

TL;DR Stand-In bolts identity control onto Wan 2.1 text-to-video backbones with only ~1% additional parameters. A conditional image branch, restricted self-attention, and conditional position mapping lock in the reference face so teams can run subject-driven, pose-guided, stylised, or face-swapped videos without retraining the full model.

What is Stand-In?

Stand-In is a lightweight identity-preserving add-on for diffusion video generators, announced on 11 August 2025 alongside an arXiv preprint and open-source repo. Rather than fine-tuning all parameters, the authors insert a small conditional branch that ingests a reference image and steers the video backbone (Wan 2.1–14B in the release) so the generated subject keeps consistent facial features.

Identity control hinges on two pieces: restricted self-attention that gates the influence of reference features, and conditional position mapping that aligns the reference embedding with frame locations. The framework learns from roughly 2,000 image–video pairs yet surpasses heavier baselines on face similarity and naturalness. Because the add-on is modular, the team shows it working with subject-driven video generation, community LoRAs, VACE pose control, stylization, and even experimental face swapping.

Links:

- Paper: <https://arxiv.org/abs/2508.07901>
 - Project page: <https://www.stand-in.tech>
 - GitHub: <https://github.com/WeChatCV/Stand-In>
 - Hugging Face weights: <https://huggingface.co/BowenXue/Stand-In>
-

Key ideas

- Conditional image branch: feeds a reference portrait through Stand-In’s adapters so the video model gets identity cues without replacing its base text-video path (paper abstract + README).
- Restricted self-attention with conditional position mapping: constrains attention to identity-relevant regions and aligns the reference to temporal positions for

stable facial structure (paper abstract).

- Tiny parameter overhead: training adds ~1% parameters relative to Wan 2.1 yet beats full-parameter methods on face similarity and naturalness metrics (README callout).
 - Data efficiency: identity adapters converge with about 2,000 paired samples, keeping compute manageable for custom subjects (paper abstract).
 - Task compatibility: the released toolkit covers subject-driven T2V, pose-referenced video generation via VACE, stylization with community LoRAs, and experimental face swapping (README usage + news log).
-

Model lineup & availability

- Stand-In v1.0 adapters (153 M parameters) targeting Wan2.1-14B-T2V.
- Hugging Face repo provides checkpoints, configs, and preprocessing scripts.
- Optional assets: official ComfyUI preprocessing node, sample inputs, and prompt presets.

The project page aggregates showcase videos; Hugging Face hosts the weights for quick download.

Quickstart (from repo docs)

Clone and set up:

```
git clone https://github.com/WeChatCV/Stand-In.git
cd Stand-In
pip install -r requirements.txt
```

Run identity-preserving generation:

```
python infer.py \
  --prompt "A man sits comfortably at a desk, chatting to camera." \
  --ip_image test/input/subject.jpg \
  --output test/output/subject.mp4
```

Load community style LoRAs alongside Stand-In:

```
python infer_with_lora.py \
  --prompt "A woman gives a studio interview in painterly lighting." \
  --ip_image test/input/subject.jpg \
  --lora_path path/to/style_lora.safetensors \
```

```
--lora_scale 1.0 \  
--output test/output/stylised.mp4
```

Face swapping (experimental):

```
python infer_face_swap.py \  
--prompt "A presenter reads tech headlines on set." \  
--ip_image test/input/identity.jpg \  
--output test/output/face_swap.mp4 \  
--denoising_strength 0.85
```

Integrate with VACE pose control:

```
python infer_with_vace.py \  
--prompt "A dancer performs a routine." \  
--vace_path checkpoints/VACE \  
--ip_image test/input/first_frame.png \  
--reference_video test/input/pose.mp4 \  
--output test/output/vace.mp4 \  
--vace_scale 0.8
```

Control & tuning tips

- Reference media: supply a high-resolution, front-facing image; the built-in preprocessing normalises size and format.
- Prompt discipline: keep subject descriptors generic (“a man”, “a woman”) if you want to preserve the reference identity; embellish the scene instead.
- Denoising strength: higher values redraw more background during face swap; lower values keep the original scene but may overfit facial texture.
- Stylization: combine Stand-In with external LoRAs via `infer_with_lora.py`; adjust scales to balance identity fidelity vs. look.
- VACE integration: match `vace_scale` and guidance weights to prevent pose control from overwhelming identity cues.

Practical production notes

- Hardware: latency mirrors Wan 2.1 inference; the extra adapters add minimal VRAM overhead, keeping workflows GPU-friendly.
- Dataset extension: fine-tune on your own 2k pair subset to adapt Stand-In to proprietary subjects or styles.

- ComfyUI support: use the official Stand-In preprocessing node for best results; third-party nodes may degrade alignment (README announcements).
 - Integration: plug adapters into existing Wan-based pipelines, including pose or depth control modules released after 18 August 2025.
 - Roadmap: upcoming releases target full ComfyUI support and cross-backbone compatibility; monitor the repo's news log.
-

References

- arXiv: <https://arxiv.org/abs/2508.07901>
- Project: <https://www.stand-in.tech>
- GitHub: <https://github.com/WeChatCV/Stand-In>
- Hugging Face: <https://huggingface.co/BowenXue/Stand-In>

Notes: Specs reflect the public repo as of 18 August 2025; check future releases for VACE presets, ComfyUI pipelines, and multi-backbone adapters.