

ToonCrafter: Generative Cartoon Interpolation

JINBO XING, The Chinese University of Hong Kong, China

HANYUAN LIU, City University of Hong Kong, China

MENGHAN XIA*, Tencent AI Lab, China

YONG ZHANG, Tencent AI Lab, China

XINTAO WANG, Tencent AI Lab, China

YING SHAN, Tencent AI Lab, China

TIEN-TSIN WONG*, The Chinese University of Hong Kong, China and Monash University, Australia

ACM Transactions on Graphics (Special issue of SIGGRAPH Asia 2024)



賴濤雨 Shih-Yu Lai

Introduction

1. 問題背景

- 卡通動畫製作極為勞力密集，需逐格繪製
- 卡通影片插值在深度神經網路的推動下已有顯著進展
- 現有真人影片補幀法無法處理卡通特性：
 - 時間間隔較大（大幅度運動）
 - 純色區域多、紋理少

ATD-12K dataset.

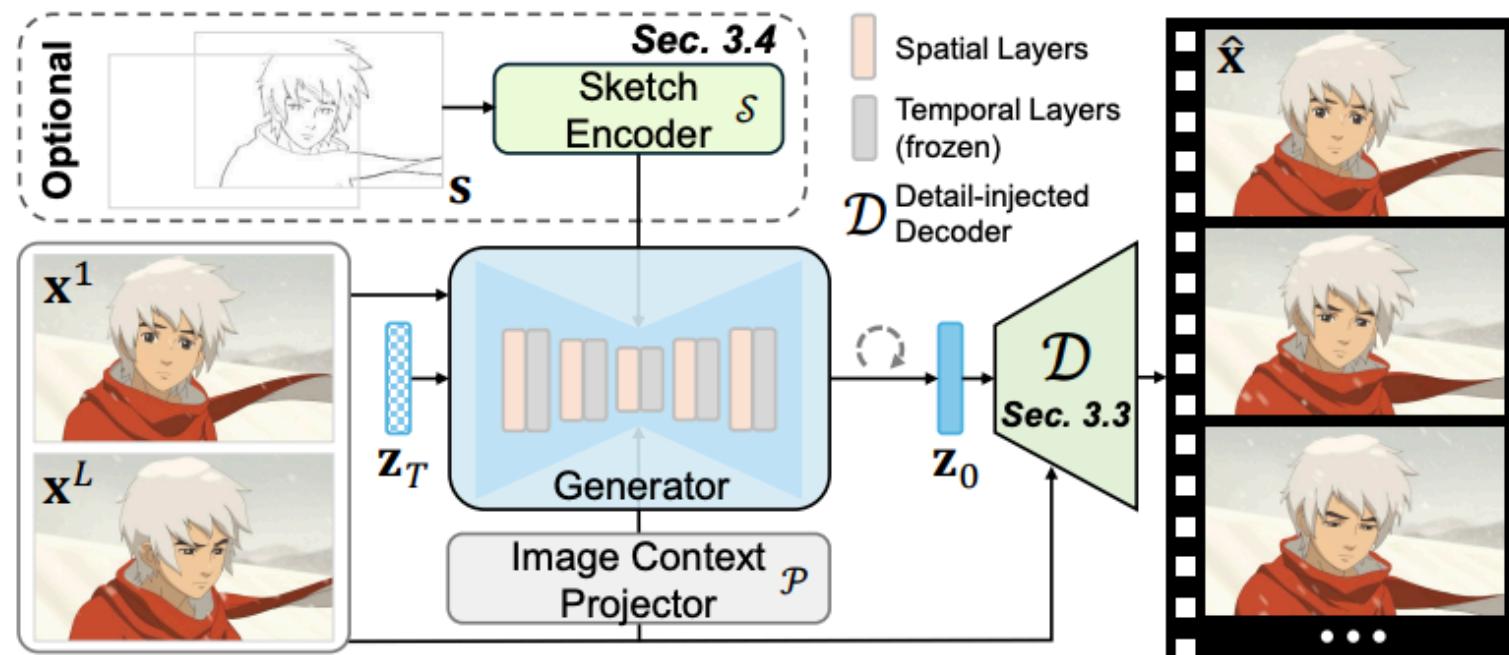


lifelike physical phenomena

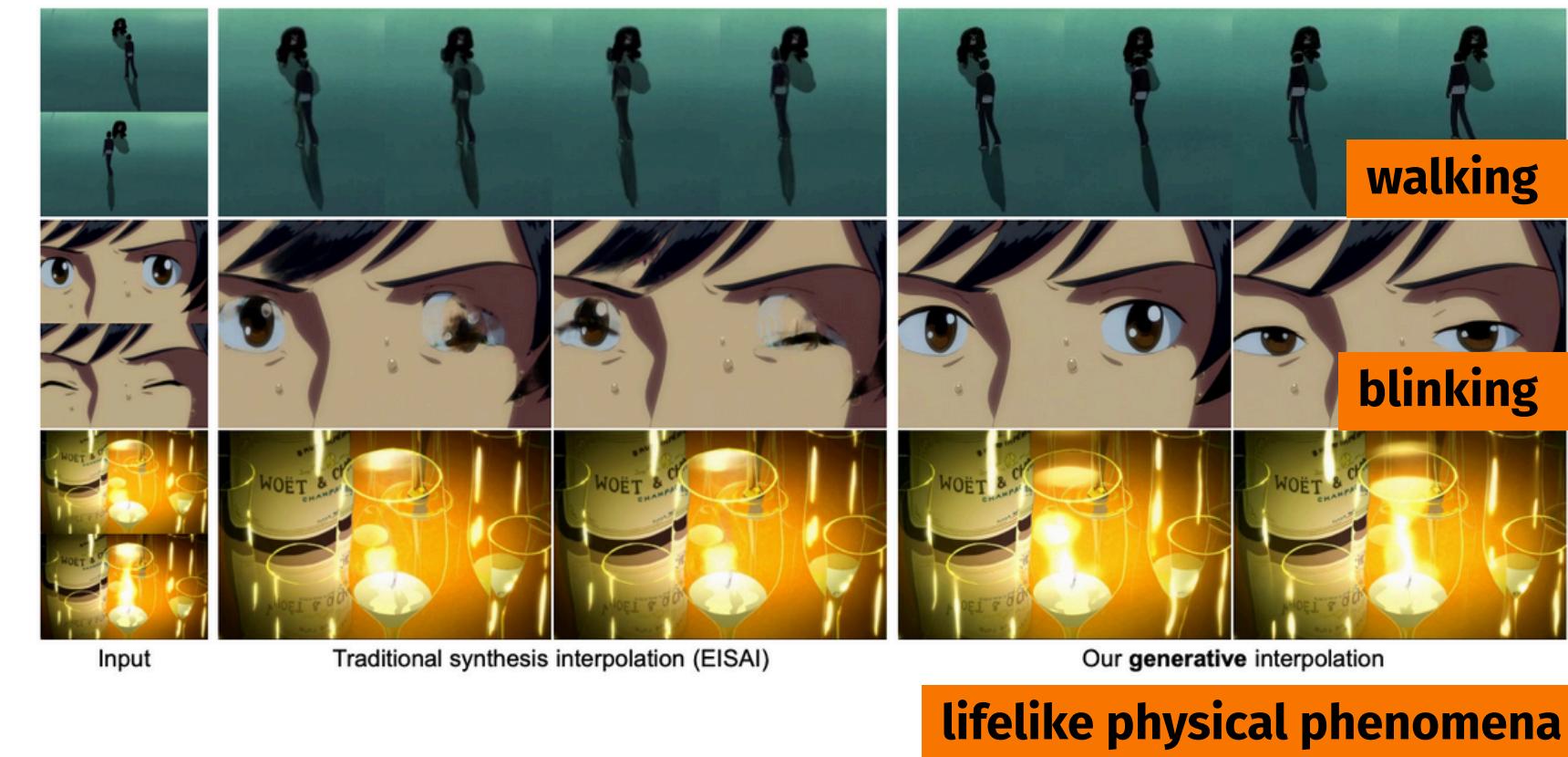
2. 現有方法限制

- 線性插值假設適用於真人影片，無法應對遮擋與誇張**非線性運動**
 - 先找出兩幀之間的對應關係，例如使用光流(Optical Flow-based)，再進行線性插值
 - 圖1（上）所示，展示了一個走路的人，顯然線性插值只能產生「漂浮」的人物，而非正確的行走動作。
 - 圖1（中）所示的**遮擋**情況則更加複雜。
- 生成式擴散模型具潛力，但面臨三大**挑戰**：
 - Domain gap**：模型對動畫情境理解不足
 - Latent Space壓縮**導致細節與畫質劣化：尤其在高對比區域、細緻結構輪廓與動畫中常見的模糊運動上更為明顯
 - 缺乏控制性**，難以指定運動方式

Introduction



ATD-12K dataset.



3. ToonCrafter 方法

針對卡通補幀的生成式擴散模型框架：

- **Toon Rectification Learning**

- 提出生成式卡通插值問題與解法，首次引入 live-action 運動先驗
- 微調生成模型中空間理解與內容生成有關的層級，保留原始模型的豐富運動先驗、適配卡通域

- **Dual-Reference 3D Decoder**

- 透過兩張輸入影像注入細節，以 **hybrid-attention-residual-learning**，補足壓縮Latent Space細節損失重建回像素空間
- 使用 **pseudo-3D 卷積**（**1D 時序卷積 + 2D 空間卷積**）增進時序一致性

- **Sketch-based Controllable Generation**

- 無需對應幀草圖的情況下，自由地透過不同數量的草圖影像，靈活控制或修改插值運動結構。
- 支援稀疏草圖輸入，自由控制插值結構，高彈性與準確度的互動式

Related Work

1. Video Frame Interpolation

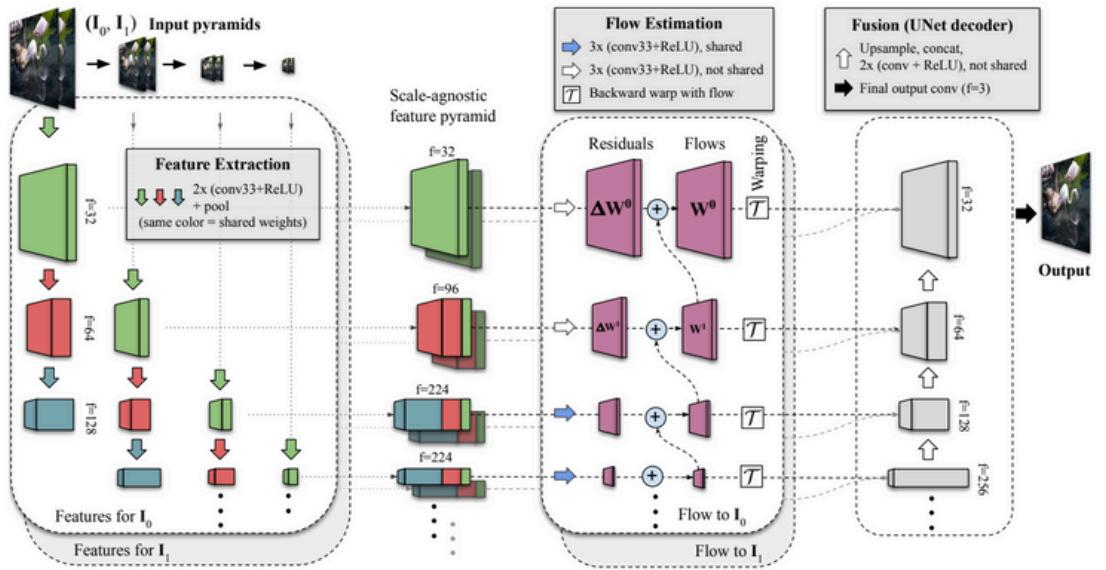
視訊幀插值的目標是合成原始影片中兩個相鄰畫格之間的多個中間畫格。這個問題近年來已被廣泛研究。

