Object-Centric Diffusion for Efficient Video Editing - Supplementary Material

Kumara Kahatapitiya*, Adil Karjauv, Davide Abati, Fatih Porikli,
Yuki M. Asano, and Amirhossein Habibian

Qualcomm AI Research†
{kkahatap,akarjauv,dabati,fporikli,asano,ahabibian}@qti.qualcomm.com

This supplementary is organized into two sections. First, in Sec. A, we present additional discussion on off-the-shelf-optimizations and benchmark settings. In Sec. B, we present additional results, both qualitative and quantitative. Project page: qualcomm-ai-research.github.io/object-centric-diffusion.

A Additional Discussion

A.1 Off-the-shelf optimizations of ToMe

Pairing token locations from inversion Many inversion-based image/video editing pipelines rely on sharing attention maps between inversion and generation stages (*e.g.* FateZero [8], Plug-and-Play [12]). As such, when applying ToMe [2, 3], it is important that locations of destination (*dst*) and unmerged (*unm*) tokens are paired in the two stages, at each corresponding attention layer and diffusion step. If that is not the case, tokens or attention maps coming from inversion are not compatible with the ones available at generation time. In practice, we compute which tokens to be merged during inversion, and merge the tokens at the same locations in generation attention maps. By doing so, we make sure the tokens that remain after merging correspond to the same locations (or, token indices), and hence, the attention maps from inversion and generation can rightly be fused. We found this strategy to be of primary importance, as testified by Fig. 3 (d-e) in the main paper.

Re-sampling destination tokens per-frame ToMe for stable diffusion [3] samples *dst* tokens randomly in a single image. When extending this to multiple frames, we initially sample the same random locations in each frame, finding this strategy to be sub-optimal. Instead, if we re-sample different random locations in each frame (or, in each temporal window in our spatio-temporal implementation), it allows us to preserve different information in each frame (or, window) after merging. We found this to be beneficial, especially at higher merging rates (*e.g.* see Fig. 3 (e to f) in the main paper).

How to search for destination match In the original ToMe for stable diffusion [3], for each source (*src*) token, we search its corresponding match within a

[*] Work completed during an internship at Qualcomm Technologies, Inc.

[†] Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

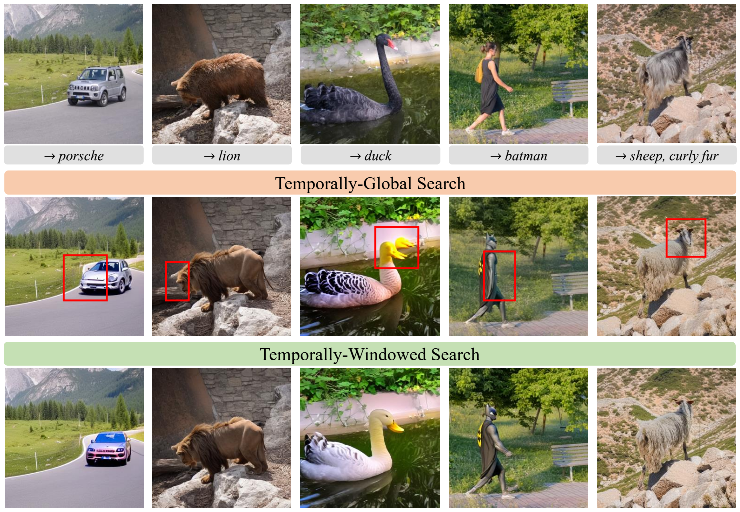

Fig. A.1: Qualitative ablation on searching *dst* match Each *src* token may search for its corresponding match within a pool of *dst* tokens. This pool can be comprised of either the whole spatio-temporal extent (i.e., all frames) as in *Temporally-Global Search*, or only the same temporal window as the corresponding *src* token as in *Temporally-Windowed Search*. Among these two strategies, the latter allows more flexibility, providing more-consistent generations with better fidelity.

pool of *dst* tokens coming from the full spatial extent of the given image $(H \times W)$. The naive transposition of this strategy to our video use-case allows, for any *src* token, its candidate match to be searched within a pool of *dst* tokens coming from the full spatio-temporal extent of the video $(T \times H \times W)$. We find that this strategy for searching *dst* match, named hereby *Temporally-Global Search*, can lead to generation artifacts. Differently, we consider restricting the temporal extent of the *dst* pool to be the same temporal-window $(s_t \times H \times W)$ as the *src* token, as in our *Temporally-Windowed Search*. As shown in Fig. A.1, the latter gives better reconstructions in general, whilst allowing more flexibility to control where merges are happening, temporally. This way, the user can also better trade-off the temporal-redundancy, smoothness and consistency by controlling the spatio-temporal window size.

Merging queries, keys or values? In our early experiments, we consider applying ToMe to all queries (with unmerging, as in [3]), keys and values. We however find that, with extreme reduction rates, merging queries can easily break the reconstructions. As such, we limit ToMe to operate on keys and values only. We also observe that in dense cross-frame attention modules, merging queries only provide a slight latency reduction.

