Appendix

This section contains supplemental material, offering further results and analysis to complement the main paper. We provide additional details on the following topics:

- Detailed Hyperparameters (Appendix A)
- Additional Ablations (Appendix B)
- Additional Dataset Details (Appendix C)
- Evaluation on Long Videos (Appendix D)
- Additional Qualitative Ablations (Appendix E)
- EVE Baseline for Videos (Appendix F)
- Broader Impact (Appendix G)

A. Detailed Hyperparameters

In Table 6 we provide comprehensive hyperparameter configurations for Video-Panda's three-stage training process.

Hyperparameter	Stage-1	Stage-2	Stage-3
Batch Size	2048	2048	1024
Learning Rate (lr)	4e-4	4e-5	2e-5
LR Schedule	cos. decay	cos. decay	cos. decay
LR Warmup Ratio	0.03	0.01	0.01
Weight Decay	0	0	0
Epoch	1	1	1
Optimizer	AdamW	AdamW	AdamW
DeepSpeed Stage	2	2	2
LLM	Frozen	Trainable	Trainable
STAB	Trainable	Trainable	Trainable

Table 6. Hyperparameter Settings

B. Additional Ablations

Training Data for Initial Alignment: Table 7 shows the impact of data scale during initial alignment. Using the full dataset (702K samples) in Stage 1 yields marginally lower performance compared to using half (351K samples). This suggests our staged training approach benefits from gradual complexity scaling, allowing the model to establish robust representations before incorporating the complete dataset in later stages.

Downsampling Position: Regarding temporal layer placement in Table 8, we find that applying LSD after LSTE improves performance on all datasets except of ActivityNet-QA. For consistency across our experiments, we maintain LSD placement after LSTE.

Downsampling Strategy: As shown in Table 9, our LSD method outperforms alternative approaches. The Perceiver Resampler (PR) performs notably poorly (21.3% lower on MSVD-QA), likely due to excessive information compression. While average pooling performs better, it still underperforms LSD by 6.7%, demonstrating the superiority of our learnable downsampling approach.

#Samples for Initial Alignment	MSVD-QA	Activity Net-QA
702K Video-Text Pairs (full)	63.7/3.8	39.7/3.3
351K Video-Text Pairs (half)	**64.7/3.8**	**40.0/3.3**

Table 7. Ablation study on amount of data for the first training stage.

Model	MSVD-QA	MSRVTT-QA	TGIF-QA	Activity Net-QA
Before LSTE	64.2/3.8	54.6/3.4	42.7/3.2	**42.3/3.3**
After LSTE (Ours)	**64.7/3.8**	**54.8/3.4**	**42.9/3.2**	40.0/3.3

Table 8. Ablation study on downsampling positions of LSD.

Model	MSVD-QA	Activity Net-QA
w/o LSD (half-resolution)	48.2/3.3	38.5/3.2
w/o LSD (avg pool)	58.0/3.6	38.1/3.2
w/o LSD (PR)	43.4/3.2	27.8/2.9
Video-Panda (LSD)	**64.7/3.8**	**40.0/3.3**

Table 9. Ablation study on downsampling methods. PR stands for Perceiver Resampler [2].

Model	MSVD-QA	Activity Net-QA
CLIP	60.3/3.5	38.6/3.2
InternVideov2	62.5/3.6	39.6/3.2
DINOv2	61.7/3.5	38.1/3.2
LanguageBind (Video-Panda)	**64.7/3.8**	**40.0/3.3**

Table 10. Ablation study on different teacher encoders.

Different Teachers: As shown in Table 10, LanguageBind consistently outperforms other teacher encoders across both datasets. While InternVideo achieves the second-best performance, it still falls short by 2.2% on MSVD-QA and 0.4% on Activity Net-QA. CLIP and DINOv2 show comparable performance to each other but lag behind LanguageBind by 3-4%, demonstrating the effectiveness of our chosen teacher encoder.

