background is a dark, starry night sky, which adds a festive atmosphere to the scene. The lighting is soft and diffused, creating a dreamy effect. The ornament is the main focus of the video, and there are no other objects or people visible in the frame. The video does not contain any text or additional information.

LLM Judge 3/5

Justification: Candidate misses key camera movement details but covers basic framing and blurring.
Included evidence: close, focus, ornament, light
Missing evidence: work, employs, series, shots, capturing



Describe the camera work, framing, and viewpoint in this video.

Reference: The camera employs a series of close-up shots that focus on the hands and the keyboard, capturing the intricate details of the typing process. The angles are slightly tilted, providing a dynamic perspective that draws the viewer's attention to the movement of the fingers. The transitions between shots are smooth, maintaining a continuous flow that mirrors the subject's concentration. There are no abrupt cuts, allowing the viewer to immerse themselves in the act of typing, while the depth of field subtly blurs the background, emphasizing the main action.

Base model

The video begins with a close-up of a person's hands typing on a laptop keyboard, with the focus on the hands and the keyboard. The camera then shifts to a wider shot, revealing the person's upper body and the laptop's screen, which is turned off. The scene transitions to a close-up of the person's hands typing on the keyboard again, with the focus on the hands and the keyboard. The video then shows a close-up of the person's hands typing on the keyboard, with the focus on the hands and the keyboard. The video concludes with a close-up of the person's hands

LLM Judge 3/5


Justification: Candidate covers basic camera work but misses key details about angles, transitions, and depth of field.
Included evidence: close, focus, hands, keyboard, typing
Missing evidence: employs, series, shots, capturing, intricate

Vid2LoRA

The video shows a person's hands typing on a laptop keyboard. The person's hands are shown in close-up shots, with the focus on the fingers as they press the keys. The background is dark, and the lighting is focused on the hands and the keyboard. The video does not show any other objects or people, and the person's actions are the main focus.

LLM Judge 4/5

Justification: Candidate misses key camera movement details but covers main framing and background blur.
Included evidence: close, shots, focus, hands, keyboard
Missing evidence: employs, series, capturing, intricate, process



Describe the camera work, framing, and viewpoint in this video.

Reference: The camera work is minimalist, focusing solely on the petals as they fall. The shots are static, with a consistent framing that captures the entire area where the petals are falling. The angle is directly overhead, providing a clear view of the petals' descent. There are no camera movements or transitions, allowing the viewer to immerse themselves in the tranquil scene without distraction. The simplicity of the camera work complements the elegance of the petals, emphasizing their delicate nature.

Base model

The video begins with a close-up of a white, leaf-like object on a light gray background, which gradually becomes more detailed and animated. The object appears to be a leaf, with a smooth, curved surface and a pointed tip. As the video progresses, the leaf-like object is shown in various stages of animation, including a close-up of a leaf with a pointed tip, a leaf with a curved edge, and a leaf with a pointed tip and a curved edge. The leaf-like object is then shown in a more detailed view, with a textured surface and a pointed tip, and is further

LLM Judge 3/5

Justification: Candidate covers basic camera style but misses key framing/angle details.
Included evidence: some coarse scene context
Missing evidence: work, minimalist, focusing, solely, petals

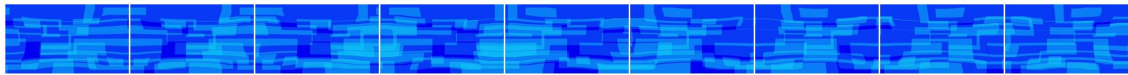
Vid2LoRA

The video begins with a plain white background, where a single pink leaf is seen floating in the air. As the video progresses, the leaf starts to move, creating a sense of motion. The leaf then begins to float in the air, and the background becomes a gradient of light and dark colors, creating a dreamy atmosphere. The leaf continues to float in the air, and the background remains a gradient of light and dark colors, creating a dreamy atmosphere. The leaf continues to float in the air, and the background remains a gradient of light and dark colors, creating a dreamy atmosphere. The leaf continues to float in the air

LLM Judge 5/5

Justification: Candidate fully covers all camera work facts with extra harmless details.
Included evidence: where
Missing evidence: work, minimalist, focusing, solely, petals

Figure 28: Qualitative examples from VDC Camera.



