

## 1 A Detailed deduction of the fixed point method

2 Below, we show the matrix formation for the fixed point iteration that solves CD-EM. To simplify  
 3 the equation, we use two constants  $c_{t-1}^1, c_{t-1}^2$  to represent the coefficients.  $c_{t-1}^1$  and  $c_{t-1}^2$  are  
 4 mathematically

$$c_{t-1}^1 = \sqrt{\frac{1 - \alpha_{t-1} - \sigma_{t-1}^2}{1 - \alpha_t}}, \quad c_{t-1}^2 = \frac{\sqrt{\alpha_{t-1}(1 - \alpha_t)} - \sqrt{1 - \alpha_{t-1} - \sigma_{t-1}^2} \cdot \sqrt{\alpha_t}}{\sqrt{1 - \alpha_t}}. \quad (1)$$

5 We aim to make the equation holds:

$$(\mathbf{x}_{t-1}^a, \mathbf{x}_{t-1}^{b,i}) - \boldsymbol{\mu}_\theta(\mathbf{x}_t) = (\mathbf{x}_{t-1}^a, \mathbf{x}_{t-1}^{b,i}) + c_{t-1}^1 \mathbf{x}_t + c_{t-1}^2 \mathbf{f}_\theta(\mathbf{x}_t). \quad (2)$$

6 We aim to find a  $\mathbf{x}_t$  that ensures

$$(\mathbf{x}_{t-1}^a, \mathbf{x}_{t-1}^{b,i}) + c_{t-1}^1 \mathbf{x}_t + c_{t-1}^2 \mathbf{f}_\theta(\mathbf{x}_t) = 0 \quad (3)$$

7 in order to maximize the expectation log-likelihood. Unluckily, this equation can not be solved  
 8 analytically since  $\mathbf{f}_\theta(\cdot)$  is a neural function. In alternative, we propose to use fixed point iteration to  
 9 get an approximated solution. We rewrite eq. (3) to matrix formation. Mathematically that is

$$- \begin{bmatrix} ((\mathbf{x}_{t-1}^a, \mathbf{x}_{t-1}^{b,1}) + c_{t-1}^2 \mathbf{f}_\theta(\mathbf{x}_t')) \\ ((\mathbf{x}_{t-1}^a, \mathbf{x}_{t-1}^{b,2}) + c_{t-1}^2 \mathbf{f}_\theta(\mathbf{x}_t')) \\ \dots \\ ((\mathbf{x}_{t-1}^a, \mathbf{x}_{t-1}^{b,N}) + c_{t-1}^2 \mathbf{f}_\theta(\mathbf{x}_t')) \end{bmatrix} = c_{t-1}^1 \begin{bmatrix} \mathbf{I} \\ \mathbf{I} \\ \dots \\ \mathbf{I} \end{bmatrix} \mathbf{x}_t \quad (4)$$

10 where we use ";" to denote concatenation in column, and  $\mathbf{x}_t'$  is  $\mathbf{x}_t$  at the previous fixed point iteration  
 11 step. We use  $\mathbb{I}$  to denote the  $Nd \times d$  matrix  $\mathbb{I} = [\mathbf{I} \quad \mathbf{I} \quad \dots \quad \mathbf{I}]^T$ . The fixed point iteration follows a  
 12 formation of least square solution. Specifically, the solution is

$$\mathbf{x}_t \leftarrow -\frac{1}{c_{t-1}^1} (\mathbb{I}^T \mathbb{I})^{-1} \mathbb{I} \begin{bmatrix} ((\mathbf{x}_{t-1}^a, \mathbf{x}_{t-1}^{b,1}) + c_{t-1}^2 \mathbf{f}_\theta(\mathbf{x}_t')) \\ ((\mathbf{x}_{t-1}^a, \mathbf{x}_{t-1}^{b,2}) + c_{t-1}^2 \mathbf{f}_\theta(\mathbf{x}_t')) \\ \dots \\ ((\mathbf{x}_{t-1}^a, \mathbf{x}_{t-1}^{b,N}) + c_{t-1}^2 \mathbf{f}_\theta(\mathbf{x}_t')) \end{bmatrix}. \quad (5)$$