C. Additional Dataset Details

Pre-training Dataset: The Valley-Pretrain-702K dataset is a large-scale pre-training dataset designed for video-language understanding tasks. It comprises 702K video-text pairs from the WebVid dataset [4], filtered by [27] using methods established by LLaVA [24] to optimize the balance between conceptual diversity and training efficiency. The dataset is structured as single-round dialogues, where each video is paired with questions about its content and corresponding caption-based answers.

Fine-tuning Dataset: The Video-ChatGPT-100K dataset was developed for fine-tuning video-language models, comprising 100K video instruction samples collected

Model	Vision Size (M)	EgoSchema	VideoMME-M
Video-ChatGPT	307	34.2	36.0
Video-LLaVA	425	36.1	**38.1**
Video-Panda	45	**36.4**	37.9

Table 11. Results on the EgoSchema and VideoMME-M datasets.

by [28]. The dataset combines human expertise with semi-automated methods to balance quality and scalability. Expert annotators provide detailed, context-rich descriptions that enhance the model's comprehension of complex video content. A semi-automatic framework leverages state-of-the-art vision-language models to generate large-scale annotations efficiently, ensuring substantial data volume while maintaining rigorous quality standards.

Fine-Grained Video QA Evaluation Dataset: We evaluate fine-grained video question answering using the Video-based Text Generation Performance Benchmarking methodology developed by Video-ChatGPT [28]. This benchmark provides a comprehensive evaluation framework for assessing text generation in video-based conversational models. Using the ActivityNet-200 dataset [8], which contains videos with descriptive captions and human-annotated question-answer pairs, the framework implements a systematic evaluation approach. The methodology utilizes GPT-3.5 to evaluate models across multiple dimensions on a scale of 1 to 5. The assessment criteria include:

(i) *Correctness of Information:* Evaluates accuracy of generated text and its alignment with video content.
(ii) *Detail Orientation:* Assesses response comprehensiveness, examining both coverage of major points and specificity of details.
(iii) *Contextual Understanding:* Measures the model's ability to interpret and respond within the video's broader context.
(iv) *Temporal Understanding:* Evaluates the model's capacity to track and articulate the chronological sequence of events.
(v) *Consistency:* Assesses the model's ability to maintain coherent responses across different questions and video segments.

D. Evaluation on Long Videos

To explore the potential of Video-Panda on long video benchmarks, we have evaluated our method on the EgoSchema [30] and Video-MME-M [13] datasets. The results presented in Table 11 confirm the results presented in the paper, i.e., we achieve similar or slightly better accuracy compared to Video-ChatGPT and Video-LLaVA, but require much less computational resources (Table 3).

E. Additional Qualitative Ablations

We present additional qualitative examples of our ablation studies in Figure 5, demonstrating Video-Panda's effectiveness across various video understanding tasks. When using the complete training dataset in Stage 1 (left-top example), the model exhibits overfitting tendencies due to data imbalance, as evidenced by the example showing dog interactions—likely influenced by the disparity between dog (7,807) and cat (5,050) instances in the Valley dataset. The right-top example reveals that placing the LSD module before LSTE impairs cliff recognition due to early token downsampling and information loss. Models using alternative approaches (average pooling, half resolution, or perceiver resampler) struggle with content recognition (e.g., cucumber, cat, pandas) compared to our learnable downsampling approach. Additionally, models using image-based teachers (CLIP and DINOv2) tend to make frame-specific predictions rather than considering global context, as demonstrated by their failure to recognize shredded potatoes across multiple frames. We also provide additional qualitative examples on each dataset in Figure 6, Figure 7, and Figure 8.

F. EVE Baseline for Videos

As the original EVE model [11] was designed for image processing, we conducted a fair comparison by re-training it (denoted as *EVE** in Table 1) using identical video data (Valley-702K and Video-ChatGPT-100K). For processing videos, each frame was treated independently as a separate image, with CLIP-ViT-L/14 [35] serving as the teacher model for distillation. While this approach enables frame-level analysis, it neglects temporal relationships. In our implementation, we employ Learnable Selective Downsampling (LSD) to process video frames efficiently, reducing each frame to a consistent token count while preserving essential information. The resulting tokens are flattened into a single sequence, with special split tokens inserted between frame representations to maintain frame boundaries and enable temporal relationship learning.