基於深度學習的方法

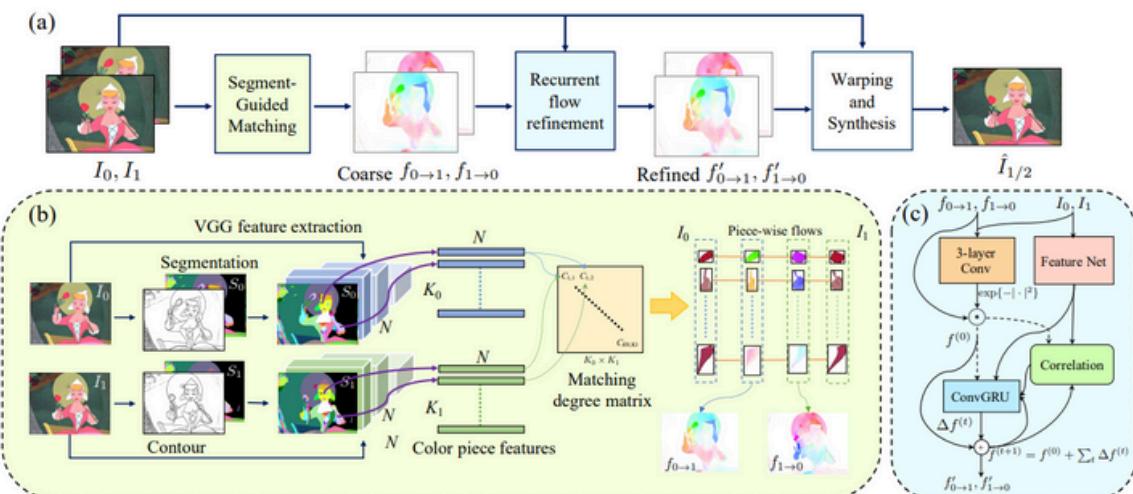
- Phase-based, Kernel-based
- **Optical/Feature Flow-based: FILM (ECCV'22)**
 - 近期最先進的方法，受益於最新的光流估計進展。常見流程是先利用光流計算兩幀間的對應關係，然後進行圖像warping與fusion。
- 儘管這些方法在真人影片的插值上表現良好，但通常難以處理卡通中的**大幅非線性運動與純色無紋理區域**。

卡通補幀挑戰與現有方法

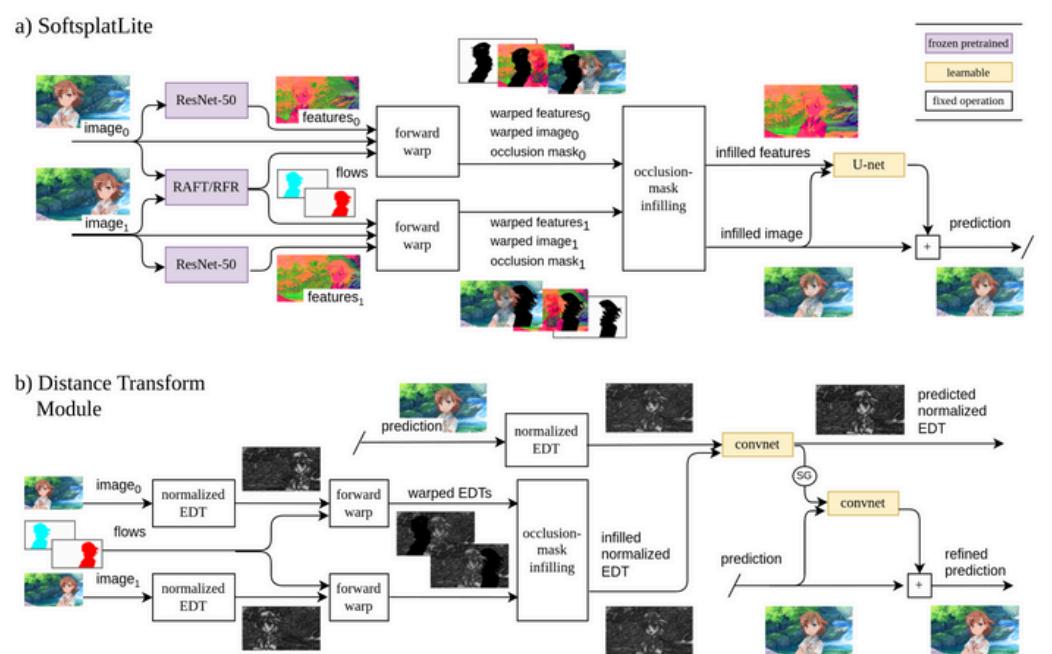
- 部分研究針對上述挑戰提出解法：
 - [Zhu et al. 2016]: 將卡通區域對應建模為network flow optimization problem。
 - **AnimelInterpp** [Li et al. 2021b]: 以segment-guided matching module based on **color piece matching**，提升對應關係識別準確度。
 - **EISAI** [Chen and Zwicker 2022]: 改善純色區域中移動物體的感知品質。
 - [Li et al. 2021a]: 引入**中介草圖**來處理大幅運動情境，但草圖資料往往難以取得，因需人工繪製。
- 雖然這些方法在卡通插值上有顯著進展，但**仍依賴顯式對應關係識別與線性運動假設**，難以建模卡通中常見的**複雜非線性運動與遮擋**現象。
- generative cartoon interpolation: 結合來自真人影片的豐富生成式運動先驗。



FILM (ECCV'22)



AnimelInterpp (CVPR'21)



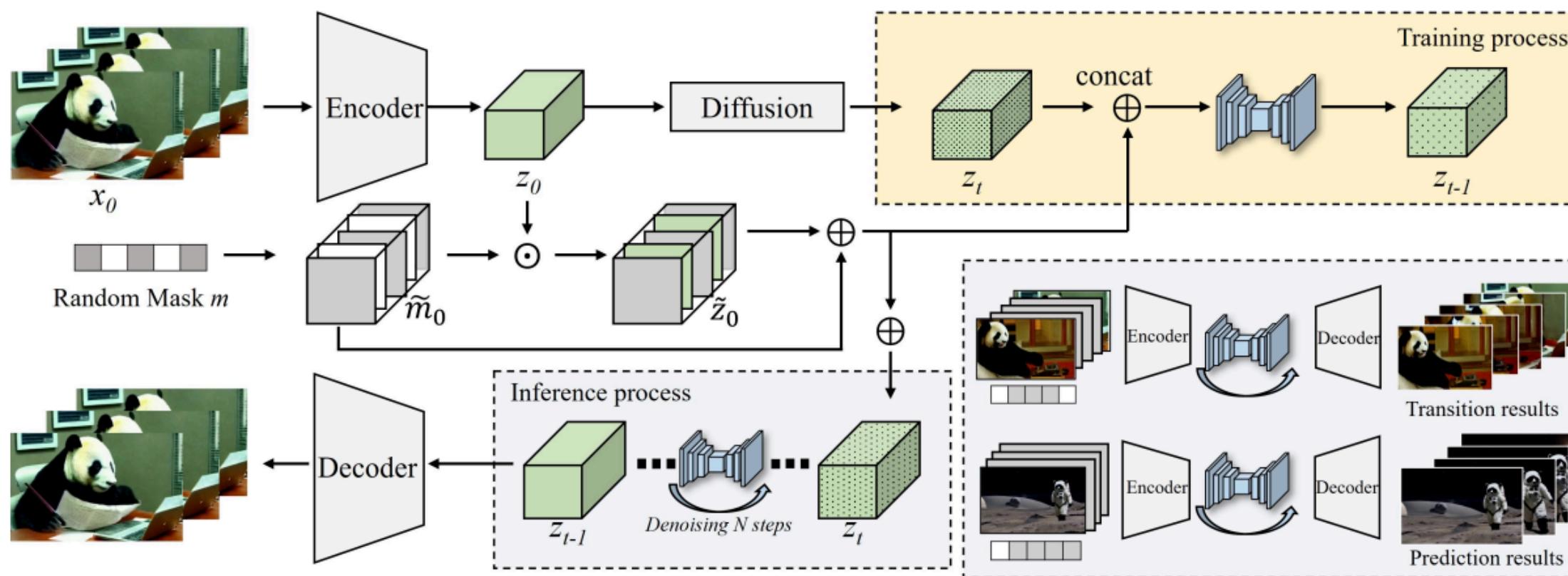
EISAI (ECCV'22)

