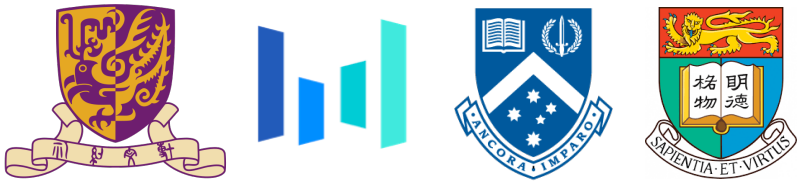
ICML
International Conference
On Machine Learning

OmniShow: Unifying Multimodal Conditions for Human-Object Interaction Video Generation

Donghao Zhou[1,*], Guisheng Liu[2,*], Hao Yang[2], Jiatong Li[2,†], Jingyu Lin[3], Xiaohu Huang[4], Yichen Liu[2],
Xin Gao[2], Cunjian Chen[3], Shilei Wen[2,§], Chi-Wing Fu[1], Pheng-Ann Heng[1,§]

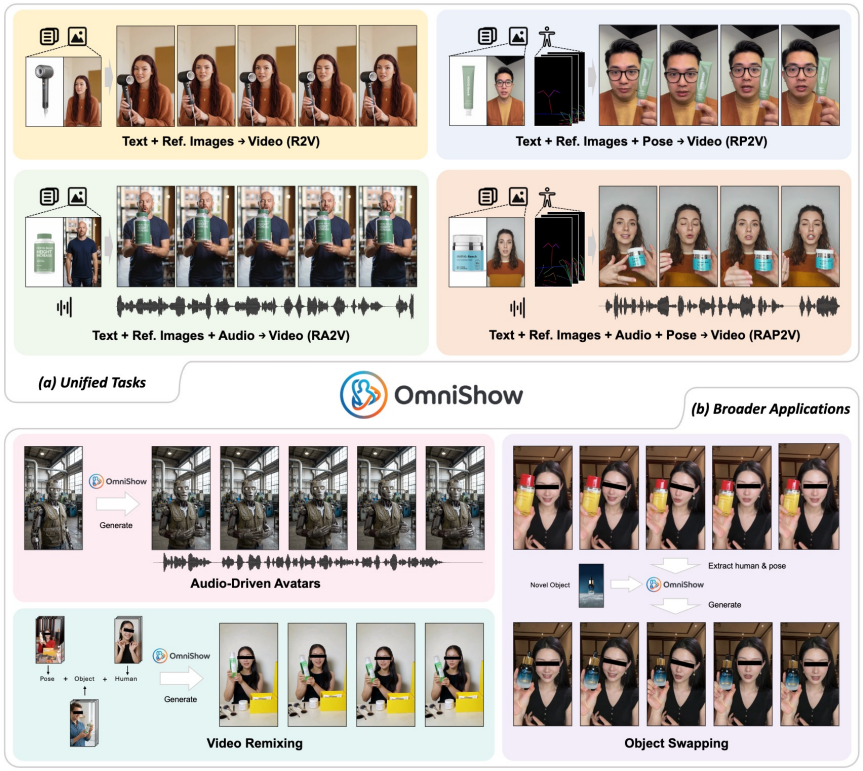
[1]The Chinese University of Hong Kong, [2]ByteDance, [3]Monash University, [4]The University of Hong Kong

([*]Equal Contribution, [†]Project Lead, [§]Corresponding Author)



