Self-alignment of Large Video Language Models with Refined Regularized Preference Optimization





Pritam Sarkar Ali Etemad

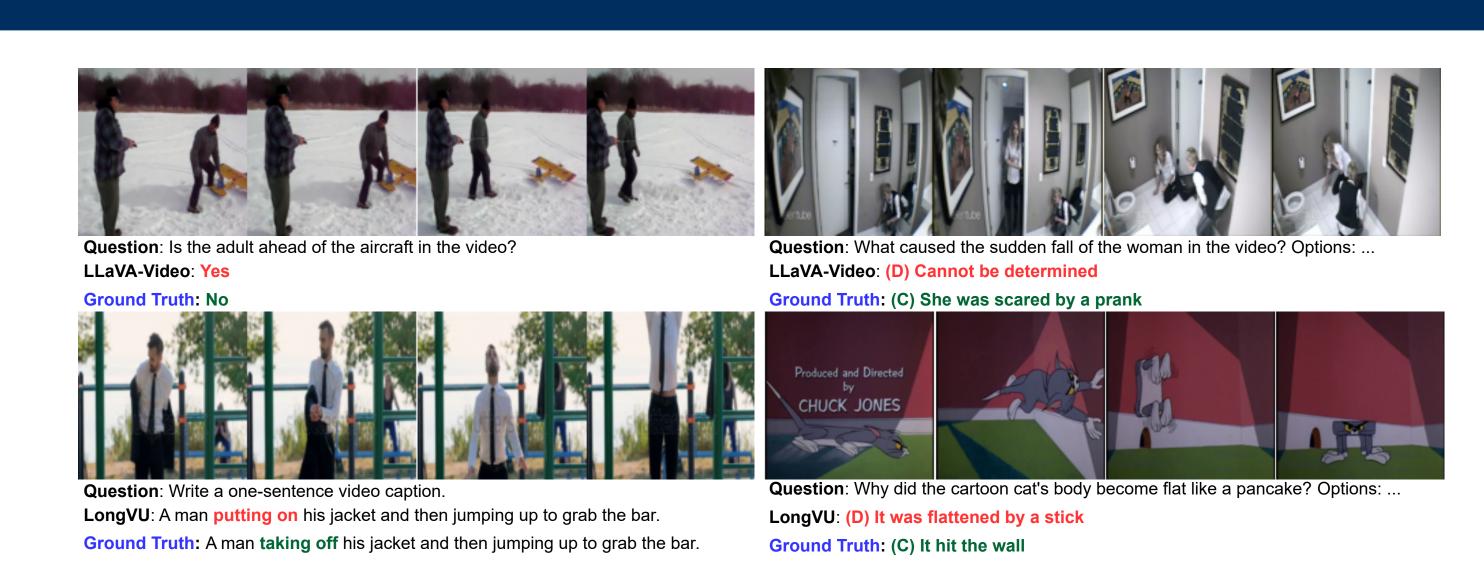
NeurlPS 2025

https://pritamsarkar.com/RRPO

Motivation

Despite recent progress, Large Video Language Models (LVLMs) still The underlying causes include:

- inadequate spatio-temporal understanding
- vision-language misalignments
- spurious correlations
- over-reliance on language over vision
- other factors



Limitations of Existing Methods

Limitation 1. Existing offline RL methods either

 use static, one-size-fits-all datasets that ignore model-specific states, or rely on costly, manually annotated model responses that do not scale.