Related Work

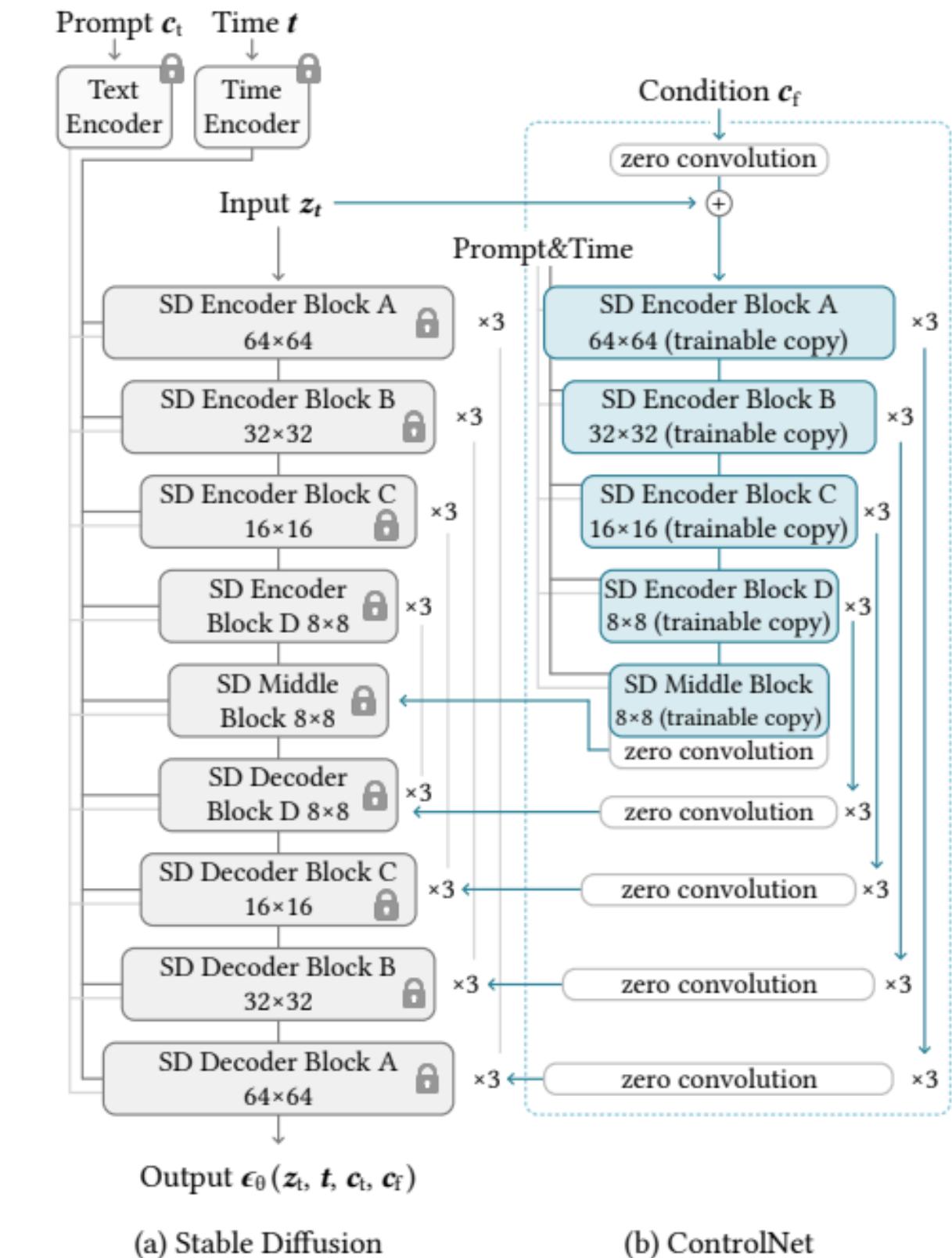
2. Image-Conditioned Video Diffusion Models

- 近期大量研究投入於訓練 large-scale **text-to-video (T2V) diffusion models**，這些模型在大量數據下能合成多樣化且自然的影片內容。
- Additional image conditions to these T2V models is well-studied for **image-to-video (I2V) synthesis**
- SENI**[Chen et al. 2024]: 透過連接兩個場景的輸入幀來生成過渡片段，將這些影像與noisy video latents共同輸入至 diffusion U-Net 模型中。
- single-image-to-video diffusion models** (如 DynamiCrafter、SparseCtrl、PixelDance) 也展示出在插值或場景過渡方面的潛力。
- 這些方法可透過拼接兩個輸入影像與latents、或使用額外編碼器進行控制（類似 **ControlNet**）。
- 然而，這些方法在應用於卡通插值時並不穩定。

因此本研究旨在將真人影片學到的運動生成先驗，適配至 I2V diffusion models的卡通生成式插值任務中，形成全新框架。



SENI (ICLR'24)



ControlNet (ICCV'23)

Method

ToonCrafter 框架

基於 I2V 的生成式 Diffusion 模型，目的是針對卡通動畫的補幀需求進行特化。

基於 SOTA 的 **DynamiCrafter**，但為了解決卡通動畫補幀的三大挑戰，設計了以下三個關鍵改進：

- Toon Rectification Learning: 跨領域適配
- Dual-reference-based 3D Decoder: 彌補潛在空間中導致的視覺劣化
- Frame-independent Sketch Encoder: 提供使用者對補幀結果的互動控制能力。

Framework

- ToonCrafter 接收兩張卡通影格 \mathbf{x}^1 和 \mathbf{x}^L ，透過 Diffusion 模型產生中間幀的潛變量 \mathbf{z}_0 ，再由 Decoder \mathcal{D} 將其還原為實際畫面。
- 使用者可輸入草圖作為補幀控制引導。

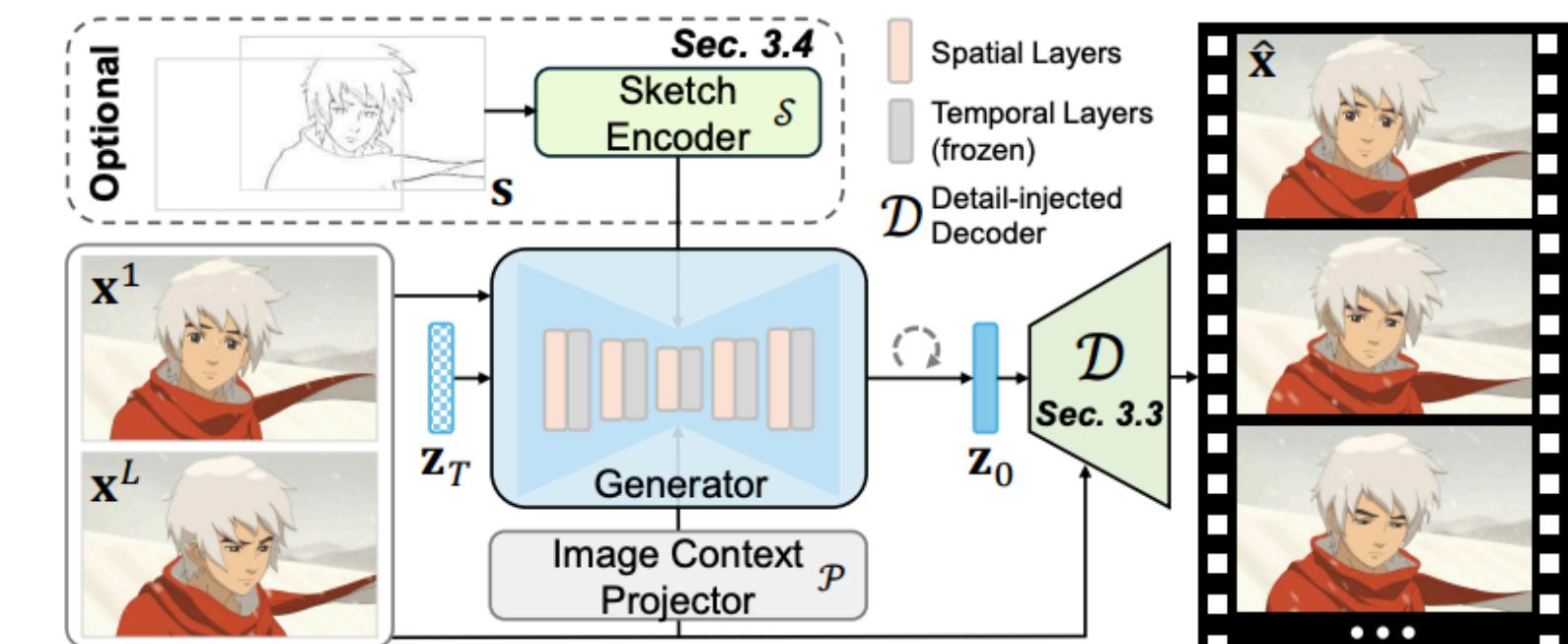


Fig. 2. Overview of the proposed *ToonCrafter*. Given two cartoon images \mathbf{x}^1 and \mathbf{x}^L , ToonCrafter leverages the image-to-video generative diffusion model as a generator to generate intermediate frame latents \mathbf{z}_0 . These latents are subsequently decoded into pixel space through the proposed detail-injected decoder with \mathbf{x}^1 and \mathbf{x}^L as detail guidance. Optionally, the interpolation can be controlled with sparse sketch guidance. ©B&T.

擴散模型 (Diffusion Models, DMs) 是一種基於得分的生成模型，會將資料 $\mathbf{x}_0 \sim p_{\text{data}}(\mathbf{x})$ 加上高斯雜訊轉為 $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ ，接著透過學習到的去雜訊網路 ϵ_θ 還原資料：

$$\min_{\theta} \mathbb{E}_{t, \mathbf{x} \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(0, \mathbf{I})} \|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2, \quad (1)$$

其中 ϵ 是真實的高斯雜訊、 t 是時間步、 \mathbf{x}_t 是前向擴散過程中的輸出。

在影片生成中，Latent Diffusion Models (LDMs) 能有效降低計算成本。基於 DynamiCrafter (Xing et al. 2023) 實作，流程如下：

- 給定影片 $\mathbf{x} \in \mathbb{R}^{L \times 3 \times H \times W}$
- 編碼為潛在表徵 $\mathbf{z} = \mathcal{E}(\mathbf{x}) \in \mathbb{R}^{L \times C \times h \times w}$
- 接著進行 forward 擴散 $\mathbf{z}_t = p(\mathbf{z}_0, t)$ 與 backward 去雜訊 $\mathbf{z}_t = p_\theta(\mathbf{z}_{t-1}, c, t)$
- 其中 c 為條件資訊 (如文字提示 c_{text} 和圖像提示 c_{img})

應用在插值上，起始與終止幀 $\mathbf{x}^1, \mathbf{x}^L$ 作為條件，留下中間幀讓模型生成。

損失函數 (引入 FPS 控制) 為：

$$\min_{\theta} \mathbb{E}_{\mathcal{E}(\mathbf{x}), t, \epsilon \sim \mathcal{N}(0, \mathbf{I})} \left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t; \mathbf{c}_{\text{img}}, \mathbf{c}_{\text{text}}, t, \text{fps})\|_2^2 \right], \quad (2)$$