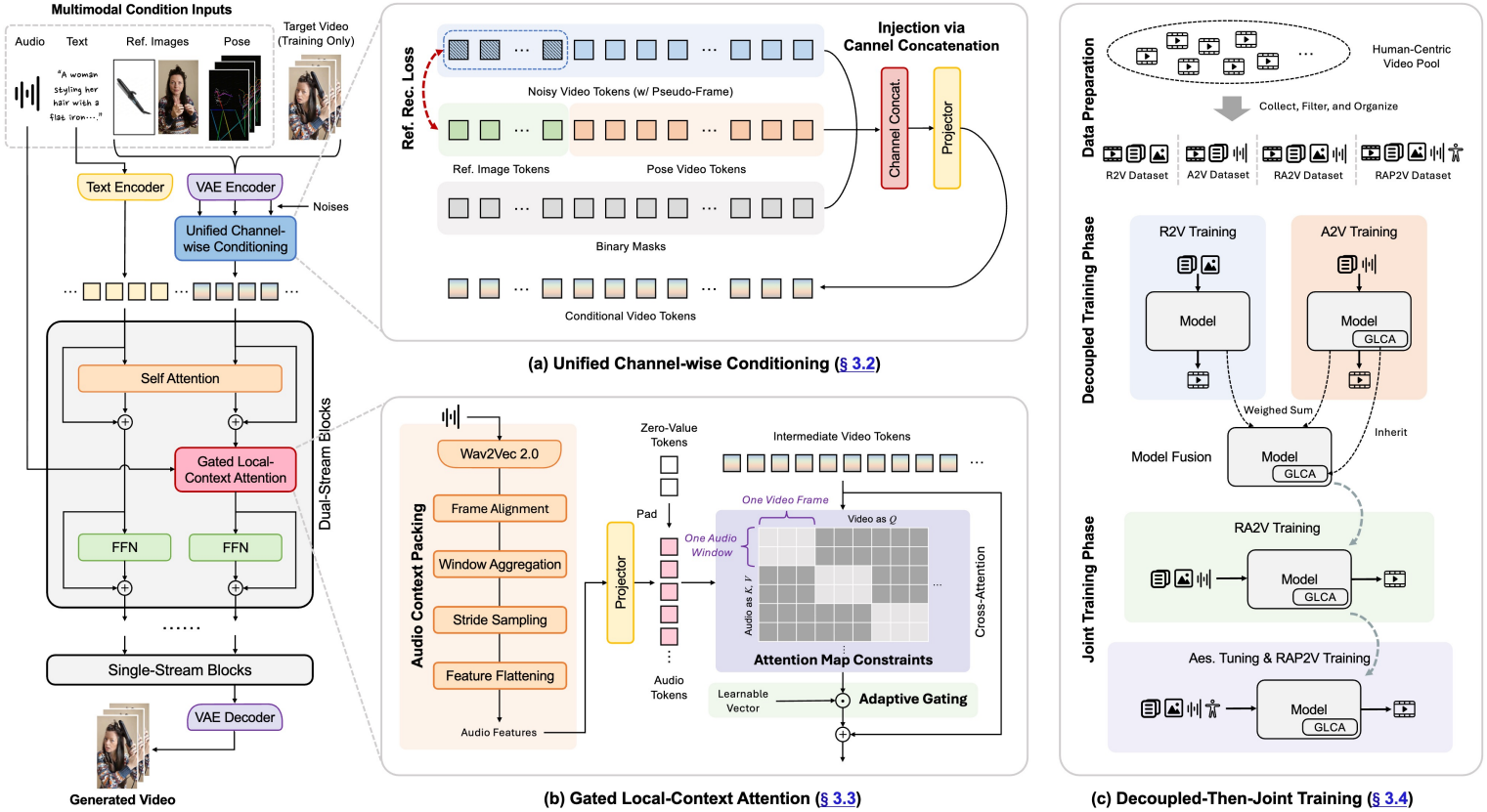
Overview

We introduce **OmniShow**, the **first all-in-one-model** that unifies **text**, **reference image**, **audio**, and **pose** conditions for **Human-Object Interaction Video Generation (HOIVG)**, which also supports broader real-world applications.



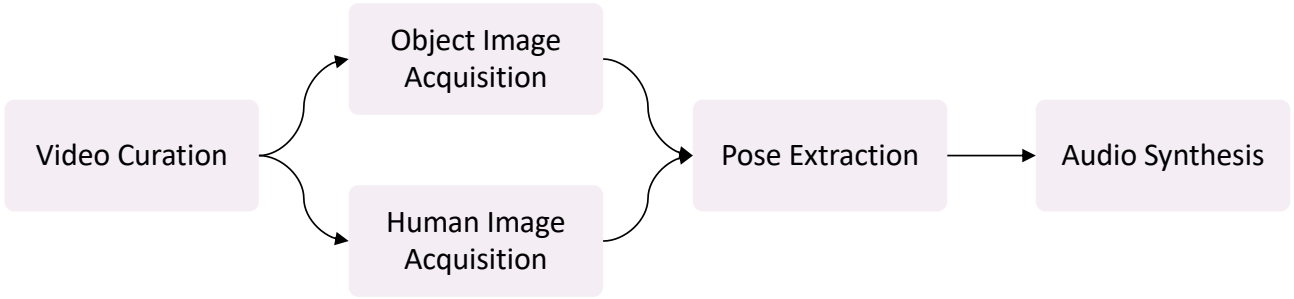
Pipeline

We introduce **OmniShow**, the **first all-in-one-model** that unifies **text**, **reference image**, **audio**, and **pose** conditions for **Human-Object Interaction Video Generation (HOIVG)**, which also supports broader real-world applications.