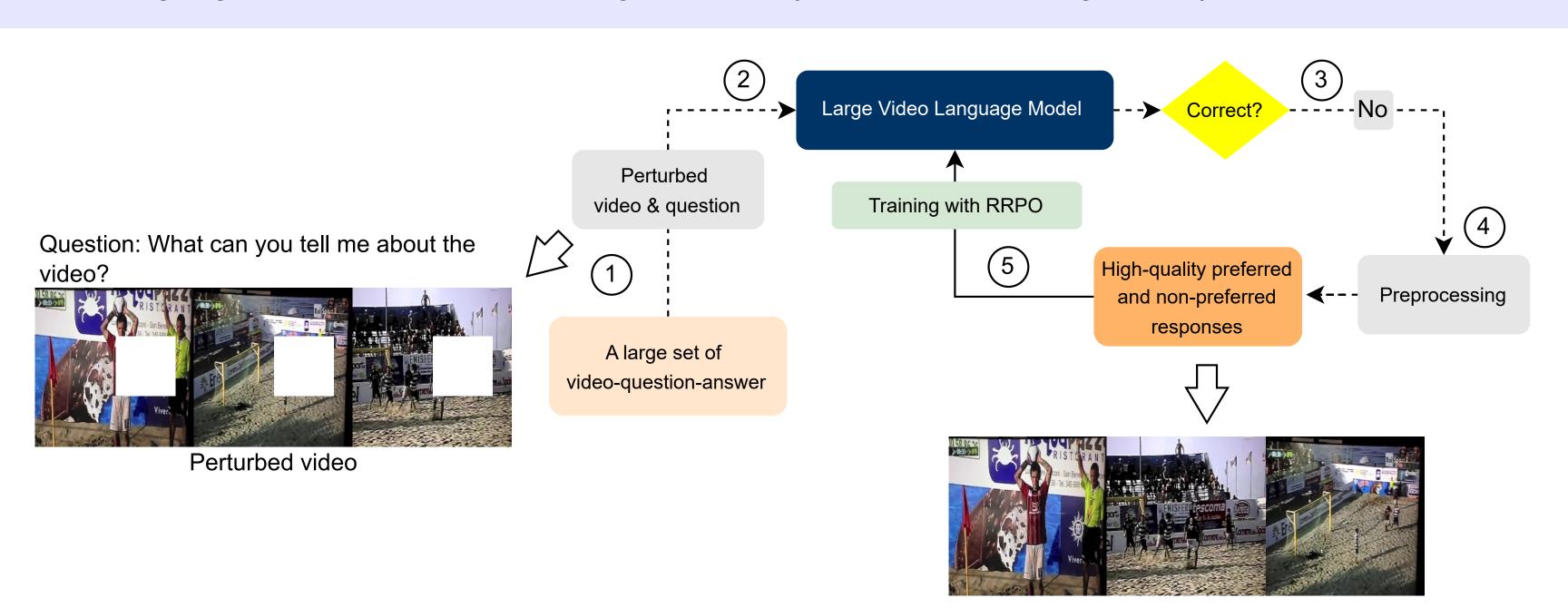
Limitation 2. Existing offline RL methods such as DPO provide a coarse-grained reward that

• penalizes all tokens in the rejected response even when only a few differ from the preferred ones,

- and do not inform the model where the error occurred.
- As a result, such objectives tend to accumulate large gradients,
- causing training instability,
- and tending to diverge and forget useful knowledge.

Our Proposed Method

To address Limitation 1, we design a self-alignment framework that allows LVLMs to learn their from own mistakes especially those arises due to vision-language misalignments, over-reliance on language over vision, lack of fine-grained temporal understanding, and spurious correlations, among others.



Question: What can you tell me about the video? Preferred (ground-truth): The video features a group of people playing soccer on a sandy field. Rejected (generated): The video shows a group of people playing volleyball on a sandy beach.

To address Limitation 2, we design a fine-grained alignment method that informs the model exactly where the error occurred. Formally, given an input x with a pair of responses $\{y^+,y^-\}$, where y_i^+ and y_i^- denote the i-th differing phrase, we aim to achieve

 $y_i^+ \mid x, y_{\le i}^+ \succ y_i^- \mid x, y_{\le i}^-$

instead of the existing formulation, e.g., DPO

$$y^+ \mid x \succ y^- \mid x$$
.

We introduce Refined Regularized Preference Optimization (RRPO) defined as:

$$\mathcal{L}_{\mathsf{RRPO}}(\pi_{\theta}; \pi_{\mathsf{ref}}) = -\mathbb{E}_{(x, y^+, y^-) \sim \mathcal{D}} \left[\log \sigma(u) + \alpha \cdot \mathbb{D}_{\mathsf{TKL}}(x, y^+) \right].$$

Refined Reward Modelling

The total reward margin $u=\sum u_i=\sum \left(r_{\theta}(x,y_i^+)-r_{\theta}(x,y_i^-)\right)$, where the finegrained reward for the i^{th} phrase $r_{\theta}(x,y_i)$ is defined as:

$$r_{ heta}(x, y_i) = \beta \log \left(rac{\prod\limits_{j=s_i}^{e_i} \pi_{ heta}(t_j \mid x, t_{< j})}{\prod\limits_{j=s_i}^{e_i} \pi_{ ext{ref}}(t_j \mid x, t_{< j})}
ight),$$

where s_i and e_i denote the start and end token indices of the i^{th} phrase.

