Qualcomm

CVPR JUNE 3-7, 2026 DENVER COLORADO

Anti-I2V: Safeguarding your photos from malicious image-to-video generation

CVPR 2026

Duc Vu

Anh Nguyen

Chi Tran

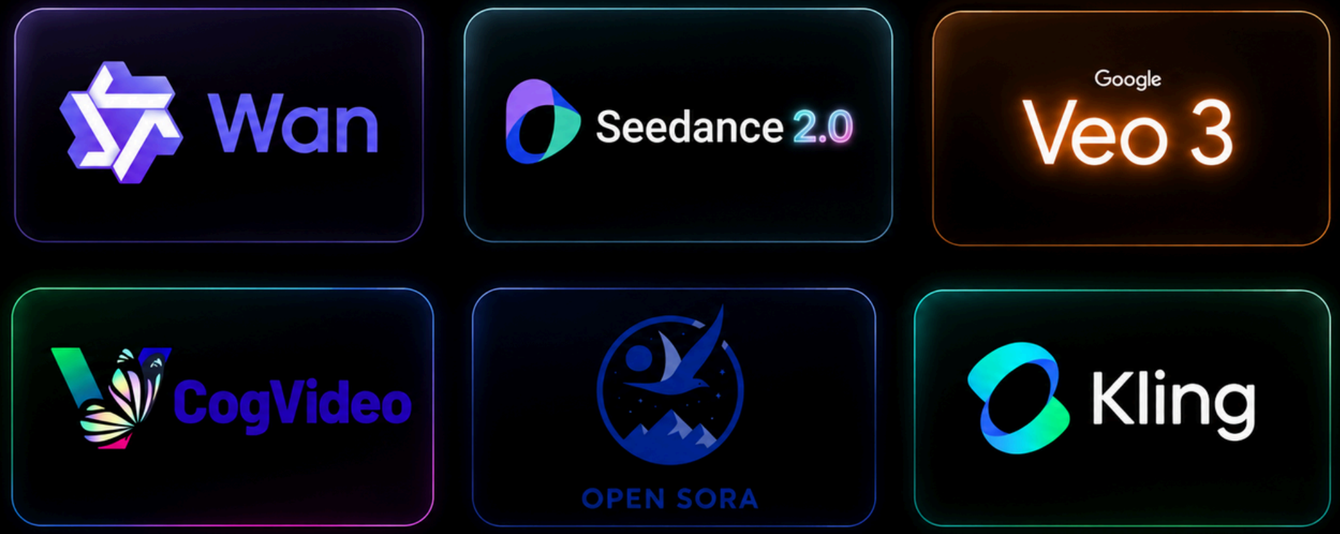
Anh Tran

Qualcomm

CVPR JUNE 3-7, 2026 DENVER COLORADO

Motivation

- **Rapid rise of Video Diffusion Models (VDMs):**
  - **Open-source:** Wan2.1, CogVideoX, OpenSora
  - **Closed-source:** Veo3, Seedance, Kling



Qualcomm

CVPR JUNE 3-7, 2026 DENVER COLORADO

Motivation

Misuse of personal image for harmful content from **Image-to-Video (I2V) models**

⚠ **Identity Theft** ⚠ **Fraud** ⚠ **Harassment**

Qualcomm

CVPR JUNE 3-7, 2026 DENVER COLORADO

Solution: Adversarial Attack

> An effective adversarial attack must craft imperceptible noise that remain invisible to humans while disrupting Image-to-Video generation

Qualcomm

CVPR JUNE 3-7, 2026 DENVER COLORADO

Limitations: Current Defenses

1 Backbone support

Most defenses target **text-to-image** or **image-to-image** generation

Current I2V defenses:

- Most designed for **UNet-based models**
- **Extra Conditions** (Reference Poses)

2 Memory

Rely on:

- **Attention-based losses** or **full-length video input**
- **External Models** (Reference Net)

=> **High-end GPUs**: 60-80 VRAM

3 Robustness

Rely on **RGB-only perturbations**
=> **Less robust** to noise removal techniques

> **Existing defenses are narrow, costly and insufficiently robust for modern Image-to-video models**