## 13 B Proof of theorem 3.1

14 We show the proof in this section.

15 *Proof.* We first study what we need  $\mathbf{f}_\theta(\cdot)$  to be like in order to ensure an unique fixed point. We use  
 16  $g(\cdot)$  to denote the iteration step

$$\mathbf{x}_t \leftarrow g(\mathbf{x}_t), \quad g(\mathbf{x}_t) = -\frac{(\mathbf{x}_{t-1}^a, \mathbf{x}_{t-1}^{b,1}) + c_{t-1}^2 \mathbf{f}_\theta(\mathbf{x}_t)}{c_{t-1}^1}. \quad (6)$$

17 According to *Banach's fixed point theorem*, we need  $g(\cdot)$  to guarantee smaller than 1 Lipschitz  
 18 constant

$$\|g(\mathbf{x}_t^1) - g(\mathbf{x}_t^2)\|_2 < 1 \cdot \|\mathbf{x}_t^1 - \mathbf{x}_t^2\|, \quad (7)$$

19 for arbitrary  $\mathbf{x}_t^1, \mathbf{x}_t^2$ . Expanding  $g(\cdot)$  to the  $\mathbf{f}_\theta(\cdot)$  formation, we derive

$$\left\| -\frac{(\mathbf{x}_{t-1}^a, \mathbf{x}_{t-1}^{b,1}) + c_{t-1}^2 \mathbf{f}_\theta(\mathbf{x}_t^1)}{c_{t-1}^1} + \frac{(\mathbf{x}_{t-1}^a, \mathbf{x}_{t-1}^{b,1}) + c_{t-1}^2 \mathbf{f}_\theta(\mathbf{x}_t^2)}{c_{t-1}^1} \right\|_2 < \|\mathbf{x}_t^1 - \mathbf{x}_t^2\|_2 \quad (8)$$

$$\iff \frac{c_{t-1}^2}{c_{t-1}^1} \|\mathbf{f}_\theta(\mathbf{x}_t^1) - \mathbf{f}_\theta(\mathbf{x}_t^2)\|_2 < \|\mathbf{x}_t^1 - \mathbf{x}_t^2\|_2 \quad (9)$$

$$\iff \|\mathbf{f}_\theta(\mathbf{x}_t^1) - \mathbf{f}_\theta(\mathbf{x}_t^2)\|_2 < \frac{c_{t-1}^1}{c_{t-1}^2} \|\mathbf{x}_t^1 - \mathbf{x}_t^2\|_2. \quad (10)$$

Therefore, we require the Lipschitz constant of  $\mathbf{f}_\theta(\cdot)$  denoted by  $\mathcal{L}_f$  to meet  $\mathcal{L}_f < \frac{c_{t-1}^1}{c_{t-1}^2}$ . Note that  $c_{t-1}^1 > 0$  and  $c_{t-1}^2 > 0$ . The former one is obvious since we pre-define  $\sigma_{t-1}$  to satisfy  $1 - \alpha_{t-1} - \sigma_{t-1}^2 > 0$ . In a diffusion model,  $\alpha_t < \alpha_{t-1}, t \in \{1, \dots, T\}$ . Therefore we also derive  $\sqrt{1 - \alpha_t} > \sqrt{1 - \alpha_{t-1} - \sigma_{t-1}^2}$  which makes the denominator of  $c_{t-1}^2$  positive. We then relate the Lipschitz constant  $\mathcal{L}_f$  to that of the real score function  $\mathcal{L}_s$ . The clean data estimant can be expressed by

$$\mathbf{f}_\theta(\mathbf{x}_t) = \frac{\mathbf{x}_t - \sqrt{1 - \alpha_t} \epsilon_\theta(\mathbf{x}_t)}{\sqrt{\alpha_t}} = \frac{\mathbf{x}_t + \mathbf{s}_\theta(\mathbf{x}_t)}{\sqrt{\alpha_t}}. \quad (11)$$