Token-wise KL Regularizer

 $\mathbb{D}_{\mathsf{KL}}(\pi_{\mathsf{ref}}||\pi_{\theta}) = \sum_{a} \pi_{\mathsf{ref}}(a) \log \frac{\pi_{\mathsf{ref}}(a)}{\pi_{\theta}(a)}$; a represents a token in the vocabulary.

Comparison between RRPO and DPO gradient:

 $\|
abla_{ heta}\mathcal{L}_{\mathsf{RRPO}}\| \ll \|
abla_{ heta}\mathcal{L}_{\mathsf{DPO}}\|.$

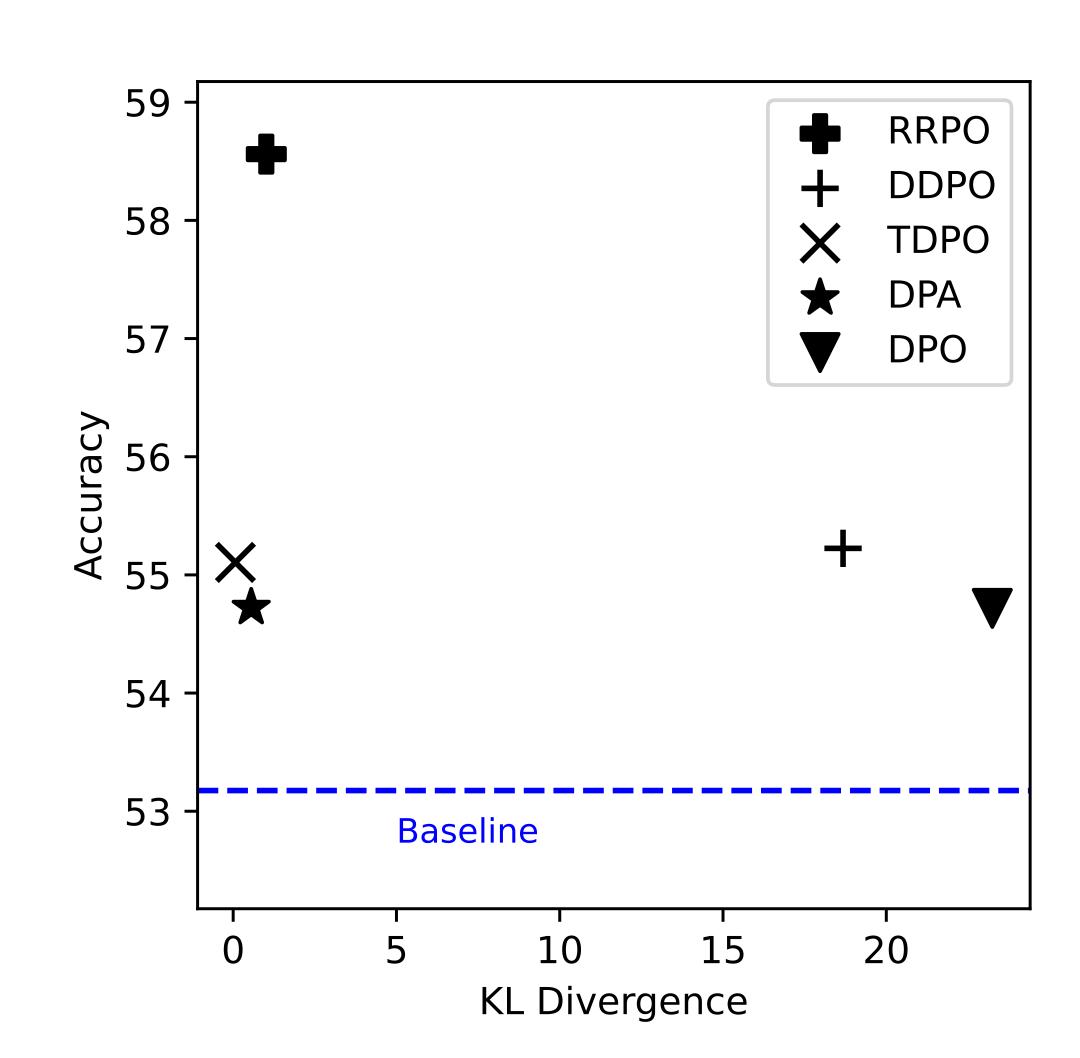
Results and Analyses

Main results. Evaluating our self-aligned LVLMs on diverse video understanding tasks including fine-grained temporal understanding, video hallucination, and comprehensive video understanding of varying video lengths.

