

# Apollo: An Exploration of Video Understanding in Large Multimodal Models

## Appendix

### Future work

Several promising directions emerge from our study on Large Multimodal Models (LMMs). First, we employed a fully unified architecture, using the video encoder for videos and images by replicating each image  $N$  times. Exploring separated architectures, where images are processed with an image encoder and videos with both image and video encoders, could allow for the unfreezing of the encoders during supervised fine-tuning.

Second, in separated architectures, training the video and image encoders during supervised fine-tuning (SFT) and evaluating their individual contributions to performance could identify optimal training strategies. Similarly, training both encoders on mixed image and video data within unified architectures may help determine which encoder influences observed performance drops, enabling targeted improvements.

Further investigation into Scaling Consistency is necessary to confirm its applicability across a broader range of model sizes, ensuring its reliability for even larger models. We did not explore memory-based LMM approaches, such as memory banks or frame retrieval methods like text-conditioned pooling in Q-Former. Evaluating these techniques could test our hypothesis that these techniques might struggle to generalize to multi-turn conversations.

Lastly, current benchmarks primarily use academic multiple-choice formats, which inadequately assess conversational abilities. Developing a dedicated conversational evaluation benchmark for LMMs is essential to more accurately measure and enhance the dialogue performance of models in real-world scenarios.

### Appendix overview

This document provides more details of our approach and additional experimental results, organized as follows:

- § [A Analyzing the benchmarks](#). We provide an in-depth analysis of the different factors affecting evaluations such as video duration and format. We then give a detailed overview of how we curated ApolloBench.
- § [B Apollo implementation details](#). We provide in-depth description of Apollo, along with all the hyperparameters needed to reproduce Apollo.
- § [C Scaling Consistency](#). We provide an in-depth analysis of the correlations between models of different sizes, compare Scaling Consistency to traditional scaling laws, and highlight their utility in future research.
- § [D Video sampling analysis](#). We expand on our Video Sampling experiments and add a per-metric breakdown.
- § [E Raw results](#). We provide all the raw data used in our study for further analysis. For Sec. 4: Tab. 12 & 13, Sec. 5.1: Tab. 9 & 10, Sec. 5.2: Tab 8, Sec. 5.4: Tab. 4, Sec. 6.3: Tab. 11.

### Table of Contents

<b>A Analyzing the benchmarks</b>	<b>2</b>
A.1 Correlations within existing benchmarks . . . . .	2
A.2 Raw evaluations . . . . .	2
A.3 ApolloBench curation . . . . .	3
<b>B Apollo implementation details</b>	<b>4</b>
B.1 Architecture . . . . .	5
B.2 Data . . . . .	6
B.3 Training . . . . .	6
<b>C Scaling Consistency: efficient model design with smaller models</b>	<b>6</b>
<b>D Effect of video sampling on the different dimensions of video perception</b>	<b>7</b>
<b>E Raw results</b>	<b>9</b>

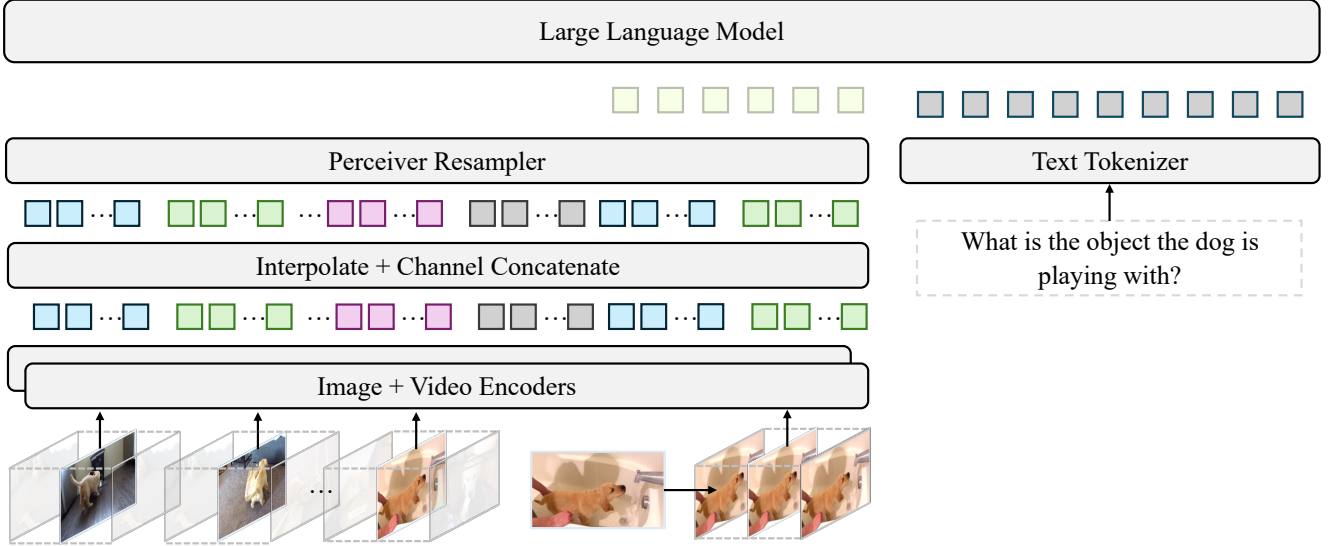


Figure 7. **Apollo architecture overview.** Apollo encodes clips of  $N$  (dependent on the video encoder) frames. Output features are interpolated and concatenated along the channel dimension before being fed to a connector. The connector up-projects the features to the Large Language Models’ hidden dimension and then resamples them into a pre-set number of  $T$  tokens/clip. Images are duplicated  $N$  times and encoded the same way as video clips.

## A. Analyzing the benchmarks

### A.1. Correlations within existing benchmarks

**Video duration.** We were interested in how the video length affected model performance to see if existing benchmarks test long video perception capabilities. In the large language model field, testing long-context has been non-trivial, where many benchmarks do not need information integration across the entire model’s context window and instead resemble needle-in-a-haystack experiments. We hypothesized that long video benchmarks may behave similarly. As such, we compared Video-MME short/medium/long and LongVideoBench’s different duration groups (see Fig. 10 and Fig. 11). We found that the two are highly correlated, where  $R^2 > 0.92$  between all duration groups in LongVideoBench (Fig. 11). On Video-MME, whether using or not using subtitles,  $R^2 > 0.83$ . When closely examining Video-MME short/medium/long in Fig. 2, one can see that the biggest difference between them is the performance in the video modality decreases, with text and image modalities being mostly unchanged. This indicates an increasing reliance on the text model’s performance rather than any vision capabilities.

**Question types.** There are currently two prevalent methods for evaluating LMMs—either open-ended questions or close-ended (multiple choice, yes/no). Scoring open-ended QA is hard because the score is ultimately subjective. The dominant way of evaluating open-ended QA is using another language model (e.g., chatGPT) to rate the prediction and decide if it is correct. As shown by Wu [45], GPT versioning strongly impacts the resulting scores that are even 10% apart. As such, recent trends show greater reliance on multiple-choice QA. However, are we losing something when evaluating methods only on multiple-choice? As seen in Fig. 9, we find these are highly correlated, with  $R^2 > 0.81$ . While multiple-choice appears to be a good option for benchmarking the video perception capabilities of video-LMMs, models overly optimized to multiple-choice will not be good conversational agents. As such, a benchmark focusing solely on a conversation is needed, ideally, one that does not suffer from high API costs and GPT versioning noise.

### A.2. Raw evaluations

We evaluated InverVL2 2B & 8 B [6], LLaVA LLaVA-OV 0.5B & 7B [18], VILA-1.5 1.5 3B & 8B [24], Qwen2-VL 2B & 7B [40], LongVA 7B [59] and XComposer-8B [57] on NExTQA [46], PerceptionTest [32], TempCompass [26], Video-MME [9], MLVU [64], and LongVideoBench [44]. All evaluations were done using LMMs-eval [56]. Full evaluations of all models on the benchmarks can be seen in Tab. 5 & 6.

