Video Motion Transfer with Diffusion Transformers

Supplementary Material

We strongly encourage readers to check the qualitative video samples in the project page at ditflow.github.io. Here, we provide additional elements for easing the understanding of our work. Specifically, we first provide implementation details in Section A and further ablations in Section B. Then, we provide additional reasoning about alternative strategies for supervision (Section C) and propose a simple experiment to justify AMF in Section D. Further MotionClone evaluation is provided in Section E. Finally, we discuss limitations (Section F).

A. Implementation

Positional embedding training details CogVideoX-5B uses a different positional embedding mechanism to CogVideoX-2B. CogVideoX-2B uses 3D sinusoidal embeddings similar to [54] and these are simply added to the tokens to provide absolute positional information. During guidance, gradients can backpropagate from the AMF loss at block 15 to these embeddings. CogVideoX-5B uses 3D rotary positional embeddings (RoPE [52]) that are embedded into all queries and keys at each attention block. Gradients still backpropagate from block 20 to the RoPE applied to all previous blocks.

Dataset We provide a sample of the dataset in Table 4. Video names are the same as those used in the DAVIS dataset [39]. Please refer to the supplementary material included in the project page for the full dataset and visuals.

B. Additional Ablation Studies

We conducted further ablation studies on temperature (τ), number of guidance blocks, and optimization algorithm in Table 3, following the experimental setup in Section 5.4. The temperature ablation reveals that setting $\tau = 5$ yields a marginal improvement in both MF and IQ metrics. Importantly, the performance varies only slightly across different temperature values, demonstrating DiTFlow's robustness to this hyperparameter. We also compare our original single-block approach (using only block 20) against a multi-block configuration (blocks 20+15+10). While DiTFlow benefits slightly from multi-block guidance, we opted for the single-block approach in our main experiments due to the additional computational overhead associated with multi-block configurations. Finally, we compare results for three different optimizers and find that Adam has the best balance between motion transfer and quality.

Temperature	MF ↑	IQ ↑
1	0.763	0.313
2	0.797	0.313
5	**0.799**	**0.317**
10	0.777	0.315

(a) Temperature

Blocks	MF ↑	IQ ↑
20	0.797	**0.313**
10,15,20	**0.804**	**0.313**

(b) Multi-block setup

Optimizer	MF ↑	IQ↑
Adam	0.797	0.313
AdamW	**0.803**	0.311
SGD	0.623	**0.320**

(c) Optimizer algorithm

Table 3. **Further ablations on CogVideoX-5B.**

C. Nearest neighbor alternatives

An alternative signal for AMF construction could have been the usage of nearest neighbors on noisy latents, as in related works [14]. In Figure 8, we visualize correspondences extracted between two frames using this technique and compare it to our AMF displacement. We demonstrate a much smoother displacement map, which can lead to better guidance on the rendered video.

D. Justification of AMF experiment

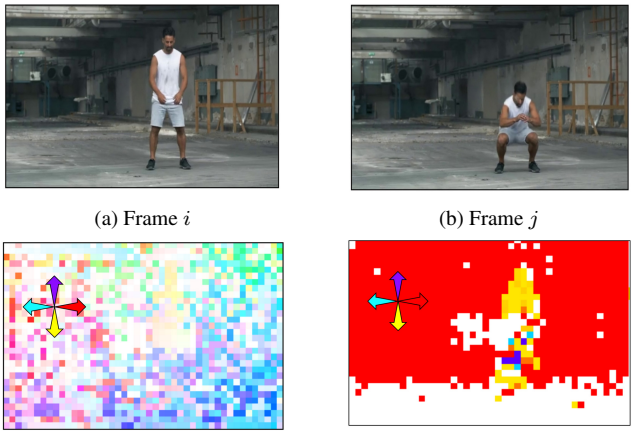
We conduct a small-scale study on 14 videos (from Section 5.4), where we move the content of each video's first frame in a random direction. We calculate the AMF of each video. If the motion of content is correctly captured, the AMF vectors should point in the direction of the introduced motion vector. We calculate the patch-wise cosine similarity between AMF and ground truth motion. We obtain 0.857 for CogVideoX-2B and 0.734 for CogVideoX-5B. The lower bound is 0.5 if random directions are predicted. This proves that *AMF is a valid signal for capturing motion*, which also aligns with the AMF visualisation in Figure 8. The superior performance of the 5B model can also be attributed to its better motion representations.

E. Injection baseline

We provide the qualitative results of our Injection baseline in Figure 9. It is able to transfer coarse subject location information while deviating significantly from the reference motion. For instance, in the Subject example, the bear walks in the opposite direction.