G. Broader Impact

We introduce Video-Panda, an encoder-free Video Language Model for video understanding. Our model addresses key practical challenges in large-scale AI deployment. While many VLMs raise concerns about data bias, privacy, and computational costs, Video-Panda mitigates these issues through two key design choices: training exclusively on publicly available datasets and eliminating the need for a pretrained encoder. This approach not only reduces ethical concerns but also significantly lowers computational requirements and deployment costs, making the model more accessible and environmentally sustainable.

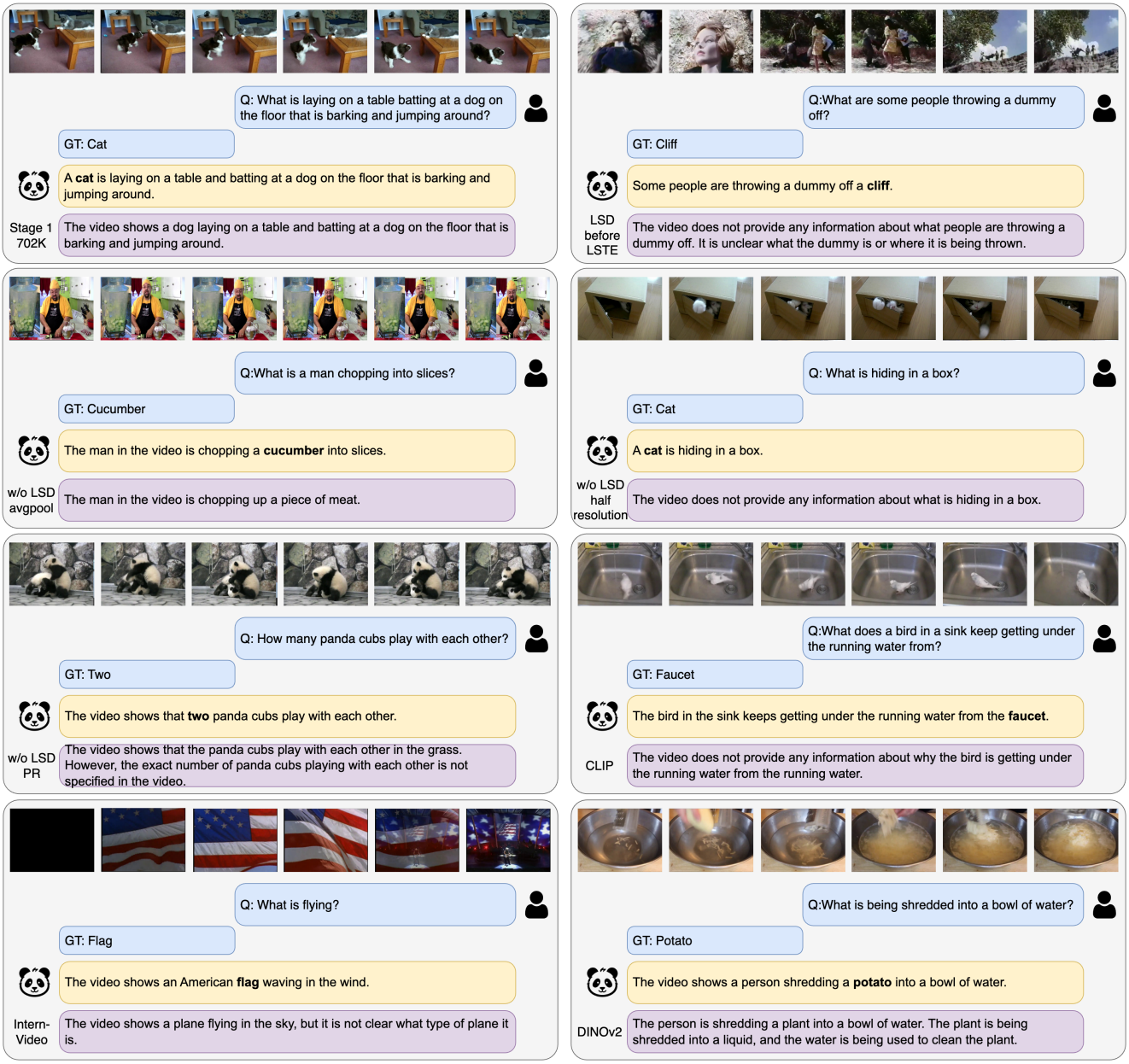

Figure 5. **Qualitative comparisons of different design choices of Video-Panda:** The figure presents eight video examples with ground truth (GT) annotations and model predictions under different training configurations. The *top row* demonstrates