Method

Toon Rectification Learning

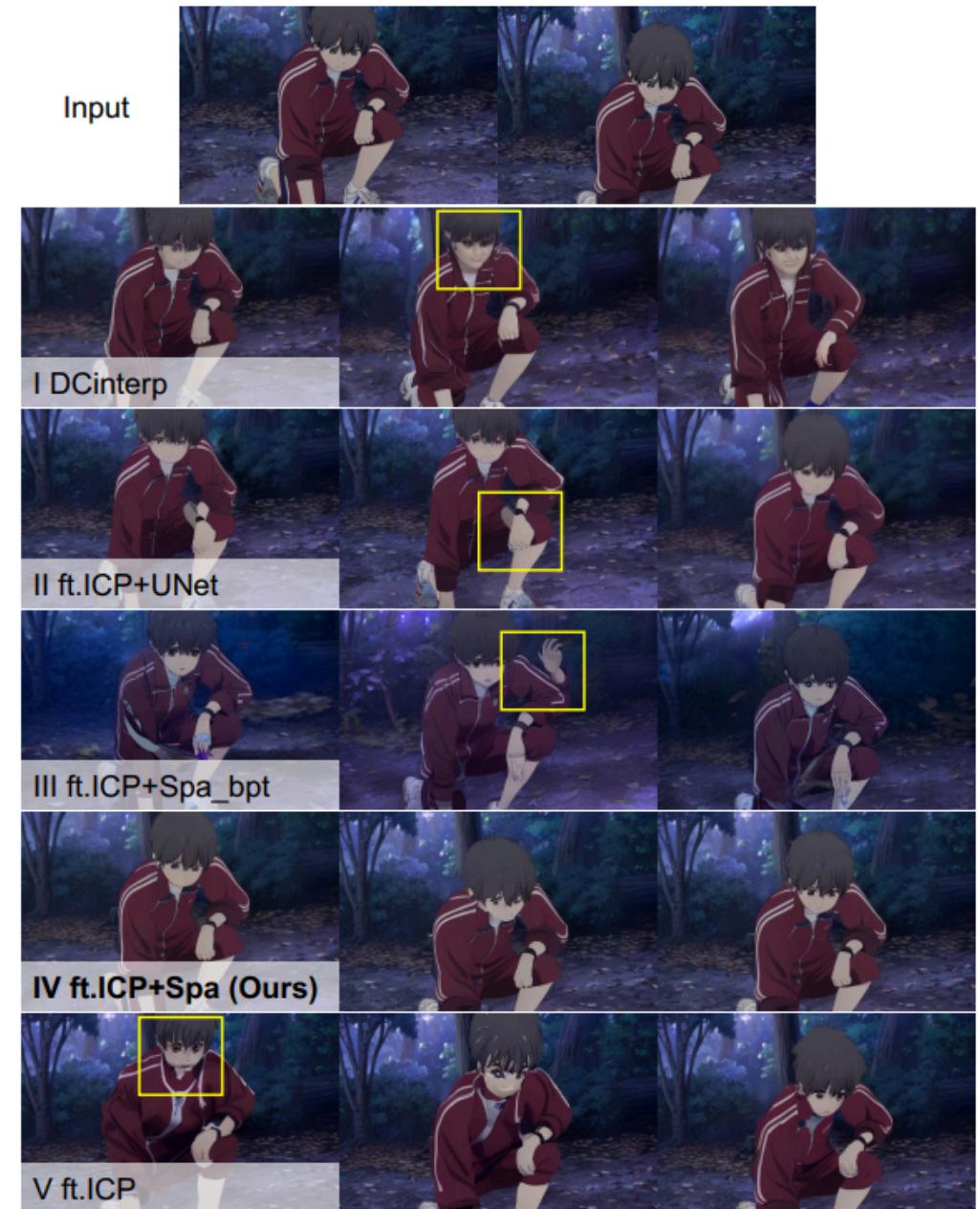
卡通與真人影片的風格差異顯著，包含誇張表情、簡化紋理與平面色塊，這使得直接套用現有擴散模型時，會出現風格錯誤或運動不合適等問題（如圖 6 左上）。

1. 蒐集卡通影片資料集

- 包含中日西動畫共 271K 幀
- 過濾字幕、多文字區塊、畫質不良、對不準等
- 使用 CRAFT 判斷文字，BLIP-2 生成標題過濾重複樣本，CLIP 度量視覺語意對齊度

2. 校正學習策略

- 若直接使用 DynamiCrafter 微調，會導致災難性遺忘
- 設計 **selective tuning**：
 - **凍結 temporal layers**以保留時序對齊能力
 - **微調 spatial layers 與 image context projector**，以適應動畫風格
 - **使用 StableDiffusion v2.1 架構**



Method

Detail Injection and Propagation in Decoding

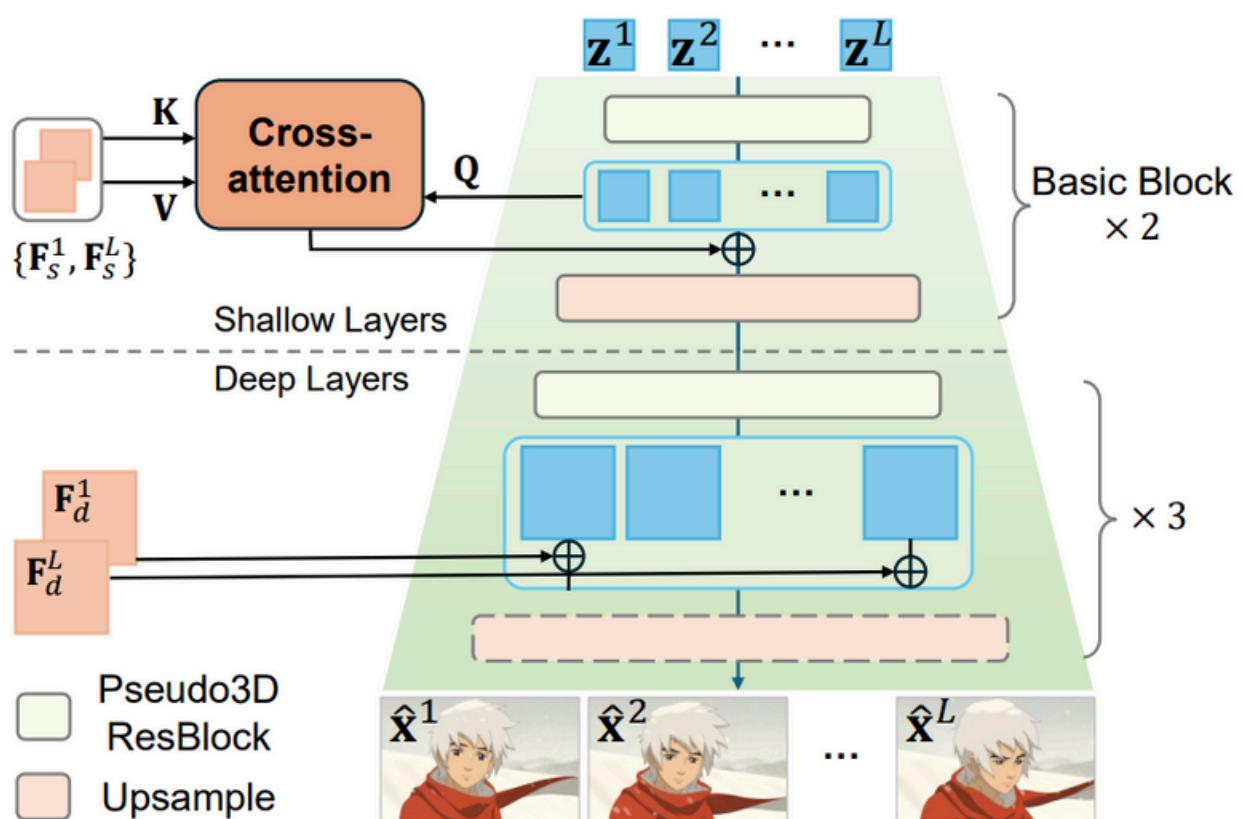
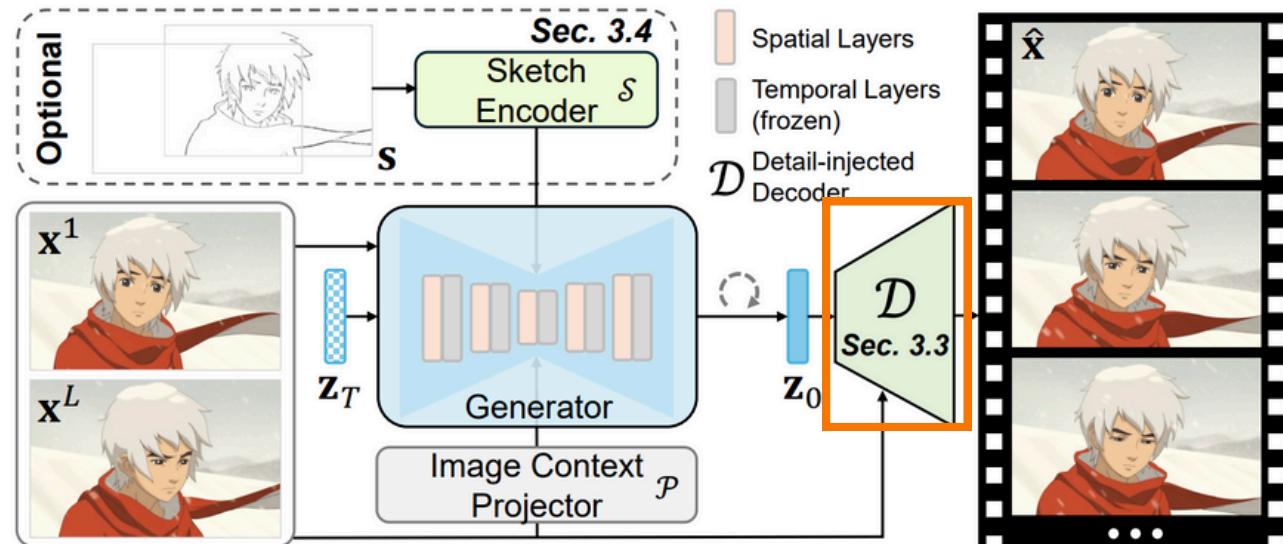


Fig. 3. Illustration of the detail-injected 3D decoder. Given frame latents z as input, we inject the intermediate features of input images x^1 and x^L from encoder \mathcal{E} through cross-attention in shallow layers, while via residual learning, i.e., addition to features of 1-st and L -th frame in deep layers.
©B&T.