Describe this video in detail.

Reference: The video immerses viewers in a captivating visual journey, where an evolving abstract design unfolds against a serene backdrop dominated by a vibrant spectrum of blue hues. The composition features an intricate arrangement of translucent rectangles, each exhibiting various shades of blue, layered against horizontal stripes that create a sense of depth and fluidity. This harmonious color theme evokes feelings of tranquility, reminiscent of gentle waves lapping at the shore, as the visual rhythm flows seamlessly throughout the piece. As the video progresses, the rectangles begin to layer, creating a fragmented effect that enhances the perception of depth. The shapes glide smoothly across the screen, their transparency allowing for a captivating layering effect where some elements recede into the background while others emerge prominently. This dynamic interplay of light and shadow creates a mesmerizing visual depth, inviting viewers to explore the intricate relationships between the shapes. The transformation of the design is marked by the dispersion of rectangles, particularly towards the right side of the frame, where they intersect distinctly with the horizontal stripes below. This intricate convergence forms complex intersections that suggest a network of interrelated configurations. As the rectangles elongate and intertwine more seamlessly, their edges soften, leading to a unified and fluid visual array that captivates the eye. The evolving shapes oscillate between fragmentation and expansion, with their boundaries increasingly blurring into one another, enhancing the sense of motion and fluidity. In the latter part of the video, the transformation subtly continues as lighter blue rectangles begin to blend more into the background, allowing the darker rectangles to gain visual prominence. This shift in focus introduces a more complex interplay between shades, as darker hues become focal towards the center while lighter hues spread towards the edges. The visual composition is enriched by the dynamic evolution of texture, where the edges of the darker rectangles blur, enhancing the overall sense of depth and spatial perception. Throughout the video, the camera work is characterized by a steady, smooth motion that captures the fluidity of the abstract shapes. Wide shots allow viewers to take in the entire composition without distraction, while seamless transitions enhance the feeling of continuity. The consistent angles focus on the interplay of shapes and colors, drawing the viewer into the immersive visual experience. Horizontal streaks within the composition stretch and merge with the background, hinting at motion and further contributing to the gentle, flowing atmosphere. As the rectangles continue to morph and shift, the overall arrangement introduces new layers and interactions among the shades of blue, suggesting a fluid, gentle motion that conveys a nuanced evolution in balance, depth, and subtle dynamism. The video culminates in a harmonious blend of colors and shapes, leaving viewers with a sense of serenity and wonder, as they are transported into this vibrant world where laughter and adventure intertwine, echoing the heartwarming moments of a journey filled with unexpected twists.

Base model

The video begins with a series of blue, rectangular shapes that gradually become more defined and interconnected, creating a complex, abstract pattern. The shapes are arranged in a seemingly random order, with some overlapping and others forming distinct lines. As the video progresses, the shapes become more defined, and the lines between them become more pronounced, eventually forming a grid-like pattern. The video then transitions to a blue, abstract background with a grid-like pattern, which gradually becomes more defined and complex, eventually forming a grid-like pattern with a grid of blue rectangles. The video concludes with a blue, abstract background with a grid of blue rectangles.

LLM Judge

1/5

Justification: Candidate describes static images with no movement, contradicting reference's dynamic evolution.
Included evidence: abstract
Missing evidence: immerses, viewers, captivating, visual, journey

Vid2LoRA

The video consists of a series of frames that display a blue background with a series of horizontal blue lines. The lines are arranged in a pattern that creates a sense of depth and movement. The video does not contain any text or other objects, and the lines remain static throughout the frames. The video does not contain any discernible actions or movements, and the lines appear to be static. The video does not contain any additional elements or objects, and the lines remain static throughout the frames.