We construct an inequality that can be the sufficient condition for to find the Lipschitz constant of  $\mathbf{f}_\theta(\cdot)$ . Assume the following inequality is satisfied

$$1 + \mathcal{L}_s < \frac{c_{t-1}^1 \sqrt{\alpha_t}}{c_{t-1}^2} \quad (12)$$

$$\iff \|\mathbf{x}_t^1 - \mathbf{x}_t^2\|_2 + \mathcal{L}_s \|\mathbf{x}_t^1 - \mathbf{x}_t^2\|_2 < \frac{c_{t-1}^1 \sqrt{\alpha_t}}{c_{t-1}^2} \|\mathbf{x}_t^1 - \mathbf{x}_t^2\|_2 \quad (13)$$

$$\stackrel{??}{\iff} \|\mathbf{x}_t^1 - \mathbf{x}_t^2\|_2 + \|\mathbf{s}_\theta(\mathbf{x}_t^1) - \mathbf{s}_\theta(\mathbf{x}_t^2)\|_2 < \|\mathbf{x}_t^1 - \mathbf{x}_t^2\|_2 + \mathcal{L}_s \|\mathbf{x}_t^1 - \mathbf{x}_t^2\|_2 < \frac{c_{t-1}^1 \sqrt{\alpha_t}}{c_{t-1}^2} \|\mathbf{x}_t^1 - \mathbf{x}_t^2\|_2 \quad (14)$$

$$\stackrel{\text{triangular inequality}}{\iff} \|(\mathbf{x}_t^1 - \mathbf{x}_t^2) + (\mathbf{s}_\theta(\mathbf{x}_t^1) - \mathbf{s}_\theta(\mathbf{x}_t^2))\|_2 < \|\mathbf{x}_t^1 - \mathbf{x}_t^2\|_2 + \|\mathbf{s}_\theta(\mathbf{x}_t^1) - \mathbf{s}_\theta(\mathbf{x}_t^2)\|_2 < \frac{c_{t-1}^1 \sqrt{\alpha_t}}{c_{t-1}^2} \|\mathbf{x}_t^1 - \mathbf{x}_t^2\|_2 \quad (15)$$

$$\iff \left\| \frac{\mathbf{x}_t^1 + \mathbf{s}_\theta(\mathbf{x}_t^1)}{\sqrt{\alpha_t}} - \frac{\mathbf{x}_t^2 + \mathbf{s}_\theta(\mathbf{x}_t^2)}{\sqrt{\alpha_t}} \right\|_2 = \|\mathbf{f}_\theta(\mathbf{x}_t^1) - \mathbf{f}_\theta(\mathbf{x}_t^2)\|_2 < \frac{c_{t-1}^1}{c_{t-1}^2} \|\mathbf{x}_t^1 - \mathbf{x}_t^2\|_2. \quad (16)$$

Therefore, if  $1 + \mathcal{L}_s < \frac{c_{t-1}^1 \sqrt{\alpha_t}}{c_{t-1}^2}$  is satisfied, we ensure an unique fixed point can be found. Then, the Lipschitz constant of the real score function  $\mathbf{s}(\cdot)$  is computed below. The score function is defined as

$$\mathbf{s}(\mathbf{x}_t | \mathbf{x}_0) \doteq \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t | \mathbf{x}_0) = -\frac{\mathbf{x}_t - \sqrt{\alpha_t} \mathbf{x}_0}{1 - \alpha_t}. \quad (17)$$