由於潛在擴散模型對影片結構的建模能力有限，還原後的幀常有失真或模糊（尤其在卡通中高對比輪廓）。我們設計雙參考解碼器：

- 從 $\mathbf{x}^1, \mathbf{x}^L$ 中提取特徵 $\{F_k^i\}_{i \in \mathcal{S}}$
- 對應至編碼器的第 k 幀，透過 cross-attention 將其注入生成幀

定義中間層輸入為：

$$\mathbf{G}_{\text{out}}^j = \text{Softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{d}}\right)\mathbf{V} + \mathbf{G}_{\text{in}}^j, j \in 1 \dots L \quad (3)$$

$$Q = G_i^{\text{in}} W_Q, \quad K = [F_1^i; F_L^i] W_K, \quad V = [F_1^i; F_L^i] W_V$$

此外，加入 ZeroConv 處理最後一層： $\mathbf{G}_{\text{out}}^1 = \text{ZeroConv}_{1 \times 1}(F_i^1) + \mathbf{G}_{\text{in}}^1$.
為提升時間一致性，我們結合 pseudo-3D 卷積 (Qiu et al. 2017) 。

損失函數設計： $\mathcal{L} = \mathcal{L}_1 + \lambda_p \mathcal{L}_p + \lambda_d \mathcal{L}_d$, (5)

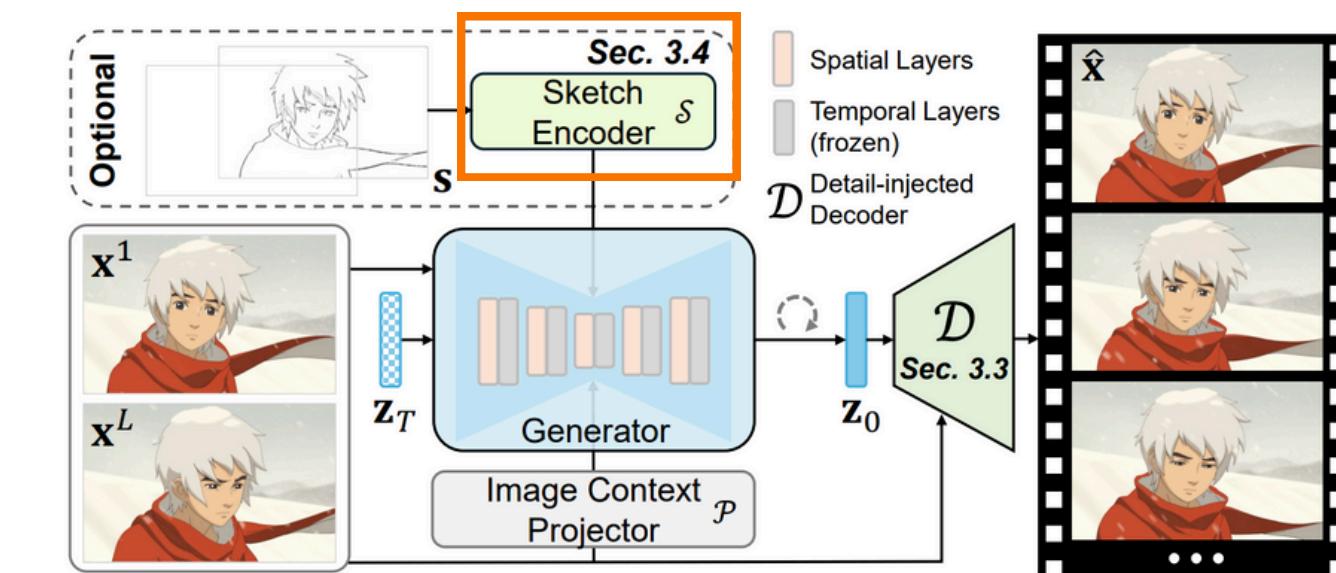
- \mathcal{L}_1 ：像素差 (MAE)
- \mathcal{L}_p ：感知損失 (LPIPS) Perceptual Similarity
- \mathcal{L}_d ：對抗損失 (GAN)
- $\lambda_p = 0.1 \cdot \lambda_d$ 自適應加權 (Esser et al., 2021)

Method

Sketch-based Controllable Generation

生成式插值具非線性與多樣性，雖能處理複雜運動，但會帶來結果變異性

- 不同使用者需求：自然多變 V.S. 可控一致
- 引入 Sketch-based Guidance：設計時間無關的草圖編碼器使用者可用稀疏草圖控制生成運動
- 基於 ControlNet 將 video diffusion model 轉為 a sketch-conditioned generative model



支援稀疏輸入

- 使用者不需為所有幀繪製草圖，只需標記部分關鍵幀（圖 4）

逐幀調整的適配器設計

- 草圖編碼器 \mathcal{S} 可依據輸入草圖 s^i 、潛變量 z^i 、時間 t 單獨調整每一幀的特徵： $F_{\text{inject}}^i = \mathcal{S}(s^i, z^i, t)$

無草圖時使用空草圖作為輸入

- 避免無梯度學習，提升草圖編碼器學習效果： $F_{\text{inject}}^i = \mathcal{S}(s^\emptyset, z^i, t)$

若省略草圖幀會導致學習偏差

- 僅有草圖幀產生梯度，導致模型只學會控制單幀的空間內容
- 損害時間序列中的整體一致性 (temporal consistency)

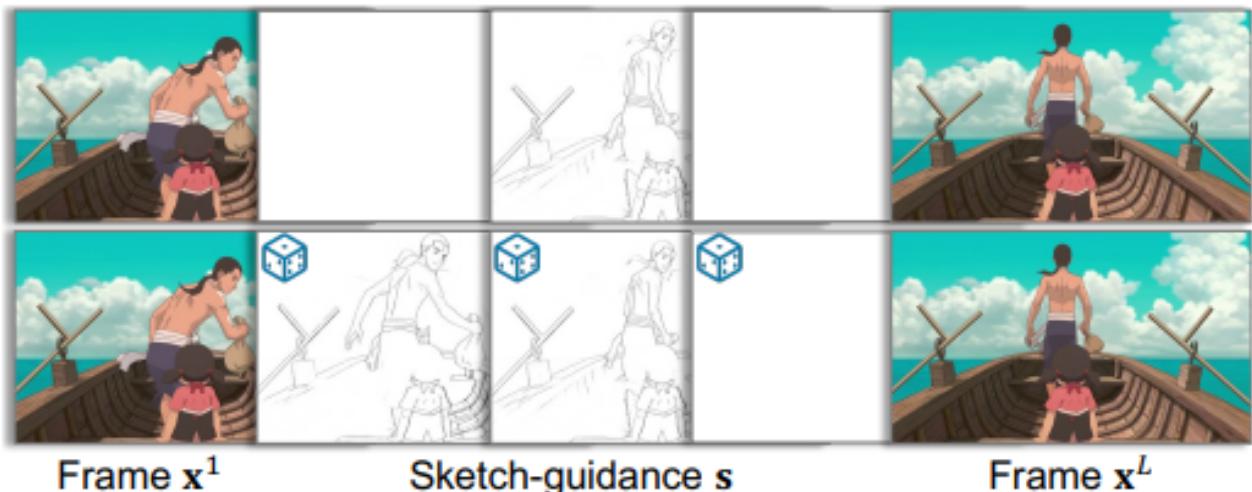


Fig. 4. Examples of different patterns of sketch-guidance: (top) bisection ($n=1$) and (bottom) random position. ©B&T.

訓練方式

將去雜訊網路 ϵ_θ 凍結，只優化草圖編碼器 \mathcal{S} 。該模組採用類似 ControlNet 架構，初始化自 StableDiffusion v2.1。

訓練目標：

$$\min_{\theta} \mathbb{E}_{\mathcal{E}(\mathbf{x}), \mathbf{s}, t, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\|\epsilon - \epsilon_{\theta}^{\mathcal{S}} (\mathbf{z}_t; \mathbf{c}_{\text{img}}, \mathbf{c}_{\text{txt}}, \mathbf{s}', t, \text{fps}) \|_2^2 \right], \quad (6)$$

- $\epsilon_{\theta}^{\mathcal{S}}$ 表示 ϵ_{θ} 與草圖編碼器 \mathcal{S} 的結合；
- s 是從 Anime2Sketch (Xiang et al. 2021) 取得的草圖；
- s' 是選取後的草圖子集 (如圖 4 所示)；
- $c_{\text{img}}, c_{\text{text}}$ 分別是影像與文字條件。

草圖選擇策略

為符合使用者習慣，80% 訓練使用 **bisection selection pattern** 決定草圖輸入幀：

- 對於插值段 (i, j) ，選擇第 $\lceil \frac{i+j}{2} \rceil$ 幀作為草圖幀。
- 遞迴選擇深度 $n \in [1, 4]$ ，將區間 $(1, L)$ 持續細分。
- 該策略模擬真實使用者行為 (習慣在等間隔處輸入草圖)。
- 其餘 20% 的訓練樣本隨機選擇草圖幀，以提升模型泛化能力。

- 基於 DynamiCrafter 的**I2V**模組（插值版本，輸入解析度為 512×320 ）。
- **Toon Rectification Learning :**
 - 微調空間層與生成層，學習率 1×10^{-5} ，批次大小 32。
- **Frame-independent Sketch Encoder :**
 - 使用 50K 幀訓練，學習率 1×10^{-5} ，批次大小 32。
- **Dual-reference-based 3D Decoder :**
 - 訓練 60K 幀，學習率 4.5×10^{-6} ，批次大小 16。
- **推論階段：**
 - 使用 **DDIM** 取樣 (Song et al. 2021)，並採用多條件分類器指導 (Ho and Salimans 2022)。

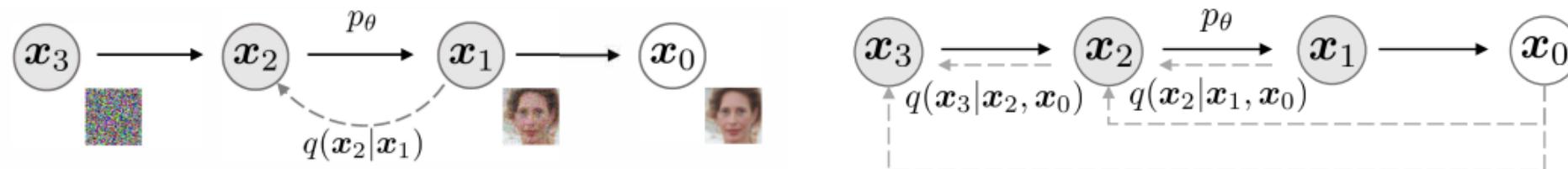


Figure 1: Graphical models for diffusion (left) and non-Markovian (right) inference models.

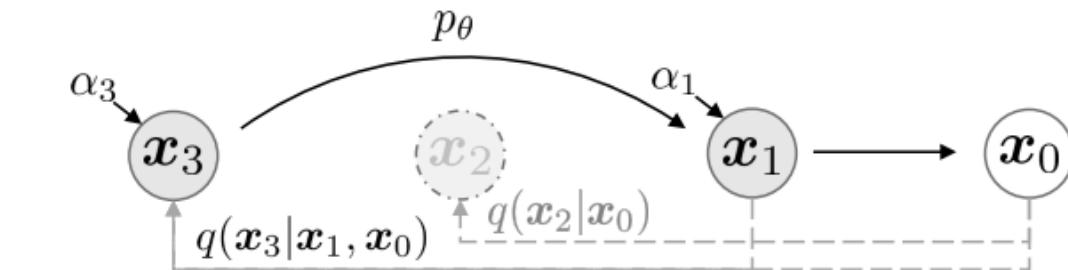


Figure 2: Graphical model for accelerated generation, where $\tau = [1, 3]$.

- 評估模型在空間與時間維度下的插值品質，指標包含：
- **FVD** : Fréchet Video Distance
- **KVD** : Kernel Video Distance
- **LPIPS** : 感知損失 (圖像相似度)
- **CLIP_{Img} / CLIP_{Txt}** : 與原始影像與文字prompt的一致性
- **CPBD** : 邊緣清晰度

Table 1. Quantitative comparison with state-of-the-art video interpolation methods on the cartoon test set (1K).