LLM Judge

1/5

Justification: Candidate describes static images while reference describes dynamic evolution.
Included evidence: some coarse scene context
Missing evidence: immerses, viewers, captivating, visual, journey



Describe this video in detail.

Reference: The video opens with an entirely black screen, enveloping the viewer in a profound darkness that is devoid of any discernible objects, movement, or details. This void persists for a brief moment, creating an atmosphere of anticipation as the viewer is left to ponder what might emerge from the obscurity. Suddenly, the scene transitions to reveal a striking circle of golden Roman numerals, elegantly arranged against the dark backdrop. The numerals, crafted with a smooth, glossy finish, reflect subtle glimmers of light, enhancing their luxurious appearance. The camera remains fixed in a static position, maintaining a focused gaze on the numeral circle, which stands out vividly against the stark black background. As the scene unfolds, golden droplets begin to cascade down from above, pouring into the upper half of the Roman numeral circle. These droplets, reminiscent of coins, vary in size and shape, creating a dynamic and visually engaging spectacle. Their random pattern and fluid descent contribute to a sense of lively activity, contrasting sharply with the stillness of the background. The sound of the droplets landing softly on the clock face adds an auditory layer to the experience, with the gentle clinking of coins enhancing the overall ambiance. As time progresses, the golden droplets accumulate, forming a dense mound at the center of the circle. This growing pile obscures parts of the Roman numerals, spilling towards the edges and creating a captivating visual interplay between the gold and the dark background. The camera's unswerving focus captures the evolution of this scene, allowing viewers to witness the transformation as the droplets spread out thinly, eventually covering the entire circular numeral array. Some of the numerals at the periphery remain partially visible beneath the scattering of gold, adding depth and intrigue to the composition. Throughout this mesmerizing sequence, the background remains unaltered, providing a consistent canvas that emphasizes the golden elements. The camera work is smooth and steady, with close-up shots that invite viewers to appreciate the intricate details of the coins and the elegant design of the clock face. The lighting is soft yet bright enough to highlight the rich golden hues, making them pop against the dark setting. There are no abrupt transitions; instead, the flow is continuous, creating a seamless viewing experience that draws the audience deeper into the unfolding narrative. The overall atmosphere of the video evokes a sense of abundance and prosperity, symbolized by the cascading golden droplets that represent wealth and the passage of time. The rhythmic movement of the coins, their shimmering descent, and the satisfying sound of their landing all contribute to a captivating portrayal of opulence. As the scene progresses, viewers are left with a lingering sense of anticipation and excitement, encapsulated in the elegant dance of gold against the timeless backdrop of darkness.

Base model

The video begins with a black screen that transitions to a circular arrangement of golden coins, each with a distinct design and color, arranged in a circular pattern. The coins are evenly spaced, and the background is a solid black color, emphasizing the circular arrangement. As the video progresses, the coins are shown to be falling, with some coins appearing to be in motion, suggesting a dynamic or possibly chaotic motion. The falling coins are scattered across the screen, with some coins closer to the center and others further away, creating a sense of movement and disorder. The falling coins are illuminated by a light source, casting a glow on the screen.

LLM Judge

3/5

Justification: Candidate misses key actions and atmosphere, but covers basic scene elements.
Included evidence: black, screen, movement
Missing evidence: opens, entirely, enveloping, viewer, profound

Vid2LoRA

The video begins with a black background featuring a gold-colored circular pattern in the center. As the video progresses, the circular pattern becomes more defined, with the gold color gradually becoming more prominent. The pattern then transforms into a series of gold coins, each with a unique design and color, arranged in a circular pattern. The video concludes with the coins still in their circular formation, with the gold color remaining the dominant color.

LLM Judge

2/5

Justification: Candidate describes coins spinning and scattering, contradicting reference's static camera and gradual droplet accumulation.
Included evidence: black
Missing evidence: opens, entirely, screen, enveloping, viewer



Describe this video in detail.