Format	ApolloBench					
	OCR	Egocentric	Spatial	Perception	Reasoning	Overall
<vid.token>	50.4	58.5	54.8	58.8	55.4	55.5
<vid.start><vid.token><vid.end>	49.2	61.7	54.8	60.2	57.9	56.7
clip from {MM:SS}-{MM:SS}: <vid.token>	50.0	61.7	54.0	60.8	57.9	<b>56.8</b>
clip from {MM:SS}-{MM:SS}: <vid.start><vid.token><vid.end>	50.0	61.2	54.2	55.7	60.6	56.2

Table 4. **Comparison of Video Token Integration Methods.** Performance of different strategies for integrating video tokens into the text sequence. Incorporating textual timestamps before each clip yields the best overall performance.

### A.3. ApolloBench curation

The creation process of ApolloBench is depicted in Fig. 8. The process begins with a collection of multiple-choice benchmarks. To eliminate the reliance on external tools like ChatGPT, we focus exclusively on multiple-choice questions, ensuring a cost-effective and consistent evaluation process [45].

We first evaluated several Large Multimodal Models (LMMs) with text-only, center-frame, and full-video inputs. Questions that could be answered correctly by more than 50% of the models using either of these modalities were filtered out, as these questions did not require video perception. Next, we categorized the remaining questions into five temporal perception categories: Temporal OCR, Egocentric, Spatial, Perception, and Reasoning. Using entropy, we identified questions with high discrimination power between models and manually verified them to ensure accuracy and quality. From this, we selected the top 400 questions with the highest entropy to form the final ApolloBench dataset. This curated benchmark is

Model	NExT-QA			Perception-Test			TempCompass (CM)			TempCompass (MC)			TempCompass (YN)		
	Video	Image	Text	Video	Image	Text	Video	Image	Text	Video	Image	Text	Video	Image	Text
InternVL2 2B [6]	68.9	61.1	42.8	49.6	46.0	38.6	67.2	63.3	51.9	53.4	47.5	35.9	62.3	59.3	51.2
InternVL2 8B [6]	70.8	72.6	49.1	57.4	52.8	41.3	77.4	66.9	58.3	65.3	54.9	43.7	68.6	62.6	52.1
LLaVA-OV 0.5B [18]	57.3	50.7	31.9	49.1	44.8	40.4	61.9	58.9	51.3	53.2	44.6	34.1	60.0	55.9	49.7
LLaVA-OV 7B [18]	79.3	70.0	48.7	57.1	49.7	41.4	73.8	60.8	56.8	64.9	51.6	41.4	69.8	57.8	53.3
LongVA 7B [59]	50.2	38.9	36.6	50.6	50.3	50.1	60.7	51.1	50.9	56.1	52.2	50.7	62.9	61.6	60.9
Qwen2-VL 2B [40]	68.7	62.1	44.0	53.1	47.5	39.8	70.9	62.5	54.3	60.6	50.4	40.1	63.7	58.6	52.3
Qwen2-VL 7B [40]	78.9	68.5	42.6	58.9	52.6	38.4	76.6	64.3	56.5	67.2	52.3	41.6	71.9	61.8	54.0
VILA-1.5 3B [24]	56.9	56.7	30.1	49.1	49.1	36.2	66.3	66.3	52.9	56.1	56.1	36.8	63.4	63.4	51.1
VILA-1.5 8B [24]	63.1	63.1	38.2	54.7	54.7	41.2	58.7	58.7	33.6	49.0	49.0	18.8	62.5	62.5	50.6
XComposer-8B [57]	71.1	47.3	41.0	55.9	45.3	39.6	72.2	59.3	49.2	61.1	39.4	31.7	64.5	57.8	52.3

Table 5. **Benchmark evaluation for different models across input modalities (1/2).** This table reports the performance of various models on the NExT-QA, Perception-Test, and TempCompass benchmarks with video, image, and text inputs.

Model	LongVideoBench			MLVU			Video-MME (Long)			Video-MME (Medium)			Video-MME (Short)		
	Video	Image	Text	Video	Image	Text	Video	Image	Text	Video	Image	Text	Video	Image	Text
InternVL2 2B [6]	44.8	37.9	32.8	48.2	41.5	32.6	33.1	30.9	31.4	38.2	32.2	28.7	51.3	39.1	32.8
InternVL2 8B [6]	51.8	45.0	40.2	50.8	40.0	37.5	42.0	40.0	38.6	50.6	39.6	38.6	62.1	48.2	39.4
LLaVA-OV 0.5B [18]	46.0	40.5	37.4	50.3	39.2	35.3	37.2	31.3	33.1	40.0	32.0	30.2	54.6	37.1	30.1
LLaVA-OV 7B [18]	56.5	45.1	41.2	65.1	50.3	45.5	49.9	36.9	39.8	54.6	39.4	38.3	70.9	47.4	40.2
LongVA 7B [59]	45.2	44.2	43.0	51.9	45.1	44.1	41.4	38.1	36.7	45.9	39.9	38.4	55.1	45.3	40.0
Qwen2-VL 2B [40]	48.5	40.8	40.4	59.5	45.1	38.4	43.2	36.9	33.3	51.0	35.0	32.3	65.3	40.4	34.8
Qwen2-VL 7B [40]	54.8	44.7	41.5	65.5	49.1	42.4	49.8	40.0	38.4	57.6	41.2	39.2	70.7	46.3	37.6
VILA-1.5 3B [24]	42.9	42.9	33.8	23.3	23.3	13.6	31.6	28.0	28.0	36.7	27.3	27.3	48.7	27.8	27.8
VILA-1.5 8B [24]	47.2	47.2	37.1	44.4	44.4	31.1	39.3	36.6	36.6	42.1	32.3	32.3	56.3	34.3	34.3
XComposer-8B [57]	47.6	30.0	32.0	37.2	8.5	7.3	46.4	28.0	35.1	50.9	26.3	35.0	66.0	28.1	36.1

Table 6. **Benchmark evaluation for different models across input modalities (2/2).** This table reports the performance of various models on the LongVideoBench, MLVU, and Video-MME benchmarks with video, image, and text inputs.

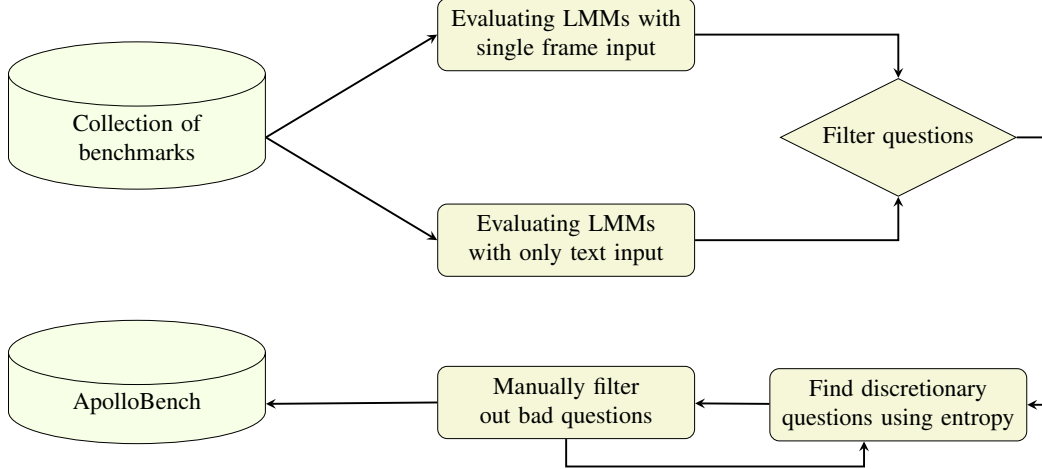


Figure 8. **Flowchart illustrating the curation process of ApolloBench.** Starting with a collection of benchmarks, we evaluate Large Multimodal Models (LMMs) using the full video, single-frame, and text inputs. Questions requiring video perception were filtered based on model performance, and discretionary questions were identified using entropy. After manual verification and categorization into five temporal perception categories, the top 400 questions were selected for the benchmark, and manually inspected.