Metric	AnimeInterp	EISAI	FILM	SEINE	Ours
FVD ↓	196.66	146.65	189.88	98.96	43.92
KVD ↓	8.44	5.55	8.01	2.93	1.52
LPIPS ↓	0.1890	0.1729	0.1702	0.2519	0.1733
CLIP _{Img} ↑	0.8866	0.9083	0.9006	0.8531	0.9221
CLIP _{Txt} ↑	0.3069	0.3097	0.3083	0.2962	0.3129
CPBD ↑	0.5974	0.6413	0.6317	0.6630	0.6723

- 評估模型在空間與時間維度下的插值品質，指標包含：

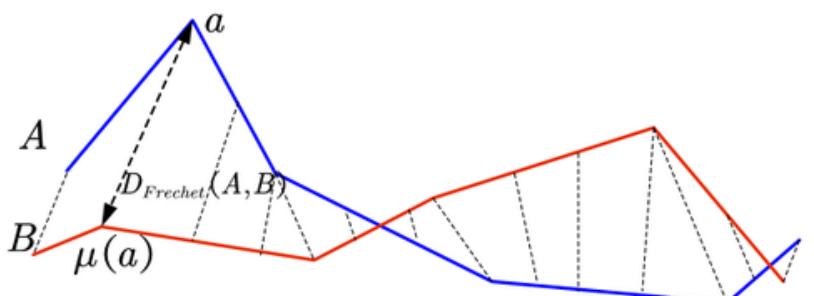
- FVD : Fréchet Video Distance**

- 簡化從真實和生成資料集中提取的影片的分佈間隔。

- 影片特徵是由影片分類器提取的。

$$d(P_R, P_G) = |\mu_R - \mu_G|^2 + \text{Tr} \left(\Sigma_R + \Sigma_G - 2(\Sigma_R \Sigma_G)^{\frac{1}{2}} \right)$$

Tr 是指 trace (跡) 運算子，也就是「矩陣對角線元素的總和」



- KVD : Kernel Video Distance**

- 影片版的Frechet Inception Distance (FID)，從影片中提取特徵

- 基於 Maximum Mean Discrepancy (MMD) 的 unbiased estimator

$$\sum_{i \neq j}^m \frac{k(x_i, x_j)}{m(m-1)} - 2 \sum_i^m \sum_j^n \frac{k(x_i, y_j)}{mn} + \sum_{j \neq i}^n \frac{k(y_i, y_j)}{n(n-1)}$$

- x_i ：從生成影片中抽出的特徵向量

- y_j ：從真實影片中抽出的特徵向量

- m, n ：分別是生成與真實影片樣本數

- $k(a, b)$ ：核函數，常用的是 Gaussian kernel : $k(a, b) = \exp \left(-\frac{\|a - b\|^2}{2\sigma^2} \right)$

- LPIPS**

- 感知損失（圖像相似度）
- 衡量兩張圖片在人類感知下的相似度，使用訓練過的 CNN 特徵距離：

$$\text{LPIPS}(x, y) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|w_l \odot (f_l^x(h, w) - f_l^y(h, w))\|_2^2$$

- f_l^x, f_l^y ：圖像 x, y 在第 l 層 CNN (如 VGG) 提取的特徵圖
- w_l ：第 l 層的通道權重 (learned)
- H_l, W_l ：該層特徵圖高與寬

- CLIPimg / CLIPtxt**

- 與原始影像與文字prompt的一致性
- 計算 CLIP 模型中影像與文字的餘弦相似度：

$$\text{CLIP}_{\text{img}} = \cos(\phi_{\text{img}}(x), \phi_{\text{img}}(x_{\text{gt}})), \quad \text{CLIP}_{\text{txt}} = \cos(\phi_{\text{img}}(x), \phi_{\text{txt}}(t))$$

- ϕ_{img} ：CLIP 影像 encoder 輸出特徵
- ϕ_{txt} ：CLIP 文本 encoder 輸出特徵
- x ：生成影像 · x_{gt} ：真實影像 · t ：對應文字 prompt

- CPBD**

- 邊緣清晰度，以人眼可接受範圍計算多少比例的區域是清晰的，值越高代表影像越清晰：

$$\text{CPBD} = \frac{1}{N} \sum_{i=1}^N P_{\text{blur}}(B_i)$$

- B_i ：第 i 條邊緣的局部對比值
- $P_{\text{blur}}(B_i)$ ：該邊緣模糊的機率 (越低代表越清晰)
- N ：總邊緣數目

Experiments

Qualitative Comparisons

- 如圖 5 所示，我們的方法能夠生成語意一致、具可理解性與自然過渡的插值幀。
- 相比之下，Animelinterp、EISAI 與 FILM 會產生漂浮人影、手掌消失等錯誤。
- SEINE 雖能處理大運動，但在結構與紋理表現上仍不穩定。



Fig. 5. Visual comparison of cartoon interpolation results from Animelinterp, EISAI, FILM, SEINE, and our ToonCrafter. The inputs are from ATD-12K dataset.

Experiments

User Study

- 招募了 24 位參與者，請他們從視覺品質角度比較結果，統計如下（表格 2）：
- 在所有主觀評估中皆勝出。
 - 運動品質 MQ
 - 時間相應性 TC
 - 幀保真度 FF

Table 2. User study statistics of the preference rate for Motion Quality (M.Q.), Temporal Coherence (T.C.), and Frame Fidelity (F.F.).

Property	AnimeInterp	EISAI	FILM	SEINE	Ours
M.Q. ↑	3.24%	6.94%	6.02%	14.81%	68.98%
T.C. ↑	6.94%	14.81%	13.43%	15.74%	49.07%
F.F. ↑	6.48%	11.57%	12.50%	18.06%	51.39%

Experiments

Ablation Study

分析以下幾種校正策略對插值品質的影響：

- I. DCInterp
 - 無微調
- II. ft.ICP+UNet
 - 微調 ICP 與 UNet
- III. ft.ICP+Spa_bpt
 - 微調 ICP 與空間層（跳過時間層）
- IV. ft.ICP+Spa (Ours)
 - 保留時間層，微調 ICP 與空間層
- V. ft.ICP
 - 僅微調 ICP

策略 IV 最佳，證明ToonCrafter方法可保留時間一致性，並維持語義正確性。

Table 3. Ablation study of different rectification learning strategies. *All these variants are evaluated without the proposed dual-reference-based 3D decoder to demonstrate the original performance of the denoising network.

Variant*	ICP	Spa.	Temp.	Bypass	Temp.	FVD↓	CLIP _{img} ↑
I DCinterp						86.62	0.8637
II ft.ICP+UNet	✓	✓	✓			70.73	0.8978
III ft.ICP+Spa_bpt	✓	✓		✓		291.45	0.7997
IV ft.ICP+Spa (Ours)	✓	✓				52.73	0.9096
V ft.ICP		✓				81.45	0.8875



Fig. 6. Visual comparison of the interpolation frames generated by variants with different rectification strategies.

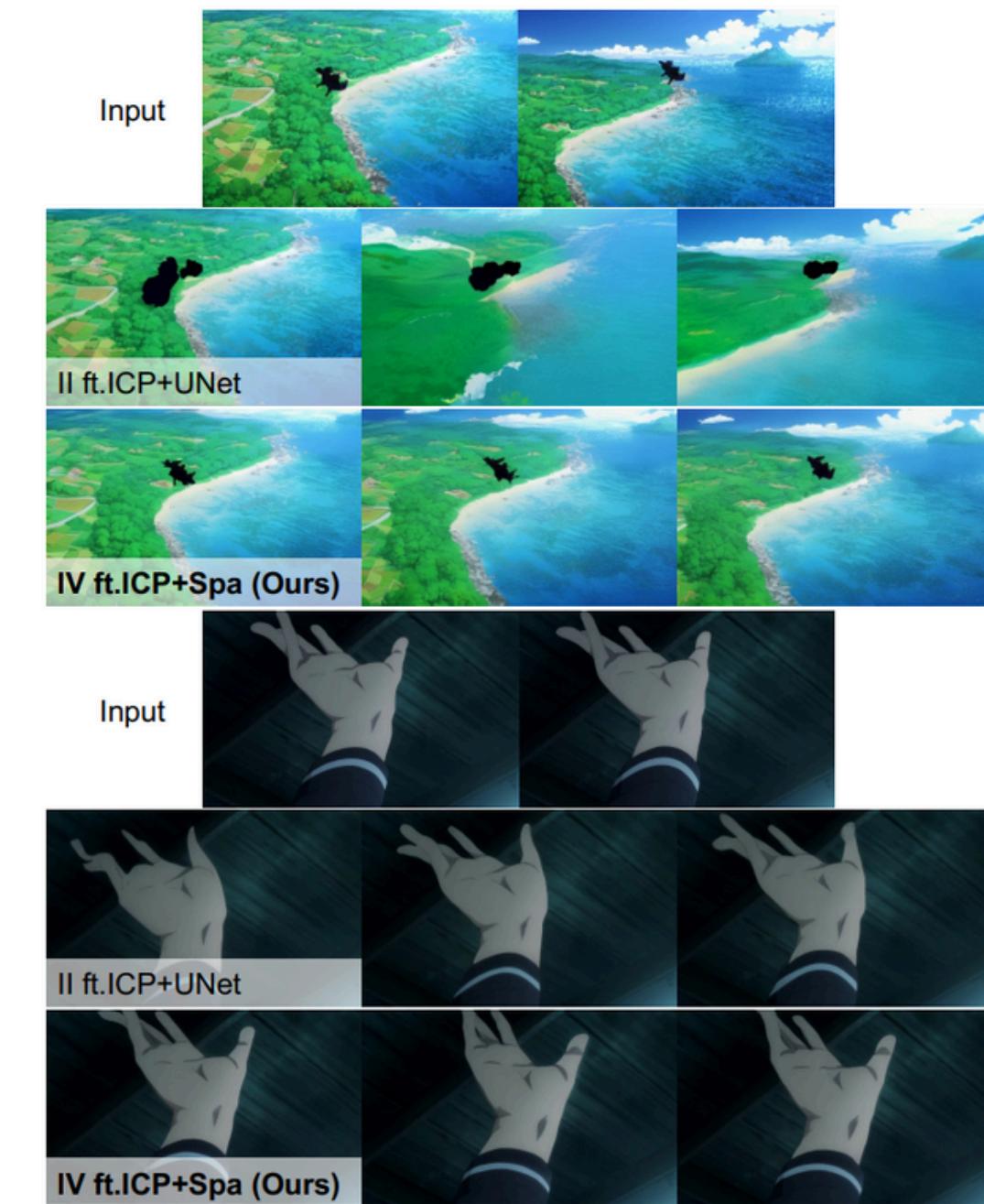


Fig. 7. Visual comparison of the interpolation frames generated by the rectification strategy variant II ft.ICP+UNet and our method ft.ICP+Spa.

