DiasR: Dual-Modal Identity-Anchored Sparse Routing for Efficient Multi-Subject Video Generation

Yangyang Li[1 2], Wu Liu[1], Jie Li[3], Xinchen Liu[4], Yongdong Zhang[1], Guoqing Jin[2]

[1]University of Science and Technology of China,

[2]The State Key Laboratory of Communication Content Cognition, People's Daily Online,

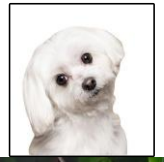
[3]Zhejiang University, [4]JD Explore Academy

Contact:

Yangyang Li: lyy1030@mail.ustc.edu.cn **Xinchen Liu**: liuxinchen1@jd.com **Guoqing Jin**: jinguoqing@people.cn

Personalized Video Generation

Inputs:

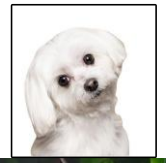
A man playing with a dog

Outputs:


A Pikachu hold a bottle

Personalized Video Generation

Inputs:

A man playing with a dog

Outputs:


A Pikachu hold a bottle

Digital Human Streamers

Films

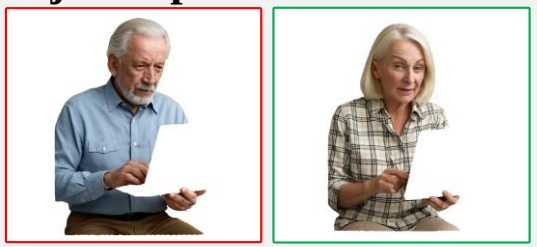
Commercials

Short Videos

Challenge in Multi-Subject Generation

Multi-Subject Inputs

Reference Image



Prompt

An elderly couple sitting in the living room. The man in a light blue shirt holding a laptop, while the woman in a checkered blouse and beige shorts, leans in closely.

Previous Work

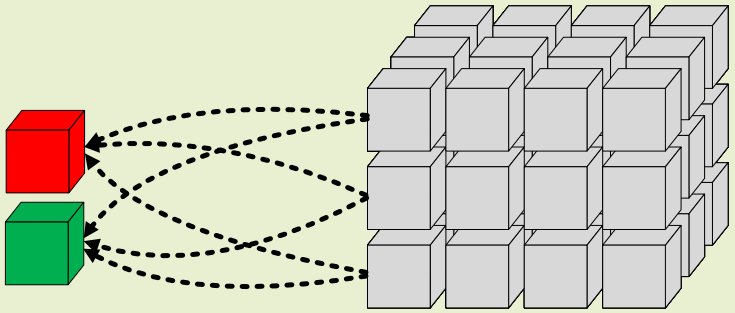


Image Tokens

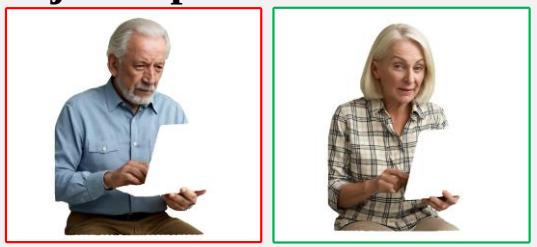
Video Tokens

Dense Attention

Challenge in Multi-Subject Generation

Multi-Subject Inputs

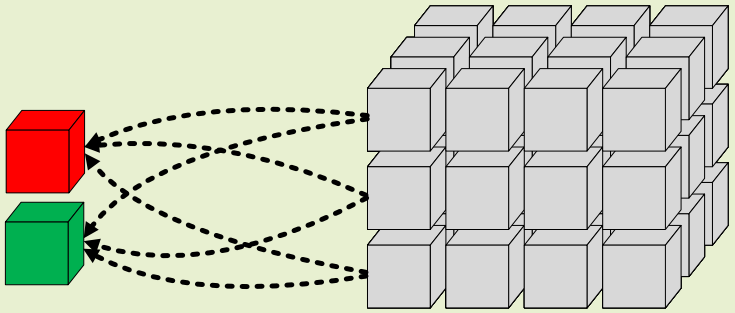
Reference Image



Prompt

An elderly couple sitting in the living room. The man in a light blue shirt holding a laptop, while the woman in a checkered blouse and beige shorts, leans in closely.

Previous Work



Dense Attention

Error attention routing

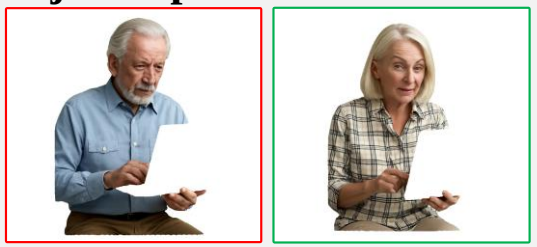
ID Entanglement ✗

ID Drift ✗

Challenge in Multi-Subject Generation

Multi-Subject Inputs

Reference Image



Prompt

An elderly couple sitting in the living room. The man in a light blue shirt holding a laptop, while the woman in a checkered blouse and beige shorts, leans in closely.