Reference: The video unfolds a breathtaking tableau of nature's beauty, showcasing a serene sunset over a tranquil body of water, likely a lake or a calm sea. As the scene progresses, the sun is initially positioned centrally just above the horizon, casting a warm and inviting glow of orange and yellow hues that dance across the water's surface. This radiant light creates a captivating pathway that extends towards the viewer, infusing them to immerse themselves in the moment. The sky is a canvas of color, transitioning from a vibrant orange near the sun to a deep, soothing blue at the top of the frame, enhancing the overall tranquility of the scene and evoking a sense of peace. In the foreground, several boats float gently on the water, their silhouettes stark against the brilliant backdrop of the sunset. These boats, seemingly at rest, provide a sense of perspective and distance, emphasizing the vastness of the lake and the beauty of the moment. In the far background, the silhouettes of a landmass—possibly a peninsula or an island—adds depth to the composition, subtly illuminated by the fading light of the setting sun. The water's surface remains calm, adorned with gentle ripples that reflect the warm colors of the sky, while a small, ambiguous object near the bottom of the frame piques curiosity—perhaps a piece of debris or a small aquatic creature, adding an element of intrigue to the serene landscape. As time elapses, the sun descends further, now gently kissing the horizon line. This descent elongates the shimmering reflection path on the water, deepening the spectacle of the day's end and enhancing the visual narrative of the day's end. The silhouettes of the boats remain relatively unchanged, suggesting a tranquil moment captured in time, where the atmosphere is undisturbed and peaceful. The sky continues to maintain its colorful gradient, transitioning smoothly from vibrant orange to deep blue, symbolizing the continuity of nature's rhythm and day gracefully giving way to evening. The camera captures this enchanting scene through a series of wide shots that emphasize the expansive sky and the calm water, creating a sense of openness and freedom. The angles are thoughtfully composed to include both the horizon and the boats, achieving a harmonious balance in the visual storytelling. Smooth transitions between shots allow viewers to fully absorb the changing colors of the sunset, while moments of slow zooming on the boats highlight the solitary figures within them, enhancing the feeling of solitude and reflection against the vastness of the lake. The use of natural lighting beautifully enhances the warm tones of the sunset, enveloping the scene in a soft, inviting atmosphere that resonates with the viewer. Overall, the video encapsulates a moment of peacefulness and stillness, inviting viewers to experience the beauty of a day's close by the water. The serene ambiance, combined with the stunning visual elements, creates a rich tapestry of nature's artistry, leaving a lasting impression of tranquility and wonder.

Base model

The video captures a serene scene of a sunset over a calm body of water, likely a lake or a bay, under a clear sky. The sun is setting, casting a warm orange glow over the water, which is calm and reflective. Several boats are visible on the water, some closer to the shore and others further out, suggesting a leisurely or fishing activity. The water's surface is slightly disturbed by the movement of the boats, creating a sense of motion. The sky is a gradient of orange and blue, with the sun positioned near the horizon, indicating the end of the day. The overall atmosphere is peaceful and calm.

LLM Judge

4/5

Justification: Covers main scene but misses key details about sun position, water reflections, and camera work.
Included evidence: sunset, over, body, water
Missing evidence: unfolds, breathtaking, tableau, nature, beauty

Vid2LoRA

The video showcases a serene and picturesque scene of a sunset over a body of water. The sky is a gradient of orange and yellow, with the sun just setting below the horizon, casting a warm glow over the water. The water is calm, with gentle ripples visible on the surface. A small boat is visible in the distance, adding a sense of depth to the scene. The overall atmosphere is one of tranquility and beauty, with the sun's light creating a beautiful contrast against the darkening sky.

LLM Judge

4/5

Justification: Covers main scene and key elements but misses some specific details like sun position and background landmarks.
Included evidence: beauty, serene, sunset, over, tranquil
Missing evidence: unfolds, breathtaking, tableau, nature, showcasing



Describe this video in detail.