We derive the Lipschitz constant through

$$\|\mathbf{s}(\mathbf{x}_t^1 | \mathbf{x}_0) - \mathbf{s}(\mathbf{x}_t^2 | \mathbf{x}_0)\|_2 = \left\| -\frac{\mathbf{x}_t^1 - \sqrt{\alpha_t} \mathbf{x}_0}{1 - \alpha_t} + \frac{\mathbf{x}_t^2 - \sqrt{\alpha_t} \mathbf{x}_0}{1 - \alpha_t} \right\|_2 \quad (18)$$

$$= \frac{1}{1 - \alpha_t} \|\mathbf{x}_t^1 - \mathbf{x}_t^2\|_2 < \mathcal{L}_s \|\mathbf{x}_t^1 - \mathbf{x}_t^2\|_2. \quad (19)$$

Therefore, the Lipschitz constant is  $\mathcal{L}_s = 1/(1 - \alpha_t)$ . Replacing this and the definition of  $c_{t-1}^1, c_{t-1}^2$  to eq. (12) we have

$$\frac{2 - \alpha_t}{1 - \alpha_t} < \frac{\sqrt{(1 - \alpha_{t-1} - \sigma_{t-1}^2) \alpha_t}}{\sqrt{\alpha_{t-1}(1 - \alpha_t)} - \sqrt{(1 - \alpha_{t-1} - \sigma_{t-1}^2) \alpha_t}} \quad (20)$$

$$\stackrel{\text{rearrange}}{\iff} (2 - \alpha_t) \left( \sqrt{\alpha_{t-1}(1 - \alpha_t)} - \sqrt{(1 - \alpha_{t-1} - \sigma_{t-1}^2) \alpha_t} \right) < (1 - \alpha_t) \sqrt{(1 - \alpha_{t-1} - \sigma_{t-1}^2) \alpha_t} \quad (21)$$

$$\iff (2 - \alpha_t) \sqrt{\alpha_{t-1}(1 - \alpha_t)} < (3 - 2\alpha_t) \sqrt{\alpha_t(1 - \alpha_{t-1} - \sigma_{t-1}^2)} \quad (22)$$

Knowing whether this inequality holds is not obvious. Instead of seeking what kind of noise schedule makes the condition holds, we study the commonly used noise schedules. We use  $m(t) =$

---

**Algorithm 1:** Conditional denoising expectation maximization for diffusion model

---

```

1 Input: Known video segment  $\mathbf{x}_0^a$ , pre-trained diffusion model  $\mathbf{f}_\theta(\cdot)$ , stopping criteria  $\tau$ ,
inference timesteps  $T$ , EM iteration number  $P$ , Monte-Carlo sample number  $N$ , fixed point
iteration step number  $K$ 
2 Output: Rendered video segment  $\mathbf{x}_0^b$ 
3  $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
4 for  $t = T$  to 0 do
5   for  $p = 1$  to  $P$  do
6      $\mathbf{x}_t^{\text{old}} \leftarrow \mathbf{x}_t$ 
7     for  $i = 1$  to  $N$  do
8        $\mathbf{x}_{t-1}^i \sim p_\theta(\mathbf{x}_{t-1}^b | \mathbf{x}_t^{\text{old}})$ 
9     end
10     $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
11     $\mathbf{x}_{t-1}^a \leftarrow \sqrt{\alpha_{t-1}} \mathbf{x}_0^a + \sqrt{1 - \alpha_{t-1}} \epsilon$ 
12    for  $k = 1$  to  $K$  do
13      Apply eq. (5) to get  $\mathbf{x}_t$ 
14    end
15  end
16   $\mathbf{x}_{t-1} \sim p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)$ 
17 end
18 return  $\mathbf{x}_0$ 