Models	# Frames	TVBench	TempCompass _{Avg}	VideoHallucer	VidHalluc	MVBench	VideoMME	$MLVU_{val}(\text{M-Avg})$	LongVideoBench _{Val}
VideoChat2 _{7B}	16	44.0	59.3	23.1	73.3	60.2	41.0	46.4	40.4
+ DPO (16f)	16	45.7	60.0	22.1	72.4	59.6	43.0	47.4	41.0
+ RRPO (16f)	16	45.8	60.2	32.9	76.4	59.0	44.3	47.9	42.8
LLaVA-Video _{7B}	64	51.0	66.0	50.0	76.6	61.1	64.0	68.6	60.1
+ DPO (16f)	64	51.9	66.4	53.3	76.5	60.6	63.1	67.4	59.4
+ RRPO (16f)	64	51.9	66.8	55.7	76.5	62.2	64.5	69.1	60.4
+ RRPO (32f)	64	52.2	67.4	55.8	76.6	62.1	64.5	69.4	60.1
LongVU _{7B}	1 fps	53.7	63.9	39.2	67.3	65.5	56.2	63.6	48.6
+ DPO (16f)	1 fps	54.3	64.3	40.9	68.5	65.9	56.6	63.6	49.4
+ RRPO (16f)	1 fps	56.5	64.5	44.0	71.7	66.8	57.7	64.5	49.7

Comparing performance and model divergence after alignment. RRPO exhibits the best performance with minimal divergence.

_		TVBench	VideoHallucer	VideoMME	MLVU	$\Delta/\%$
	Baseline	53.7	39.2	56.2	63.6	
	+ DPO	54.3	40.9	56.6	63.6	0.7/1
	+ DPA	54.6	40.3	56.9	63.9	0.7/1
	+ TDPO	53.9	41.4	57.0	63.8	0.8/1
	+ DDPO	54.2	41.7	56.7	63.6	0.9/2
	+ RRPO (Ours)	56.5	44.0	57.7	64.5	2.5/5



Impact of data size. Our data-generation pipeline is effective in producing high-quality training samples and RRPO scales well with more data.

	TVBench	VideoHallucer	VideoMME	MLVU	$\Delta/\%\Delta$
Baseline	51.0	50.0	64.0	68.6	
+ 5 K	50.9	53.7	64.0	69.0	1.0/1.9
+ 10 K	51.2	53.8	64.3	69.0	1.2/2.3
+ 15 K	51.8	54.4	64.2	68.9	1.4/2.8
+ 20 K	51.9	55.7	64.5	69.1	1.9/3.7

Impact of varying number of frames. RRPO consistently improves performance over the base model across all setups.

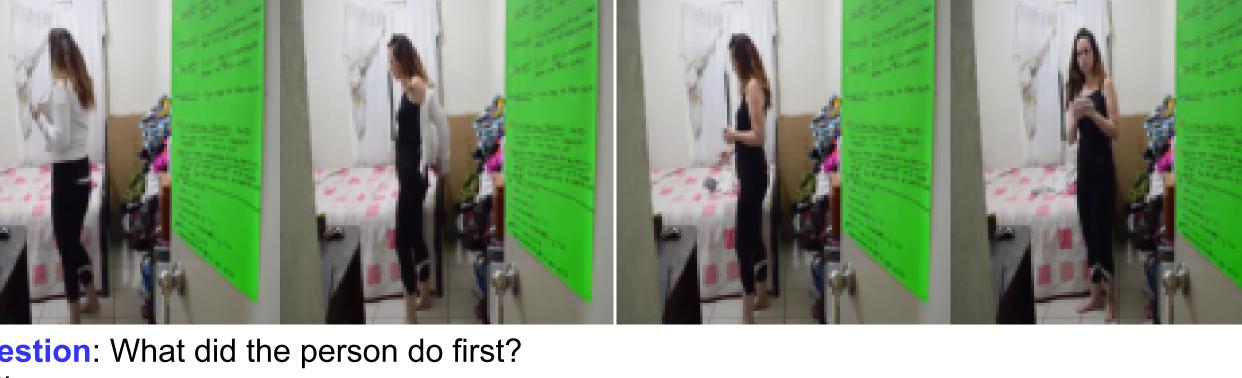
Frames Models TVBench MVBench LongVideoBench