Experiments

Ablation Study

Dual-reference-based 3D VAE decoder

- 移除 hybrid-attention-residual (**HAR**)或 pseudo-3D convolutions (**P3D**)皆導致品質下降，證明雙參考 + 殘差機制關鍵。
- 圖 8：視覺比較顯示，無參考注入或去除 P3D/HAR 時，結構與細節明顯劣化。

Table 4. Quantitative comparison of reconstruction by different decoders on the 1K cartoon video evaluation set (256×256). HAR: Hybrid-attention-residual. P3D: Pseudo-3D Convolution.

Variant	Ref.	Temp.	PSNR ↑	SSIM ↑	LPIPS ↓
Ours	✓	✓	33.83	0.9450	0.0204
Ours _{w/o P3D}	✓	✗	32.51	0.9270	0.0326
Ours _{w/o HAR & P3D}	✗	✗	29.49	0.8670	0.0426

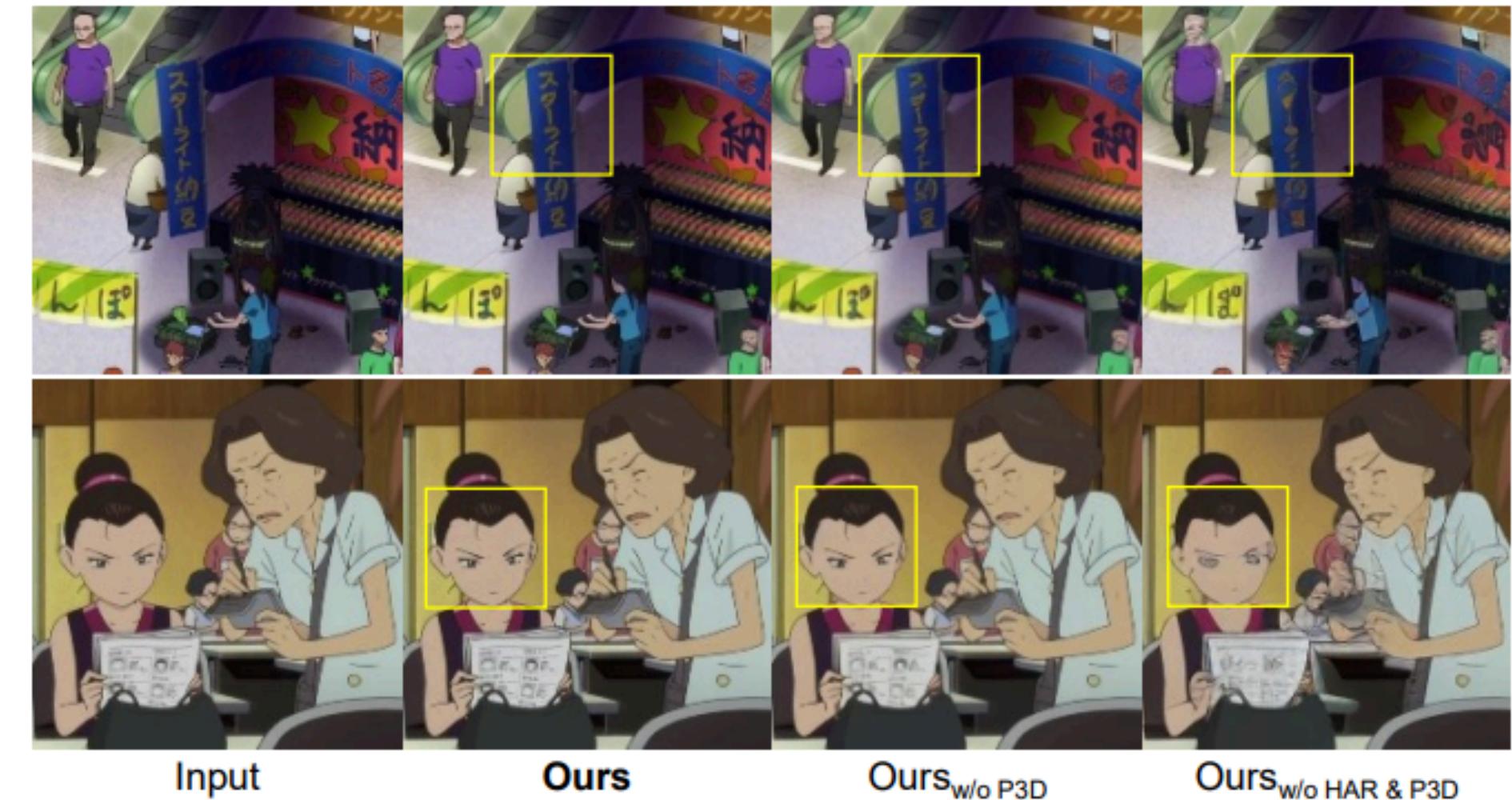


Fig. 8. Visual comparison of the reconstructed video frames produced by different decoder variants. Only the middle frames are shown above.

Experiments

Ablation Study

- 訓練一個 **ZeroGate** 變體：推論時將草圖全置零。
- 與我們的 FrameInd.Enc. (frameindependent encoder) 比較：
 - **ZeroGate** 難以維持一致性；
 - **FrameInd.Enc.** 在有/無草圖下均表現穩定（圖 9）。
- 圖 10：展示了不同幀位置的重建品質，ToonCrafter顯著優於其他版本。

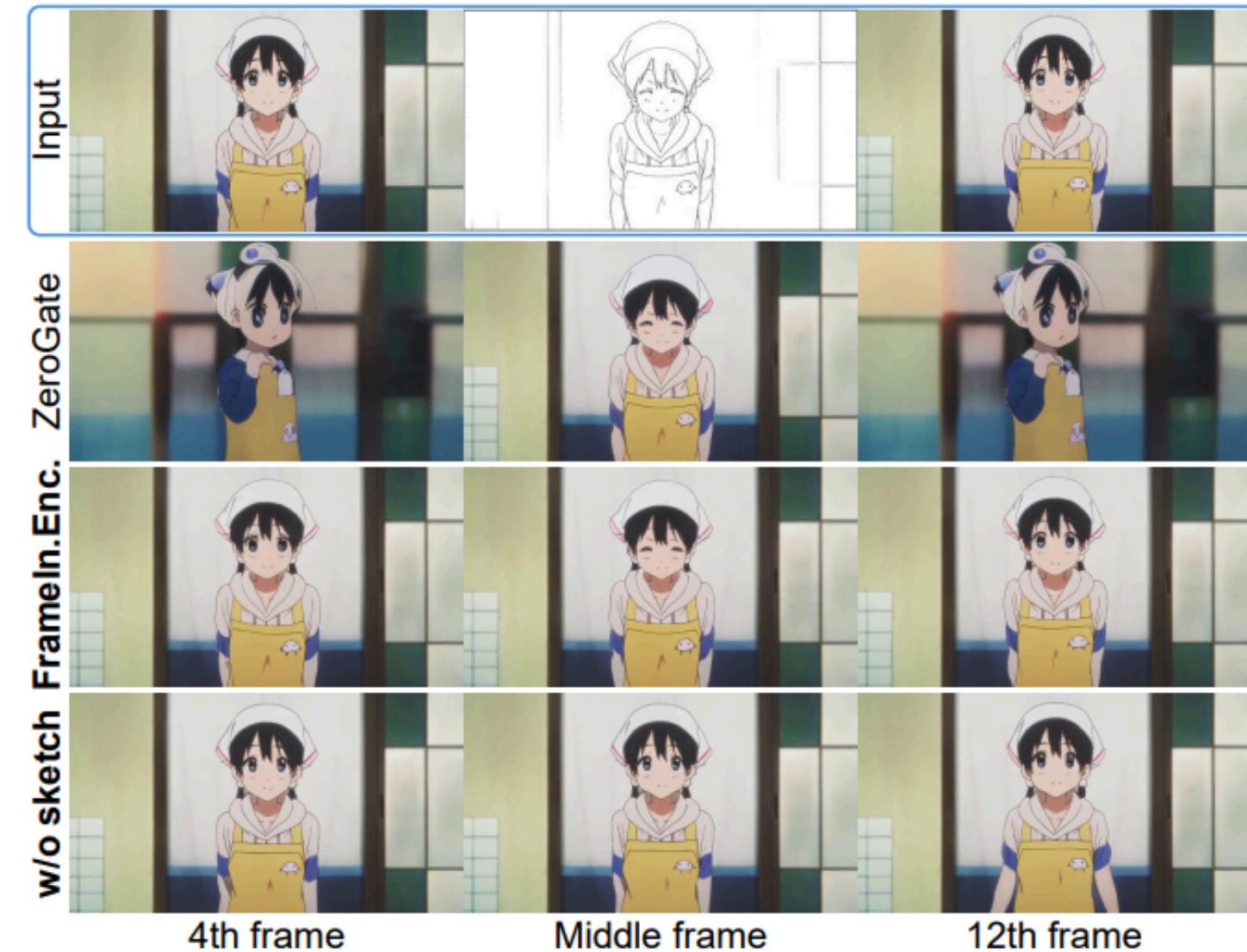


Fig. 9. Visual comparison of intermediate frames generated by different variants, with or without sparse sketch guidance (middle frame only).