Reference: The video presents a breathtaking view of a grand mosque, a stunning architectural masterpiece that commands attention with its pristine white stone facade. This magnificent structure is adorned with multiple towering minarets and large, elegant domes, each crowned with gleaming golden crescent moons that glisten in the sunlight. The mosque's design features pointed arches over some windows, while others are embellished with intricate geometric patterns, showcasing the artistry and craftsmanship that went into its construction. Set against a backdrop of a brilliant blue sky, the scene is further enhanced by soft, fluffy clouds that drift lazily, creating a serene and tranquil atmosphere. The surrounding landscape is a lush tapestry of well-manicured greenery, with neatly trimmed hedges and vibrant gardens that frame the path leading to the mosque. This natural beauty complements the architectural splendor of the mosque, inviting viewers to immerse themselves in the peaceful ambiance of the location. The gentle rustling of leaves in the breeze adds a soothing auditory element, reinforcing the sense of calm that permeates the scene. As the camera glides smoothly through the video, it captures the mosque from various angles, allowing viewers to appreciate its grandeur and intricate details. Wide-angle shots encompass the entire structure, emphasizing its impressive height and the harmonious relationship it shares with the surrounding landscape. The camera then transitions to closer shots, focusing on the delicate craftsmanship of the arches and the ornate decorative elements that adorn the facade. The sunlight reflects off the white stone, creating a radiant glow that highlights the mosque's beauty and invites a sense of reverence and awe. Throughout the video, the environment remains consistently tranquil, with no noticeable changes in the mosque's appearance or the surrounding scenery. This stability enhances the viewer's experience, allowing them to fully absorb the architectural beauty and spiritual significance of the mosque. The dynamic camera work, characterized by smooth pans and tilts, creates a fluid viewing experience that draws the audience into this vibrant world, where the grandeur of the mosque and the serenity of its setting intertwine seamlessly. The overall mood is one of appreciation and wonder, inviting viewers to reflect on the artistry and spirituality embodied in this magnificent structure.

Base model

The video showcases a grand, white mosque with multiple domes and minarets, set against a backdrop of a clear blue sky with scattered clouds. The mosque is surrounded by lush green vegetation, and the sky is filled with a soft, light blue hue, suggesting a sunny day. The architectural design of the mosque is characterized by its symmetrical and symmetrical design, with a central tower and multiple smaller domes. The mosque is surrounded by a well-maintained garden, enhancing its serene and peaceful atmosphere. The video captures the mosque from various angles, providing a comprehensive view of its architectural features and the surrounding environment.

LLM Judge

4/5

Justification: Covers main elements but misses key architectural details and atmospheric elements.
Included evidence: view, grand, mosque, architectural, white
Missing evidence: presents, breathtaking, stunning, masterpiece, commands

Vid2LoRA


The video showcases a grand, white mosque with a tall minaret, set against a backdrop of a clear blue sky. The mosque is surrounded by lush green grass and trees, creating a serene and peaceful atmosphere. The architecture of the mosque is intricate, with multiple domes and minarets, and the overall design is symmetrical and symmetrical. The video captures the mosque from various angles, highlighting its architectural details and the tranquility of the setting.

LLM Judge

4/5

Justification: Covers main elements but misses key architectural details and atmospheric elements.
Included evidence: grand, mosque, architectural, white
Missing evidence: presents, breathtaking, view, stunning, masterpiece


Figure 29: Qualitative examples from VDC Detailed.



Describe the main subject and its actions in this video.

Reference: The main subjects, two butterflies, are depicted in various frames as they hover and land on the white flowers. One butterfly, adorned with striking black and yellow patterns, delicately lands on a flower, extending its proboscis to extract nectar. The other butterfly, slightly smaller, flutters nearby, occasionally landing on the same flower, creating a dance of interaction. Their wings move gracefully, revealing intricate details and colors as they shift positions, showcasing their delicate nature. The butterflies exhibit a range of movements, from hovering in place to quick, darting flights, embodying a sense of freedom and playfulness.