Capped merging in low-res UNet stages As observed in [3], the high resolution UNet [11] stages are the most expensive in terms of self-attention (or, cross-frame attention) modules, and the ones that can benefit the most by applying ToMe. Contrarily to the original formulation which does not optimize low-resolution layers, we do apply ToMe in all layers as we observe it has a meaningful impact on latency. We however cap the minimum #tokens preserved after merging in low-resolution layers, in order to avoid degenerate bottlenecks (*e.g.* collapsing to a single representation). Specifically, we maintain at-least 4 and 16 tokens per-frame after merging at 8×8 and 16×16 resolutions, respectively.

A.2 Benchmark settings

Evaluation metrics We consider two metrics for quantitative evaluation: CLIP-score and Temporal-consistency, similar to prior work [8,15]. CLIP-score is computed as the cosine similarity between CLIP [9] visual embedding of each frame and CLIP text embedding of the corresponding edit prompt, aggregated over all frames and sequences. It measures the semantic fidelity of the generated video. Temporal-consistency is computed as the cosine similarity between the CLIP visual embeddings of each consecutive pairs of frames, aggregated over all pairs and sequences. It conveys the visual quality of the generated video, measuring how temporally coherent frames are. We highlight that, despite their use is due in fair comparisons due to their popularity, both these fidelity metrics are far from perfect. For instance, we find the CLIP score to be sensitive to global semantic discrepancies, yet it often overlooks generation artifacts and smaller pixel-level details. Furthermore, Temporal-Consistency can be simply exploited by a fake video repeating a frame over time. For these reasons, extensive visual comparisons are still required to assess different models, and future research should be encouraged towards more informative quantitative protocols for video editing.
Sequence-prompt pairs We present the sequence-prompt pairs considered in our evaluation of inversion-based pipelines in Table A.1. Most sequences here are from DAVIS [7] dataset, with the exception of a few in-the-wild videos introduced in [8]. The Benchmark setting corresponds to the original quantitative evaluation of FateZero [8], which includes 9 sequence-prompt pairs. We also present the sequence-prompt pairs used to evaluate our Object-Centric Sampling (see Table 5 in the main paper), categorized based on the foreground object size: Large, Medium and Small. In Table A.2, we show the 125 sequence-prompt pairs used in ControlNet-based pipelines, provided by the authors of ControlVideo [15].

Table A.1: Sequence-prompt pairs used to evaluate inversion-based pipelines: Most sequences here are from DAVIS [7], except for a few in-the-wild videos used in [8]. The Benchmark pairs correspond to the original FateZero [8] quantitative evaluation setting. We also show the sequence-prompt pairs used to evaluate our Object-Centric Sampling, separately for Large, Medium and Small objects.

	Sequence	Source prompt	Target prompts
FateZero [8] Benchmark	blackswan	a black swan with a red beak swimming in a river near a wall and bushes.	– a white duck with a yellow beak swimming in a river near a wall and bushes. – a pink flamingo with a red beak walking in a river near a wall and bushes. – a Swarovski crystal swan with a red beak swimming in a river near a wall and bushes. – cartoon photo of a black swan with a red beak swimming in a river near a wall and bushes.
	car-turn	a silver jeep driving down a curvy road in the countryside.	– a Porsche car driving down a curvy road in the countryside. – watercolor painting of a silver jeep driving down a curvy road in the countryside.
	kite-surf	a man with round helmet surfing on a white wave in blue ocean with a rope.	– a man with round helmet surfing on a white wave in blue ocean with a rope in the Ukiyo-e style painting.
	train (in-the-wild)	a train traveling down tracks next to a forest filled with trees and flowers and a man on the side of the track.	– a train traveling down tracks next to a forest filled with trees and flowers and a man on the side of the track in Makoto Shinkai style.
	rabbit (in-the-wild)	a rabbit is eating a watermelon.	– pokemon cartoon of a rabbit is eating a watermelon.
Large object	blackswan	a black swan with a red beak swimming in a river near a wall and bushes.	– a white duck with a yellow beak swimming in a river near a wall and bushes. – a pink flamingo with a red beak walking in a river near a wall and bushes. – a Swarovski crystal swan with a red beak swimming in a river near a wall and bushes.
	bear	a brown bear walking on the rock against a wall.	– a red tiger walking on the rock against a wall. – a yellow leopard walking on the rock against a wall.
Medium object	breakdance	a man wearing brown tshirt and jeans doing a breakdance flare on gravel.	– a woman with long-hair wearing green-sweater and jeans doing a breakdance flare on gravel. – a spiderman wearing red-blue spidersuit doing a breakdance flare on gravel. – a chimpanzee wearing a black jeans doing a breakdance flare on gravel.
	boat	a white color metal boat cruising in a lake near coast.	– a heavily-rusted metal boat cruising in a lake near coast. – a light-brown color wooden boat cruising in a lake near coast.
	car-turn	a silver jeep driving down a curvy road in the countryside.	– a Porsche car driving down a curvy road in the countryside.
Small object	mallard	a brown mallard running on grass land close to a lake.	– a white duck running on grass land close to a lake. – a golden chicken running on grass land close to a lake. – a gray goose running on grass land close to a lake.
	lucia	a woman wearing a black dress with yellow handbag walking on a pavement.	– a woman wearing a white pant-suit with yellow handbag walking on a pavement. – a woman wearing a black dress and a hat with red handbag walking on a pavement. – a batman wearing a black bat-suit walking on a pavement.
	soapbox	two men driving a soapbox over a ramp.	– two robots driving a mars-rover over a ramp.