Qualcomm

CVPR JUNE 3-7, 2026 DENVER COLORADO

Contribution

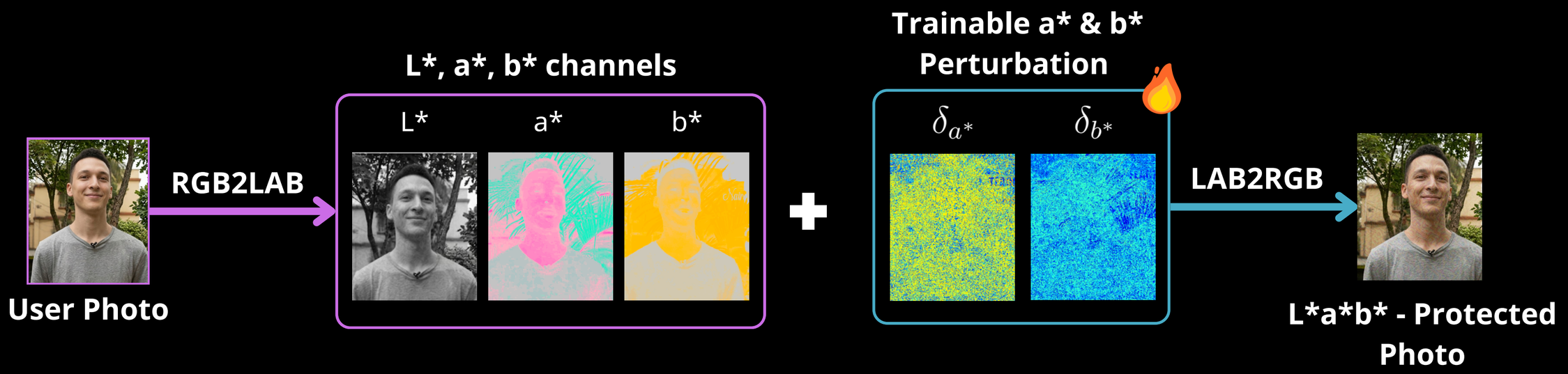
We propose **Anti-I2V**, a robust and effective adversarial attack on I2V models with **3 main conponents**

1. **Dual-Space Perturbation (DSP)**
2. **Inter Representation Collapse (IRC)**
3. **Internal Representation Anchor (IRA)**