```

---

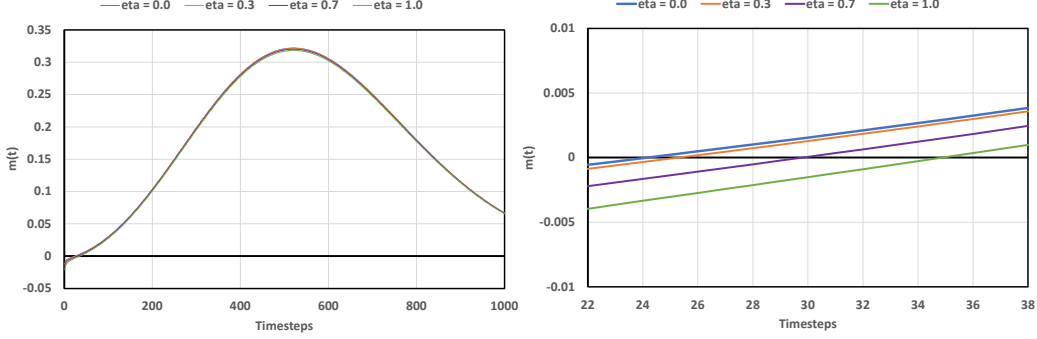


Figure 1: The curve of  $m(t)$  under different settings of  $\eta$ , The right curve zoom in the regions that curves hit zero line.

35  $(3-2\alpha_t)\sqrt{\alpha_t(1-\alpha_{t-1}-\sigma_{t-1}^2)} - (2-\alpha_t)\sqrt{\alpha_{t-1}(1-\alpha_t)}$  for simplification. Then the convergency  
 36 and uniqueness hold when  $m(t) > 0$ . We plot  $m(t)$  curve for a series of diffusion schedules. The  
 37 results are shown in fig. 1. We plot the situation of CogVideoX scheduler with different choices of  $\eta$ .  
 38 In most sampling timestep, we can ensure uniqueness and convergency.  $\square$

## 39 **C CD-EM on flow matching generators**

40 Our method is not only applicable to diffusion based generators. Instead, CD-EM is also applicable  
 41 to flow matching generative models. Consider a model that uses the following flows and velocity  
 42 field in  $t \in [0, T)$ .

$$\mathbf{x}_t = \phi_t(\mathbf{x}_0) = \sigma(t)\mathbf{x}_0 + (1 - \sigma(t))\epsilon, \quad u_t(\phi_t(\mathbf{x}_0)) = \frac{\partial}{\partial t} \phi_t(\mathbf{x}_0) = \frac{\partial \sigma(t)}{\partial t} \cdot (\mathbf{x}_0 - \epsilon), \quad (23)$$

43 where  $\sigma(\cdot)$  is the noise schedule, and  $\epsilon$  is a random gaussian noise. The velocity field  $u_t(\cdot)$  will  
 44 support a probability path

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t | \sigma(t)\mathbf{x}_0, (1 - \sigma(t))^2 \mathbf{I}). \quad (24)$$

45 Solving the ODE in Eq.(4.1.1) produces the sampling result. Euler sampler is commonly used in  
 46 which the iteration is

$$\mathbf{x}_s = \mathbf{x}_t + (s - t)u_t(\mathbf{x}_t), \quad 0 \leq s < t \leq T \quad (25)$$

47 To optimize the model through flow matching, one usually use a neural network  $\mathbf{v}_\theta(\cdot)$  to approximate  
 48 the scaled velocity field  $\mathbf{v}_\theta(\mathbf{x}_t) \simeq \epsilon - \mathbf{x}_0$ ,  $u_t(\mathbf{x}_t) \simeq -\partial\sigma(t)/\partial t \cdot \mathbf{v}_\theta(\mathbf{x}_t)$ . Replace the velocity in  
 49 Eq.(4.1.3) with the model estimation, we derive

$$\mathbf{x}_s = \mathbf{x}_t + (\sigma(t) - \sigma(s))\mathbf{v}_\theta(\mathbf{x}_t). \quad (26)$$

50 To optimize the trajectory with expectation maximization method, we need to know the sampling  
 51 posterior distribution  $q(\mathbf{x}_s|\mathbf{x}_t, \mathbf{x}_0)$ ,  $0 \leq s < t \leq T$  of the flow model. Applying *bayes formula*, we  
 52 get

$$q(\mathbf{x}_s|\mathbf{x}_t, \mathbf{x}_0) = \frac{q(\mathbf{x}_t|\mathbf{x}_s, \mathbf{x}_0)q(\mathbf{x}_s|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)}. \quad (27)$$