Table A.2: Sequence-prompt pairs used to evaluate ControlNet-based pipelines: All sequences are from DAVIS [7]. These pairs correspond to the original ControlVideo [15] quantitative evaluation setting. [continued...]

Sequence	Source prompt	Target prompts
blackswan	a black swan moving on the lake.	– A black swan moving on the lake – A white swan moving on the lake. – A white swan moving on the lake, cartoon style. – A crochet black swan swims in a pond with rocks and vegetation. – A yellow duck moving on the river, anime style.
boat	a boat moves in the river.	– A sleek boat glides effortlessly through the shimmering river, van gogh style. – A majestic boat sails gracefully down the winding river. – A colorful boat drifts leisurely along the peaceful river. – A speedy boat races furiously across the raging river. – A rustic boat bobs gently on the calm and tranquil river.
breakdance-flare	a man dances on the road.	– A young man elegantly dances on the deserted road under the starry night sky. – The handsome man dances enthusiastically on the bumpy dirt road, kicking up dust as he moves. – A man gracefully dances on the winding road, surrounded by the picturesque mountain scenery. – The athletic man dances energetically on the long and straight road, his sweat glistening under the bright sun. – The talented man dances flawlessly on the busy city road, attracting a crowd of mesmerized onlookers.
bus	a bus moves on the street.	– A big red bus swiftly maneuvers through the crowded city streets. – A sleek silver bus gracefully glides down the busy urban avenue. – A colorful double-decker bus boldly navigates through the bustling downtown district. – A vintage yellow bus leisurely rolls down the narrow suburban road. – A modern electric bus silently travels along the winding coastal highway.
camel	a camel walks on the desert.	– A majestic camel gracefully strides across the scorching desert sands. – A lone camel strolls leisurely through the vast, arid expanse of the desert. – A humpbacked camel plods methodically across the barren and unforgiving desert terrain. – A magnificent camel marches stoically through the seemingly endless desert wilderness. – A weathered camel saunters across the sun-scorched dunes of the desert, its gaze fixed on the horizon.
car-roundabout	a jeep turns on a road.	– A shiny red jeep smoothly turns on a narrow, winding road in the mountains. – A rusty old jeep suddenly turns on a bumpy, unpaved road in the countryside. – A sleek black jeep swiftly turns on a deserted, dusty road in the desert. – A modified green jeep expertly turns on a steep, rocky road in the forest. – A gigantic yellow jeep slowly turns on a wide, smooth road in the city.
car-shadow	a car moves to a building.	– A sleek black car swiftly glides towards a towering skyscraper. – A shiny silver vehicle gracefully maneuvers towards a modern glass building. – A vintage red car leisurely drives towards an abandoned brick edifice. – A luxurious white car elegantly approaches a stately colonial mansion. – A rusty blue car slowly crawls towards a dilapidated concrete structure.
car-turn	a jeep on a forest road.	– A shiny silver jeep was maneuvering through the dense forest, kicking up dirt and leaves in its wake. – A dusty old jeep was making its way down the winding forest road, creaking and groaning with each bump and turn. – A sleek black jeep was speeding along the narrow forest road, dodging trees and rocks with ease. – A massive green jeep was lumbering down the rugged forest road, its powerful engine growling as it tackled the steep incline. – A rusty red jeep was bouncing along the bumpy forest road, its tires kicking up mud and gravel as it went.
cows	a cow walks on the grass.	– A spotted cow leisurely grazes on the lush, emerald-green grass. – A contented cow ambles across the dewy, verdant pasture. – A brown cow serenely strolls through the sun-kissed, rolling hills. – A beautiful cow saunters through the vibrant, blooming meadow. – A gentle cow leisurely walks on the soft, velvety green grass, enjoying the warm sunshine.

Table A.2: Sequence-prompt pairs used to evaluate ControlNet-based pipelines: All sequences are from DAVIS [7]. These pairs correspond to the original ControlVideo [15] quantitative evaluation setting. [continued...]