<p>Base model</p> <p>The video showcases a series of close-up shots of a white and yellow flower with a black and yellow striped pattern, which appears to be a daisy, set against a blurred background of green foliage. Initially, the flower is shown in a close-up, with a bee hovering near it, indicating a moment of interaction between the bee and the flower. The bee is in focus, with its wings spread wide, and the flower is in sharp detail, with its petals fully open and a yellow center. The bee's position suggests it is either feeding on nectar or pollinating the flower. As the video progresses</p> <p>LLM Judge 2/5</p> <p><small>Justification: Candidate describes one butterfly with wrong wing colors and misses key actions and multiple butterflies. Included evidence: hover, white, black, yellow Missing evidence: subjects, butterflies, depicted, various, frames</small></p>	<p>Vid2LoRA</p> <p>The video features a close-up view of a bee on a flower, with the bee's wings and body clearly visible. The bee is black and yellow, with a long tongue extended towards the flower. The flower has a white center and is surrounded by green leaves. The background is blurred, but it appears to be a natural setting with other flowers and foliage. The video captures the bee's movement as it moves towards the flower, with the bee's wings and body moving in a rhythmic pattern. The video also shows the bee's wings and body moving in a rhythmic pattern</p> <p>LLM Judge 1/5</p> <p><small>Justification: Candidate only mentions one butterfly and states it's stationary, contradicting reference's multiple active butterflies. Included evidence: white, flowers, black, yellow Missing evidence: subjects, butterflies, depicted, various, frames</small></p>
--	---



Describe the main subject and its actions in this video.

Reference: The main subject, a person wearing a pink shirt, is engaged in food preparation. They first wash a piece of produce under the running water, their hands skillfully maneuvering to ensure it is clean. After rinsing, they place the item on a white cutting board and begin to chop it with a knife, demonstrating focus and precision. Their posture is relaxed yet attentive, with occasional glances towards the stove, indicating a multitasking approach to cooking. The subject's hands move fluidly, showcasing a familiarity with the kitchen tasks.


<p>Base model</p> <p>The video showcases a person in a kitchen setting, likely preparing food, as indicated by the presence of a microwave, a toaster, and various kitchen utensils. The person is seen holding a smartphone, possibly recording or taking a photo, and is engaged in the process of preparing food. The kitchen is equipped with a sink, a stove, and a microwave, suggesting a well-organized and functional kitchen environment. The person is seen using a knife to cut a piece of food, indicating a focus on food preparation. Throughout the video, the person's actions and the kitchen setting remain consistent, with no significant changes in the camera</p> <p>LLM Judge 3/5</p> <p><small>Justification: Candidate covers basic food prep but misses key actions and details from reference. Included evidence: engaged, food, preparation, piece Missing evidence: pink, shirt, first, wash, produce</small></p>	<p>Vid2LoRA</p> <p>A person is preparing food in a kitchen. They are using a microwave to heat up a pan on the stove. The person is also using a bowl to scoop out food from the pan.</p> <p>LLM Judge 2/5</p> <p><small>Justification: Candidate describes different actions than reference, missing key details while adding conflicting ones. Included evidence: food Missing evidence: pink, shirt, engaged, preparation, first</small></p>
--	--



Describe the main subject and its actions in this video.

Reference: The main focus of the video is the water as it flows energetically over the rocks. The water cascades in a series of small rapids, creating white frothy bubbles as it tumbles over the edges of the stones. The movement is fluid and rhythmic, with the water appearing to dance as it navigates the rocky terrain. The surface of the water reflects the sunlight, creating a shimmering effect that adds to the visual appeal. Occasionally, the water splashes against the rocks, sending droplets into the air, which catch the light and create a sparkling effect. The overall motion is continuous and soothing, evoking a sense of calm and connection to nature.