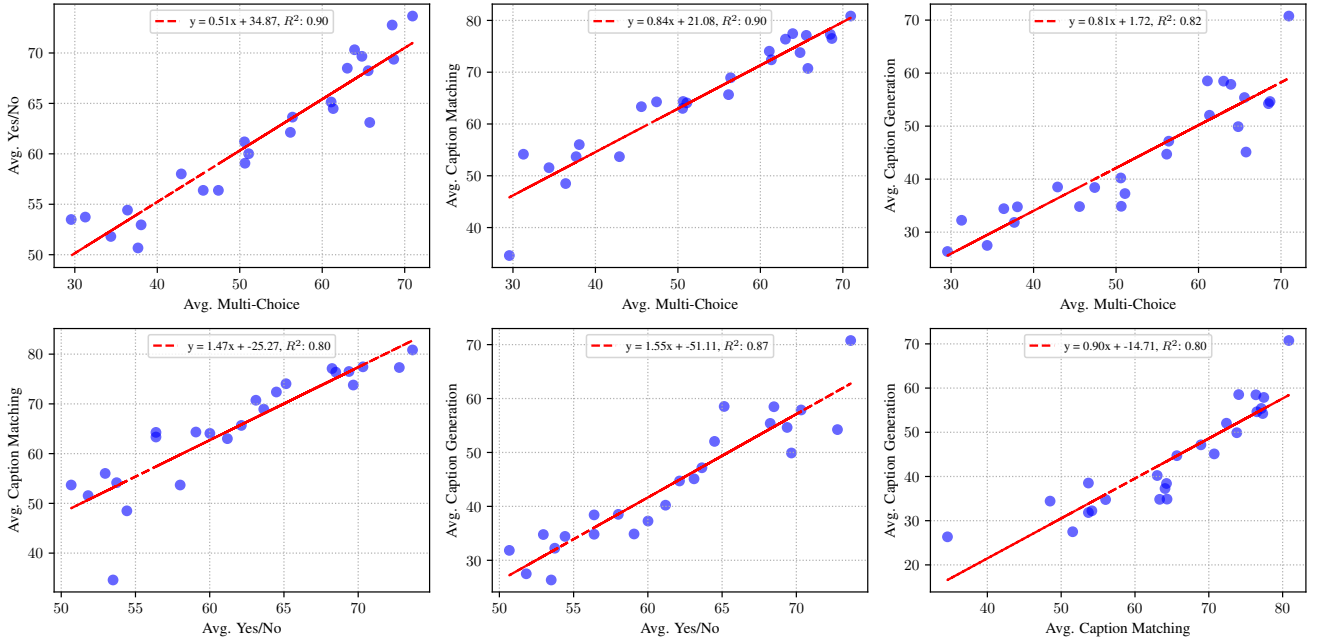


Figure 9. **Effect of question type on model performance.** Correlations between different question types (multiple-choice, yes/no) on the TempCompass benchmark are shown. The high correlation indicates consistency in evaluating model performance across various question formats, indicating that multiple choice is a reasonable option in existing benchmarks.

41× faster to evaluate compared to existing benchmarks while maintaining a high correlation with their results (see Fig. 2, right). Additionally, ApolloBench emphasizes video perception, as shown in Fig. 2, left.

## B. Apollo implementation details

In this section, we provide detailed descriptions of all the design decisions in Apollo, including implementation specifics, hyperparameters, and other relevant details.



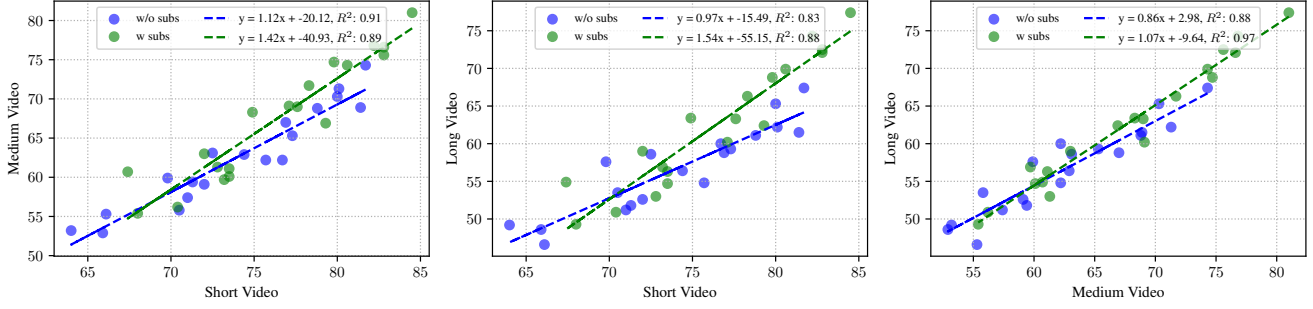


Figure 10. **Correlation between Video-MME duration groups.** The correlations between short, medium, and long video duration groups on the Video-MME benchmark. The analysis highlights how model performance scales with video length, emphasizing the reliance on text and image modalities as video duration increases.

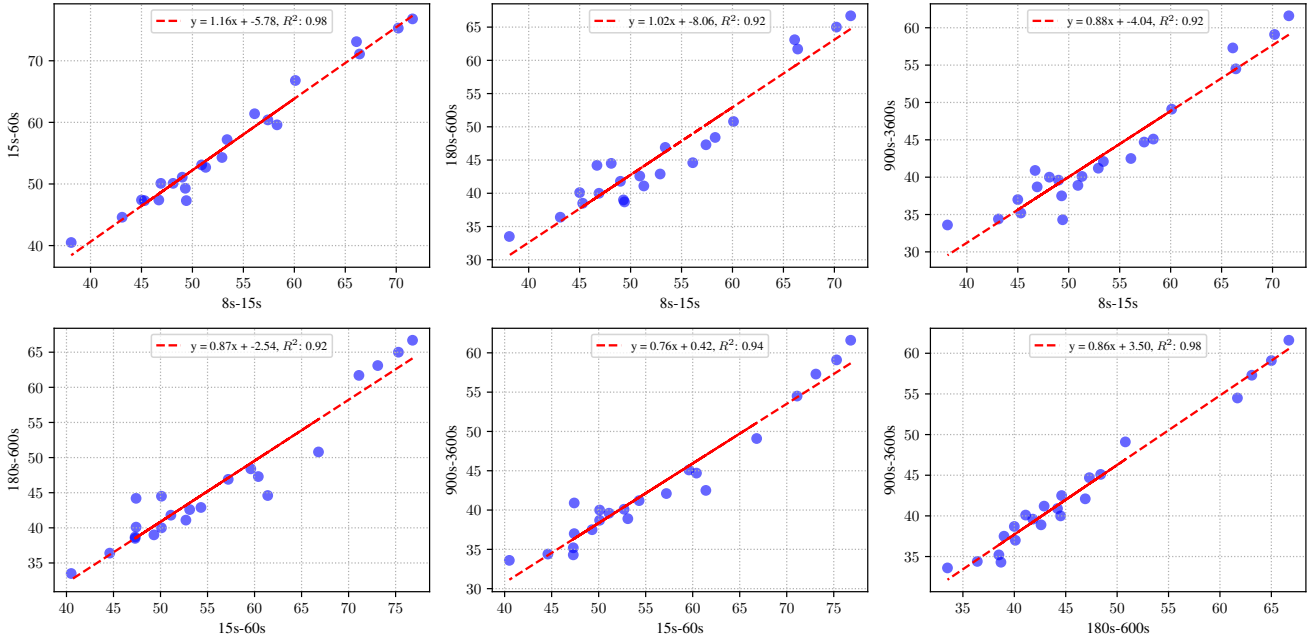


Figure 11. **Correlation between LongVideoBench duration groups.** Correlations between different video duration categories on LongVideoBench are depicted, with  $R^2 > 0.92$  across groups. This consistency suggests that performance trends remain stable across varying video lengths.