53 The distribution  $q(\mathbf{x}_s|\mathbf{x}_0)$  and  $q(\mathbf{x}_t|\mathbf{x}_0)$  are already known. However, since a flow model is deter-  
 54 ministic, the term in the nominator  $q(\mathbf{x}_t|\mathbf{x}_s, \mathbf{x}_0)$  is specifically Dirac delta. Therefore, the sampling  
 55 distribution will also be Dirac delta that

$$q(\mathbf{x}_s|\mathbf{x}_t, \mathbf{x}_0) = \delta\left(\mathbf{x}_s - \frac{1 - \sigma(s)}{1 - \sigma(t)} \cdot \mathbf{x}_t + \frac{\sigma(t) - \sigma(s)}{1 - \sigma(t)} \cdot \mathbf{x}_0\right). \quad (28)$$

56 Meanwhile, we also derive the modeled sampling distribution

$$p_\theta(\mathbf{x}_s|\mathbf{x}_t) = \delta(\mathbf{x}_s - \mathbf{x}_t - (\sigma(t) - \sigma(s))\mathbf{v}_\theta(\mathbf{x}_t)). \quad (29)$$

57 Given this, we are able to maximize the expectation in Eq.(2.2.3)

$$Q(\mathbf{x}_t|\mathbf{x}_t^{\text{old}}) \simeq \sum_i^N \log p_\theta(\mathbf{x}_s^a, \mathbf{x}_s^{b,i}|\mathbf{x}_t) d\mathbf{x}_s^b, \quad (30)$$

58 which requires  $p_\theta(\mathbf{x}_s^a, \mathbf{x}_s^{b,i}|\mathbf{x}_t)$  to be maximized. In this case, we expect the following equation set to  
 59 hold

$$\begin{cases} (\mathbf{x}_s^a, \mathbf{x}_s^{b,1}) = \mathbf{x}_t + (\sigma(t) - \sigma(s))\mathbf{v}_\theta(\mathbf{x}_t^{\text{old}}) \\ (\mathbf{x}_s^a, \mathbf{x}_s^{b,2}) = \mathbf{x}_t + (\sigma(t) - \sigma(s))\mathbf{v}_\theta(\mathbf{x}_t^{\text{old}}) \\ \dots \\ (\mathbf{x}_s^a, \mathbf{x}_s^{b,N}) = \mathbf{x}_t + (\sigma(t) - \sigma(s))\mathbf{v}_\theta(\mathbf{x}_t^{\text{old}}) \end{cases} \iff \begin{bmatrix} (\mathbf{x}_s^a, \mathbf{x}_s^{b,1}) + (\sigma(s) - \sigma(t))\mathbf{v}_\theta(\mathbf{x}_t^{\text{old}}); \\ (\mathbf{x}_s^a, \mathbf{x}_s^{b,2}) + (\sigma(s) - \sigma(t))\mathbf{v}_\theta(\mathbf{x}_t^{\text{old}}); \\ \dots \\ (\mathbf{x}_s^a, \mathbf{x}_s^{b,N}) + (\sigma(s) - \sigma(t))\mathbf{v}_\theta(\mathbf{x}_t^{\text{old}}) \end{bmatrix} = \begin{bmatrix} \mathbf{I} \\ \mathbf{I} \\ \dots \\ \mathbf{I} \end{bmatrix} \mathbf{x}_t. \quad (31)$$