Previous Work

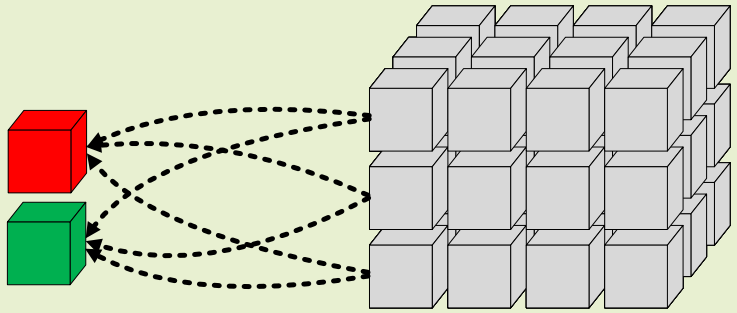


Image Tokens Video Tokens

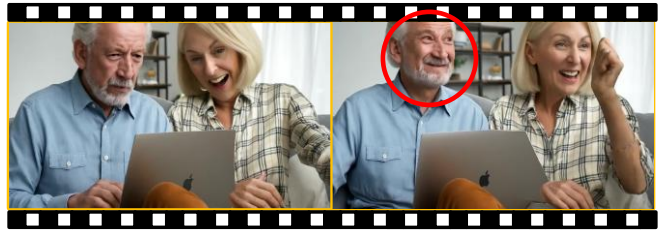
Dense Attention

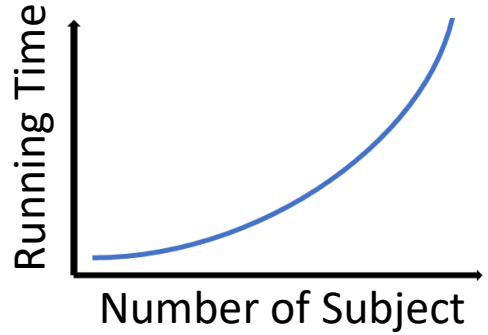
Error attention routing

Extra Context

ID Entanglement ✗

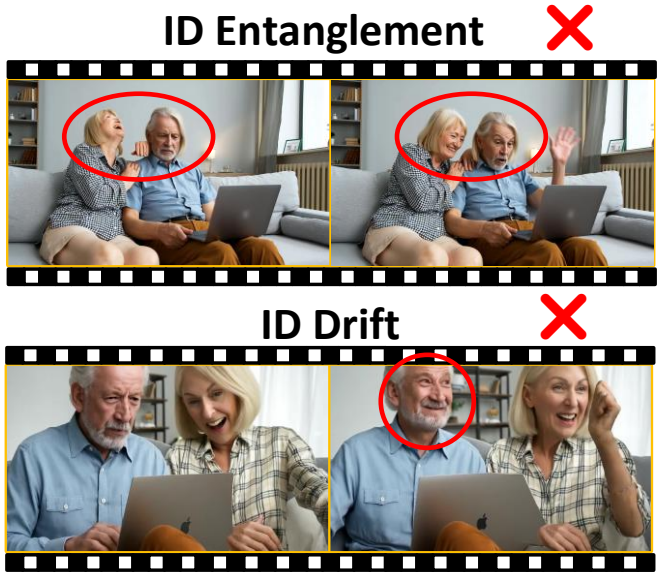
ID Drift ✗



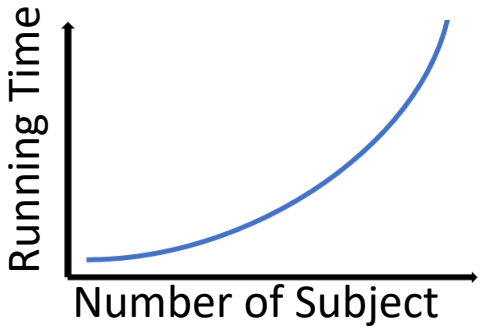


Motivations

Error Attention Routing

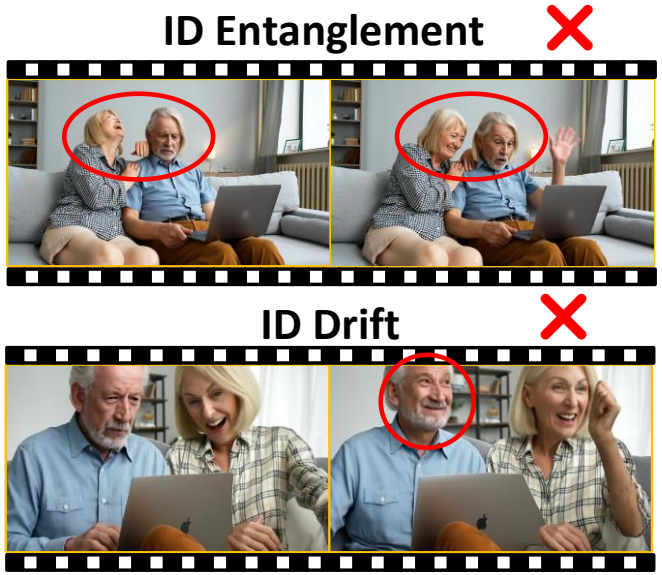


Extra Context



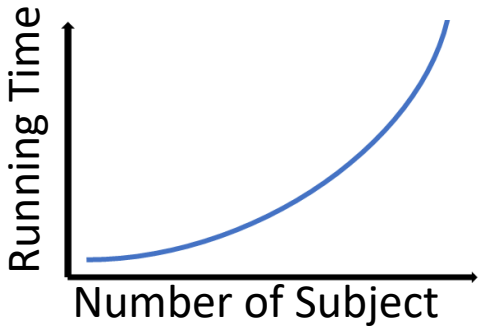
Motivations

Error Attention Routing



➡ Subject-level Attention Routing

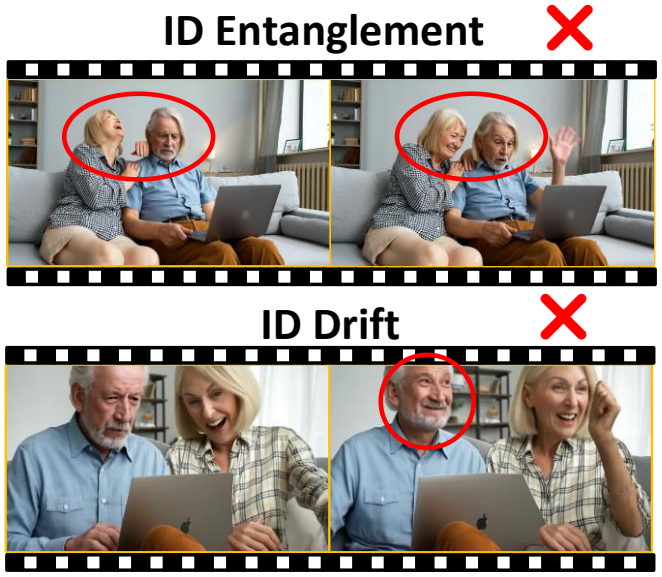
Extra Context