## B.1. Architecture

Apollo encodes clips consisting of  $N$  frames, where  $N$  depends on the video encoder used ( $= 4$  for InternVideo2+SigLIP-SO400M). We opted for a fully shared pipeline for both images and videos, so when encoding images, we replicate the image  $N$  times to match the clip length. The frames are then encoded independently with the InternVideo2 and SigLIP-SO400M encoders. The output features are interpolated and concatenated along the channel dimension before being fed into a connector module. The connector projects the features to match the hidden dimension of the Large Language Model, and the resampler resamples them into a predetermined number of  $T$  tokens per clip using the Perceiver Resampler. An overview of Apollo is shown in Fig. 7. For vision-text token integration, we utilize the clip from  $\{\text{MM:SS}\}-\{\text{MM:SS}\}:\langle \text{vid.token} \rangle$  token integration strategy.

Apollo effectively samples videos as a series of independent clips. By keeping the clip sampling frames per second (fps) constant, the model learns to reason about fine-grained temporal aspects, such as the speed of objects. Many previous methods employ uniform frame sampling, especially when handling long videos, effectively changing the “playback speed” between

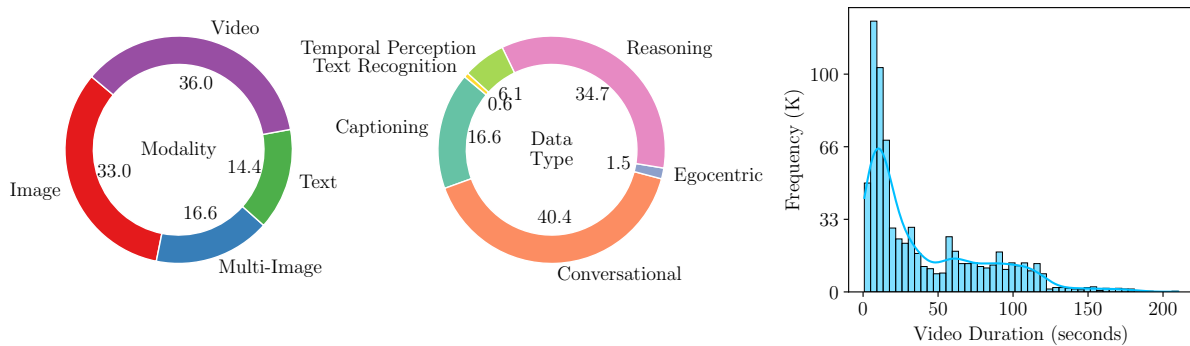


Figure 12. **Data statistics of the fine-tuning dataset.** (Left) Breakdown of data modalities, including text, image, multi-image, and video, illustrating the composition of the fine-tuning dataset. (Middle) Distribution of video annotation types, highlighting the proportions of Conversational, Reasoning, Egocentric, Temporal Perception, Text Recognition, and Captioning annotations. (Right) Histogram of video durations, showing the frequency distribution of video lengths in seconds.

iterations. In contrast, we sample clips uniformly spaced throughout the video, and if the video is too long, we distribute the individual clips uniformly rather than adjusting the frame sampling rate. We, therefore, sample clips concurrently until reaching the maximum number of clips (see Tab. 7), at which point we start uniformly distancing the clips.

## B.2. Data

We utilized a diverse mix of publicly available and licensed datasets across text, image-text, multi-image, and video modalities. Due to licensing restrictions, we excluded non-permissive datasets—such as those leveraging ChatGPT—which limited our inclusion of some commonly used datasets. We generated multi-turn conversations to enrich our training data by leveraging Large Multimodal Models (LMMs), such as Qwen2VL-7B, for captioning. Then, we used LLaMA 3.1 70B [39] to convert these captions into conversations. Detailed data statistics are presented in Fig. 6. It is possible that performance could be further improved without such restrictions and by training on larger datasets like those introduced in LLaVA-OneVision [18] and Cambrian1 [37]. Our training process comprised three distinct stages:

1. **Alignment:** In this phase, we trained on a 50/50 mixture of image and video captions, totaling 198K samples.
2. **Vision Pretraining:** We tuned the encoders using a video-only caption dataset of 396K samples.
3. **Supervised Fine-tuning (SFT):** We trained on a mixture of text, image, multi-image, and video data, with a total of 3.2 million samples.

## B.3. Training

We trained our models using 128 NVIDIA A100 GPUs. Due to the large-scale nature of this study, we automated model training to be spawned from csv files, which would automatically update with the final evaluations. Most experiments were done with ZeRO2 optimization, as full model sharding was unnecessary for our models, but ZeRO3 optimization is supported for researchers interested in training larger models. We utilized the AdamW optimizer for all training stages with a gradient clipping threshold of 1. We applied a warm-up ratio of 0.03 and a cosine learning rate schedule. The training objective was the cross-entropy loss for autoregressive text generation only. We adjusted the learning rates of the Large Language Model (LLM) components proportionally to the square root of their relative model sizes. We found that employing a higher learning rate for the connector module yielded the best performance.

## C. Scaling Consistency: efficient model design with smaller models

Developing Large Multimodal Models (LMMs) with billions of parameters is computationally intensive. A key question is whether smaller models can reliably inform design decisions for larger ones. We introduce Scaling Consistency, a phenomenon where design choices evaluated on moderately sized models (approximately 2–4 billion parameters) correlate highly with those on larger models, enabling efficient model development.

To investigate Scaling Consistency, we conducted extensive experiments varying key aspects of LMM design, such as architecture, video sampling, training strategies, and data mixtures. We selected 21 model variations exploring different

		Align			Vision Pretraining			SFT		
		1.5B	3B	7B	1.5B	3B	7B	1.5B	3B	7B
Sampling	<b>Max clips</b>	25	25	25	25	25	25	200	200	150
	fps	2	2	2	2	2	2	2	2	2
	tps	32	32	32	32	32	32	32	32	32
	tpf	16	16	16	16	16	16	16	16	16
Data	<b>Dataset</b>	A	A	A	VpT	VpT	VpT	SFT	SFT	SFT
	#Samples	198K	198K	198K	396K	396K	396K	3.2M	3.2M	3.2M
	Type	I+V	I+V	I+V	V	V	V	T+I+MI+V	T+I+MI+V	T+I+MI+V
Model	<b>Trainable</b>	38.4M	63.6M	177M	1.4B	1.5B	1.6B	1.6B	3.2B	7.8B
	$\psi_{\text{vision}}$	—	—	—	1.4B	1.4B	1.4B	—	—	—
	$\theta_{\text{connector}}$	38.4M	63.6M	177M	38.4M	63.6M	177M	38.4M	63.6M	177M
	$\phi_{\text{LLM}}$	—	—	—	—	—	—	1.54B	3.09B	7.62B
Training	<b>Batch Size</b>	256	256	256	256	256	256	256	256	256
	<b>LR: <math>\psi_{\text{vision}}</math></b>	0	0	0	$5 \times 10^{-6}$	$5 \times 10^{-6}$	$5 \times 10^{-6}$	0	0	0
	<b>LR: <math>\theta_{\text{connector}}</math></b>	$1 \times 10^{-4}$	$1 \times 10^{-4}$	$1 \times 10^{-4}$	$1 \times 10^{-4}$	$1 \times 10^{-4}$	$1 \times 10^{-4}$	$1 \times 10^{-4}$	$1 \times 10^{-4}$	$1 \times 10^{-4}$
	<b>LR: <math>\phi_{\text{LLM}}</math></b>	0	0	0	0	0	0	$5 \times 10^{-5}$	$2 \times 10^{-5}$	$1 \times 10^{-5}$
	<b>Epoch</b>	1	1	1	1	1	1	1	1	1

Table 7. **Detailed configuration for each training stage of Apollo.** The table summarizes the maximum clips per video, frames per second (fps), dataset information, trainable parameters, and training hyperparameters across different stages of training (**Alignment**, **Vision pretraining**, **SFT**) for Apollo models of varying sizes (1.5B, 3B, and 7.6B).

design dimensions. Each variation was trained using four different Large Language Models (LLMs): Qwen2-0.5B, Qwen2-1.5B, Qwen1.5-4B, and Qwen2-7B, resulting in a total of 84 models.