F. Limitations

As seen in previous methods [62], generations are still limited to the pre-trained video generator, so it has difficulty transferring motion with prompts or motions that are out of distribution. For example, complex body movements (*e.g. backflips*) still remain a difficult task for these models. Moreover, we highlight that motion transfer is inherently ambiguous if not associated to prompts. For example,



(a) Frame i (b) Frame j

(c) Latent nearest neighbour (d) AMF displacement $\Delta_{i,j}$

Figure 8. **Displacement maps of squat motion.** We visualise the displacement map between frames (a) and (b) computed on latents. The displacements are mapped to colours according to the colour wheel arrows shown. Taking the latent nearest neighbour [14] in (c) results in very noisy displacements with poor matching of content between frames. The AMF displacement in (d) captures the downwards (yellow) motion of the person and rightwards (red) motion of the panning camera better.

transferring the motion of a dog to a plane may risk to map motion features of other elements in the scene to the plane in the rendered video, even with KV injection. For future work, we believe it will be important to associate specific semantic directions (e.g. dog $\mapsto$ plane) to constrain editing, similar to what happens in inversion-based editing [35].

The pairwise nature of AMF does lead to slightly more memory consumption compared to previous methods. However, we found performance to be consistent across setups with different number of frames. Long video generation approaches that generate a smaller set of frames at a time may readily be applied to our motion transfer approach.

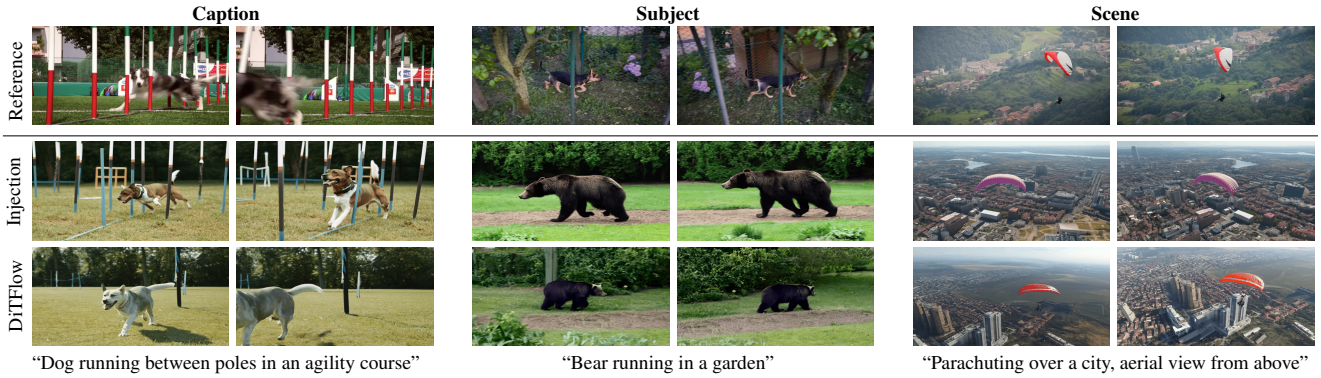


Figure 9. **Injection baseline.** Results are provided in the same setup as Figure 4.

Video	Caption	Subject	Scene
blackswan	A black swan swimming in a river	A duck with a tophat swimming in a river	A paper boat floating in a bathtub
car-turn	A gray car with black tires driving on a road in a forest	A man with a black top running on a road in a forest, camera shot from a distance	Black suv with tinted windows driving through a roundabout in a bustling city, surrounded by tall buildings and bright lights
car-roundabout	A gray mini cooper driving around a roundabout in a town	A man riding a unicycle around a roundabout in a town	A lion walking through a bustling roundabout, surrounded by vibrant city life
libby	Dog running in a garden	Bear running in a garden	Plane flying through the sky above the clouds
bus	Aerial view of bus driving on a street	Aerial view of red ferrari driving on a street	Closeup aerial view of an ant crawling in a desert
camel	A camel walking in a zoo	A giraffe walking in a zoo	A blue Sedan car turning into a driveway
bear	A bear walking on the rocks	A giraffe walking on the rocks	A giraffe walking in the zoo
bmx-bumps	BMX rider biking up a sandy hill	Black Jeep driving up a sandy hill	Black Jeep driving up a hill in a bustling city
bmx-trees	Kid with white shirt riding a bike up a hill, seen from afar, long-distance view	Leopard running up a grassy hill	Leopard running up a snowy hill in a forest
boat	Fishing boat sails through the sea in front of an island, close-up, medium shot, elevated camera angle, wide angle view	Black yacht sails through the sea in front of an island	Black yacht sails through the sea in front of a bustling city

Table 4. **Dataset snippet.** Sample of DAVIS videos chosen with associated prompts from each category described in Section 5.