60  $\mathbf{I} \in \mathbb{R}^{n \times n}$  is an identity matrix. We solve this equation set with least square and produces the fixed  
 61 point iteration step

$$\mathbf{x}_t \leftarrow (\mathbb{I}^T \mathbb{I})^{-1} \mathbb{I} \begin{bmatrix} (\mathbf{x}_s^a, \mathbf{x}_s^{b,1}) + (\sigma(s) - \sigma(t))\mathbf{v}_\theta(\mathbf{x}_t^{\text{old}}); \\ (\mathbf{x}_s^a, \mathbf{x}_s^{b,2}) + (\sigma(s) - \sigma(t))\mathbf{v}_\theta(\mathbf{x}_t^{\text{old}}); \\ \dots \\ (\mathbf{x}_s^a, \mathbf{x}_s^{b,N}) + (\sigma(s) - \sigma(t))\mathbf{v}_\theta(\mathbf{x}_t^{\text{old}}) \end{bmatrix}, \quad (32)$$

62 where we use  $\mathbb{I}$  to denote the  $Nd \times d$  matrix  $\mathbb{I} = [\mathbf{I} \quad \mathbf{I} \quad \dots \quad \mathbf{I}]^T$ . Similarly, when  $N = 1$ , the  
 63 iteration step is implied to

$$\mathbf{x}_t \leftarrow (\mathbf{x}_s^a, \mathbf{x}_s^b) + (\sigma(s) - \sigma(t))\mathbf{v}_\theta(\mathbf{x}_t^{\text{old}}). \quad (33)$$

64 To uniqueness and convergency of the fixed point iteration can be found using a similar way in  
 65 appendix B. We try this method on a flow matching video generator named HunyuanVideo (4) to  
 66 perform video editing. Please see our supplementary material.

## 67 D Computation and memory cost

68 **Ablation experiment** The computation and memory cost of ZeroPathcer is shown in table 1.  
 69 The theoretical complexity considers the computation required for each denoising step. We let the  
 70 complexity for every function evaluation be  $O(1)$ .  $P$  is the outer loop step number of the expectation  
 71 maximization.  $N$  denotes how many  $\mathbf{x}_{t-1}^b$  are sampled in order to compute the likelihood expectation.  
 72  $K$  is the step number of the fixed point iteration. With a reasonable set of hyperparameters (i.e.  
 73  $P = 2, N = 1, K = 1$ ), ZeroPathcer will not introduce enormous computation compared to the  
 74 direct text-to-video inference. Meanwhile, we show that the ZeroPathcer does not use considerable  
 75 addition memory which is usually required for attention manipulation methods like Video-P2P (2)  
 76 and FateZero (3).

Model	Theoretical complexity	Memory	Speed
Inference params	$P = 2, N = 1, K = 1$		
t2v model	$O(1)$	22,489MB	2.61 s/it
+ BP	$O(1)$	22,512MB	2.63 s/it
+ CD-EM	$O(1 + PN + P(K - 1))$	22,567MB	7.41 s/it
+ LMF	$O(1 + PN + P(K - 1))$	22,893MB	7.63 s/it

Table 1: Theoretical and practical computation cost analysis of our method. Theoretical complexity and speed show the computation required for each denoising step. We use LMF to represent the latent mask fuser. The speed is computed under seconds per iteration.

Method	Resolution	Memory	NFE	Speed
Text-to-video				
CogVideoX	$49 \times 720 \times 480$	22,488MB	30	78.3 s/sample
Inpainting				
SDEdit	$49 \times 720 \times 480$	22,503MB	100	240.3 s/sample
DDNM	$49 \times 720 \times 480$	22,512MB	100	241.7 s/sample
Ours	$49 \times 720 \times 480$	22,567MB	100	250.9 s/sample
Editing				
PF	$8 \times 512 \times 512$	4,279 MB	50	81.2 s/sample
VideoComposer	$16 \times 256 \times 256$	7,394MB	50	101.4 s/sample
VideoP2P	$8 \times 512 \times 512$	19,453MB	100	137.8 s/sample
DDNM	$49 \times 720 \times 480$	22,522MB	100	241.5 s/sample
Ours	$49 \times 720 \times 480$	22,569MB	100	249.4 s/sample

Table 2: A comparison over computation and memory usage.

