DiTFlow

- □□□□□Video Motion Transfer with Diffusion Transformers
- □□□□□https://arxiv.org/abs/2412.07776
- CVPR 2025

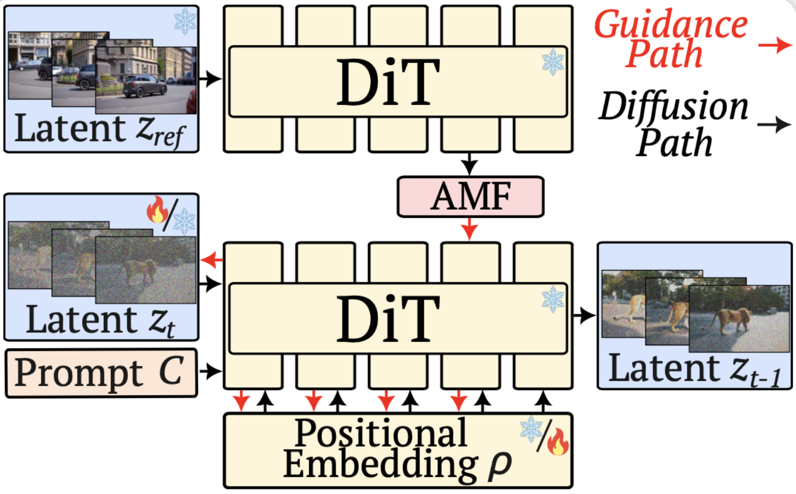

Figure 2. **Core idea of DiTFlow.** We extract the AMF from a reference video and we use that to guide the latent representation z_t towards the motion of the reference video. In our experiments, we also tested optimizing positional embeddings for improved zero-shot performance.

□□□□□□□□□□DiT□□□□□□□□DiTFlow□□□□□□□□□□□□□□□□□□□□□□□□□□□t=0□□□□DiT□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□DDIM□□□□□□□□□□□□□□□□□□□□□DiT block□□□□□□□□□□attention□□□□□□□□□□□□□□patch□□□□□□□□□□□□□□□□□□□□□□□□□□□□□AMF□□□□□□□□□□□□□□□□□□□attention□□□□□argmax□□□□□□□□□patch□□□□□□□□□□□□□AMF□□gt□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□AMF□□□attention□□□□□□□□□□□□□□□□□□□□□□□□zt□□□□□□□□□□□□□□□□□□□□□

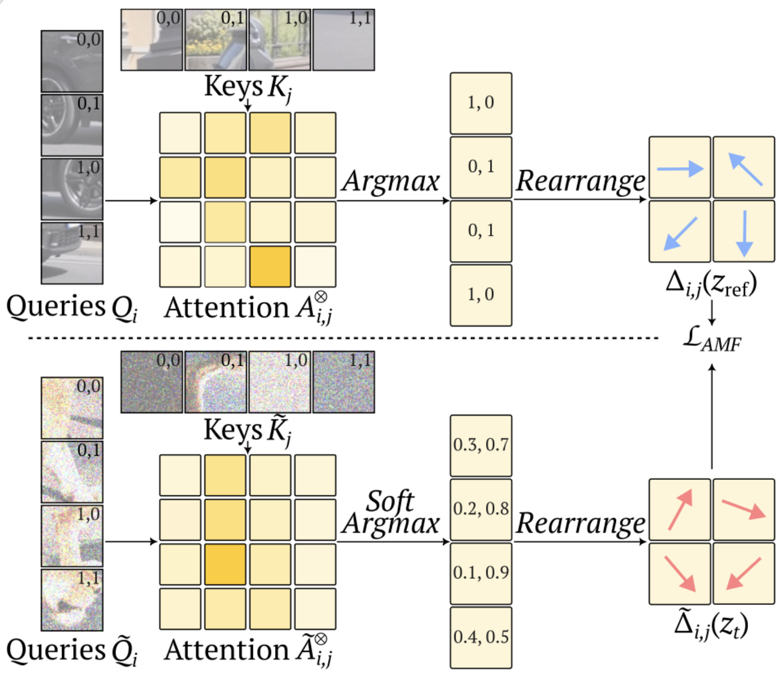

Figure 3. **Guidance.** We compute the reference displacement by processing cross-frame attentions with an argmax operation and rearranging them into displacement maps, identifying patch-aware cross-frame relationships. For video synthesis, we do the same operation with a soft argmax to preserve gradients, and impose

reconstruction with the reference displacement.

□□□□□zt□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□prompt□□□□□□□□□□□□□□□□□

- □□□□CogVideoX
- □□□□□□□□
- □□□MF□□□□□□□□□□□□□□□□□□IQ□□□CLIP□□□□□□□□□□□□□□
- □□□1 A40
- □□□https://ditflow.github.io

Newer | Older

2025-06-05
AnalysisAttentionVDiT

2025-06-04
FreeTraj