1 Dual-Space Perturbation (DSP)

CVPR JUNE 3-7, 2026 DENVER COLORADO

L∗a∗b∗ color space



Apply noise only to **a*** and **b*** channels

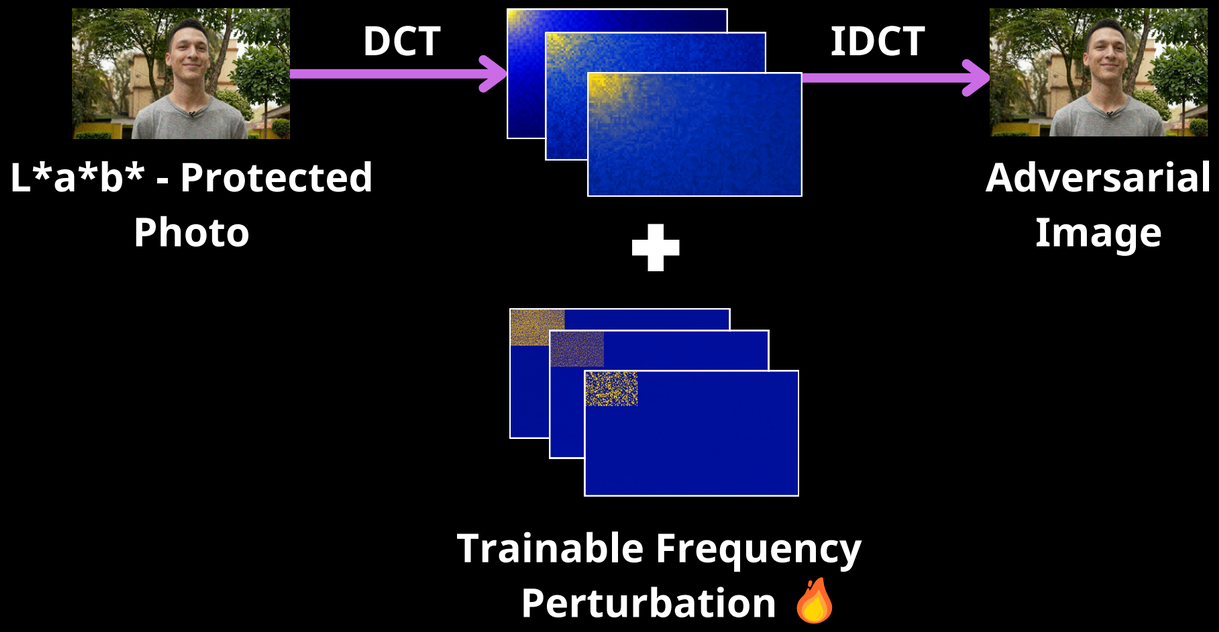
Better aligns with human perception

Less perceptible to observers

Improved robustness

1 Dual-Space Perturbation (DSP)

Frequency Space via DCT



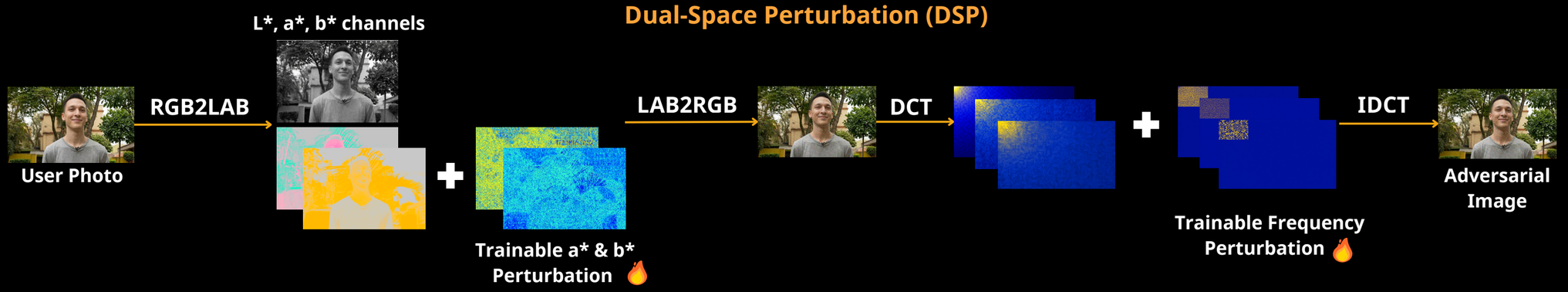
- Perturb **top-left DCT coefficients** (Low-frequency components)
- Inject adversarial noise directly in the DCT domain

=> Directly **alter core structural & textural information** encoded in low-frequency components

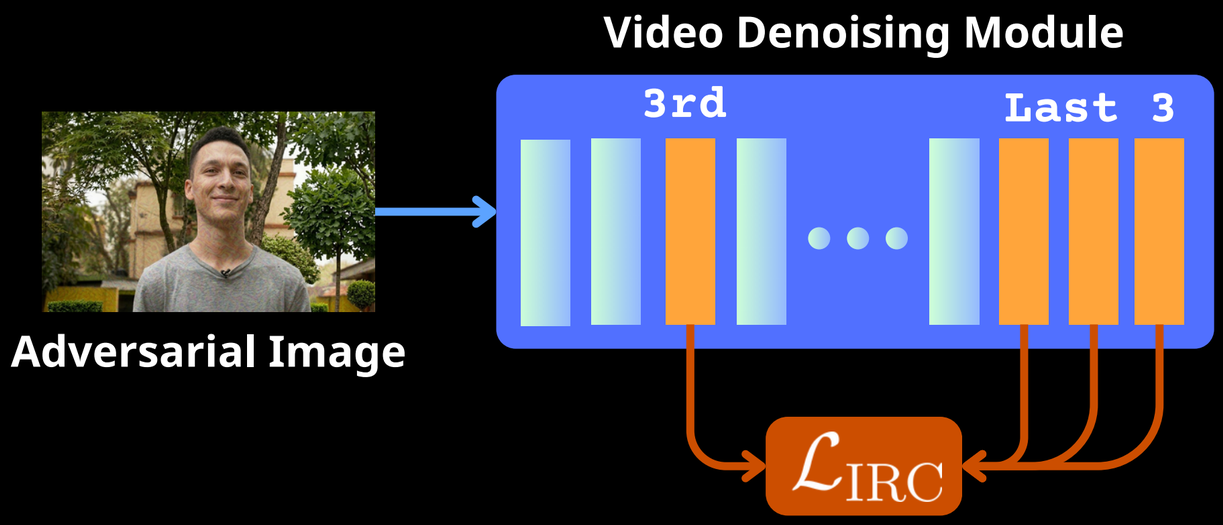
1 Dual-Space Perturbation (DSP)