<p>Base model</p> <p>The video captures a serene and dynamic scene of a rocky river, where the water flows over a large, rounded rock. The water's movement is gentle, creating a soothing rhythm that contrasts with the rocky texture. The river's surface is a mix of white foam and clear water, indicating recent or ongoing water flow. The rocks surrounding the river are of varying sizes and shapes, with some appearing to be partially submerged in the water. The overall scene is one of tranquility and natural beauty, with the water's movement and the rocks' textures creating a harmonious balance.</p> <p>LLM Judge 3/5</p> <p><small>Justification: Candidate covers basic water-rock interaction but misses key details about water's movement and effects. Included evidence: water, flows, over, rocks, creating Missing evidence: focus, energetically, cascades, series, small</small></p>	<p>Vid2LoRA</p> <p>The video showcases a serene and natural scene of a waterfall cascading down a rocky cliff. The water, a deep blue, is captured in motion, creating a powerful and mesmerizing visual effect. The rocks, a mix of gray and white, are interspersed with rocks of varying sizes, adding to the natural beauty of the scene. The water appears to be flowing at a steady pace, with no visible signs of human activity or disturbance. The overall atmosphere is one of tranquility and natural beauty, with the waterfall serving as a focal point in the video.</p> <p>LLM Judge 4/5</p> <p><small>Justification: Covers main water actions but misses some descriptive details about movement and light effects. Included evidence: water, over, rocks, creating, white Missing evidence: focus, flows, energetically, cascades, series</small></p>
--	--




Describe the main subject and its actions in this video.

Reference: The main focus is the sun as it descends towards the horizon, gradually transforming the sky into a canvas of warm colors. As the sun lowers, its light creates a shimmering path across the water, intensifying the golden reflections. The trees in the foreground stand still, their branches reaching towards the sky, adding a touch of nature's beauty to the scene. The overall mood is peaceful and contemplative, inviting viewers to pause and appreciate the beauty of nature.

<p>Base model</p> <p>The video captures a serene and picturesque scene of a sunset over a body of water, featuring a solitary tree standing in the foreground. The sky is painted in hues of orange and yellow, with the sun just beginning to set, casting a warm glow over the landscape. The water, calm and still, reflects the vibrant colors of the sky, creating a mirror-like effect. The tree, devoid of leaves, stands alone, its silhouette stark against the sky. The scene is devoid of any human activity, allowing the viewer to fully immerse themselves in the tranquil beauty of the sunset.</p> <p>LLM Judge 4/5</p> <p><small>Justification: Candidate covers main subject and key actions but misses some specific details about reflections and branches. Included evidence: warm, colors, water Missing evidence: focus, descends, towards, horizon, gradually</small></p>	<p>Vid2LoRA</p> <p>The video showcases a serene and picturesque scene of a sunset over a body of water. The sky is painted with hues of orange and yellow, reflecting off the water's surface. The water is calm, with gentle ripples adding to the tranquility of the scene. The sun is setting behind a line of trees, casting a warm glow on the water and the surrounding landscape. The video captures the beauty of nature in its natural state, with no human activity or man-made objects present. The overall atmosphere is one of peace and tranquility, with the sun's rays creating a beautiful and calming effect.</p> <p>LLM Judge 4/5</p> <p><small>Justification: Candidate covers main subject and key actions but misses minor details about trees and reflection intensity. Included evidence: warm, water Missing evidence: focus, descends, towards, horizon, gradually</small></p>
--	---

Figure 30: Qualitative examples from VDC Main Object.



Summarize this video in one detailed sentence.

Reference: The video is a cinematic and introspective portrait of daily life, seamlessly weaving together snapshots of work, leisure, and learning in a series of vignettes that explore the human experience through subtle shifts in environment and expression.

Base model

The video captures a person in a modern office environment, beginning with a close-up of a book titled "About the Action" by John C. Maxwell. The scene transitions to a wider view of the office, revealing a gray couch with red and white pillows, a desk with a lamp, and a bookshelf filled with books and magazines. The person then moves to a different part of the office, where they are seen holding a book titled "The Power of Now" by Eckhart Tolle. The video concludes with a view of the office from a different angle, showing a person in a red shirt working at a

LLM Judge 1/5


Justification: Candidate describes a single scene with a man, missing the reference's broader scope of daily life vignettes.
Included evidence: work
Missing evidence: cinematic, introspective, portrait, daily, life

Vid2LoRA

A man in a red shirt is standing in a modern office space, talking to the camera. He is then seen sitting at a table in a restaurant, reading a book. The video also shows a close-up of a book on a table, and a man in a white shirt sitting at a table in a restaurant.