Unlike traditional scaling laws—which typically require training multiple models from within the same model family to understand how performance scales with size—Scaling Consistency allows us to transfer design insights without such extensive efforts. In scaling laws, researchers train around 3–5 models of different sizes to establish scaling relationships, and only then can they determine which design decisions are beneficial at larger scales. In contrast, Scaling Consistency shows that design decisions on moderately sized models transfer well to larger ones, even across different model families. Our primary goal is to show that design decisions transfer reliably, reducing computational burden and accelerating research.

In Fig. 13, we present all the correlation plots from our study. When comparing the 7B model to smaller ones (first row), we observe that the  $R^2$  correlation progressively increases with model size. A similar pattern is seen when comparing the 4B model to smaller models. For the 1.5B model, however, the  $R^2$  decreases when compared to larger models, and with the 0.5B model, the  $R^2$  is essentially random. We find that the  $R^2$  behaves log-linearly with model size. This suggests that at approximately 3 billion parameters, we expect an  $R^2$  correlation greater than 0.9 when compared with the 7B model. Since the behavior is log-linear, models above the 3–4 billion parameter range can be expected to have high correlation even with much larger models, such as 32B ( $> R^2 \simeq 0.86$ ) or 72B parameters ( $> R^2 \simeq 0.84$ ).

## D. Effect of video sampling on the different dimensions of video perception

Fig. 14 presents a detailed analysis of how varying frames per second (fps) and tokens per second (tps) impact our model’s performance across different video perception tasks: Optical Character Recognition (OCR), Spatial Understanding, Egocentric Understanding, Perception, and Reasoning. Our findings indicate that OCR and Spatial Understanding tasks show consistent performance decline with fewer tokens per frame when tps is reduced, particularly noticeable at lower values of 2–4 tps, regardless of fps settings. This suggests that these tasks are highly sensitive to the amount of visual information encoded per frame, significantly affecting performance by the number of tokens per frame.

In contrast, Egocentric Understanding and Reasoning tasks show a less severe performance drop when tps is reduced, especially at lower fps values. This implies that these tasks are less sensitive to the number of tokens per frame and are more influenced by the temporal resolution provided by fps, with the ability to capture temporal dynamics being more critical than the density of visual information per frame. The Perception metric behaves as an outlier; apart from an anomalous data point at 1 fps, perception performance tends to favor lower fps values and is less affected by variations in tps. This indicates that

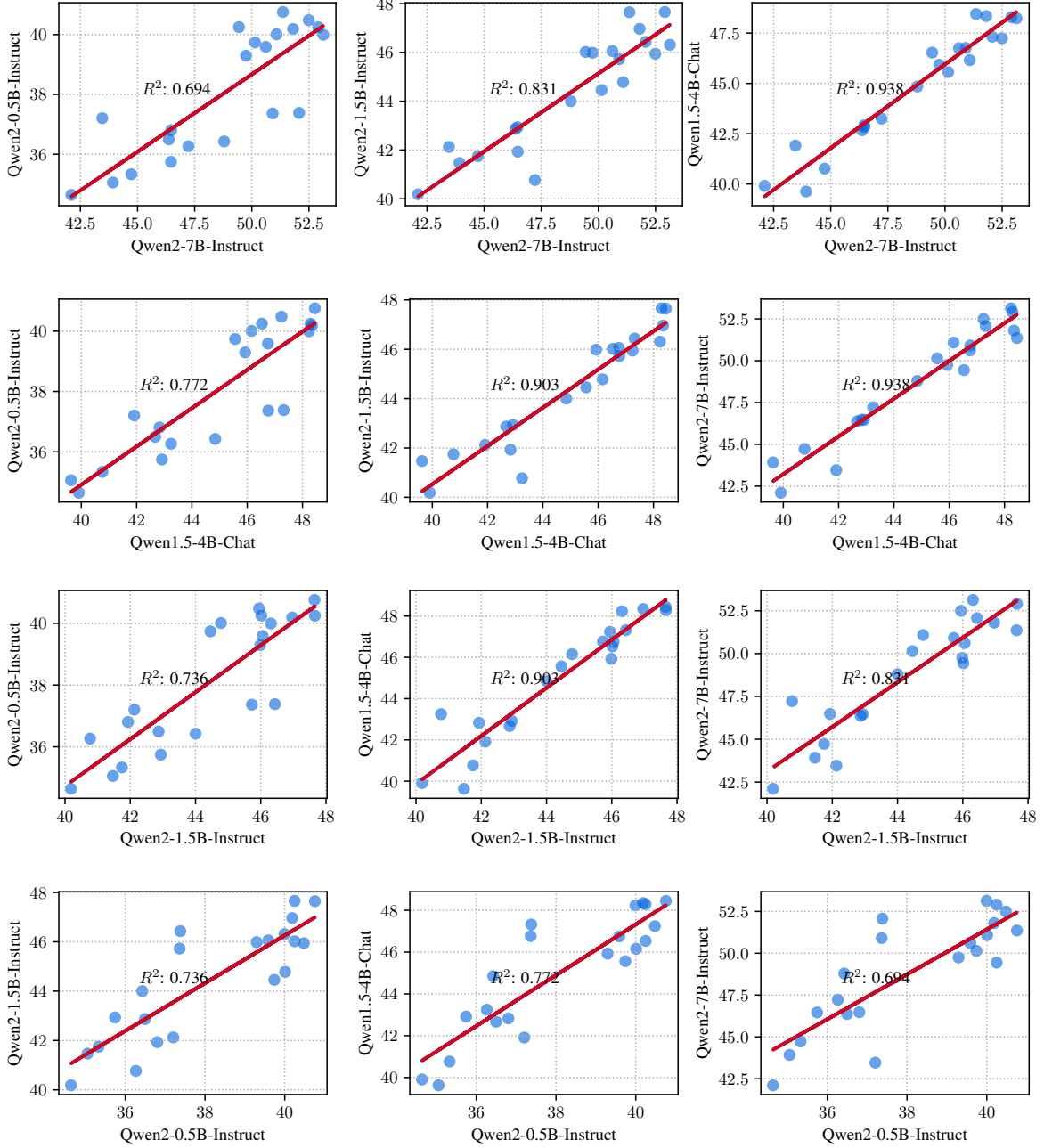


Figure 13. **Scaling Consistency.** Average accuracy for each one for each design variation, we can tell model’s correlation gets progressively better. When comparing two small models (1.5B and 0.5B), we do not see a good correlation, confirming that the Scaling Consistency is not due to the models being of similar size but larger than a certain size.

for specific perceptual tasks, increased temporal sampling does not always improve performance, and effective performance can be achieved with fewer frames and tokens.