CVPR JUNE 3-7, 2026 DENVER COLORADO

Final DSP Pipeline



2 Inter Representation Collapse (IRC)



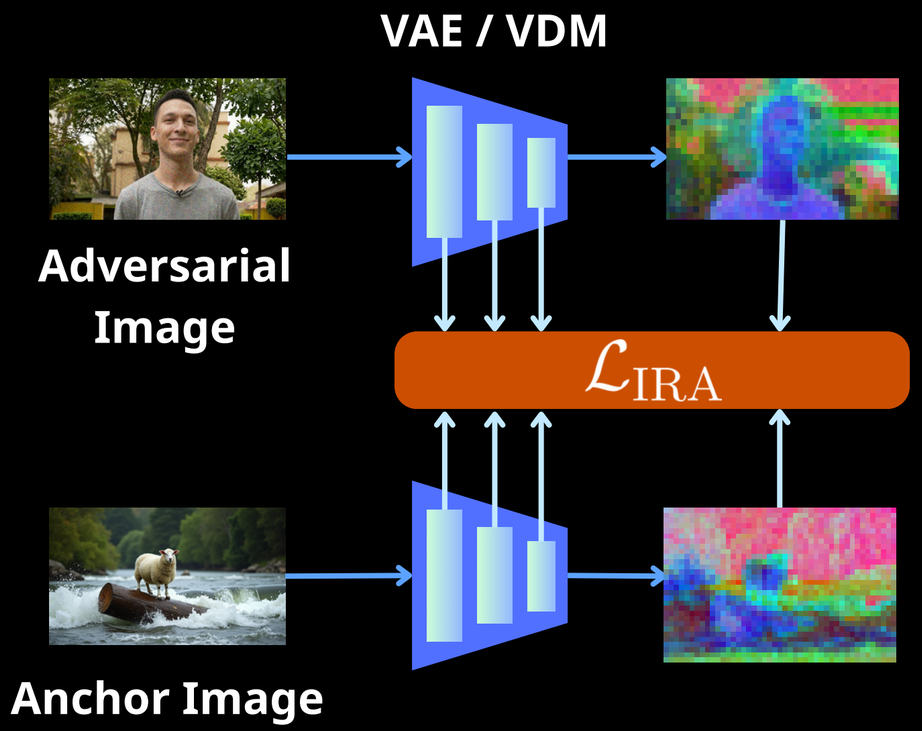
- **Goal:** Disrupt feature propagation => Suppress rich semantic features into low-information representation
- **How:** Align rich semantic layers w/ layers that contain minimal semantic information

MSE loss between **features of the last three layers** to **features of the third layer**

$$\mathcal{L}_{\text{IRC}}^{i,j} = \mathbb{E} \left\| e_\theta^j(z_t, z_\xi, t, y) - \epsilon_\theta^i(z_t, z_\xi, t, y) \right\|_2^2$$

3 Internal Representation Anchor (IRA)

- IRC does not explicitly interfere with feature extraction at individual layers
- **Previous works**: Only focus on the **final output** of VAEs



IRA Loss

- Explicitly disrupts layer-wise feature representations
- Minimizes Euclidean distance between hidden features at each layers

$$\mathcal{L}^{m}_{\text{IRA},\epsilon_\theta} = \mathbb{E}\left\|\epsilon^{m}_{\theta}(z_t, z_\xi, t, y) - \epsilon^{m}_{\theta}(z_t, z_\psi, t, y)\right\|_2^2$$