LLM Judge 1/5

Justification: Candidate describes specific scene details but misses main reference focus on human experience through vignettes.
Included evidence: some coarse scene context
Missing evidence: cinematic, introspective, portrait, daily, life



Summarize this video in one detailed sentence.

Reference: A young child engages in joyful and creative play, using modeling clay to express themselves, showcasing acael imagination and dexterity in a warm and inviting domestic setting.

Base model

A young child, dressed in a light blue polo shirt and dark pants, is engaged in a creative activity on a rug in a living room. The child is using a colorful crayon to draw on a white piece of paper, surrounded by various toys and a soccer ball.

LLM Judge 2/5


Justification: Candidate misses key reference elements like modeling clay and creative play.
Included evidence: young, child, creative, using
Missing evidence: engages, joyful, play, modeling, clay

Vid2LoRA

A young girl is sitting on the floor in a living room, drawing on a piece of paper with a marker. She is wearing a light blue shirt and dark pants.

LLM Judge 2/5

Justification: Candidate describes different activity and setting than reference.
Included evidence: young
Missing evidence: child, engages, joyful, creative, play



Summarize this video in one detailed sentence.

Reference: A serene and breathtaking video captures a tranquil sunset over a calm body of water, with vibrant colors and gentle ripples creating a peaceful atmosphere that invites viewers to reflect and unwind.

Base model

The video captures a serene sunset over a calm body of water, with the sun setting behind a distant shoreline. The scene is peaceful, with the sun casting a warm glow on the water and creating a beautiful reflection on the surface. The camera slowly pans across the water, revealing the horizon and the distant shoreline, with the sun's reflection adding depth to the scene. The video concludes with a close-up of the sun's reflection on the water, emphasizing the tranquil and picturesque setting.

LLM Judge 4/5


Justification: Captures main scene but misses vibrant colors and ripples.
Included evidence: serene, captures, tranquil, sunset, over
Missing evidence: breathtaking, vibrant, colors, gentle, ripples

Vid2LoRA

The video shows a serene sunset over a body of water, with a small boat visible in the distance. The sky is a gradient of orange and yellow, and the water is calm. The sun is setting behind the horizon, casting a warm glow over the scene. The boat is small and appears to be a motorboat, and there are no people visible in the video. The video captures the beauty of the sunset and the tranquility of the water.

LLM Judge 4/5

Justification: Captures main scene but misses 'breathtaking' and reflective invitation.
Included evidence: serene, captures, tranquil, sunset, over
Missing evidence: breathtaking, vibrant, colors, gentle, ripples



Summarize this video in one detailed sentence.

Reference: The video presents a mesmerizing, symmetrical composition of concentric hexagonal shapes in vibrant red hues, creating a hypnotic tunnel effect that invites viewers to embark on a meditative and captivating visual journey.

Base model

The video presents a series of concentric red circles, creating a hypnotic visual effect, with a subtle change in the background color from red to black.

LLM Judge 5/5

Justification: Candidate captures all key reference elements: hexagons, red color, tunnel effect, and hypnotic quality.
Included evidence: presents, concentric, creating, hypnotic, effect
Missing evidence: mesmerizing, symmetrical, composition, hexagonal, shapes

Vid2LoRA

The video displays a series of frames featuring a red, swirling pattern that appears to be a representation of a wave or a pattern of lines. The pattern is dynamic and appears to be moving or changing in some way, as the frames are not consecutive and the pattern changes slightly between each frame. The video does not contain any text or other discernible objects, and the focus is solely on the visual representation of the red, swirling pattern.

LLM Judge 5/5

Justification: Candidate captures all key reference elements with minor extra details that don't conflict.
Included evidence: some coarse scene context
Missing evidence: presents, mesmerizing, symmetrical, composition, concentric

Figure 31: Qualitative examples from VDC Short.