➡ Irrelevant Context Pruning

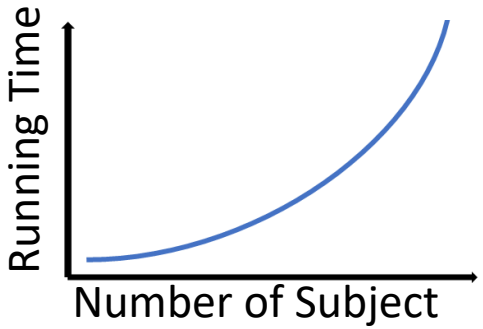
Motivations

Error Attention Routing

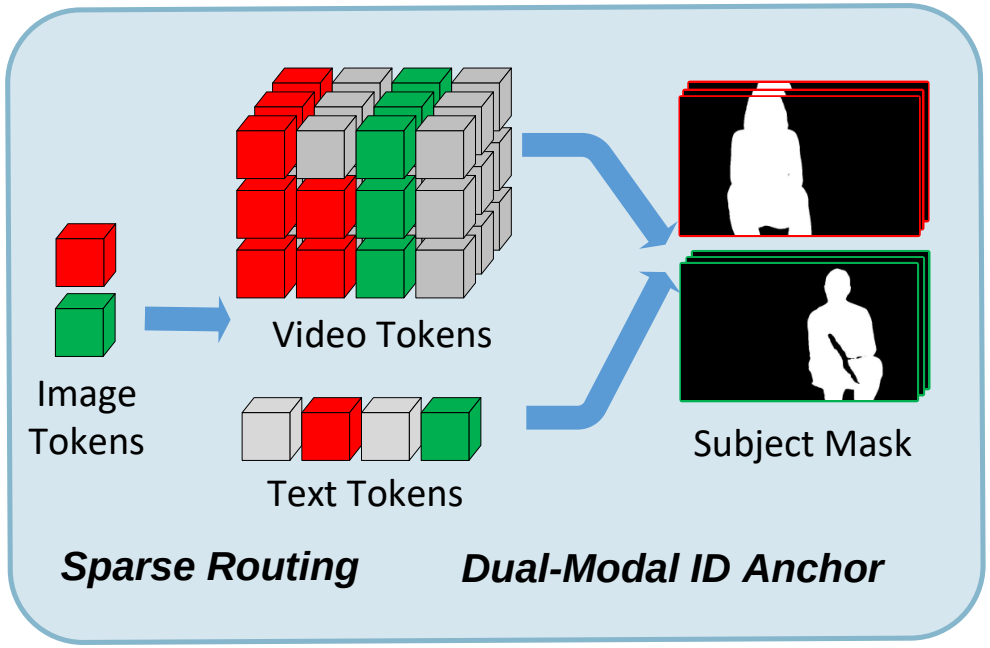


Subject-level Attention Routing

Extra Context



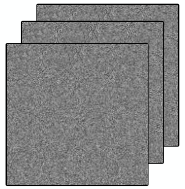
Irrelevant Context Pruning



Framework

Dual-Modal Identity-Anchored Alignment

Noisy Latent



Reference Subjects

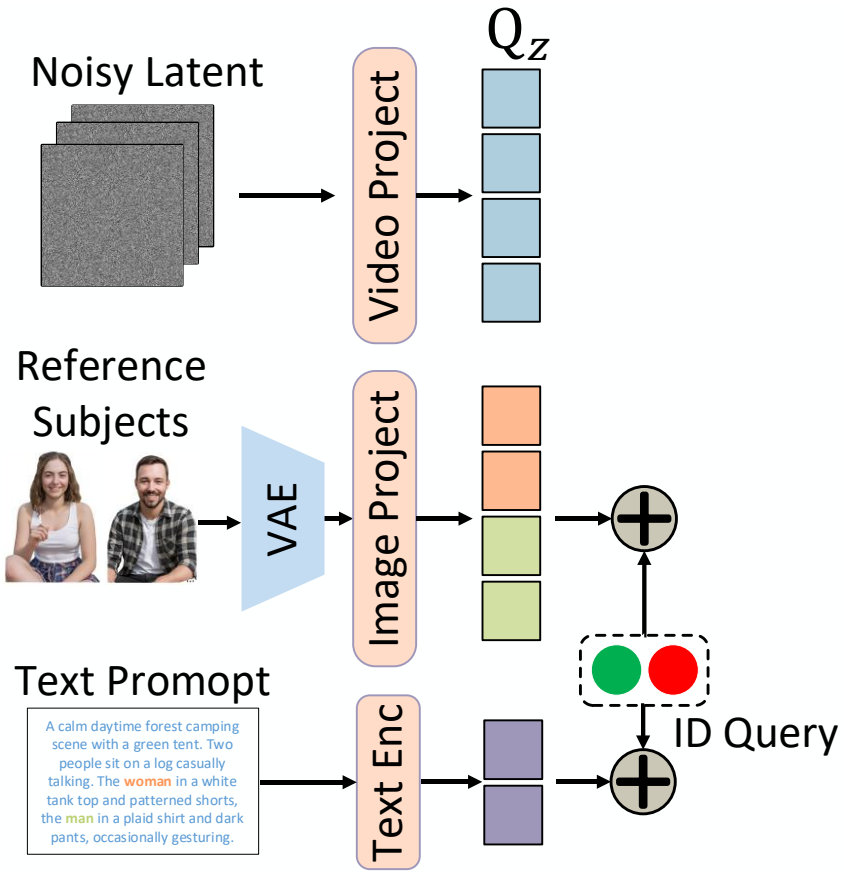
Text Promopt

A calm daytime forest camping scene with a green tent. Two people sit on a log casually talking. The woman in a white tank top and patterned shorts, the man in a plaid shirt and dark pants, occasionally gesturing.