Sequence	Source prompt	Target prompts
dog	a dog walks on the ground.	– A fluffy brown dog leisurely strolls on the grassy field. – A scruffy little dog energetically trots on the sandy beach. – A majestic black dog gracefully paces on the polished marble floor. – A playful spotted dog joyfully skips on the leaf-covered path. – A curious golden dog curiously wanders on the rocky mountain trail.
elephant	an elephant walks on the ground.	– A massive elephant strides gracefully across the dusty savannah. – A majestic elephant strolls leisurely along the lush green fields. – A mighty elephant marches steadily through the rugged terrain. – A gentle elephant ambles peacefully through the tranquil forest. – A regal elephant parades elegantly down the bustling city street.
flamingo	a flamingo wanders in the water.	– A graceful pink flamingo leisurely wanders in the cool and refreshing water, its slender legs elegantly stepping on the soft sand. – A vibrant flamingo casually wanders in the clear and sparkling water, its majestic wings spread wide in the sunshine. – A charming flamingo gracefully wanders in the calm and serene water, its delicate neck curving into an elegant shape. – A stunning flamingo leisurely wanders in the turquoise and tranquil water, its radiant pink feathers reflecting the shimmering light. – A magnificent flamingo elegantly wanders in the sparkling and crystal-clear water, its striking plumage shining brightly in the sun.
gold-fish	golden fishers swim in the water.	– Majestic golden fishers glide gracefully in the crystal-clear waters. – Brilliant golden fishers swim serenely in the shimmering blue depths. – Glittering golden fishers dance playfully in the glistening aquamarine waves. – Gleaming golden fishers float leisurely in the peaceful turquoise pools. – Radiant golden fishers meander lazily in the tranquil emerald streams.
hike	a man hikes on a mountain.	– A rugged man is trekking up a steep and rocky mountain trail. – A fit man is leisurely hiking through a lush and verdant forest. – A daring man is scaling a treacherous and jagged peak in the alpine wilderness. – A seasoned man is exploring a remote and rugged canyon deep in the desert. – A determined man is trudging up a snowy and icy mountain slope, braving the biting cold and fierce winds.
hockey	a player is playing hockey on the ground.	– A skilled player is furiously playing ice hockey on the smooth, glistening rink. – A young, agile player is energetically playing field hockey on the lush, green grass. – An experienced player is gracefully playing roller hockey on the sleek, polished pavement. – A determined player is passionately playing street hockey on the gritty, urban asphalt. – A talented player is confidently playing air hockey on the fast-paced, neon-lit table.
kite-surf	a man is surfing on the sea.	– A muscular man is expertly surfing the gigantic waves of the Pacific Ocean. – A handsome man is gracefully surfing on the crystal clear waters of the Caribbean Sea. – A daring man is fearlessly surfing through the dangerous, choppy waters of the Atlantic Ocean. – An athletic man is skillfully surfing on the wild and untamed waves of the Indian Ocean. – A young man is confidently surfing on the smooth, peaceful waters of a serene lake.
lab-coat	three women stands on the lawn.	– Three stunning women are standing elegantly on the lush green lawn, chatting and laughing. – Three young and vibrant women are standing proudly on the well-manicured lawn, enjoying the sunshine. – Three fashionable women in colorful dresses are standing gracefully on the emerald green lawn, taking selfies. – Three confident women with radiant smiles are standing tall on the soft, green lawn, enjoying the fresh air. – Three beautiful women, each dressed in their own unique style, are standing on the lush and verdant lawn, admiring the scenery.

Table A.2: Sequence-prompt pairs used to evaluate ControlNet-based pipelines: All sequences are from DAVIS [7]. These pairs correspond to the original ControlVideo [15] quantitative evaluation setting.

Sequence	Source prompt	Target prompts
longboard	a man is playing skateboard on the alley.	– A young man is skillfully skateboarding on the busy city street, weaving in and out of the crowds with ease. – An experienced skateboarder is fearlessly gliding down a steep, curvy road on his board, executing impressive tricks along the way. – A daring skater is performing gravity-defying flips and spins on his board, effortlessly navigating through a challenging skatepark course. – A talented skateboarder is carving up the smooth pavement of an empty parking lot, creating beautiful patterns with his board and body. – A passionate skater is practicing his moves on a quiet neighborhood street, with the sound of his board echoing through the peaceful surroundings.
mallard-water	a mallard swims on the water.	– A vibrant mallard glides gracefully on the shimmering water. – A beautiful mallard paddles through the calm, blue water. – A majestic mallard swims elegantly on the tranquil lake. – A striking mallard floats effortlessly on the sparkling pond. – A colorful mallard glides smoothly on the rippling surface of the water.
mbike-trick	a man riding motorbike.	– A young man riding a sleek, black motorbike through the winding mountain roads. – An experienced man effortlessly riding a powerful, red motorbike on the open highway. – A daring man performing gravity-defying stunts on a high-speed, blue motorbike in an empty parking lot. – A confident man cruising on a vintage, yellow motorbike along the picturesque coastal roads. – A rugged man maneuvering a heavy, dusty motorbike through the rugged terrain of a desert.
rhino	a rhino walks on the rocks.	– A massive rhino strides confidently across the jagged rocks. – A majestic rhino gracefully navigates the rugged terrain of the rocky landscape. – A powerful rhino marches steadily over the rough and rocky ground. – A colossal rhino plods steadily through the craggy rocks, undeterred by the challenging terrain. – A sturdy rhino confidently traverses the treacherous rocks with ease.
surf	a sailing boat moves on the sea.	– A graceful sailing boat glides smoothly over the tranquil sea. – A sleek sailing boat cuts through the shimmering sea with ease. – A majestic sailing boat cruises along the vast, azure sea. – A vintage sailing boat bobs gently on the calm, turquoise sea. – A speedy sailing boat races across the glistening, open sea.
swing	a girl is playing on the swings.	– A young girl with pigtails is joyfully swinging on the colorful swings in the playground. – The little girl, giggling uncontrollably, is happily playing on the old-fashioned wooden swings. – A blonde girl with a big smile on her face is energetically playing on the swings in the park. – The girl, wearing a flowery dress, is gracefully swaying back and forth on the swings, enjoying the warm breeze. – A cute little girl, dressed in a red coat, is playfully swinging on the swings, her hair flying in the wind.
tennis	a man is playing tennis.	– The skilled man is effortlessly playing tennis on the court. – A focused man is gracefully playing a game of tennis. – A fit and agile man is playing tennis with precision and finesse. – A competitive man is relentlessly playing tennis with his opponent. – The enthusiastic man is eagerly playing a game of tennis, sweat pouring down his face.
walking	a selfie of walking man.	– A stylish young man takes a selfie while strutting confidently down the busy city street. – An energetic man captures a selfie mid-walk, showcasing his adventurous spirit. – A happy-go-lucky man snaps a selfie as he leisurely strolls through the park, enjoying the sunny day. – A determined man takes a selfie while briskly walking towards his destination, never breaking stride. – A carefree man captures a selfie while wandering aimlessly through the vibrant cityscape, taking in all the sights and sounds.