**Comparison with other methods** We compare the memory and throughput among the methods. The result is shown in table 2. Our method achieves remarkable improvements over the training-free methods under similar computation cost. To achieve faster inference, one can choose to trade performance for speed by letting  $N = 1, K = 1, P = 1$ .

## E Algorithmic description of CD-EM

We show an algorithmic description of CD-EM in Algo.1. It provides the clean CD-EM without back projection and latent mask fuser.

## F Details of latent mask fuser

In this sector, we briefly introduce the architecture of our latent mask fuser. It consists of 6 residual blocks, a convolutional input layer, and a convolutional output layer. After the 3-th block, we use an adaptive instance normalization layer (1) to perform mask fusing. Each residual block is made up of two convolution blocks (a concatenation of a GroupNorm layer, a SiLU layer, and a 3D convolution layer) and a skip connection layer achieved by a  $1 \times 1 \times 1$  3D convolution. During inference, we concatenate three latent features from the two source video and the mask in channel dimension, and then the feature is feed into the input layer. No upsampling and downsampling modules are used in latent mask fuser. The model uses a channel size of 1536.

## G Inference details and hyperparameters

The inference hyperparameters are shown in table 3. We use DDIM diffusion sampler with sampling stochasticity factor  $\eta = 1.0$ . Our experiment shows using  $\eta = 1.0$  can greatly increase video

Hyperparameter	CogVideoX
EM outer loop steps $P$	2
Number of samples in expectation $N$	1
Fixed point iteration steps $K$	1
Sampler	DDIM
Inference steps	50
CD-EM stop step	25
NFEs	100
Classifier-free guidance scale	6
eta	1.0

Table 3: Hyperparameters use in inference.

consistency since the model will not follow the deterministic sampling trajectory. CD-EM is not always required in all denoising timesteps. We find using CD-EM in early sampling stages can already produce plausible results. Therefore, we add a stoping step for CD-EM at 25 to save computation without losing noticeable performance. With CD-EM hypermarameters set to  $P = 2, N = 1, K = 1$ , we ensure every denoising step with CD-EM will only cost 3 NFEs. With the remaining 25 steps using only 1 NFE, sampling through ZeroPatcher requires 100 NFEs in total. We use 8 NVIDIA A100 80G GPUs to run in parallel during inference to get results faster.

## H Additional visual results

We attach result videos in our supplementary material. We would appreciate it if you check them.

## I Broader impact

The integration of pre-trained text-to-video foundation models into video inpainting and editing workflows presents transformative opportunities alongside critical ethical challenges. By enabling high-quality video manipulation without additional training, this approach democratizes access to advanced editing tools, empowering independent creators and small studios to achieve professional-grade results—potentially revitalizing archival restoration efforts and lowering costs for cultural heritage preservation. However, the same capabilities raise significant concerns: the efficiency of dynamic object removal and latent-space masking could streamline the creation of convincing deepfakes, exacerbating misinformation risks in an era already plagued by synthetic media distrust. The model-agnostic nature of the method further amplifies these risks, as it could be applied to any foundation model, including those with fewer ethical safeguards.

## References

- [1] Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2017)
- [2] Liu, S., Zhang, Y., Li, W., Lin, Z., Jia, J.: Video-p2p: Video editing with cross-attention control (2023)
- [3] Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: Fatezero: Fusing attentions for zero-shot text-based video editing. arXiv:2303.09535 (2023)
- [4] Weijie Kong, Qi Tian, Z.Z.R.M.Z.D.J.Z.J.X.X.L.B.W.J.Z.K.W.Q.L.A.W.A.W.C.L.D.H.F.Y.H.T.H.W.J.S.J.B.J.W.J.X.J.W.J.Y.K.W.M.L.P.L.S. Jie Jiang, a.w.C.Z.: Hunyuanvideo: A systematic framework for large video generative models (2024), <https://arxiv.org/abs/2412.03603>