> TL;DR HuMo targets human-centric video generation with collaborative multi-modal conditioning. It pairs text, reference images, and audio; trains progressively across sub-tasks (subject consistency and audio-visual sync); and uses a time-adaptive CFG during inference for flexible control.

What is HuMo?

HuMo is a research framework for generating controllable human videos from combinations of text, images, and audio. It focuses on two hard sub-tasks: keeping the subject consistent (face, clothing, identity) and aligning generated motion and lip dynamics with audio.

Links:

- Paper (arXiv): https://arxiv.org/abs/2509.08519
- Project: https://phantom-video.github.io/HuMo/
- Repo: https://github.com/Phantom-video/HuMo

Key ideas

- Paired tri-modal data: Curates paired text, reference images, and audio to supervise collaborative multimodal control.
- Progressive multimodal training: Two-stage scheme that first builds subject preservation, then introduces audio-visual sync on top of it.
- Minimal-invasive image injection: Preserves the base model's prompt-following and visual fidelity while injecting identity/appearance.
- Focus-by-predicting for audio: Beyond audio cross-attention, the model is guided to associate audio with facial regions for improved sync.
- Time-adaptive CFG (inference): Dynamically adjusts guidance weights across denoising steps for finer per-modality control.

These design choices aim to unify separate sub-tasks under one model rather than maintaining specialized models per task.

Models and availability

- HuMo-17B: research-grade quality; 480p and 720p supported (heavier compute).

- HuMo-1.7B: lighter; 480p in ~8 minutes on a 32G GPU (per project README), with audio-visual sync largely retained vs 17B.

Weights and example code are available from the project's Hugging Face hub and GitHub repo.

Quickstart (from repo docs)

Environment setup:

```
conda create -n humo python=3.11
conda activate humo
pip install torch==2.5.1 torchvision==0.20.1 torchaudio==2.5.1 \
  --index-url https://download.pytorch.org/whl/cu124
pip install flash_attn==2.6.3
pip install -r requirements.txt
conda install -c conda-forge ffmpeg
```

Model prep (abbrev.):

```
huggingface-cli download bytedance-research/HuMo --local-dir ./weights/HuMo
huggingface-cli download Wan-AI/Wan2.1-T2V-1.3B --local-dir ./weights/Wan2.1-T2V-1.3B
huggingface-cli download openai/whisper-large-v3 --local-dir ./weights/whisper-large-v3
```

Run inference:

```
# Text + Audio (TA)
bash scripts/infer_ta.sh       # 17B
bash scripts/infer_ta_1_7B.sh  # 1.7B

# Text + Image + Audio (TIA)
bash scripts/infer_tia.sh       # 17B
bash scripts/infer_tia_1_7B.sh  # 1.7B
```

Config highlights (inference)

Key knobs (see `humo/configs/inference/generate.yaml` in the repo):

```
generation:
  frames: 97 # video length; trained at 97 frames @ 25 FPS
  height: 720 # 720p preferred; 480p faster
  width: 1280
  mode: "TA" # TA (text+audio) or TIA (text+image+audio)
  scale_a: 1.0 # audio guidance strength
  scale_t: 1.0 # text guidance strength

dit:
```

```
  sp_size: 2 # sequence parallelism size (match num GPUs)

diffusion:
  timesteps:
    sampling:
      steps: 50 # denoising steps (30–40 is faster)
```

Practical notes

- Compute: 17B needs high-VRAM GPUs; 1.7B is friendlier (32G for 480p per README). Multi-GPU inference uses FSDP + sequence parallel.
- Length: trained at 97 frames; longer generations may degrade without a longer-video checkpoint (noted in repo roadmap).
- Sync vs. style: increase `scale_a` for tighter lip-sync/body motion; adjust `scale_t` for prompt adherence and style.
- Inputs: for TIA, prepare a JSON case file with text, image references, and audio paths (see `examples/test_case.json`).

References

- arXiv: https://arxiv.org/abs/2509.08519
- Project: https://phantom-video.github.io/HuMo/
- Code: https://github.com/Phantom-video/HuMo

Notes: Details and figures reflect the public paper/README at publish time. Validate performance on your own prompts and hardware; check the repo for updates (e.g., longer-video checkpoints).