B Additional results

B.1 Qualitative comparisons

We show additional qualitative comparisons for inversion-based pipelines in Fig. B.1. Here, we mainly focus on shape editing, and present multiple edited frames of each sequence using FateZero [8], Tune-A-Video [13], TokenFlow [4] and Frame SDEdit [6]. Tune-A-Video requires 1-shot finetuning on a given sequence, whereas TokenFlow and Frame SDEdit are based on stable diffusion [10] checkpoints. FateZero and our implementation rely on Tune-A-Video checkpoints for shape editing, without needing any further finetuning for their respective proposed improvements. Frame SDEdit shows no consistency among frames, being an image editing pipeline. Among video editing pipelines, ours show the best fidelity and temporal-consistency, while also generating outputs faster (see latency measurements given in Table 1 and Fig. 5 in the main paper). Notably, thanks to Object-Centric Sampling, our pipeline gives more-faithful background reconstructions, as such regions are expected to be un-edited based on the given shape editing prompts.

In Fig. B.2, we show additional qualitative comparisons for ControlNet-based pipelines sucha as ControlVideo [15] and Text2Video-Zero [5]. Here, all methods are conditioned on Depth-maps, while using SD [10] checkpoints without further finetuning. OCD shows comparable performance with its baseline ControlVideo, while being significantly faster (see latency measurements given in Table 2 in the main paper). It is also more temporally-consistent compared to Text2Video-Zero which uses sparse instead of dense cross-frame attention, while having comparable latency.

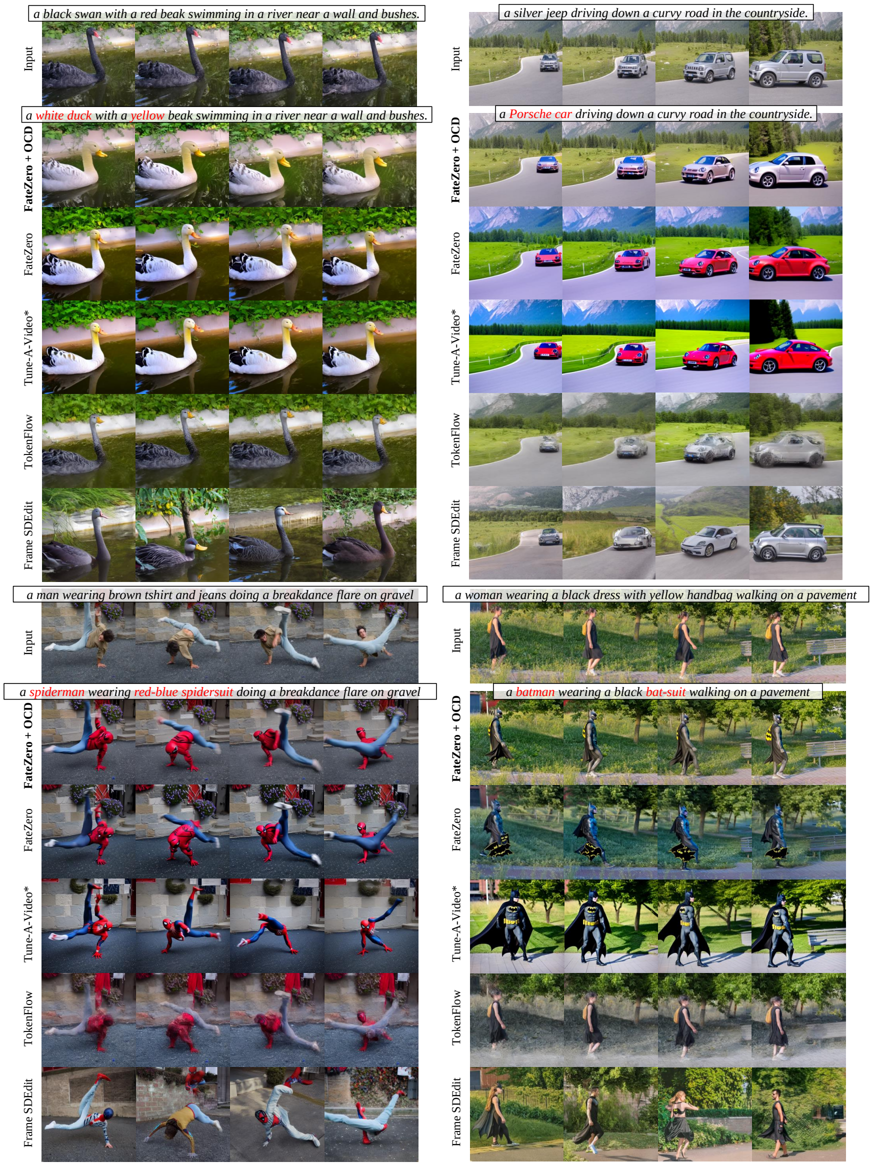
a black swan with a red beak swimming in a river near a wall and bushes.
Input
a white duck with a yellow beak swimming in a river near a wall and bushes.
FateZero + OCD
FateZero
Tune-A-Video*
TokenFlow
Frame SDEdit
a silver jeep driving down a curvy road in the countryside.
Input
a Porsche car driving down a curvy road in the countryside.
FateZero + OCD
FateZero
Tune-A-Video*
TokenFlow
Frame SDEdit
a man wearing brown tshirt and jeans doing a breakdance flare on gravel
Input
a spiderman wearing red-blue spidersuit doing a breakdance flare on gravel
FateZero + OCD
FateZero
Tune-A-Video*
TokenFlow
Frame SDEdit
a woman wearing a black dress with yellow handbag walking on a pavement
Input
a batman wearing a black bat-suit walking on a pavement
FateZero + OCD
FateZero
Tune-A-Video*
TokenFlow
Frame SDEdit