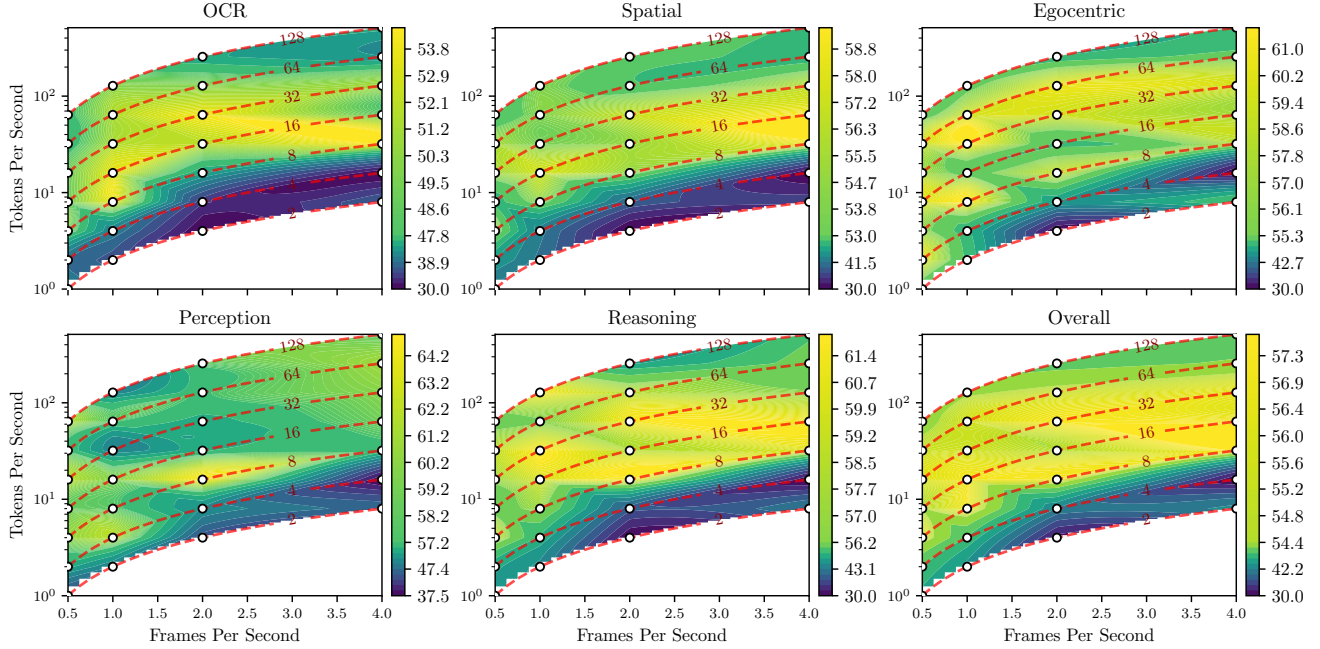


Figure 14. **Video fps sampling analysis.** Full analysis on the effect of frames per second (fps, x-axis), tokens per second (tps, y-axis), and tokens per frame (tpf, dotted red lines) on each of ApolloBench’s dimensions. The number of tokens/frames is highlighted via the dotted red lines.

Hyperparameters			ApolloBench					
LLM	Vision Encoders		OCR	Spatial	Egocentric	Perception	Reasoning	Overall
1	Qwen2.5-3B-Instruct	DINOv2	36.6	40.5	55.9	48.0	46.3	45.5
2	Qwen2.5-3B-Instruct	LanguageBind-Image	41.2	46.2	49.5	51.0	51.7	47.9
3	Qwen2.5-3B-Instruct	SigLIP SO400M	41.9	52.2	57.4	52.0	60.0	52.7
4	Qwen2.5-3B-Instruct	VideoMAE	35.6	35.5	47.9	47.0	40.0	41.2
5	Qwen2.5-3B-Instruct	V-JEPA	39.4	35.2	44.1	52.0	44.6	43.1
6	Qwen2.5-3B-Instruct	LanguageBind-Video	41.2	47.8	54.8	53.2	46.3	48.7
7	Qwen2.5-3B-Instruct	InternVideo2	43.7	46.5	56.4	55.2	58.1	52.0
8	Qwen2.5-3B-Instruct	VideoMAE + DINOv2	40.1	43.2	57.4	59.5	47.5	49.6
9	Qwen2.5-3B-Instruct	VideoMAE + LanguageBind-Image	39.8	49.8	55.9	57.5	49.8	50.5
10	Qwen2.5-3B-Instruct	VideoMAE + SigLIP SO400M	45.8	54.8	55.9	63.0	55.6	55.0
11	Qwen2.5-3B-Instruct	V-JEPA + DINOv2	41.5	43.2	56.4	55.2	48.5	49.0
12	Qwen2.5-3B-Instruct	V-JEPA + LanguageBind-Image	43.3	49.2	50.5	59.2	52.9	51.1
13	Qwen2.5-3B-Instruct	V-JEPA + SigLIP SO400M	48.6	53.2	59.0	57.8	58.1	55.3
14	Qwen2.5-3B-Instruct	LanguageBind-Video + DINOv2	41.5	44.9	54.6	57.6	51.0	50.0
15	Qwen2.5-3B-Instruct	LanguageBind-Video + LanguageBind-Image	41.2	48.5	53.2	62.7	54.7	52.1
16	Qwen2.5-3B-Instruct	LanguageBind-Video + SigLIP SO400M	45.4	50.5	59.6	56.8	54.9	53.4
17	Qwen2.5-3B-Instruct	InternVideo2 + DINOv2	43.0	48.2	50.0	58.0	57.1	51.3
18	Qwen2.5-3B-Instruct	InternVideo2 + LanguageBind-Image	45.8	48.0	51.6	62.3	56.9	52.9
19	Qwen2.5-3B-Instruct	InternVideo2 + SigLIP SO400M	46.8	55.0	60.1	63.2	64.5	57.9

Table 8. **Raw results for vision encoders experiment.** The table presents performance scores on ApolloBench at a tokens-per-second (TPS) rate of 32. Metrics include OCR, spatial understanding, egocentric reasoning, perception, reasoning, and overall performance. The encoders are grouped and ordered as follows: single image encoders, single video encoders, and dual encoder configurations.

## E. Raw results

We provide the raw evaluations of all the models utilized in our study. Many investigations required multiple experiments to test whether design decisions hold under multiple design decisions. We provide all the raw data used in our study for further analysis. For Sec. 4: Tab. 12 & 13, Sec. 5.1: Tab. 9 & 10, Sec. 5.2: Tab 8, Sec. 5.4: Tab. 4, Sec. 6.3: Tab. 11.

Hyperparameters						ApolloBench					
LLM	Vision Encoders	tps	fps	tpf	OCR	Spatial	Egocentric	Perception	Reasoning	Overall	
1	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	512.0	4.0	128.0	46.0	51.0	52.1	59.0	54.0	52.4
2	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	256.0	2.0	128.0	45.5	53.5	51.5	59.0	49.0	51.7
3	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	128.0	1.0	128.0	51.0	55.0	55.3	51.0	62.5	54.9
4	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	64.0	0.5	128.0	48.0	52.0	54.2	63.0	56.0	54.6
5	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	256.0	4.0	64.0	43.5	50.0	55.3	62.0	56.0	53.3
6	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	128.0	2.0	64.0	51.0	52.0	61.6	58.0	59.5	56.4
7	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	64.0	1.0	64.0	52.5	55.0	60.6	58.5	57.0	56.7
8	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	32.0	0.5	64.0	47.5	56.5	60.0	58.0	60.0	56.4
9	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	128.0	4.0	32.0	52.0	57.5	60.6	61.0	57.5	57.7
10	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	64.0	2.0	32.0	55.0	58.0	60.6	55.5	62.5	58.3
11	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	32.0	1.0	32.0	52.5	54.5	62.7	51.0	63.0	56.7
12	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	16.0	0.5	32.0	50.0	56.0	58.4	63.0	58.0	57.1
13	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	64.0	4.0	16.0	49.5	60.5	58.4	60.0	62.5	58.2
14	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	32.0	2.0	16.0	53.0	56.0	53.1	56.0	59.5	55.6
15	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	16.0	1.0	16.0	54.5	58.5	55.3	61.0	61.0	58.1
16	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	8.0	0.5	16.0	50.0	50.5	61.1	59.5	55.5	55.3
17	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	32.0	4.0	8.0	55.5	59.5	59.0	57.5	61.5	58.6
18	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	16.0	2.0	8.0	45.5	55.5	60.0	66.0	62.5	57.9
19	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	8.0	1.0	8.0	54.5	55.0	62.7	59.0	58.0	57.8
20	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	4.0	0.5	8.0	50.5	56.0	57.4	61.5	60.0	57.1
21	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	16.0	4.0	4.0	29.5	25.0	1.0	38.5	12.5	21.5
22	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	8.0	2.0	4.0	35.0	40.5	48.9	52.0	40.0	43.2
23	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	4.0	1.0	4.0	41.5	43.5	52.1	63.0	51.5	50.3
24	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	2.0	0.5	4.0	39.5	47.0	61.6	55.0	50.0	50.5
25	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	8.0	4.0	2.0	38.5	36.5	54.2	47.5	44.5	44.1
26	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	4.0	2.0	2.0	26.5	23.5	30.8	44.9	27.5	32.4
27	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	2.0	1.0	2.0	37.3	41.8	53.1	50.3	45.9	44.8
28	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	1.0	0.5	2.0	41.0	42.0	54.2	45.0	48.2	47.3

Table 9. **Raw results of video sampling experiment.** ApolloBench breaks down metrics to OCR, spatial understanding, egocentric reasoning, perception, reasoning, and overall performance. The table highlights the impact of frames per second (fps), tokens per second (tps), and tokens per frame (tpf).