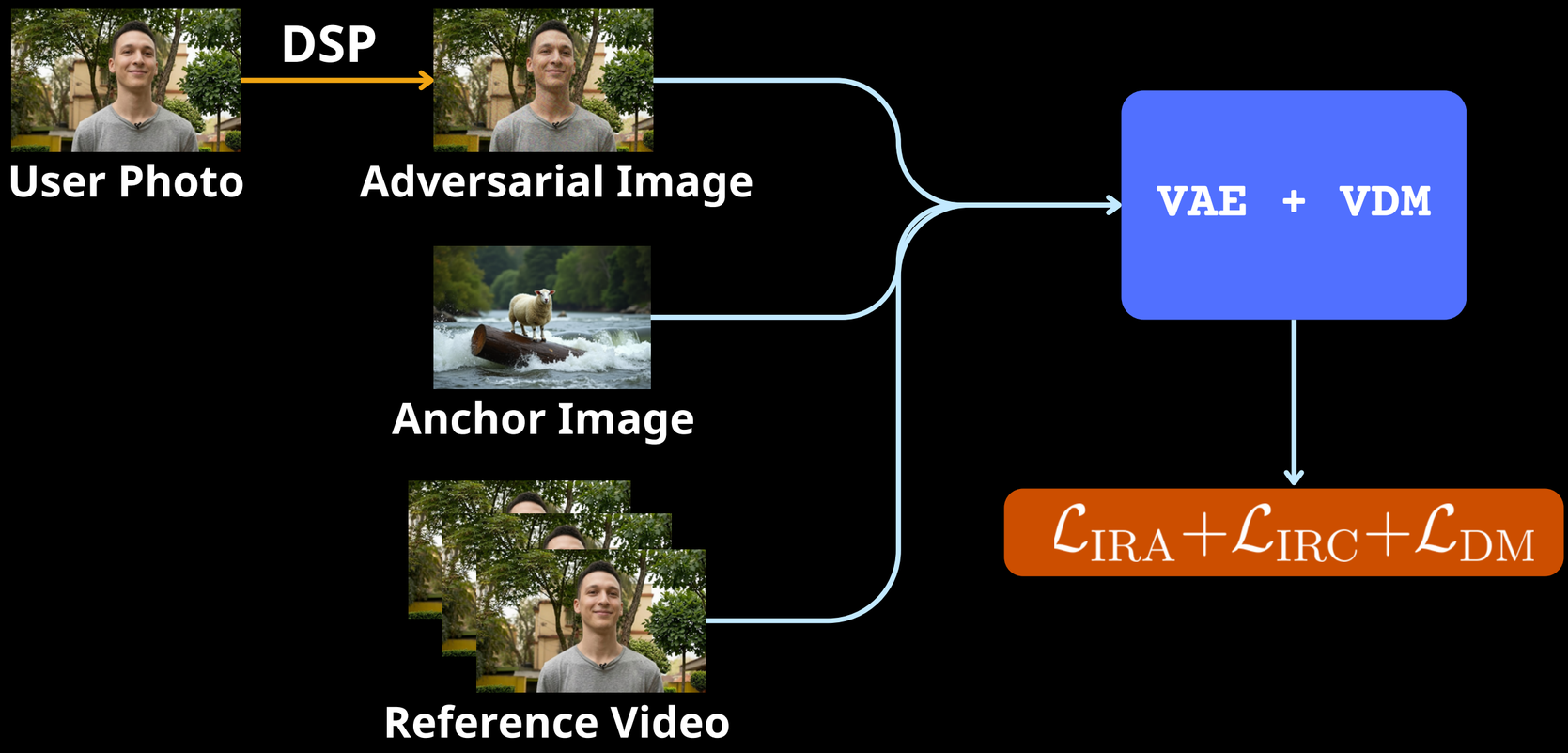
$$\mathcal{L}^{n}_{\text{IRA},E} = \mathbb{E}\left\|E^{n}(z_\xi) - E^{n}(z_\psi)\right\|_2^2$$

$$\mathcal{L}_{\text{IRA}} = \mathcal{L}_{\text{IRA},\epsilon_\theta} + \mathcal{L}_{\text{IRA},E}$$