Multi-Modal Inputs

Framework

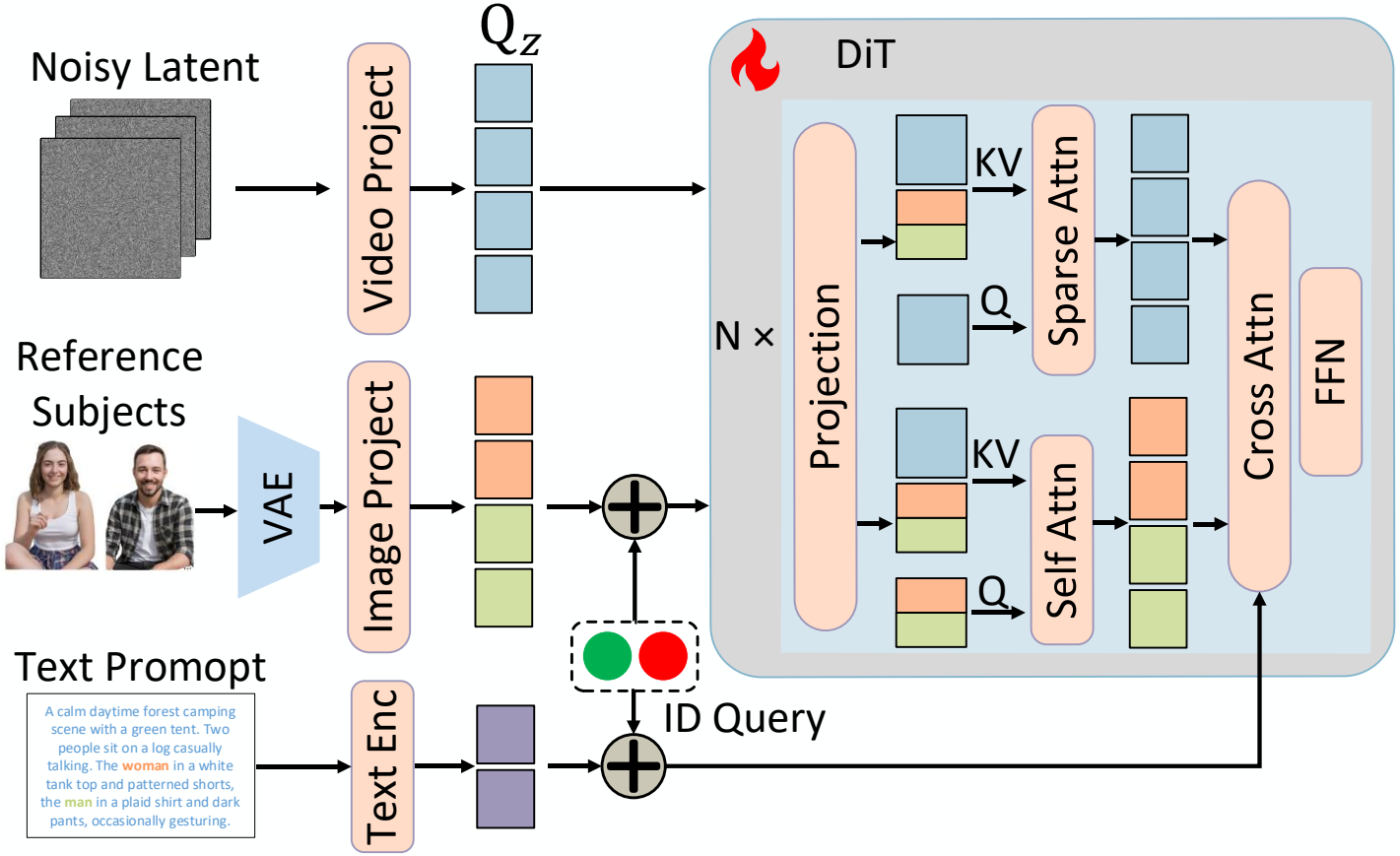
Dual-Modal Identity-Anchored Alignment



Projection & Identity Queries Injection

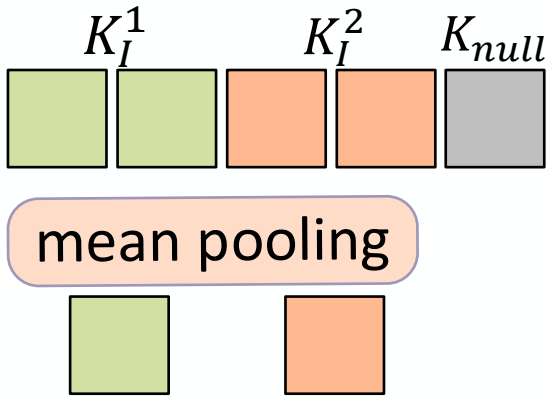
Framework

Dual-Modal Identity-Anchored Alignment



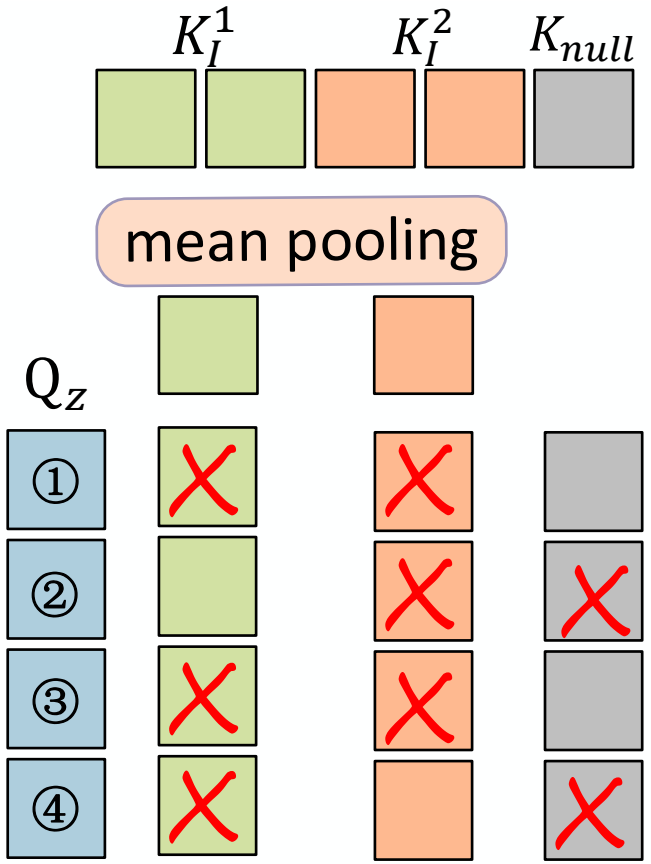
Sparse Attention

Sparse Routing Strategy



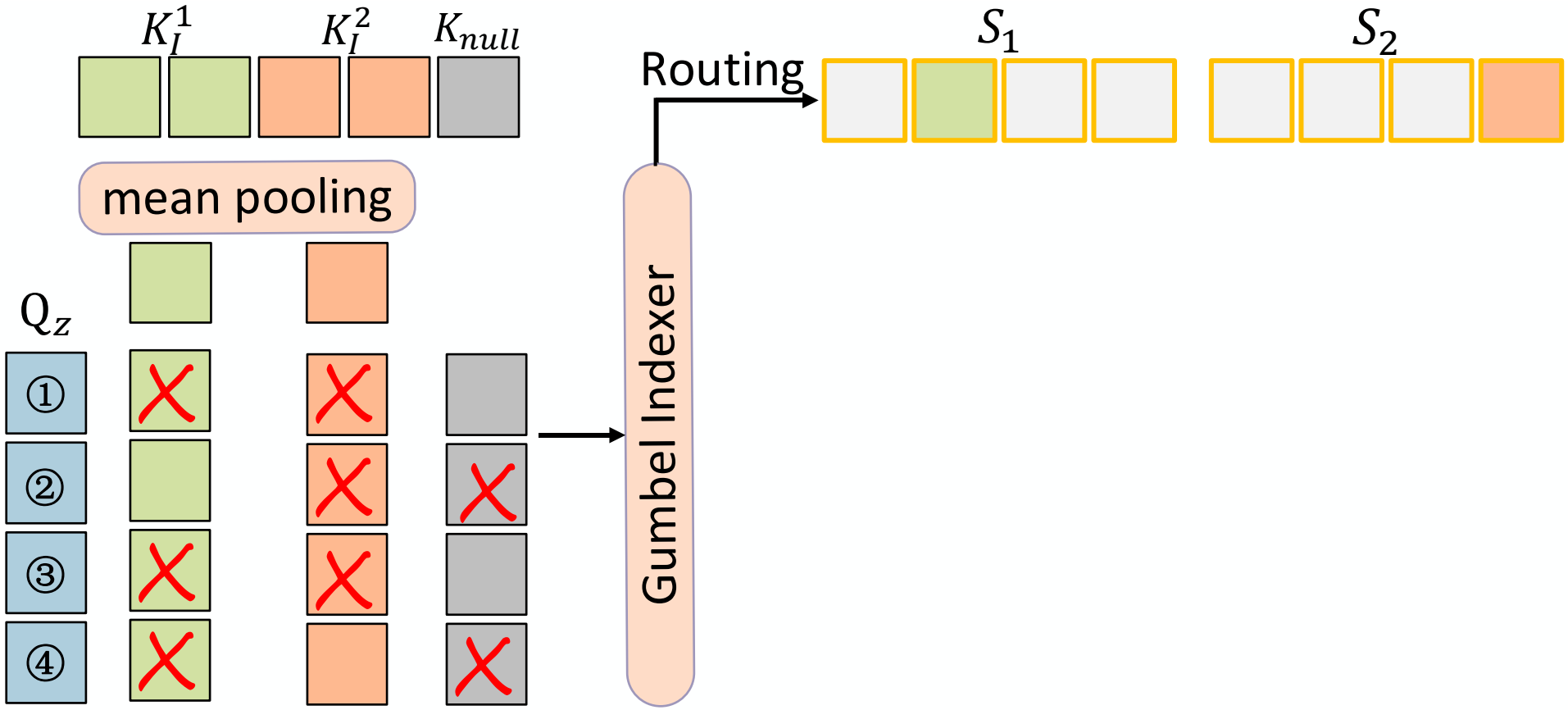
Subject Tokens & Null Tokens

Sparse Routing Strategy



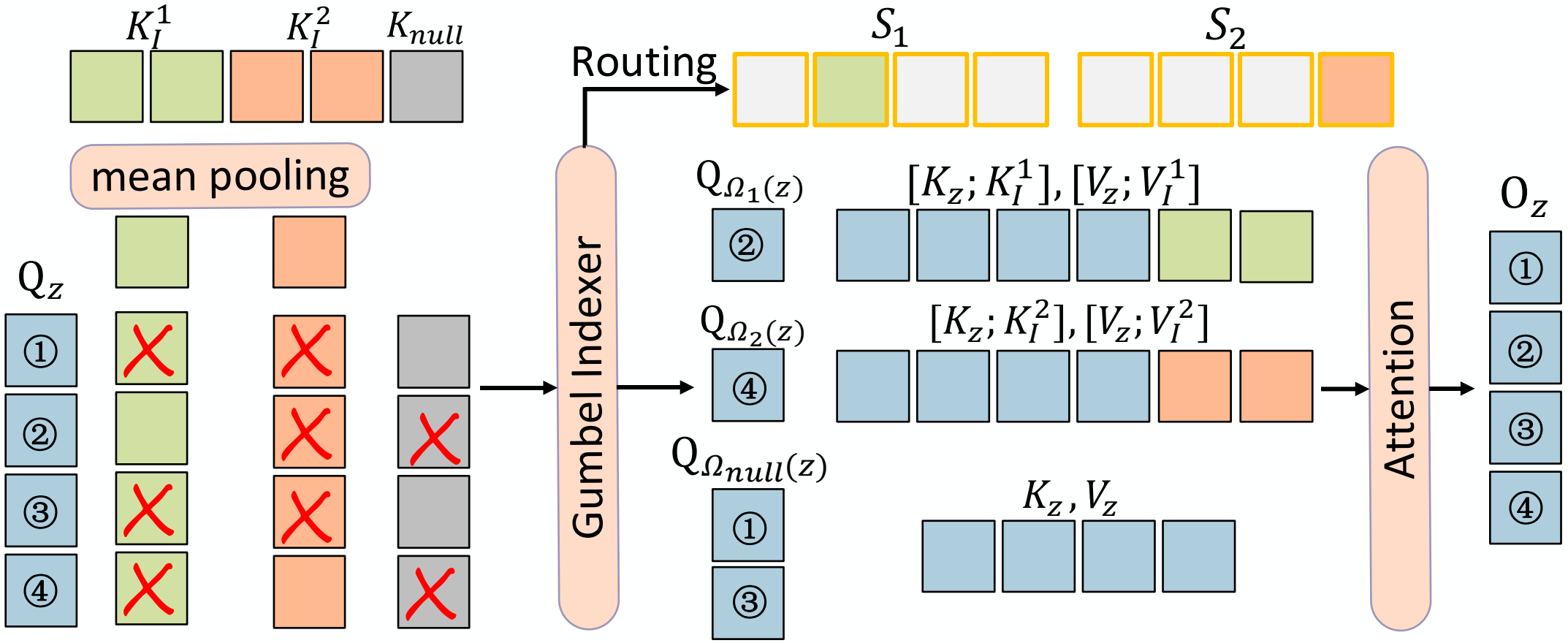
Similarity Matching

Sparse Routing Strategy



Dynamic Subject-level Routing

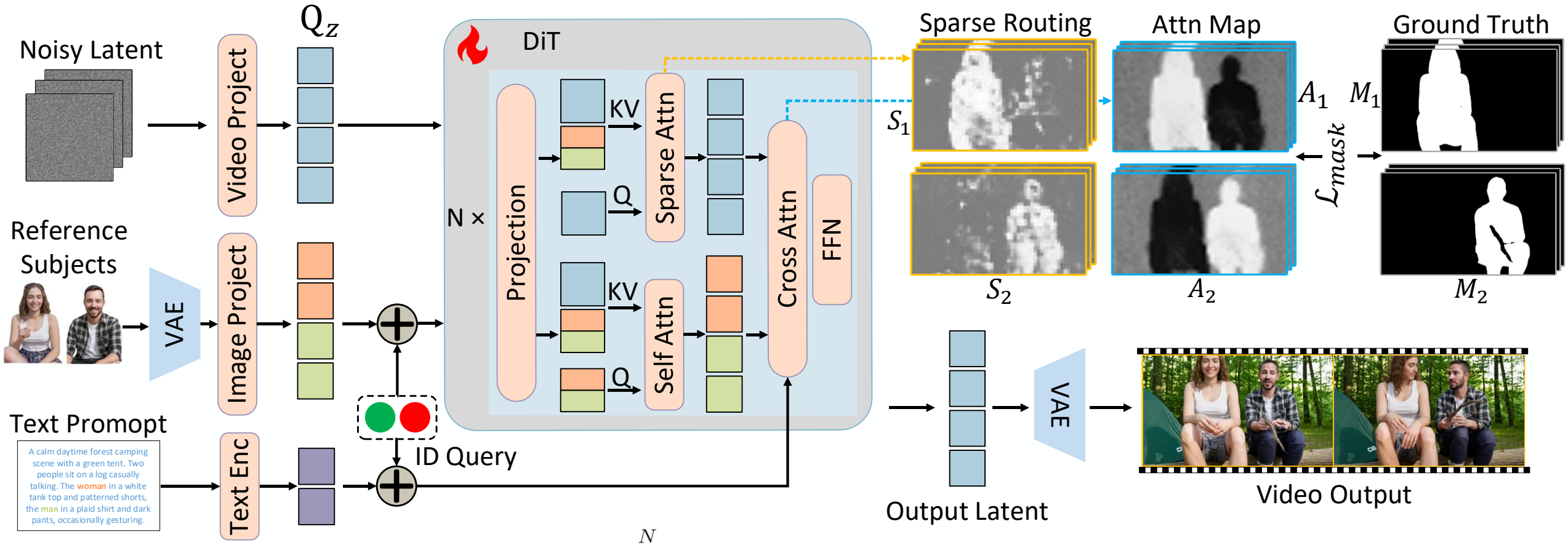
Sparse Routing Strategy



Subject-Sparse Bucketed Attention