Fig. B.1: Qualitative comparison on *blackswan*, *car-turn*, *breakdance-flare* and *lucia* sequences [7]: We show shape editing results of our method (Optimized-FateZero + OCD), in comparison with FateZero [8], Tune-A-Video [13], TokenFlow [4] and SDEdit [6]. Our results show better semantic quality (*e.g.* alignement with target prompt) and visual fidelity (*e.g.* temporal consistency, faithful background), while also being more efficient (Table 1 in the main paper). Best viewed zoomed-in.

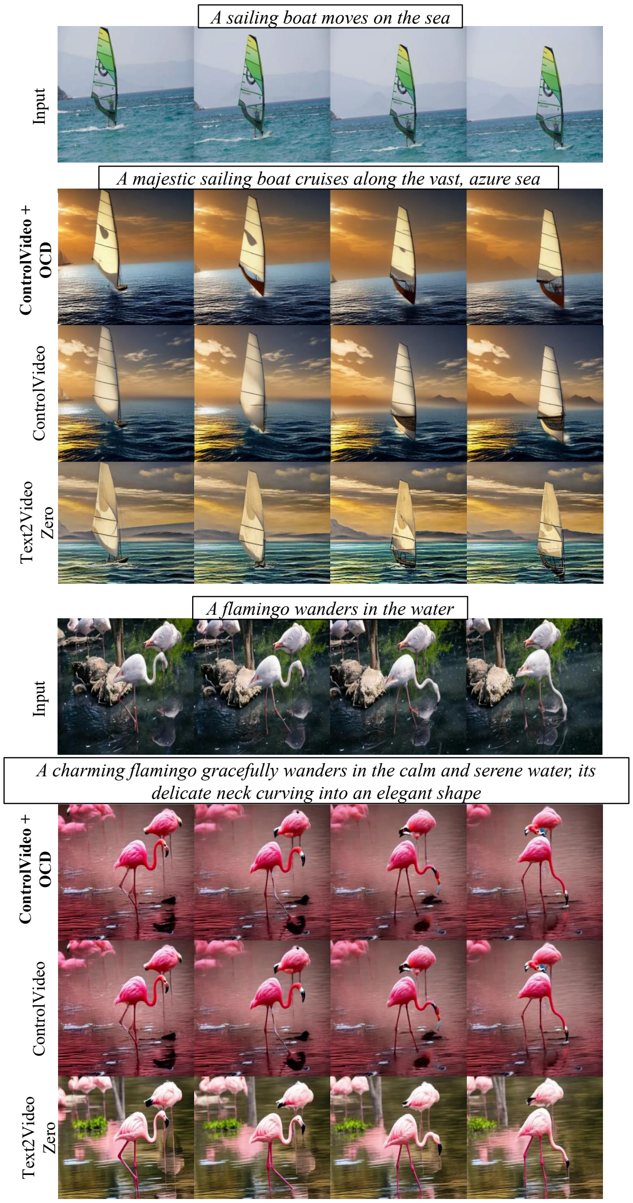
A sailing boat moves on the sea
Input
A majestic sailing boat cruises along the vast, azure sea
ControlVideo + OCD
ControlVideo
Text2Video Zero
A flamingo wanders in the water
Input
A charming flamingo gracefully wanders in the calm and serene water, its delicate neck curving into an elegant shape
ControlVideo + OCD
ControlVideo
Text2Video Zero

Fig. B.2: Qualitative comparison on *surf* and *flamingo* sequences [7]: We show shape editing results of our method (Optimized-ControlVideo + OCD), in comparison with ControlVideo [15] and Text2Video-Zero [5]. All methods use Depth conditioning. Our results show comparable quality with baseline ControlVideo while being significantly faster (Tab 2 in the main paper), and better temporal consistency compared to Text2Video-Zero. Best viewed zoomed-in.

Table B.1: Additional FateZero [8] baselines: We report CLIP metrics of fidelity (Temporal-consistency, CLIP-score) and latency (inversion, generation, UNet [11] time). The difference between inversion and UNet time corresponds to other overheads, dominated by memory access. Fewer diffusion steps (*e.g.* 5) and pooling operations can also gain significant speed-ups, but break reconstructions (not always visible in fidelity metrics).