Qualcomm

CVPR JUNE 3-7, 2026 DENVER COLORADO

Final Pipeline



Qualcomm

CVPR JUNE 3-7, 2026 DENVER COLORADO

Benchmark

No existing I2V benchmark simultaneously provides **realistic human videos** and **rich text annotations** for human animation evaluation

CelebV-Text

Facial Realism - Identity consistency

- **Facial-centric** human videos
- Rich text annotations
- 1,000 videos of unique identities
- 5 generations per pair
  => **5,000 videos for evaluation**

UCF101

Motion - Action Coherence

- **Action-centric videos**
- **D**iverse full-body motions
- Detailed text annotations from LVLM
- 200 videos, 5 generations per pair
  => **1,000 videos for evaluation**

All experiments are run on **a single A100 40GB GPUs**

Qualcomm

CVPR JUNE 3-7, 2026 DENVER COLORADO

Metrics

- **Identity Score Matching (ISM)**: ArcFace between detected faces and the clean reference faces

- **CLIP-FIQA (C-FIQA): Advanced metric** for detected facial images

- **Q-Align:** State-of-the-art image and video evaluation metric
  - **Q-Align Frame (QA-F):** Average scores for individual frames
  - **Q-Align Video (QA-V)**: Scores for entire videos

- **DINO:** Cosine similarity between DINO-extracted features of generated frames and clean reference image.