Hyperparameters					ApolloBench					
LLM		Vision Encoders	Uniform Frames (Train)	Uniform Frames (Test)	OCR	Spatial	Egocentric	Perception	Reasoning	Overall
1	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	8	8	38.0	41.0	43.1	50.3	44.0	44.2
2	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	16	16	40.5	46.7	55.9	55.3	46.1	48.1
3	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	32	32	49.5	52.0	51.1	58.5	48.5	51.9
4	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	64	64	46.5	52.0	61.2	56.5	59.5	55.1
5	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	8	No	42.5	44.5	54.8	52.0	51.5	49.0
6	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	16	No	48.0	43.5	58.5	60.5	53.0	52.6
7	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	32	No	46.0	50.0	52.1	57.5	57.5	52.6
8	Qwen2.5-3B-Instruct	InternVideo2-1B + SigLIP SO400M	64	No	48.5	53.5	59.0	54.5	54.0	53.8

Table 10. **Raw results of uniform sampling experiment.** ApolloBench evaluates metrics including OCR, spatial understanding, egocentric reasoning, perception, reasoning, and overall performance. Top half are results when models are both trained and tested with uniform frame sampling. The bottom half is when the models are trained with uniform frame sampling but tested at an fps of 2.



	Data Compisiton				ApolloBench					
	Text	Image	Multi-Image	Video	OCR	Spatial	Egocentric	Perception	Reasoning	Overall
1	25.0	25.0	25.0	25.0	41.0	49.5	59.0	57.0	59.5	53.1
2	15.0	25.0	20.0	40.0	47.5	59.0	60.6	66.0	62.0	59.0
3	15.0	32.5	20.0	32.5	45.0	57.0	52.1	60.0	61.5	55.2
4	15.0	40.0	20.0	25.0	46.5	52.0	58.0	65.5	63.5	57.1
5	7.0	38.7	20.0	34.3	44.5	53.0	54.3	58.0	55.5	53.0
6	7.0	55.0	20.0	18.0	45.5	51.0	52.7	62.0	60.0	54.3
7	7.0	0.0	0.0	93.0	37.5	33.5	52.7	40.5	45.5	41.8
8	7.0	0.0	20.0	73.0	37.0	44.0	51.1	45.0	49.0	45.1
9	5.0	20.0	20.0	55.0	51.0	56.0	53.7	60.5	62.5	57.2
10	5.0	10.0	40.0	45.0	37.0	44.5	50.5	54.0	48.5	46.9
11	2.0	30.0	30.0	38.0	30.5	43.0	50.0	51.0	44.5	43.7
12	0.0	38.7	20.0	41.3	33.0	41.0	48.9	50.5	46.0	43.8

Table 11. **Raw results of data composition experiments.** Performance outcomes of video-based Large Multimodal Models (LMMs) trained with varying proportions of Text, Image, Multi-Image, and Video data mixtures. The table presents benchmark scores across OCR, Spatial, Egocentric, Perception, Reasoning, and Overall performance metrics for each distinct data composition. These results emphasize the critical role of balanced data mixtures in optimizing model performance (see Sec. 6.3 for details).













































	LLM	Vision Towers	Vision Freeze	Clip Duration	Tokens /Clip	fps	tps	Tokens /Frame	Data Mixture	Average
1	Qwen2-7B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M		5	32	1.6	6.4	4	A	46.37
2	Qwen2-7B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M		5	64	1.6	12.8	8	A	46.46
3	Qwen2-7B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M		5	64	1.6	12.8	8	A	48.79
4	Qwen2-7B-Instruct	VJEPa-H@384 + SigLIP SO400M		5	32	3.2	6.4	2	A	47.22
5	Qwen2-7B-Instruct	VJEPa-H@384 + SigLIP SO400M		10	64	1.6	6.4	4	A	43.46
6	Qwen2-7B-Instruct	VJEPa-H@384 + SigLIP SO400M		5	64	3.2	12.8	2	A	46.47
7	Qwen2-7B-Instruct	VJEPa-H@384 + SigLIP SO400M		5	64	3.2	12.8	2	A	42.11
8	Qwen2-7B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M		5	32	1.6	6.4	4	B	49.75
9	Qwen2-7B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M		5	64	1.6	12.8	8	B	50.61
10	Qwen2-7B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M		5	64	1.6	12.8	8	B	50.91
11	Qwen2-7B-Instruct	VJEPa-H@384 + SigLIP SO400M		5	32	3.2	6.4	2	B	49.44
12	Qwen2-7B-Instruct	VJEPa-H@384 + SigLIP SO400M		10	64	1.6	6.4	4	B	50.14
13	Qwen2-7B-Instruct	VJEPa-H@384 + SigLIP SO400M		5	64	3.2	12.8	2	B	51.08
14	Qwen2-7B-Instruct	VJEPa-H@384 + SigLIP SO400M		5	64	3.2	12.8	2	B	43.92
15	Qwen2-7B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M		5	32	1.6	6.4	4	C	51.80
16	Qwen2-7B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M		5	64	1.6	12.8	8	C	52.91
17	Qwen2-7B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M		5	64	1.6	12.8	8	C	52.07
18	Qwen2-7B-Instruct	VJEPa-H@384 + SigLIP SO400M		5	32	3.2	6.4	2	C	51.36
19	Qwen2-7B-Instruct	VJEPa-H@384 + SigLIP SO400M		10	64	1.6	6.4	4	C	52.49
20	Qwen2-7B-Instruct	VJEPa-H@384 + SigLIP SO400M		5	64	3.2	12.8	2	C	53.13
21	Qwen2-7B-Instruct	VJEPa-H@384 + SigLIP SO400M		5	64	3.2	12.8	2	C	44.73
22	Qwen1.5-4B-Chat	LanguageBind-Video-v1.5 + SigLIP SO400M		5	32	1.6	6.4	4	A	42.67
23	Qwen1.5-4B-Chat	LanguageBind-Video-v1.5 + SigLIP SO400M		5	64	1.6	12.8	8	A	42.92
24	Qwen1.5-4B-Chat	LanguageBind-Video-v1.5 + SigLIP SO400M		5	64	1.6	12.8	8	A	44.85
25	Qwen1.5-4B-Chat	VJEPa-H@384 + SigLIP SO400M		5	32	3.2	6.4	2	A	43.25
26	Qwen1.5-4B-Chat	VJEPa-H@384 + SigLIP SO400M		10	64	1.6	6.4	4	A	41.91
27	Qwen1.5-4B-Chat	VJEPa-H@384 + SigLIP SO400M		5	64	3.2	12.8	2	A	42.83
28	Qwen1.5-4B-Chat	VJEPa-H@384 + SigLIP SO400M		5	64	3.2	12.8	2	A	39.91
29	Qwen1.5-4B-Chat	LanguageBind-Video-v1.5 + SigLIP SO400M		5	32	1.6	6.4	4	B	45.90
30	Qwen1.5-4B-Chat	LanguageBind-Video-v1.5 + SigLIP SO400M		5	64	1.6	12.8	8	B	46.75
31	Qwen1.5-4B-Chat	LanguageBind-Video-v1.5 + SigLIP SO400M		5	64	1.6	12.8	8	B	46.76
32	Qwen1.5-4B-Chat	VJEPa-H@384 + SigLIP SO400M		5	32	3.2	6.4	2	B	46.53
33	Qwen1.5-4B-Chat	VJEPa-H@384 + SigLIP SO400M		10	64	1.6	6.4	4	B	45.56
34	Qwen1.5-4B-Chat	VJEPa-H@384 + SigLIP SO400M		5	64	3.2	12.8	2	B	46.16
35	Qwen1.5-4B-Chat	VJEPa-H@384 + SigLIP SO400M		5	64	3.2	12.8	2	B	39.63
36	Qwen1.5-4B-Chat	LanguageBind-Video-v1.5 + SigLIP SO400M		5	32	1.6	6.4	4	C	48.34
37	Qwen1.5-4B-Chat	LanguageBind-Video-v1.5 + SigLIP SO400M		5	64	1.6	12.8	8	C	48.29
38	Qwen1.5-4B-Chat	LanguageBind-Video-v1.5 + SigLIP SO400M		5	64	1.6	12.8	8	C	47.32
39	Qwen1.5-4B-Chat	VJEPa-H@384 + SigLIP SO400M		5	32	3.2	6.4	2	C	48.45
40	Qwen1.5-4B-Chat	VJEPa-H@384 + SigLIP SO400M		10	64	1.6	6.4	4	C	47.24
41	Qwen1.5-4B-Chat	VJEPa-H@384 + SigLIP SO400M		5	64	3.2	12.8	2	C	48.24
42	Qwen1.5-4B-Chat	VJEPa-H@384 + SigLIP SO400M		5	64	3.2	12.8	2	C	40.76