the effect of 702K training samples in stage 1 (*left*) and the impact of performing Local Spatial Downsampling (LSD) before Local Spatial-Temporal Encoding (LSTE) (*right*). The *second row* shows results from removing LSD while using average pooling (*left*), half-resolution (*right*), and perceiver resampler (*third row left*). The *third row right* and *fourth row* illustrate the effects of different teacher models for knowledge distillation: CLIP (*third row right*), Intern-Video (*left*), and DINOv2 (*right*). Each example includes the original model prediction (yellow) and an ablated version (purple), highlighting how architectural and training choices affect Video-Panda's ability to interpret dynamic visual scenes and answer questions. The qualitative examples are from the MSVD-QA dataset.

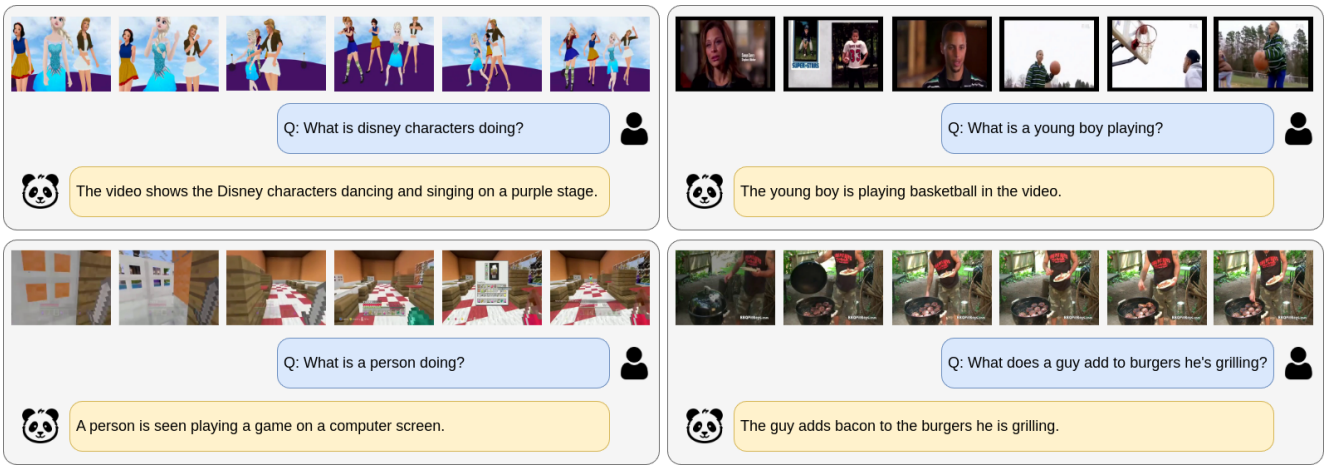

Figure 6. Qualitative examples from the MSRVTT-QA dataset.