J <u>Z</u>	+ RRPO	51.3	61.7	58.9	
64	Baseline + RRPO	51.0 52.2	61.1 62.1	60.1 60.1	
128	Baseline + RRPO	49.4 51.3	60.5 61.2	60.3 61.3	-

Impact of using subtitles along with videos. Our method maintains its benefits over base models even when subtitles are included.

	Without	With
VideoChat2 _{7B}	41.0	48.0
+ RRPO	44.3	49.4
LLaVA-Video _{7B}	63.8	67.4
+ RRPO	64.5	68.0
LongVU _{7B}	56.2	62.0
+ RRPO	57.7	63.1

Qualitative Results



Question: What did the person do first? (A) Held the phone/camera. (B) Threw the clothes. LLaVA-Video: (A) Held the phone/camera.

LLaVA-Video + DPO: (A) Held the phone/camera.

LLaVA-Video + RRPO: (B) Threw the clothes.



Question: Choose the caption that closely matches the video. LLaVA-Video: The man above is jumping rope at a steady pace. LLaVA-Video + DPO: The man above is jumping rope at a steady pace. LLaVA-Video + RRPO: The man above is jumping rope at the same speed.



Qualitative Results (Contd...)

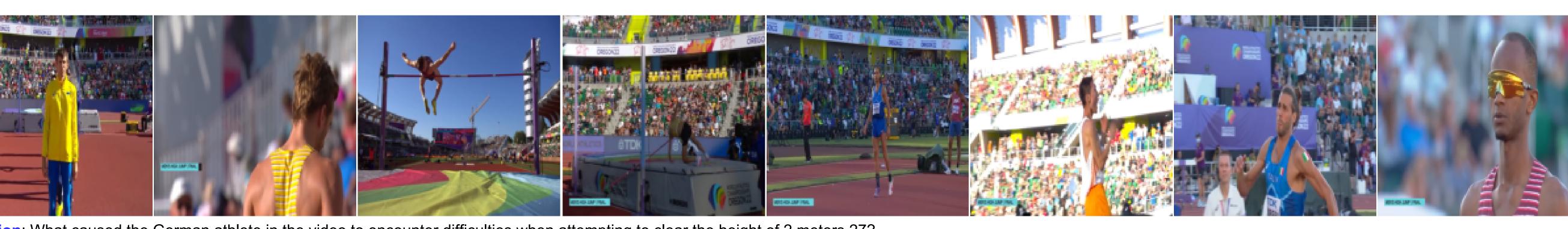


Question: What's the right option for how the scenes in the video change? Options: (A) From the elevator to the secretary's office. (B) From the secretary's office to the elevator.

LongVU: (A) From the elevator to the secretary's office. LongVU + DPO: (A) From the elevator to the secretary's office.

LongVU + RRPO: (B) From the secretary's office to the elevator.

Question: Choose the caption that closely matches the video. LongVU: A man putting on his jacket and then jumping up to grab the bar. LongVU + DPO: A man putting on his jacket and then jumping up to grab the bar. LongVU + RRPO: A man taking off his jacket and then jumping up to grab the bar.



Question: What caused the German athlete in the video to encounter difficulties when attempting to clear the height of 2 meters 27?

(A) The athlete slipped on the takeoff and lost balance, causing him to miss the jump. (B) The athlete's coach gave him incorrect instructions, leading to a failed attempt.

(C) The athlete suddenly stopped when he was approaching the crossbar frame, leaving little time for another high jump attempt.

(D) The athlete misjudged the distance and jumped too early, resulting in a failed attempt. LLaVA-Video: (A) The athlete slipped on the takeoff and lost balance, causing him to miss the jump.

LLaVA-Video + DPO: (A) The athlete slipped on the takeoff and lost balance, causing him to miss the jump. LLaVA-Video + RRPO: (C) The athlete suddenly stopped when he was approaching the crossbar frame, leaving little time for another high jump attempt.