Framework

Dual-Modal Identity-Anchored Alignment



$$\mathcal{L}_{\text{mask}} = \sum_{i=1}^{N} \left[\|\mathbf{S}_i - \mathbf{M}_i\|_2^2 + \|\mathbf{A}_i - \mathbf{M}_i\|_2^2 \right],$$

Dual-Modal Perceptual Supervision & Outputs

Multi-Subject Annotated Video Dataset

MuSA-2M:

- Fully annotated video–text–subject–mask quadruples
- Human and object interaction

① *Data Source & Filter*

OpenS2V Nexus

Scene Detect	Aesthetic
Resolution	Motion
FPS	OCR
Duration	WaterMark

② *Video Caption & Subject Annotation*

Input Video

Subject–level Caption

Caption Expert

Subjects Entities

Annotation expert

Main Subjects Filter

striped-shirt man, sitting woman
white handbag

tree

Precise Description Filter

striped-shirt man

man

③ *Subject & Mask Extarction*

Input Video

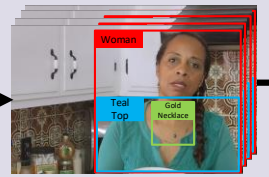
Random Start Frame Sample

Middle-aged Woman
Gold necklace
Teal top

Subject Descriptions

Segmentation

Video Segmentation



Max Area Frame

Cross–Pair Retrieval

Video Mask

FaceCrop

Size Filter

Qualitative Comparison

A girl with blue eyes stands in a sun-drenched, vibrant urban alleyway, her hand reaching out to gently grasp the cool, translucent blue water bottle. She lifts it slightly, her fingers wrapping around its smooth, curved body.