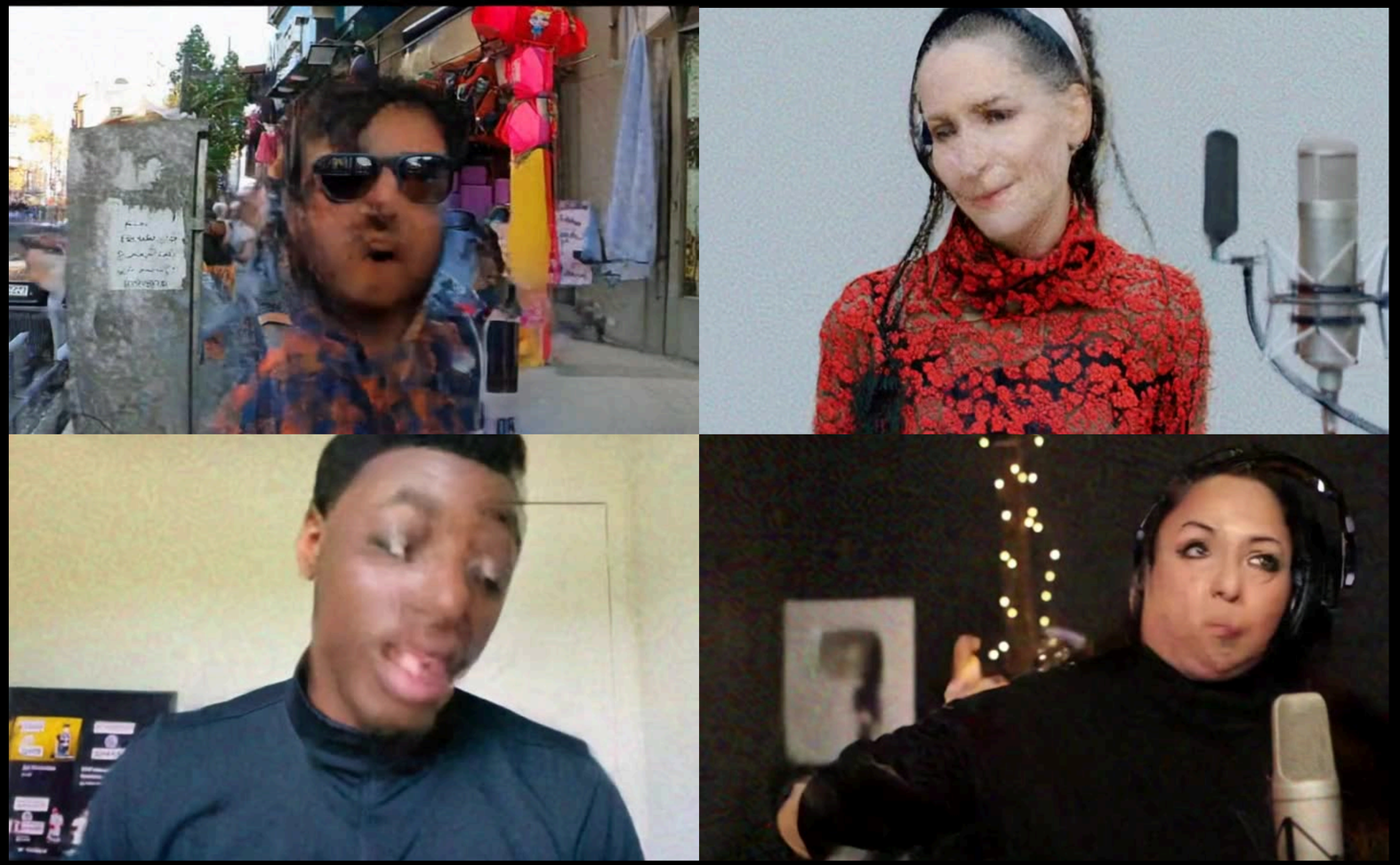
Results

Table 1. **Quantitative comparison of Anti-I2V against baseline protections.** ↓ indicates that lower values correspond to poorer video quality and thus stronger protection. The best scores are highlighted in **bold**, while the second-best results are underlined.

		CelebV-Text [54]					UCF101 [41]				
Model	**Method**	**ISM ↓**	**C-FIQA ↓**	**Q-A(F) ↓**	**Q-A(V) ↓**	**DINO ↓**	**ISM ↓**	**C-FIQA ↓**	**Q-A(F) ↓**	**Q-A(V) ↓**	**DINO ↓**
CogVideoX-5B	Clean	0.721	0.522	0.746	0.802	0.828	0.466	0.373	0.361	0.436	0.801
	SDS+ [51]	0.591	0.482	0.473	0.563	0.754	0.381	0.291	0.286	0.344	0.754
	SDS- [51]	0.607	0.497	0.511	0.594	0.747	0.386	0.298	0.313	0.393	0.760
	AdvDM [26]	0.583	0.473	0.463	0.543	0.748	0.370	0.292	0.271	0.342	0.753
	MIST [25]	0.561	0.463	0.476	0.577	0.750	0.355	0.290	0.262	0.340	0.751
	VGMShield [34]	0.554	0.461	0.464	0.557	0.745	0.361	0.292	0.265	0.343	0.753
	Anti-I2V	**0.448**	**0.433**	**0.447**	**0.532**	**0.722**	**0.346**	**0.283**	**0.251**	**0.323**	**0.734**
DynamiCrafter	Clean	0.528	0.467	0.724	0.794	0.622	0.384	0.345	0.501	0.562	0.709
	SDS+ [51]	0.264	0.372	0.213	0.245	0.389	0.122	0.305	0.157	0.194	0.433
	SDS- [51]	0.278	0.418	0.223	0.250	0.392	0.128	0.338	0.164	0.226	0.439
	AdvDM [26]	0.269	0.370	0.167	0.207	0.397	0.110	0.335	0.162	0.201	0.451
	MIST [25]	0.262	0.379	0.232	0.269	0.386	0.100	0.335	0.322	0.381	0.493
	VGMShield [34]	0.286	0.431	0.243	0.289	0.401	0.108	0.336	0.318	0.374	0.486
	Anti-I2V	**0.151**	**0.303**	**0.032**	**0.047**	**0.167**	**0.068**	**0.268**	**0.057**	**0.084**	**0.164**
Open-Sora	Clean	0.598	0.508	0.712	0.782	0.811	0.400	0.382	0.409	0.437	0.750
	SDS+ [51]	0.502	0.481	0.494	**0.548**	0.730	0.355	0.293	0.360	0.389	0.692
	SDS- [51]	0.508	0.494	0.514	0.591	0.731	0.333	0.311	0.351	0.387	0.687
	AdvDM [26]	0.506	0.478	0.496	0.561	0.725	0.346	0.309	0.327	0.362	0.686
	MIST [25]	0.493	0.475	0.497	0.594	**0.710**	0.339	0.309	0.338	0.392	0.677
	VGMShield [34]	0.500	0.476	0.497	0.578	0.716	0.341	0.312	0.335	0.369	0.680
	Anti-I2V	**0.461**	**0.453**	**0.478**	0.554	0.713	**0.318**	**0.248**	**0.311**	**0.347**	**0.642**