Model	CLIP metrics ↑		Latency (s) ↓		
	Tem-con	Cl-score	Inv	Gen	UNet
FateZero [8]	0.961	0.344	135.80	41.34	20.63
+ diff. steps=5	0.968	0.306	14.84	4.98	2.17
+ diff. steps=20	0.961	0.341	61.82	18.41	8.03
+ avg. pool (4,4)	0.958	0.335	9.91	11.80	7.08
+ max pool (4,4)	0.959	0.275	9.96	11.91	6.91
Optimzed-FateZero	0.966	0.334	9.54	10.14	7.79
+ OCD	0.967	0.331	8.22	9.29	6.38

B.2 Quantitative comparisons

Other baselines for latency reduction We discuss simple baselines for reducing latency in Table B.1 and Table B.2. We report both fidelity (Temporal-consistency, CLIP-score) and latency (inversion, generation, UNet [11] time). Here, UNet time corresponds to just running UNet inference without any additional overheads (*e.g.* memory access), which we use to highlight the cost of such overheads. For ControlVideo [15] baselines, we show results with either Depth or Canny-edge conditioning.

In both inversion and ControlNet-based settings, we devise our optimized-baselines by reducing diffusion steps 50→20 and applying token merging, which give reasonable reconstructions. This is our default starting point for implementing OCD. Going further, we also consider diff. steps=5, which fails to retain details in reconstructions. Instead of token merging, we can also apply pooling strategies on key-value tokens. Despite giving similar speed-ups, these result in sub-par performance compared to ToMe, especially in shape editing (although not always captured in quantitative numbers). In ControlVideo setup, we can choose to do merging on both UNet and ControlNet [14] models, resulting in further speed-ups with a minimal drop in fidelity. We further observe that we can re-use the same control signal for multiple diffusion steps, allowing us to run ControlNet at a reduced rate (Reduced/Single inference in Table B.2).

Cost of memory access vs. computations Inversion-based editing pipelines rely on guidance from the inversion process during generation (*e.g.* based on latents [12] or attention maps [8]). When running inference, such features need to be stored (which may include additional GPU→RAM transfer), accessed and reused. This cost can be considerable, especially for resource-constrained hardware. This cost measured in latency, is shown in Table B.1, as the difference between inversion and UNet times. Alternatively, it can also be seen as the stor-

Table B.2: Additional ControlVideo [15] baselines: We report CLIP metrics of fidelity (Temporal-consistency, CLIP-score) with either Depth or (Canny-edge) conditioning, and latency (generation, UNet [11] time). The difference between generation and UNet time corresponds to other overheads, dominated by ControlNet [14]. Fewer diffusion steps (*e.g.* 5) and pooling operations can also gain significant speed-ups, but break reconstructions (not always visible in fidelity metrics). We also observe that ControlNet inference need not be done at the same frequency as denoising, which can lead to further speed-ups.

Model	CLIP metrics ↑		Latency (s) ↓	
	Tem-con	Cl-score	Gen	UNet
ControlVideo [15]	0.972 (0.968)	0.318 (0.308)	152.64	137.68
+ diff. steps=5	0.978 (0.971)	0.309 (0.295)	19.58	13.58
+ diff. steps=20	0.978 (0.971)	0.316 (0.304)	64.61	54.83
+ avg.pool (2,2)	0.977 (0.968)	0.309 (0.295)	30.53	20.56
+ max.pool (2,2)	0.972 (0.973)	0.225 (0.212)	30.32	20.53
Optimized-ControlVideo	0.978 (0.972)	0.314 (0.303)	31.12	21.42
+ OCD	0.977 (0.967)	0.313 (0.302)	25.21	15.61
+ OCD (UNet, ControlNet)	0.976 (0.969)	0.306 (0.297)	25.13	15.41
+ ControlNet Red. Inf.	0.977 (0.968)	0.313 (0.301)	23.62	15.47
+ ControlNet Sin. Inf.	0.973 (0.964)	0.307 (0.293)	22.35	15.48

Table B.3: Memory requirement for attention maps: In FateZero [8] setting, we show additional baselines and the corresponding storage requirements which directly affect the memory-access overhead. FateZero stores attention maps of all UNet [11] blocks for all diffusion steps. Our contributions help reduce this cost. It can potentially enable attention maps to be kept on GPU memory itself (w/o having to move between GPU and RAM), further improving latency. Each float is stored in 16bits.

Model	Disk-space (GB) ↓
FateZero	74.54
+ diff. steps=5	7.45
+ diff. steps=20	29.82
+ pool (4,4)	3.06
Optimized-FateZero	5.05
+ OCD	4.22

age requirement as given in Table B.3. On FateZero [8], we observe that the storage cost is indeed significant, and affects the latency more than the computations. With OCD, we directly reduce the cost for attention computation, storage and access.

Expected savings of Object-Centric Sampling We run a control experiment to observe the expected latency reductions when using our Object-Centric Sampling, at different object sizes ($\varDelta$) and Blending step ratios (γ), given in Table B.4. Here, we consider hypothetical inputs, so that we can ablate the settings more-clearly. The baseline sees inputs of size $\varDelta = 64 \times 64$, and runs all diffusion steps at full-resolution ($\gamma = 1.0$). In our Object-Centric Sampling, we

Table B.4: Control experiment on Object-Centric Sampling: We evaluate the latency savings at different *hypothetical* object sizes ($\varDelta$) and blending step ratios (γ). The baseline is with $\varDelta = 64 \times 64$ and $\gamma = 1.0$ (with total 20 diffusion steps). We can get the most savings at a smaller object size and blending step ratio. It is worth noting that this control experiment does not correspond to actual sequence-prompt pairs, and is just intended to give the reader an idea about expected savings.