Reference

Phantom

VACE

HuMo

Kaleido

Ours

Qualitative Comparison

The video captures a scene inside a cozy cafes with large windows offering. A man and a woman are seated at a table, engaged in conversation. The woman, dressed in a dark top, leans forward slightly. The man, wearing a gray shirt, listens attentively, nodding his head occasionally. Outside the window, the cityscape is visible, featuring modern buildings and a busy street with cars passing by.

Reference

Phantom

VACE

HuMo

Kaleido

Ours

Experiments

Quantitative comparison

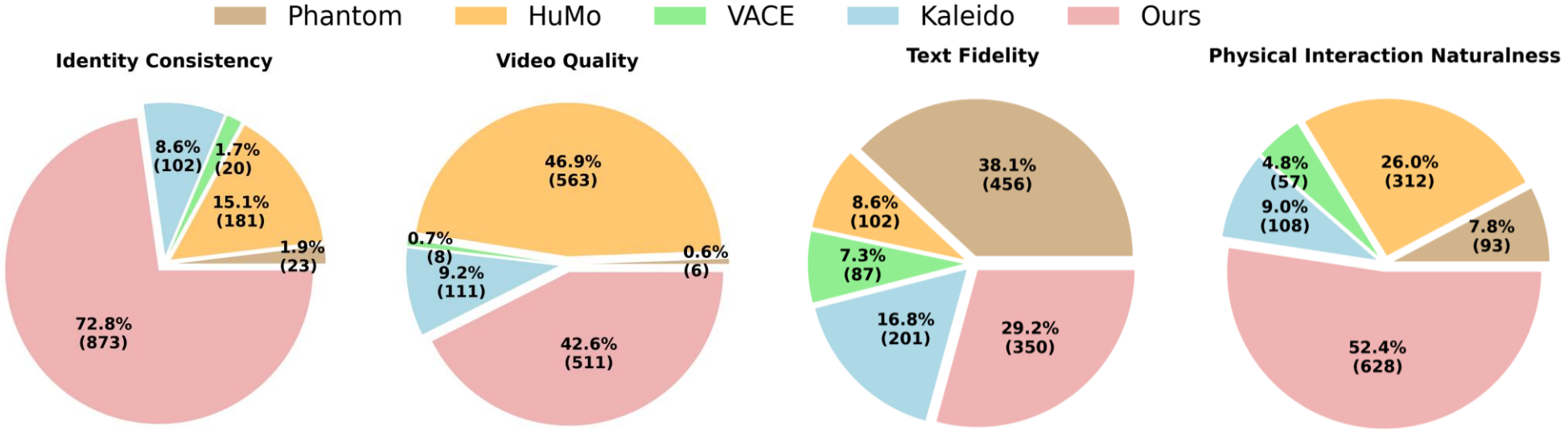
Method	TotalScore ↑	Aesthetics ↑	Motion Smoothness ↑	Motion Amplitude ↑	FaceSim ↑	GmeScore ↑	NexusScore ↑	NaturalScore ↑
Pika-2.1	49.38%	46.88%	87.06%	24.71%	30.38%	69.19%	45.40%	63.32%
Vidu-2.0	48.87%	41.48%	90.45%	13.52%	35.11%	67.57%	43.37%	65.88%
kling-1.6	56.32%	44.59%	86.93%	**41.60%**	40.1%	66.2%	45.89%	**82.52%**
Phantom-1.3B	53.78%	46.67%	93.30%	14.29%	48.56%	69.43%	42.48%	71.06%
Phantom-14B	55.38%	46.39%	96.31%	33.42%	51.46%	70.65%	37.43%	74.03%
VACE-1.3B	48.67%	48.24%	97.20%	18.83%	20.57%	**71.26%**	37.91%	73.47%
VACE-14B	55.96%	47.21%	94.97%	15.02%	55.09%	67.27%	44.08%	73.75%
HuMo-1.7B	50.49%	38.54%	95.64%	13.23%	57.53%	68.56%	42.16%	56.06%
HuMo-17B	56.22%	48.39%	**97.97%**	20.10%	55.37%	66.19%	41.29%	75.15%
Kaleido-14B	55.83%	**48.66%**	97.57%	13.40%	47.97%	69.24%	41.09%	79.86%
Ours	**57.79%**	45.87%	97.47%	16.48%	**60.66%**	66.60%	**47.40%**	74.60%

Experiments

Quantitative comparison

Method	TotalScore ↑	Aesthetics ↑	Motion Smoothness ↑	Motion Amplitude ↑	FaceSim ↑	GmeScore ↑	NexusScore ↑	NaturalScore ↑
Pika-2.1	49.38%	46.88%	87.06%	24.71%	30.38%	69.19%	45.40%	63.32%
Vidu-2.0	48.87%	41.48%	90.45%	13.52%	35.11%	67.57%	43.37%	65.88%
kling-1.6	56.32%	44.59%	86.93%	**41.60%**	40.1%	66.2%	45.89%	**82.52%**
Phantom-1.3B	53.78%	46.67%	93.30%	14.29%	48.56%	69.43%	42.48%	71.06%
Phantom-14B	55.38%	46.39%	96.31%	33.42%	51.46%	70.65%	37.43%	74.03%
VACE-1.3B	48.67%	48.24%	97.20%	18.83%	20.57%	**71.26%**	37.91%	73.47%
VACE-14B	55.96%	47.21%	94.97%	15.02%	55.09%	67.27%	44.08%	73.75%
HuMo-1.7B	50.49%	38.54%	95.64%	13.23%	57.53%	68.56%	42.16%	56.06%
HuMo-17B	56.22%	48.39%	**97.97%**	20.10%	55.37%	66.19%	41.29%	75.15%
Kaleido-14B	55.83%	**48.66%**	97.57%	13.40%	47.97%	69.24%	41.09%	79.86%
Ours	**57.79%**	45.87%	97.47%	16.48%	**60.66%**	66.60%	**47.40%**	74.60%