Methodology

Existing benchmarks often focus on limited-modality control (*e.g.*, text+pose or text+images). We construct **HOIVG-Bench** to bridge this gap by providing an evaluation suite comprising 135 **carefully curated samples** and **dedicated metrics**.



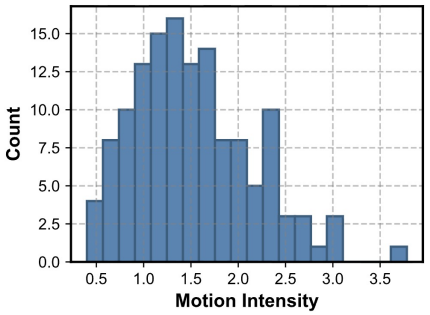
Text Alignment:
- TA from VideoReward

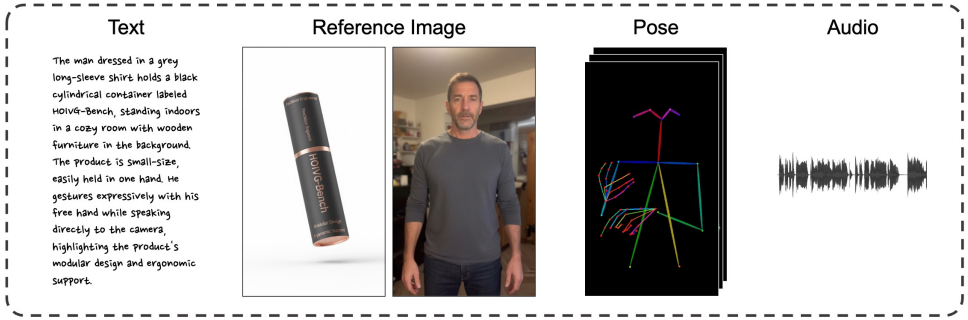
Reference Consistency:
- Face Sim from OpenS2V
- NexusScore from OpenS2V

Pose Accuracy:
- AKD based on DWPose
- PCK based on DWPose

Audio-Visual Synchronization:
- Sync-C based on SyncNet
- Sync-D based on SyncNet

Video Quality:
- AES from VBench
- IQA from VBench
- VQ from VideoReward
- MQ from VideoReward





Experiments

We compare our OmniShow with existing state-of-the-art methods on **HOIVG-Bench**, showing that OmniShow can achieve **superior overall performance** across **diverse multimodal task settings**.

Method	Text Align.	Reference Consistency		Audio-Visual Sync.		Pose Accuracy		Video Quality			
	TA↑	FaceSim↑	NexusScore↑	Sync-C↑	Sync-D↓	AKD↓	PCK↑	AES↑	IQA↑	VQ↑	MQ↑
Text+Reference-to-Video (R2V)											
HunyuanCustom [21]	7.523	0.440	0.359	-	-	-	-	0.452	0.697	10.11	5.286
HuMo-1.7B [4]	7.087	0.647	0.333	-	-	-	-	0.441	0.723	9.76	3.406
HuMo-17B [4]	7.949	0.843	0.346	-	-	-	-	0.448	0.726	9.97	3.685
VACE [31]	8.413	0.759	0.368	-	-	-	-	0.457	0.722	10.72	5.442
Phantom-1.3B [42]	8.342	0.708	0.351	-	-	-	-	0.459	0.722	10.90	5.637
Phantom-14B [42]	**8.609**	**0.876**	0.366	-	-	-	-	0.449	**0.741**	10.93	5.517
OmniShow (Ours)	7.746	0.874	**0.389**	-	-	-	-	**0.468**	0.740	**11.12**	**5.885**
Text+Reference+Audio-to-Video (RA2V)											
HunyuanCustom [21]	7.289	0.457	0.350	6.072	10.08	-	-	0.439	0.715	9.15	3.658
HuMo-1.7B [4]	7.489	0.575	0.329	7.234	9.117	-	-	0.428	0.731	9.97	4.182
HuMo-17B [4]	**8.146**	0.805	0.344	8.013	8.316	-	-	0.439	0.739	10.27	4.269
OmniShow (Ours)	8.093	**0.810**	**0.369**	**8.612**	**7.608**	-	-	**0.465**	**0.742**	**10.86**	**5.554**
Text+Reference+Pose-to-Video (RP2V)											
AnchorCrafter [64]	2.669	0.404	0.215	-	-	0.229	0.176	**0.499**	0.673	8.95	4.241
VACE [31]	**7.690**	**0.600**	0.352	-	-	0.206	0.336	0.450	0.712	10.14	**5.393**
OmniShow (Ours)	6.526	0.474	**0.418**	-	-	**0.174**	**0.460**	0.447	**0.722**	**10.28**	4.937

Quantitative Comparison

Experiments

OmniShow can achieve **industry-graded quality** while achieve **consistent controllability** across **multiple tasks**. Note that OmniShow is the **first model** that supports four multimodal inputs at one time.