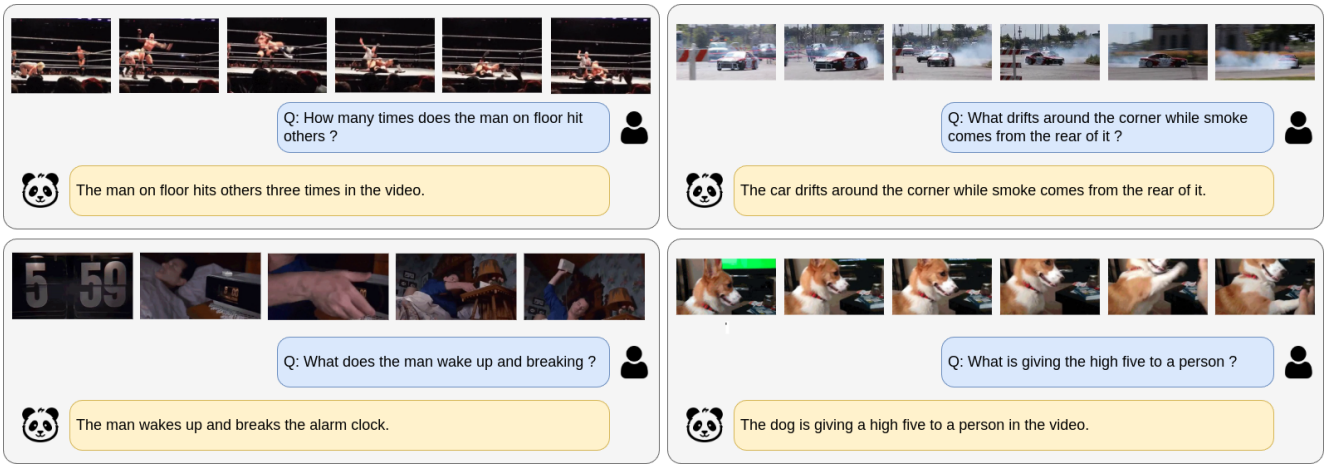

Figure 7. Qualitative examples from the TGIF-QA dataset.

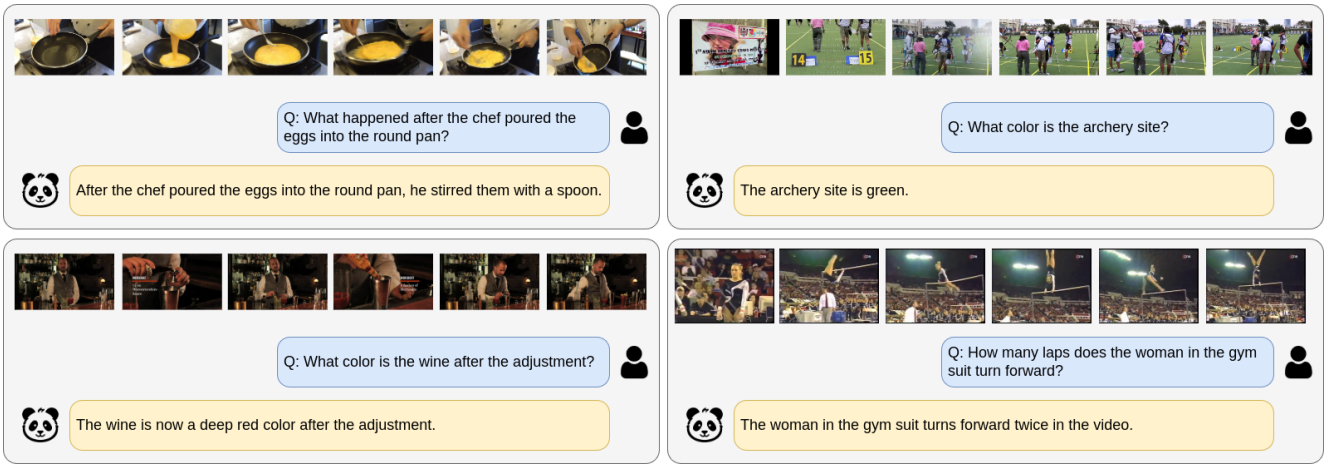

Figure 8. Qualitative examples from the ActivityNet-QA dataset.