LongVU: The video captures a sequence of events in a snowy landscape, beginning with a person standing on the snow, holding a long, thin object, possibly a stick or rod, with a small, colorful object attached to it. The person is dressed in winter clothing, including a jacket, pants, and a beanie, suggesting cold weather conditions. The background features leafless trees and a clear sky, indicating a winter setting.

As the video progresses, the person's position changes slightly, and the camera angle shifts to reveal more of the snowy landscape, including a small, colorful object on the ground, which is later identified as a toy airplane. The person's actions suggest they are preparing to launch the airplane, as they are seen holding the rod and the airplane is positioned on the ground, ready for takeoff.

The person then moves closer to the airplane, indicating they are about to launch it. The camera angle changes to provide a clearer view of the person's actions and the airplane's position. The person's stance and the positioning of the airplane suggest they are ready for takeoff.

The video then shows the airplane in motion, with the person holding the rod and the airplane airborne, indicating that it has been launched. The person's posture and the airplane's position suggest they are actively controlling its

Finally, the airplane is seen descending towards the ground, with the person still holding the rod and the airplane in motion. The person's posture and the airplane's position suggest they are actively controlling its descent. The snowy landscape and leafless trees in the background remain consistent throughout the video, providing a continuous winter setting.

LongVU + DPO: The video captures a sequence of events in a snowy landscape, beginning with a person standing on the snow-covered ground, holding a long, thin object, possibly a stick or rod, with a small, colorful object attached to it. The person is dressed in winter clothing, including a jacket, pants, and a beanie, suggesting cold weather conditions. The background features leafless trees and a clear sky, indicating a winter setting.

As the video progresses, the person's position changes slightly, and the camera angle shifts to reveal more of the snowy landscape, including a small, colorful object on the ground, which appears to be a toy or model airplane. The person's posture and the positioning of the object suggest they might be preparing for an activity involving the airplane.

Further into the video, the person is seen holding a remote control device, indicating they are controlling the airplane is now visible in the air, suggesting it has been launched or is being flown. The person's actions and

the airplane's position imply they are actively engaged in flying the aircraft.

The video continues to show the person adjusting their stance and the remote control device, possibly to maintain control of the airplane. The airplane's position in the air remains consistent, indicating stable flight conditions. The snowy landscape and leafless trees in the background remain unchanged, providing a consistent winter setting throughout the video.

Towards the end of the video, the person is seen holding the remote control device with both hands, possibly preparing for a maneuver or adjusting the airplane's flight path. The airplane is no longer visible in the frame, suggesting it may have moved out of the camera's field of view or is flying at a higher altitude.

In the final scene, the person is standing alone in the snowy landscape, holding the remote control device in one hand, with the other hand raised, possibly signaling or adjusting the flight of the airplane. The person's posture and the positioning of the remote control device suggest they are actively engaged in flying the airplane. The snowy landscape and leafless trees in the background remain consistent, indicating that the video captures a continuous sequence of events in the same winter setting.

LongVU + RRPO: The video captures a sequence of events centered around the operation of a colorful remote-controlled airplane in a snowy outdoor setting. Initially, a person is seen holding a remote control device, presumably controlling the airplane, which is not visible in the frame. The individual is dressed in winter clothing, including a yellow jacket, and is standing on a snowy surface with a backdrop of leafless trees and a clear sky. The ground is marked with tire tracks, suggesting recent activity.

As the video progresses, the person is shown walking towards the camera, indicating movement across the snowy landscape. The remote control device is no longer in view, implying it might have been set aside or is being operated from a distance. The individual's attire and the snowy environment remain consistent throughout these frames.

Subsequently, the focus shifts to the remote-controlled airplane itself, which is now visible on the snow, suggesting it has been recently operated or is being prepared for flight. The individual who was previously controlling the airplane is no longer in the frame, and the camera has zoomed in to capture the airplane's details, including its colorful design and the presence of a propeller.

The video concludes with the airplane in motion, as indicated by the blurred propeller and the disturbed snow around it. This suggests that the airplane has been activated and is either taking off or landing. The camera maintains its

focus on the airplane, capturing the dynamic action against the unchanged snowy backdrop and leafless trees, emphasizing the contrast between the moving airplane and the stillness of the surrounding environment.

Acknowledgment

We thank the Bank of Montreal and Mitacs for funding this research, and the Vector Institute for providing computational resources.