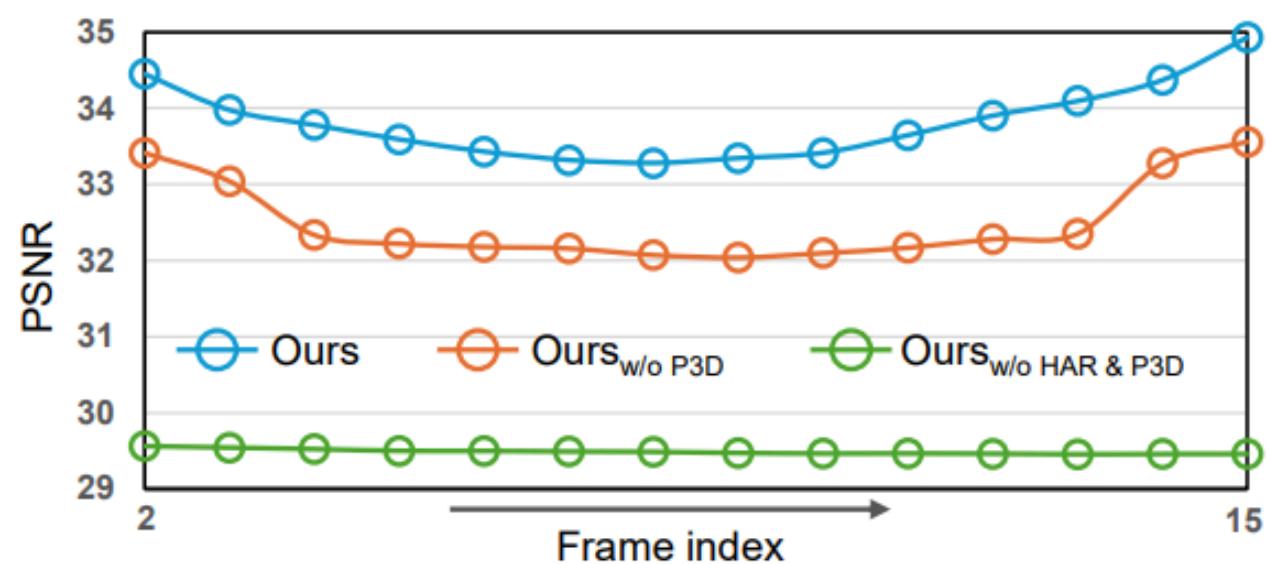
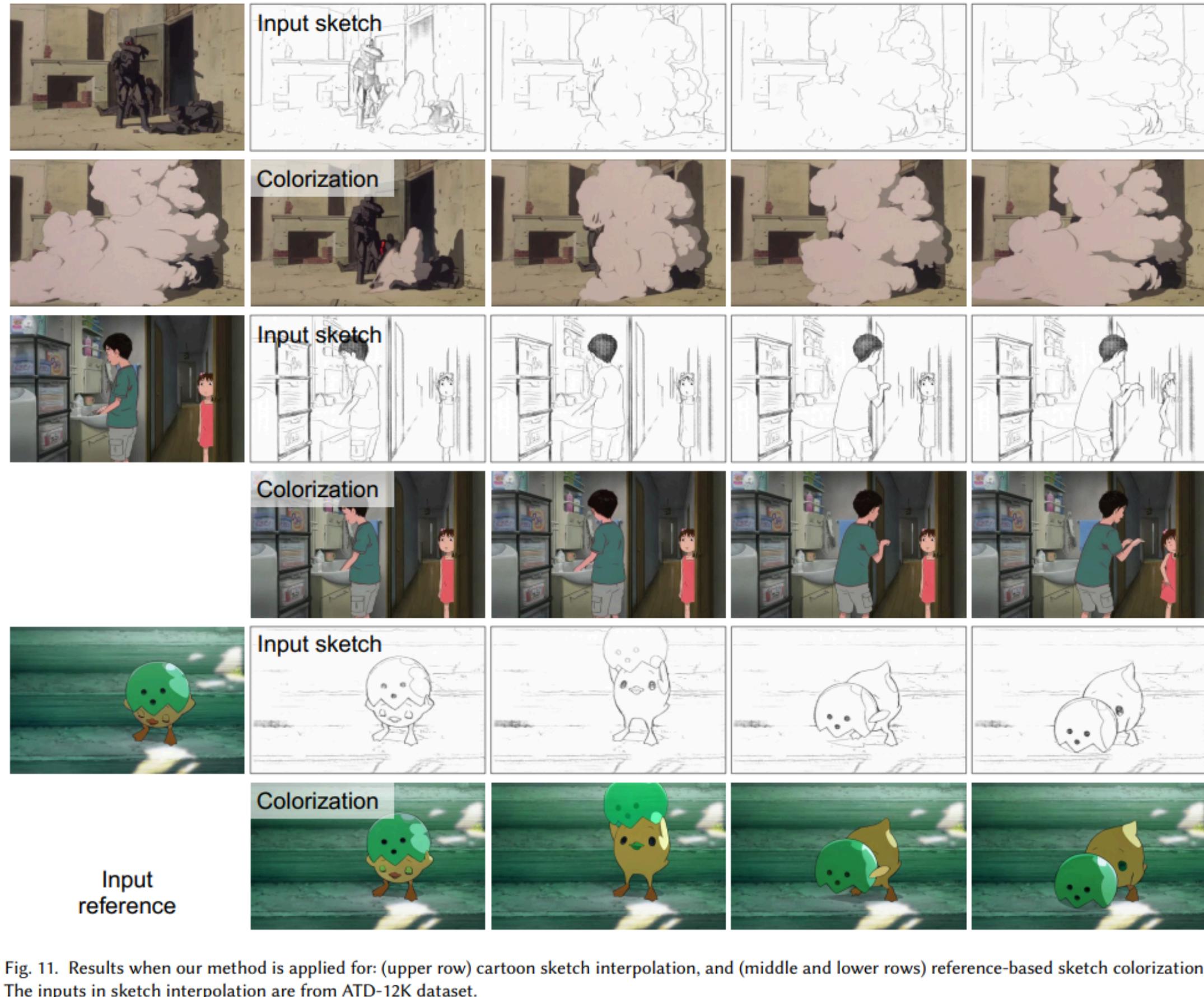
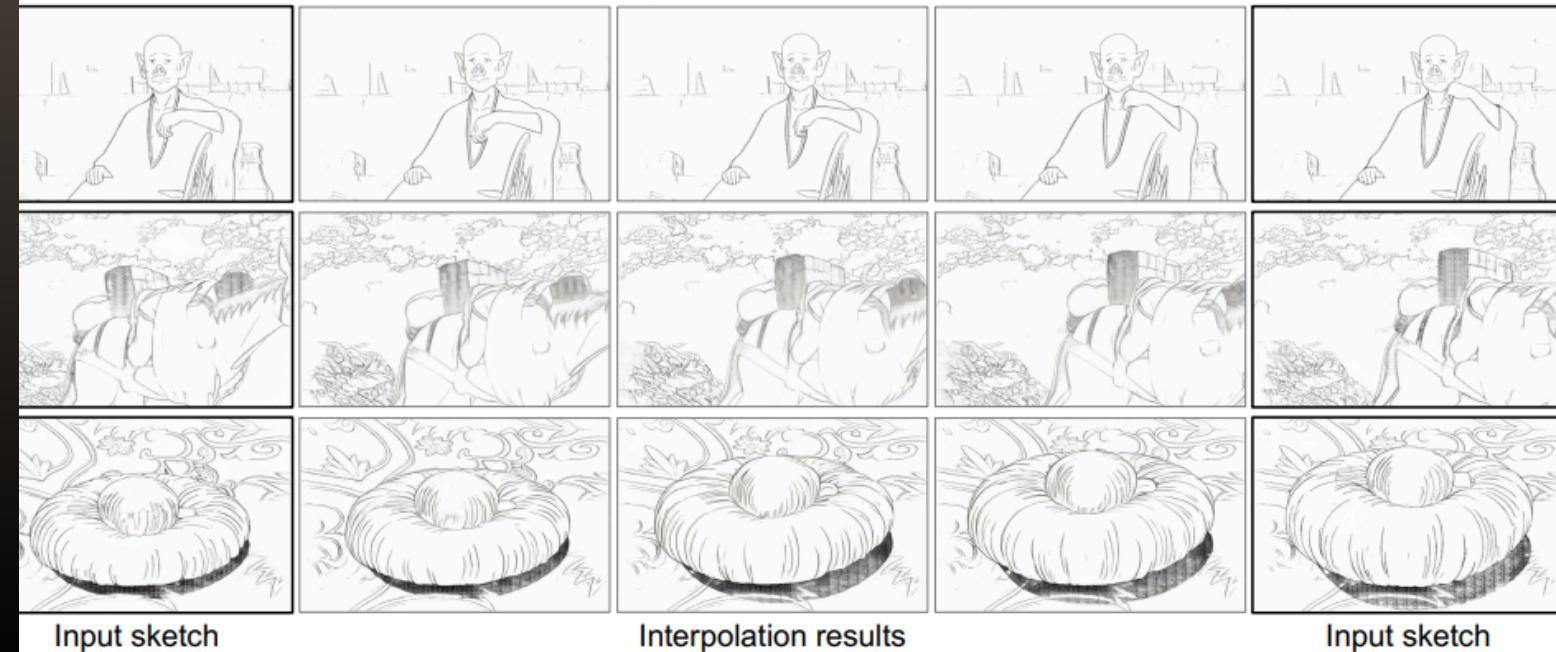


Fig. 10. Comparison of reconstruction quality against the frame index.

Applications

- 卡通草圖插值是一項極具挑戰的任務。儘管輸入簡略，ToonCrafter的方法仍能生成高品質插值幀。
- 進一步應用包含：
 - 參考草圖著色（圖 11）；
 - 僅使用一至兩張草圖即能完成補幀與著色。



Limitations

Our model may not correctly and semantically understand the image contents. (e.g., the black part should be the rigid body of the aircraft, which cannot sway with the wind.)

Input starting frame



Input ending frame



Our failure case



Our model may struggle to generate convincing transition motions when objects appear or disappear in the frame.

Input starting frame



Input ending frame



Our failure case



Conclusion

本研究提出 ToonCrafter：一個用於卡通補幀的生成式框架。

關鍵技術包括：

- **Toon rectification learning**：有效適配真人運動先驗至卡通領域；
- **Dual-reference 3D decoder**：解決細節損失；
- **Sketch-Based Controllable Generation**：提升可控性與使用者互動性。

定量與定性實驗皆證明本方法顯著優於現有方法，並展現出良好的泛化與應用潛力。

1. 場景切換、變遷的細節

- **限制**：ToonCrafter 的描述聚焦在單一連續場景內的幀插值，若場景中出現明顯跳接 (**scene transition**)，如背景突變或場景切換，可能會導致不連貫的插值結果。
- **突破**：可探索場景分割與轉場建模 (e.g., detect scene boundaries + apply different interpolation models)，或學習跨場景的風格與語義轉換。

2. 角色間互動與複雜場景佈局 (Scene Layout)

- **限制**：目前方法偏向擅長處理單一角色、簡單背景的動畫，對多角色互動、遮擋、前後景深關係複雜的 layout仍存在挑戰，且插值生成物件會變形（車尾扭曲）。
- **突破**：設計更細粒度的 scene layout-aware interpolation model，顯式建模角色的 spatial hierarchy，例如用 layout parsing 或 object-centric representations。



3. 視覺敘事與時間連貫性 (Temporal Narrative Coherence)

- **限制**：插值模型主要關注在幀與幀間的 motion/appearance 過渡，較少考慮到動畫中的敘事邏輯或情感變化的時間連貫性。
- **突破**：引入更高層次的 **story understanding**，例如用 LLM 或 VQA 模型分析動畫敘事，再輔助控制 interpolation 的語義發展。

4. 可擴展性與使用場景

- **限制**：當前模型可能主要針對經過精簡處理的動畫素材，對於高解析度長片、複雜鏡頭語言 (如快速推拉、搖攝) 適應性未知。
- **突破**：探索可擴展的 **hierarchical interpolation 系統** (scene-level → shot-level → frame-level)，或支援用戶導向的動畫生成工作流 (e.g., 插值+音效+敘事引導)。