Table 12. **Raw results of Scaling Consistency experiments (1/2).** This table presents the raw performance data of 42 model configurations used in the Scaling Consistency experiments. Each configuration explores the effect of various parameters, including the LLM size (Qwen variants), vision tower configurations, freezing or training vision encoders, clip duration, tokens per clip, frames per second (fps), tokens per second (tps), tokens per frame, and data mixture. The “Average” column reports the overall performance score. These results support the investigation into how smaller models can serve as proxies for larger models in determining effective design decisions.

	LLM	Vision Towers	Vision Freeze	Clip Duration	Tokens /Clip	fps	tps	Tokens /Frame	Data Mixture	Average
43	Qwen2-1.5B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M	❄️	5	32	1.6	6.4	4	A	42.87
44	Qwen2-1.5B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M	❄️	5	64	1.6	12.8	8	A	42.94
45	Qwen2-1.5B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M	🔥	5	64	1.6	12.8	8	A	44.00
46	Qwen2-1.5B-Instruct	VJEPa-H@384 + SigLIP SO400M	❄️	5	32	3.2	6.4	2	A	40.77
47	Qwen2-1.5B-Instruct	VJEPa-H@384 + SigLIP SO400M	❄️	10	64	1.6	6.4	4	A	42.13
48	Qwen2-1.5B-Instruct	VJEPa-H@384 + SigLIP SO400M	❄️	5	64	3.2	12.8	2	A	41.93
49	Qwen2-1.5B-Instruct	VJEPa-H@384 + SigLIP SO400M	🔥	5	64	3.2	12.8	2	A	40.18
50	Qwen2-1.5B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M	❄️	5	32	1.6	6.4	4	B	45.98
51	Qwen2-1.5B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M	❄️	5	64	1.6	12.8	8	B	46.06
52	Qwen2-1.5B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M	🔥	5	64	1.6	12.8	8	B	45.73
53	Qwen2-1.5B-Instruct	VJEPa-H@384 + SigLIP SO400M	❄️	5	32	3.2	6.4	2	B	46.02
54	Qwen2-1.5B-Instruct	VJEPa-H@384 + SigLIP SO400M	❄️	10	64	1.6	6.4	4	B	44.46
55	Qwen2-1.5B-Instruct	VJEPa-H@384 + SigLIP SO400M	❄️	5	64	3.2	12.8	2	B	44.78
56	Qwen2-1.5B-Instruct	VJEPa-H@384 + SigLIP SO400M	🔥	5	64	3.2	12.8	2	B	41.47
57	Qwen2-1.5B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M	❄️	5	32	1.6	6.4	4	C	46.96
58	Qwen2-1.5B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M	❄️	5	64	1.6	12.8	8	C	47.66
59	Qwen2-1.5B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M	🔥	5	64	1.6	12.8	8	C	46.43
60	Qwen2-1.5B-Instruct	VJEPa-H@384 + SigLIP SO400M	❄️	5	32	3.2	6.4	2	C	47.65
61	Qwen2-1.5B-Instruct	VJEPa-H@384 + SigLIP SO400M	❄️	10	64	1.6	6.4	4	C	45.94
62	Qwen2-1.5B-Instruct	VJEPa-H@384 + SigLIP SO400M	❄️	5	64	3.2	12.8	2	C	46.31
63	Qwen2-1.5B-Instruct	VJEPa-H@384 + SigLIP SO400M	🔥	5	64	3.2	12.8	2	C	41.76
64	Qwen2-0.5B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M	❄️	5	32	1.6	6.4	4	A	36.50
65	Qwen2-0.5B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M	❄️	5	64	1.6	12.8	8	A	35.75
66	Qwen2-0.5B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M	🔥	5	64	1.6	12.8	8	A	36.43
67	Qwen2-0.5B-Instruct	VJEPa-H@384 + SigLIP SO400M	❄️	5	32	3.2	6.4	2	A	36.27
68	Qwen2-0.5B-Instruct	VJEPa-H@384 + SigLIP SO400M	❄️	10	64	1.6	6.4	4	A	37.21
69	Qwen2-0.5B-Instruct	VJEPa-H@384 + SigLIP SO400M	❄️	5	64	3.2	12.8	2	A	36.80
70	Qwen2-0.5B-Instruct	VJEPa-H@384 + SigLIP SO400M	🔥	5	64	3.2	12.8	2	A	34.64
71	Qwen2-0.5B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M	❄️	5	32	1.6	6.4	4	B	39.29
72	Qwen2-0.5B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M	❄️	5	64	1.6	12.8	8	B	39.59
73	Qwen2-0.5B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M	🔥	5	64	1.6	12.8	8	B	37.36
74	Qwen2-0.5B-Instruct	VJEPa-H@384 + SigLIP SO400M	❄️	5	32	3.2	6.4	2	B	40.25
75	Qwen2-0.5B-Instruct	VJEPa-H@384 + SigLIP SO400M	❄️	10	64	1.6	6.4	4	B	39.74
76	Qwen2-0.5B-Instruct	VJEPa-H@384 + SigLIP SO400M	❄️	5	64	3.2	12.8	2	B	40.01
77	Qwen2-0.5B-Instruct	VJEPa-H@384 + SigLIP SO400M	🔥	5	64	3.2	12.8	2	B	35.05
78	Qwen2-0.5B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M	❄️	5	32	1.6	6.4	4	C	40.19
79	Qwen2-0.5B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M	❄️	5	64	1.6	12.8	8	C	40.25
80	Qwen2-0.5B-Instruct	LanguageBind-Video-v1.5 + SigLIP SO400M	🔥	5	64	1.6	12.8	8	C	37.38
81	Qwen2-0.5B-Instruct	VJEPa-H@384 + SigLIP SO400M	❄️	5	32	3.2	6.4	2	C	40.76
82	Qwen2-0.5B-Instruct	VJEPa-H@384 + SigLIP SO400M	❄️	10	64	1.6	6.4	4	C	40.48
83	Qwen2-0.5B-Instruct	VJEPa-H@384 + SigLIP SO400M	❄️	5	64	3.2	12.8	2	C	39.99
84	Qwen2-0.5B-Instruct	VJEPa-H@384 + SigLIP SO400M	🔥	5	64	3.2	12.8	2	C	35.33

Table 13. **Raw results of Scaling Consistency experiments (2/2).** This table presents the raw performance data of 42 model configurations used in the Scaling Consistency experiments. Each configuration explores the effect of various parameters, including the LLM size (Qwen variants), vision tower configurations, freezing or training vision encoders, clip duration, tokens per clip, frames per second (fps), tokens per second (tps), tokens per frame, and data mixture. The “Average” column reports the overall performance score. These results support the investigation into how smaller models can serve as proxies for larger models in determining effective design decisions.