Blending steps ratio (γ)	Latency (s) @ #tokens ($\varDelta$) ↓							
	64×64		48×48		32×32		16×16	
	Inv	Gen	Inv	Gen	Inv	Gen	Inv	Gen
1.00	12.31	10.39	-	-	-	-	-	-
0.50	-	-	9.56	8.67	9.05	7.60	8.25	7.01
0.25	-	-	8.03	7.94	7.11	6.00	5.90	4.96
0.05	-	-	6.82	7.17	5.85	4.99	4.43	3.80

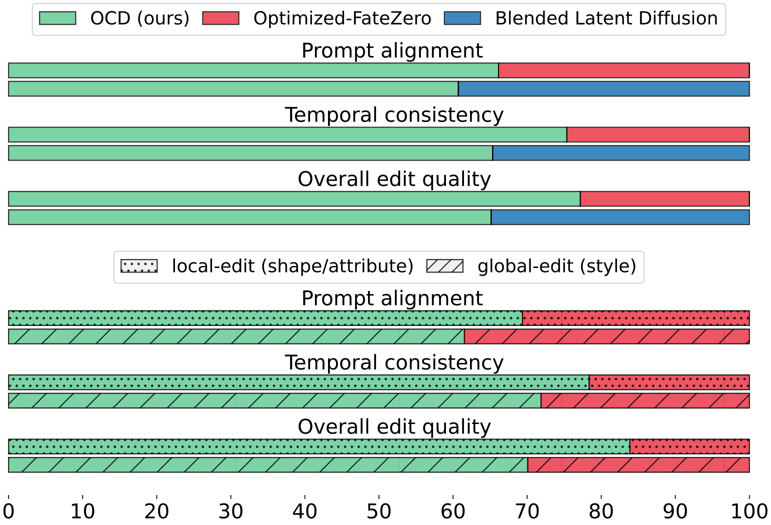

Fig. B.3: User study: (Top) Preferences for Object-Centric Diffusion (ours) w.r.t. Optimized-FateZero or Blended Latent Diffusion [1] on FateZero benchmark (Bottom) Preferences for local vs. global edits. In both cases, OCD is better preferred.

use $\gamma = 0.25$ by default, whereas $\varDelta$ depends on objects in a given sequence. As expected, we can obtain the most savings with fewer blending steps and when foreground objects are smaller in size. A user may refer to this guide to get an idea about the expected savings.

User preference study We conduct a user study to measure the editing quality of OCD w.r.t. (1) Optimized-FateZero and (2) Blended Latent Diffusion [1]. The study consists of randomized A/B preferences tests, where observers were asked to assess video edits in overall quality as well as temporal consistency and alignment with edit prompt. Based on 1143 responses collected from 37 participants, we find that OCD is likely preferred by users in all assessments. More specifically, this study testifies the benefit of OCD w.r.t. the optimized baseline beyond computational savings, as its edits are preferred 77% of the time, and rewarded by users in temporal consistency (75%) and prompt alignment (66%).

Other methods with disentangled diffusion sampling Both OCD and Blended Latent Diffusion (BLD) [1] use saliency masks to disentangle foreground and background during diffusion sampling. Although BLD focuses on local edits, its design does not imply any efficiency gains (as it processes background, but discards during blending). On the contrary, our proposal trades-off computational cost in the background, allowing for a better foreground edit at a reduced latency. Such a change in scope results in several key differences. In OCD, background and foreground latents undergo two *separate* (even, parallelizable) sampling processes, operating at different resolutions and sampling rates, before being blended at a certain pre-defined step. Differently, latents in BLD are blended at every step of the *same* diffusion process, and its latency is by design lower bounded by that of a standard diffusion process. To compare it with our approach, we included BLD edits in the user study in Fig. B.3, where we observe OCD is preferred most of the times in terms of temporal consistency (65.35%) and prompt alignment (60.71%) while also being faster (17.51s vs 19.68s). In metrics, OCD is comparable to BLD both in temporal consistency (0.967 vs 0.968) and clip score (0.331 vs 0.329) respectively.

References

1. Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. ACM Transactions on Graphics (TOG) (2023)
2. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. International Conference on Learning Representations (2023)
3. Bolya, D., Hoffman, J.: Token merging for fast stable diffusion. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2023)
4. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. International Conference on Learning Representations (2024)
5. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. IEEE International Conference on Computer Vision (2023)
6. Meng, C., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Image synthesis and editing with stochastic differential equations. International Conference on Learning Representations (2022)
7. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv:1704.00675 (2017)

8. Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: Fatezero: Fusing attentions for zero-shot text-based video editing. IEEE International Conference on Computer Vision (2023)
9. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021)
10. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2022)
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) (2015)
12. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2023)
13. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: IEEE International Conference on Computer Vision (2023)
14. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: IEEE International Conference on Computer Vision (2023)
15. Zhang, Y., Wei, Y., Jiang, D., Zhang, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation. International Conference on Learning Representations (2024)