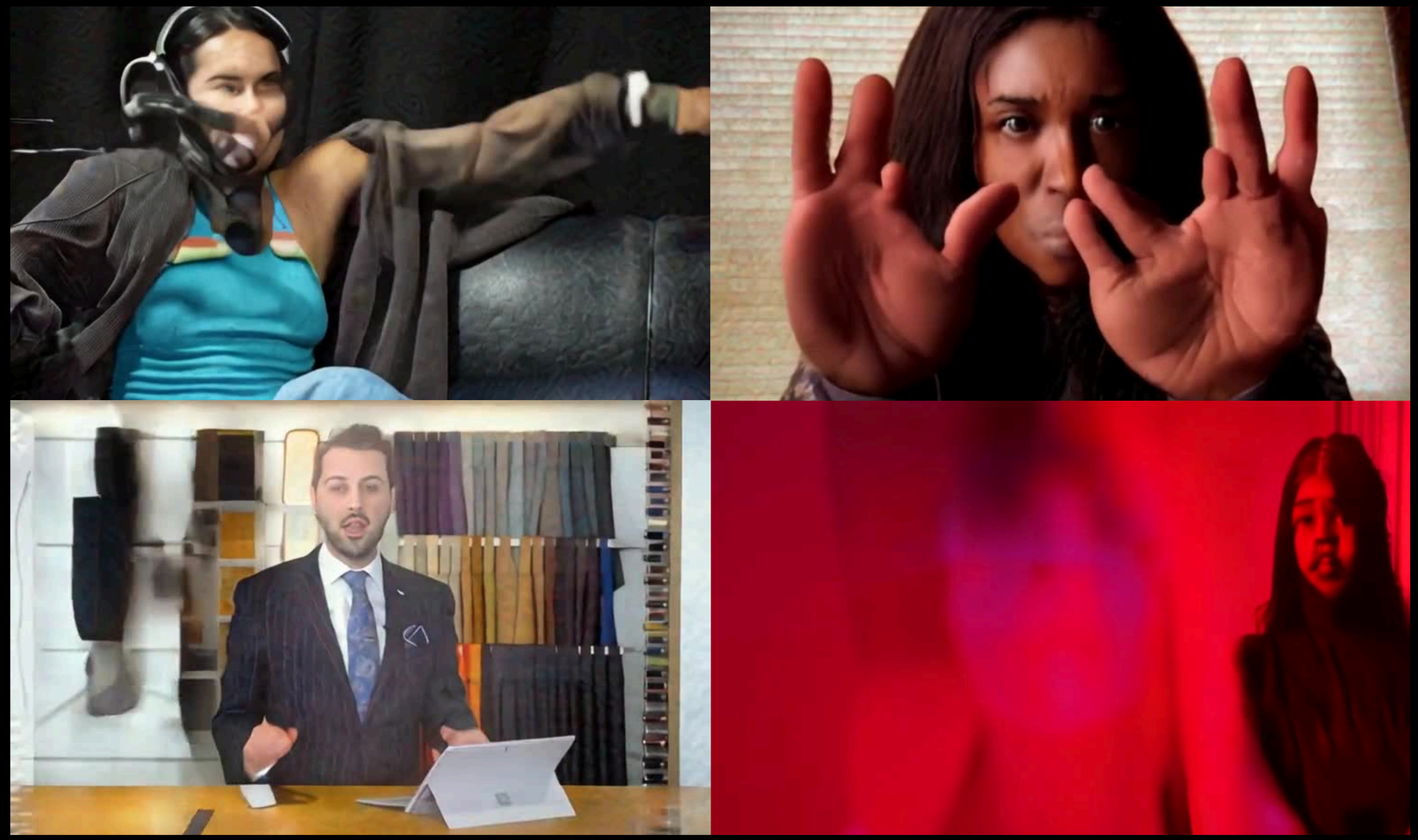
Qualitative Demo

CogVideoX-5B



DynamiCrafter

OpenSora 1.2



Thank you!!!