Qualitative Comparison

Experiments

The model derived from the A2V training, denoted as **OmniShow-A2V**, also achieve **top-tier performance of audio-driven avatars**.

Method	*IQA↑	*AES↑	Sync-C↑	Sync-D↓
FantasyTalking [58]	2.11	1.12	1.11	12.88
HunyuanVideo-Avatar [7]	1.76	1.18	4.89	9.37
Hallo3 [10]	**2.31**	1.48	4.26	10.22
MultiTalk [34]	2.07	1.30	6.34	**8.47**
OmniAvatar [18]	2.16	1.31	5.40	9.13
OmniShow-A2V (Ours)	2.26	**1.51**	**6.49**	8.97

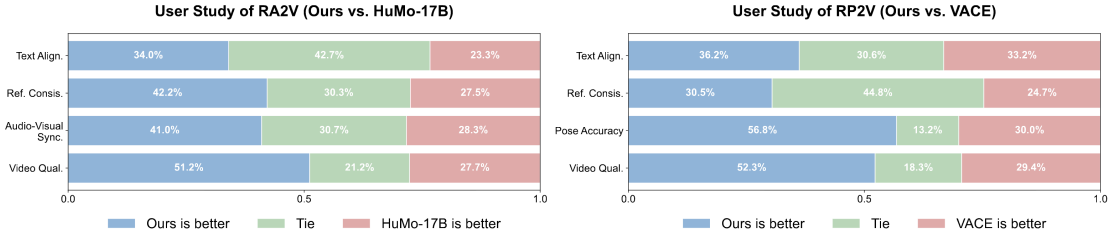
Quantitative Comparison on the EMTD Benchmark

Compared with a representative cascaded baseline (VACE + LatentSync), OmniShow achieve **best results across all evaluation metrics**, **avoiding the artifacts and blur** often introduced by isolated lip-synchronization process.

Method	Text Align.	Reference Consistency		Audio-Visual Sync.		Pose Accuracy		Video Quality			
	TA↑	FaceSim↑	NexusScore↑	Sync-C↑	Sync-D↓	AKD↓	PCK↑	AES↑	IQA↑	VQ↑	MQ↑
Cascaded Baseline	6.885	0.591	0.341	7.016	7.823	0.198	0.340	0.417	0.709	10.05	3.911
OmniShow (Ours)	**7.134**	**0.645**	**0.353**	**7.699**	**7.674**	**0.172**	**0.478**	**0.424**	**0.725**	**11.06**	**5.880**

Qualitative Comparison with VACE + LatentSync

Judged by human evaluators, our OmniShow achieves superior performance in overall quality, showing great alignment with **human preference**.



Qualitative Comparison with VACE + LatentSync

We conducted detailed ablation studies for our key techniques and design choices, to show the **effectiveness** and **robustness** of our method.

(a)

Variant	FaceSim↑	NexusScore↑	AES↑
Token-Concat	0.601	0.344	0.466
w/o Ref. Rec. Loss	0.678	0.352	0.466
Ours	**0.707**	**0.353**	**0.471**

(b)

Variant	Sync-C↑	Sync-D↓	AES↑
w/o Audio Context	8.872	7.878	0.533
w/o Attn.-Map Constraints	2.201	13.01	**0.545**
w/o Adaptive Gating	8.872	7.819	0.529
Ours	**9.023**	**7.419**	0.540

(c)

Variant	NexusScore↑	Sync-D↓	AES↑
Single-Stage (Only RA2V)	0.345	13.11	0.453
Multi-Stage (R2V → RA2V)	0.360	13.23	0.473
Multi-Stage (A2V → RA2V)	0.342	**7.38**	0.456
Ours	**0.364**	8.14	**0.474**

(a)

RoPE Strategy	FaceSim↑	NexusScore↑	AES↑
Temporal Shift	0.675	0.351	0.468
Spatiotemporal Shift	0.279	0.339	0.456
Native (Ours)	**0.707**	**0.353**	**0.471**

(b)

Context Setup	Sync-C↑	Sync-D↓	AES↑
Context Window = 1	8.872	7.878	0.533
Context Window = 11	7.020	9.588	0.527
Context Window = 5 (Ours)	**9.023**	**7.419**	**0.540**

Quantitative Ablation Studies

Thank you very much for your listening!