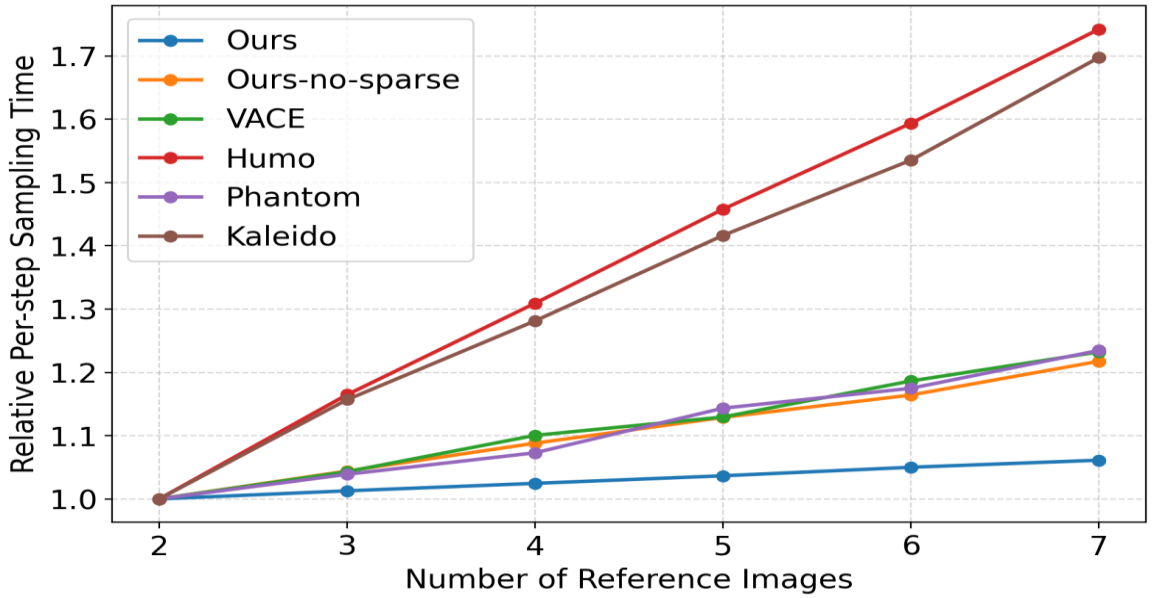
Human Preference Study



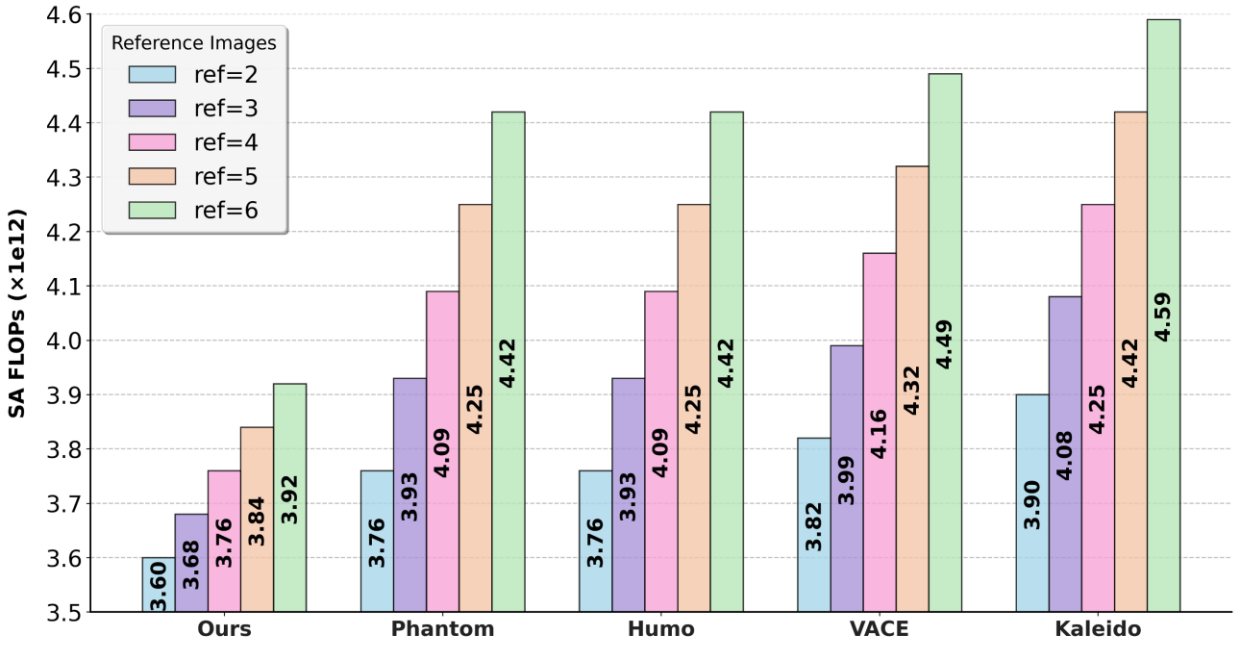
Experiments

Computational Performance Comparison:

- One-step Sampling time
- One-step floating-point operations (FLOPs)



One-step Sampling time



One-step FLOPs

Conclusion

Contributions:

- MuSA-2M, a novel large-scale dataset with subject-level mask annotations
- Dual-Modal Identity-Anchored Alignment that enhances identity consistency
- Sparse Routing Strategy that reduces computational overhead

Future Work:

- Extend MuSA-2M to include videos with 4+ salient subject to support more subjects interaction
- Adapt DiasR to autoregressive architecture for minute-level multi-subject interaction videos

Thank You for Listening

Project Page: https://tale17.github.io/diasr/

Contact:
Yangyang Li: lyy1030@mail.ustc.edu.cn **Xinchen Liu**: liuxinchen1@jd.com **Guoqing Jin**: jinguoqing@people.cn